

2021 MIRI Conversations

1. [Ngo and Yudkowsky on alignment difficulty](#)
2. [Ngo and Yudkowsky on AI capability gains](#)
3. [Yudkowsky and Christiano discuss "Takeoff Speeds"](#)
4. [Soares, Tallinn, and Yudkowsky discuss AGI cognition](#)
5. [Christiano, Cotra, and Yudkowsky on AI progress](#)
6. [Biology-Inspired AGI Timelines: The Trick That Never Works](#)
7. [Reply to Eliezer on Biological Anchors](#)
8. [Shulman and Yudkowsky on AI progress](#)
9. [More Christiano, Cotra, and Yudkowsky on AI progress](#)
10. [Conversation on technology forecasting and gradualism](#)
11. [Ngo's view on alignment difficulty](#)
12. [Ngo and Yudkowsky on scientific reasoning and pivotal acts](#)
13. [Christiano and Yudkowsky on AI predictions and human intelligence](#)
14. [Shah and Yudkowsky on alignment failures](#)
15. [Late 2021 MIRI Conversations: AMA / Discussion](#)

Ngo and Yudkowsky on alignment difficulty

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This post is the first in a series of transcribed Discord conversations between Richard Ngo and Eliezer Yudkowsky, moderated by Nate Soares. We've also added Richard and Nate's running summaries of the conversation (and others' replies) from Google Docs.

Later conversation participants include Ajeya Cotra, Beth Barnes, Carl Shulman, Holden Karnofsky, Jaan Tallinn, Paul Christiano, Rob Bensinger, and Rohin Shah.

The transcripts are a complete record of several Discord channels MIRI made for discussion. We tried to edit the transcripts as little as possible, other than to fix typos and a handful of confusingly-worded sentences, to add some paragraph breaks, and to add referenced figures and links. We didn't end up redacting any substantive content, other than the names of people who would prefer not to be cited. We swapped the order of some chat messages for clarity and conversational flow (indicated with extra timestamps), and in some cases combined logs where the conversation switched channels.

Color key:

Chat by Richard and Eliezer	Other chat	Google Doc content	Inline comments
-----------------------------	------------	--------------------	-----------------

0. Prefatory comments

[Yudkowsky][8:32] (Nov. 6 follow-up comment)

(At Rob's request I'll try to keep this brief, but this was an experimental format and some issues cropped up that seem large enough to deserve notes.)

Especially when coming in to the early parts of this dialogue, I had some backed-up hypotheses about "What might be the main sticking point? and how can I address that?" which from the standpoint of a pure dialogue might seem to be causing me to go on digressions, relative to if I was just trying to answer Richard's own questions. On reading the dialogue, I notice that this looks evasive or like point-missing, like I'm weirdly not just directly answering Richard's questions.

Often the questions are answered later, or at least I think they are, though it may not be in the first segment of the dialogue. But the larger phenomenon is that I came in with some things I wanted to say, and Richard came in asking questions, and there was a minor accidental mismatch there. It would have looked better if we'd both stated positions first without question marks, say, or if I'd just confined myself to answering questions from Richard. (This is not a huge catastrophe, but it's something for the reader to keep in mind as a minor hiccup that showed up in the early parts of experimenting with this new format.)

[Yudkowsky][8:32] (Nov. 6 follow-up comment)

(Prompted by some later stumbles in attempts to summarize this dialogue. Summaries seem plausibly a major mode of propagation for a sprawling dialogue like this, and the following request seems like it needs to be very prominent to work - embedded requests later on didn't work.)

Please don't summarize this dialogue by saying, "and so Eliezer's MAIN idea is that" or "and then Eliezer thinks THE KEY POINT is that" or "the PRIMARY argument is that" etcetera. From my perspective, everybody comes in with a different set of sticking points versus things they see as obvious, and the conversation I have changes drastically depending on that. In the old days this used to be the Orthogonality Thesis, Instrumental Convergence, and superintelligence being a possible thing at all; today most OpenPhil-adjacent folks have other sticking points instead.

Please transform:

- "Eliezer's main reply is..." -> "Eliezer replied that..."
- "Eliezer thinks the key point is..." -> "Eliezer's point in response was..."
- "Eliezer thinks a major issue is..." -> "Eliezer replied that one issue is..."
- "Eliezer's primary argument against this is..." -> "Eliezer tried the counterargument that..."
- "Eliezer's main scenario for this is..." -> "In a conversation in September of 2021, Eliezer sketched a hypothetical where..."

Note also that the transformed statements say what you *observed*, whereas the untransformed statements are (often incorrect) *inferences* about my latent state of mind.

(Though "distinguishing relatively unreliable inference from more reliable observation" is not necessarily *the key idea here or the one big reason* I'm asking for this. That's just one point I tried making - one argument that I hope might help drive home the larger thesis.)

1. September 5 conversation

1.1. Deep vs. shallow problem-solving patterns

[Ngo][11:00]

Hi all! Looking forward to the discussion.

[Yudkowsky][11:01]

Hi and welcome all. My name is Eliezer and I think alignment is really actually quite extremely difficult. Some people seem to not think this! It's an important issue so ought to be resolved somehow, which we can hopefully fully do today. (I will however want to take a break after the first 90 minutes, if it goes that far and if Ngo is in sleep-cycle shape to continue past that.)

[Ngo][11:02]

A break in 90 minutes or so sounds good.

Here's one way to kick things off: I agree that humans trying to align arbitrarily capable AIs seems very difficult. One reason that I'm more optimistic (or at least, not confident that we'll have to face the full very difficult version of the problem) is that at a certain point AIs will be doing most of the work.

When you talk about alignment being difficult, what types of AIs are you thinking about aligning?

[Yudkowsky][11:04]

On my model of the Other Person, a lot of times when somebody thinks alignment shouldn't be that hard, they think there's some particular thing you can do to align an AGI, which isn't that hard, and their model is missing one of the foundational difficulties for why you can't do (easily or at all) one step of their procedure. So one of my own conversational processes might be to poke around looking for a step that the other person doesn't realize is hard. That said, I'll try to directly answer your own question first.

[Ngo][11:07]

I don't think I'm confident that there's any particular thing you can do to align an AGI. Instead I feel fairly uncertain over a broad range of possibilities for how hard the problem turns out to be.

And on some of the most important variables, it seems like evidence from the last decade pushes towards updating that the problem will be easier.

[Yudkowsky][11:09]

I think that after AGI becomes possible at all and then possible to scale to dangerously superhuman levels, there will be, in the best-case scenario where a lot of other social difficulties got resolved, a 3-month to 2-year period where only a very few actors have AGI, meaning that it was socially possible for those few actors to decide to *not* just scale it to where it automatically destroys the world.

During this step, if humanity is to survive, somebody has to perform some feat that causes the world to *not* be destroyed in 3 months or 2 years when too many actors have access to AGI code that will destroy the world if its intelligence dial is turned up. This requires that the first actor or actors to build AGI, be able to do *something* with that AGI which prevents the world from being destroyed; if it didn't require superintelligence, we could go do that thing right now, but no such human-doable act apparently exists so far as I can tell.

So we want the least dangerous, most easily aligned thing-to-do-with-an-AGI, but it does have to be a pretty powerful act to prevent the automatic destruction of Earth after 3 months or 2 years. It has to "flip the gameboard" rather than letting the suicidal game play out. We need to align the AGI that performs this pivotal act, to perform that pivotal act without killing everybody.

Parenthetically, no act powerful enough and gameboard-flipping enough to qualify is inside the Overton Window of politics, or possibly even of effective altruism, which presents a separate social problem. I usually dodge around this problem by picking an exemplar act which is powerful enough to actually flip the gameboard, but not the most alignable act because it would require way too many aligned details: Build self-replicating open-air nanosystems and use them (only) to melt all GPUs.

Since any such nanosystems would have to operate in the full open world containing lots of complicated details, this would require tons and tons of alignment work, is not the pivotal act easiest to align, and we should do some other thing instead. But the other thing I have in mind is also outside the Overton Window, just like this is. So I use "melt all GPUs" to talk about the requisite power level and the Overton Window problem level, both of which seem around the right levels to me, but the actual thing I have in mind is more alignable; and this way, I can reply to anyone who says "How dare you?!" by saying "Don't worry, I don't actually plan on doing that."

[Ngo][11:14]

One way that we could take this discussion is by discussing the pivotal act "make progress on the alignment problem faster than humans can".

[Yudkowsky][11:15]

This sounds to me like it requires extreme levels of alignment and operating in extremely dangerous regimes, such that, if you could do that, it would seem much more sensible to do some other pivotal act first, using a lower level of alignment tech.

[Ngo][11:16]

Okay, this seems like a crux on my end.

[Yudkowsky][11:16]

In particular, I would hope that - in unlikely cases where we survive at all - we were able to survive by operating a superintelligence only in the lethally dangerous, but still less dangerous, regime of "engineering nanosystems".

Whereas "solve alignment for us" seems to require operating in the even more dangerous regimes of "write AI code for us" and "model human psychology in tremendous detail".

[Ngo][11:17]

What makes these regimes so dangerous? Is it that it's very hard for humans to exercise oversight?

One thing that makes these regimes seem less dangerous to me is that they're broadly in the domain of "solving intellectual problems" rather than "achieving outcomes in the world".

[Yudkowsky][11:19][11:21]

Every AI output *effectuates* outcomes in the world. If you have a powerful unaligned mind hooked up to outputs that can start causal chains that effectuate dangerous things, it doesn't matter whether the comments on the code say "intellectual problems" or not.

The danger of "solving an intellectual problem" is when it requires a powerful mind to think about domains that, when solved, render very cognitively accessible strategies that can do dangerous things.

I expect the first alignment solution you can actually deploy in real life, in the unlikely event we get a solution at all, looks like 98% "don't think about all these topics that we do not absolutely need and are adjacent to

the capability to easily invent very dangerous outputs" and 2% "actually think about this dangerous topic but please don't come up with a strategy inside it that kills us".

[Ngo][11:21][11:22]

Let me try and be more precise about the distinction. It seems to me that systems which have been primarily trained to make predictions about the world would by default lack a lot of the cognitive machinery which humans use to take actions which pursue our goals.

Perhaps another way of phrasing my point is something like: it doesn't seem implausible to me that we build AIs that are significantly more intelligent (in the sense of being able to understand the world) than humans, but significantly less agentic.

Is this a crux for you?

(obviously "agentic" is quite underspecified here, so maybe it'd be useful to dig into that first)

[Yudkowsky][11:27][11:33]

I would certainly have learned very new and very exciting facts about intelligence, facts which indeed contradict my present model of how intelligences liable to be discovered by present research paradigms work, if you showed me... how can I put this in a properly general way... that problems I thought were about searching for states that get fed into a result function and then a result-scoring function, such that the input gets an output with a high score, were in fact not about search problems like that. I have sometimes given more specific names to this problem setup, but I think people have become confused by the terms I usually use, which is why I'm dancing around them.

In particular, just as I have a model of the Other Person's Beliefs in which they think alignment is easy because they don't know about difficulties I see as very deep and fundamental and hard to avoid, I also have a model in which people think "why not just build an AI which does X but not Y?" because they don't realize what X and Y have in common, which is something that draws deeply on having deep models of intelligence. And it is hard to convey this deep theoretical grasp.

But you can also see powerful practical hints that these things are much more correlated than, eg, Robin Hanson was imagining during the [FOOM debate](#), because Robin did not think something like GPT-3 should exist; Robin thought you should need to train lots of specific domains that didn't generalize. I argued then with Robin that it was something of a hint that humans had visual cortex and cerebellar cortex but not Car Design Cortex, in order to design cars. Then in real life, it proved that reality was far to the Eliezer side of Eliezer on the [Eliezer-Robin axis](#), and things like GPT-3 were built with *less* architectural complexity and generalized *more*

than I was arguing to Robin that complex architectures should generalize over domains.

The metaphor I sometimes use is that it is very hard to build a system that drives cars painted red, but is not at all adjacent to a system that could, with a few alterations, prove to be very good at driving a car painted blue. The "drive a red car" problem and the "drive a blue car" problem have too much in common. You can maybe ask, "Align a system so that it has the capability to drive red cars, but refuses to drive blue cars." You can't make a system that is very good at driving red-painted cars, but lacks the basic capability to drive blue-painted cars because you never trained it on that. The patterns found by gradient descent, by genetic algorithms, or by other plausible methods of optimization, for driving red cars, would be patterns very close to the ones needed to drive blue cars. When you optimize for red cars you get the blue car *capability* whether you like it or not.

[Ngo][11:32]

Does your model of intelligence rule out building AIs which make dramatic progress in mathematics without killing us all?

[Yudkowsky][11:34][11:39]

If it were possible to perform some pivotal act that saved the world with an AI that just made progress on proving mathematical theorems, without, eg, needing to explain those theorems to humans, I'd be *extremely* interested in that as a potential pivotal act. We wouldn't be out of the woods, and I wouldn't actually know how to build an AI like that without killing everybody, but it would immediately trump everything else as the obvious line of research to pursue.

Parenthetically, there is very very little which my model of intelligence *rules out*. I think we all die because we cannot do certain dangerous things correctly, *on the very first try in the dangerous regimes where one mistake kills you*, and do them *before* proliferation of much easier technologies kills us. If you have the Textbook From 100 Years In The Future that gives the simple robust solutions for everything, that actually work, you can write a superintelligence that thinks $2 + 2 = 5$ because the Textbook gives the methods for doing that which are simple and actually work in practice in real life.

(The Textbook has the equivalent of "use ReLUs instead of sigmoids" everywhere, and avoids all the clever-sounding things that will work at subhuman levels and blow up when you run them at superintelligent levels.)

[Ngo][11:36][11:40]

Hmm, so suppose we train an AI to prove mathematical theorems when given them, perhaps via some sort of adversarial setter-solver training process.

By default I have the intuition that this AI could become extremely good at proving theorems - far beyond human level - without having goals about real-world outcomes.

It seems to me that in your model of intelligence, being able to do tasks like mathematics is closely coupled with trying to achieve real-world outcomes. But I'd actually take GPT-3 as some evidence against this position (although still evidence in favour of your position over Hanson's), since it seems able to do a bunch of reasoning tasks while still not being very agentic.

There's some alternative world where we weren't able to train language models to do reasoning tasks without first training them to perform tasks in complex RL environments, and in that world I'd be significantly less optimistic.

[Yudkowsky][11:41]

I put to you that there is a predictable bias in your estimates, where you don't know about the Deep Stuff that is required to prove theorems, so you imagine that certain cognitive capabilities are more disjoint than they actually are. If you knew about the things that humans are using to reuse their reasoning about chipped handaxes and other humans, to prove math theorems, you would see it as more plausible that proving math theorems would generalize to chipping handaxes and manipulating humans.

GPT-3 is a... complicated story, on my view of it and intelligence. We're looking at an interaction between tons and tons of memorized shallow patterns. GPT-3 is very unlike the way that natural selection built humans.

[Ngo][11:44]

I agree with that last point. But this is also one of the reasons that I previously claimed that AIs could be more intelligent than humans while being less agentic, because there are systematic differences between the way in which natural selection built humans, and the way in which we'll train AGIs.

[Yudkowsky][11:45]

My current suspicion is that Stack More Layers alone is not going to take us to GPT-6 which is a true AGI; and this is because of the way that GPT-3 is, in your own terminology, "not agentic", and which is, in my terminology, not having gradient descent on GPT-3 run across sufficiently deep problem-solving patterns.

[Ngo][11:46]

Okay, that helps me understand your position better.

So here's one important difference between humans and neural networks: humans face the genomic bottleneck which means that each individual has to rederive all the knowledge about the world that their parents already had. If this genetic bottleneck hadn't been so tight, then individual humans would have been significantly less capable of performing novel tasks.

[Yudkowsky][11:50]

I agree.

[Ngo][11:50]

In my terminology, this is a reason that humans are "more agentic" than we otherwise would have been.

[Yudkowsky][11:50]

This seems indisputable.

[Ngo][11:51]

Another important difference: humans were trained in environments where we had to run around surviving all day, rather than solving maths problems etc.

[Yudkowsky][11:51]

I continue to nod.

[Ngo][11:52]

Supposing I agree that reaching a certain level of intelligence will require AIs with the "deep problem-solving patterns" you talk about, which lead AIs to try to achieve real-world goals. It still seems to me that there's likely a lot of space between that level of intelligence, and human intelligence.

And if that's the case, then we could build AIs which help us solve the alignment problem before we build AIs which instantiate sufficiently deep problem-solving patterns that they decide to take over the world.

Nor does it seem like the reason *humans* want to take over the world is because of a deep fact about our intelligence. It seems to me that

humans want to take over the world mainly because that's very similar to things we evolved to do (like taking over our tribe).

[Yudkowsky][11:57]

So here's the part that I agree with: If there were one theorem only mildly far out of human reach, like proving the ABC Conjecture (if you think it hasn't already been proven), and providing a machine-readable proof of this theorem would immediately save the world - say, aliens will give us an aligned superintelligence, as soon as we provide them with this machine-readable proof - then there would exist a plausible though not certain road to saving the world, which would be to try to build a *shallow* mind that proved the ABC Conjecture by memorizing tons of relatively shallow patterns for mathematical proofs learned through self-play; without that system ever abstracting math as deeply as humans do, but the sheer width of memory and sheer depth of search sufficing to do the job. I am not sure, to be clear, that this would work. But my model of intelligence does not rule it out.

[Ngo][11:58]

(I'm actually thinking of a mind which understands maths more deeply than humans - but perhaps only understands maths, or perhaps also a range of other sciences better than humans.)

[Yudkowsky][12:00]

Parts I disagree with: That "help us solve alignment" bears any significant overlap with "provide us a machine-readable proof of the ABC Conjecture without thinking too deeply about it". That humans want to take over the world only because it resembles things we evolved to do.

[Ngo][12:01]

I definitely agree that humans don't *only* want to take over the world because it resembles things we evolved to do.

[Yudkowsky][12:02]

Alas, eliminating 5 reasons why something would go wrong doesn't help much if there's 2 remaining reasons something would go wrong that are much harder to eliminate!

[Ngo][12:02]

But if we imagine having a human-level intelligence which *hadn't* evolved primarily to do things that reasonably closely resembled taking over the

world, then I expect that we could ask that intelligence questions in a fairly safe way.

And that's also true for an intelligence that is noticeably above human level.

So one question is: how far above human level could we get before a system which has only been trained to do things like answer questions and understand the world will decide to take over the world?

[Yudkowsky][12:04]

I think this is one of the very rare cases where the intelligence difference between "village idiot" and "Einstein", which I'd usually see as very narrow, makes a structural difference! I think you can get some outputs from a village-idiot-level AGI, which got there by training on domains exclusively like math, and this will probably not destroy the world (*if* you were right about that, about what was going on inside). I have more concern about the Einstein level.

[Ngo][12:05]

Let's focus on the Einstein level then.

Human brains have been optimised very little for doing science.

This suggests that building an AI which is Einstein-level at doing science is significantly easier than building an AI which is Einstein-level at taking over the world (or other things which humans evolved to do).

[Yudkowsky][12:08]

I think there's a certain broad sense in which I agree with the literal truth of what you just said. You will systematically overestimate *how much* easier, or how far you can push the science part without getting the taking-over-the-world part, for as long as your model is ignorant of what they have in common.

[Ngo][12:08]

Maybe this is a good time to dig into the details of what they have in common, then.

[Yudkowsky][12:09][12:11]][12:13]

I feel like I haven't had much luck with trying to explain that on previous occasions. Not to you, to others too.

There are shallow topics like why p-zombies can't be real and how quantum mechanics works and why science ought to be using likelihood functions instead of p-values, and I can *barely* explain those to *some* people, but then there are some things that are apparently much harder to explain than that and which defeat my abilities as an explainer.

That's why I've been trying to point out that, even if you don't know the specifics, there's an estimation bias that you can realize should exist in principle.

Of course, I also haven't had much luck in saying to people, "Well, even if you don't know the truth about X that would let you see Y, can you not see by abstract reasoning that knowing *any* truth about X would predictably cause you to update in the direction of Y" - people don't seem to actually internalize that much either. Not you, other discussions.

[Ngo][12:10][12:11][12:13]

Makes sense. Are there ways that I could try to make this easier? E.g. I could do my best to explain what I think your position is.

Given what you've said I'm not optimistic about this helping much.

But insofar as this is the key set of intuitions which has been informing your responses, it seems worth a shot.

Another approach would be to focus on our predictions for how AI capabilities will play out over the next few years.

I take your point about my estimation bias. To me it feels like there's also a bias going the other way, which is that as long as we don't know the mechanisms by which different human capabilities work, we'll tend to lump them together as one thing.

[Yudkowsky][12:14]

Yup. If you didn't know about visual cortex and auditory cortex, or about eyes and ears, you would assume much more that any sentience ought to both see and hear.

[Ngo][12:16]

So then my position is something like: human pursuit of goals is driven by emotions and reward signals which are deeply evolutionarily ingrained, and without those we'd be much safer but not that much worse at pattern recognition.

[Yudkowsky][12:17]

If there's a pivotal act you can get just by supreme acts of pattern recognition, that's right up there with "pivotal act composed solely of math" for things that would obviously instantly become the prime direction of research.

[Ngo][12:18]

To me it seems like maths is *much more* about pattern recognition than, say, being a CEO. Being a CEO requires coherence over long periods of time; long-term memory; motivation; metacognition; etc.

[Yudkowsky][12:18][12:23]

(One occasionally-argued line of research can be summarized from a certain standpoint as "how about a pivotal act composed entirely of predicting text" and to this my reply is "you're trying to get fully general AGI capabilities by predicting text that is *about* deep / 'agentic' reasoning, and that doesn't actually help".)

Human math is very much about goals. People want to prove subtheorems on the way to proving theorems. We might be able to make a *different* kind of mathematician that works more like GPT-3 in the dangerously inscrutable parts that are all noninspectable vectors of floating-point numbers, but even there you'd need some Alpha-Zero-like outer framework to supply the direction of search.

That outer framework might be able to be powerful enough without being reflective, though. So it would plausibly be *much easier* to build a mathematician that was capable of superhuman formal theorem-proving but not agentic. The reality of the world might tell us "lolnope" but my model of intelligence doesn't mandate that. That's why, if you gave me a pivotal act composed entirely of "output a machine-readable proof of this theorem and the world is saved", I would pivot there! It actually does seem like it would be a lot easier!

[Ngo][12:21][12:25]

Okay, so if I attempt to rephrase your argument:

Your position: There's a set of fundamental similarities between tasks like doing maths, doing alignment research, and taking over the world. In all of these cases, agents based on techniques similar to modern ML which are very good at them will need to make use of deep problem-solving patterns which include goal-oriented reasoning. So while it's possible to beat humans at some of these tasks without those core competencies, people usually overestimate the extent to which that's possible.

[Yudkowsky][12:25]

Remember, a lot of my concern is about what happens *first*, especially if it happens soon enough that future AGI bears any resemblance whatsoever to modern ML; not about what can be done in principle.

[Soares][12:26]

(Note: it's been 85 min, and we're planning to take a break at 90min, so this seems like a good point for a little bit more clarifying back-and-forth on Richard's summary before a break.)

[Ngo][12:26]

I'll edit to say "plausible for ML techniques"?

(and "extent to which that's plausible")

[Yudkowsky][12:28]

I think that obvious-to-me future outgrowths of modern ML paradigms are *extremely* liable to, if they can learn how to do sufficiently superhuman X, generalize to taking over the world. How fast this happens does depend on X. It would plausibly happen relatively slower (at higher levels) with theorem-proving as the X, and with architectures that carefully stuck to gradient-descent-memorization over shallow network architectures to do a pattern-recognition part with search factored out (sort of, this is not generally safe, this is not a general formula for safe things!); rather than imposing anything like the genetic bottleneck you validly pointed out as a reason why humans generalize. Profitable X, and all X I can think of that would actually save the world, seem much more problematic.

[Ngo][12:30]

Okay, happy to take a break here.

[Soares][12:30]

Great timing!

[Ngo][12:30]

We can do a bit of meta discussion afterwards; my initial instinct is to push on the question of how similar Eliezer thinks alignment research is to theorem-proving.

[Yudkowsky][12:30]

Yup. This is my lunch break (actually my first-food-of-day break on a 600-calorie diet) so I can be back in 45min if you're still up for that.

[Ngo][12:31]

Sure.

Also, if any of the spectators are reading in real time, and have suggestions or comments, I'd be interested in hearing them.

[Yudkowsky][12:31]

I'm also cheerful about spectators posting suggestions or comments during the break.

[Soares][12:32]

Sounds good. I declare us on a break for 45min, at which point we'll reconvene (for another 90, by default).

Floor's open to suggestions & commentary.

1.2. Requirements for science

[Yudkowsky][12:50]

I seem to be done early if people (mainly Richard) want to resume in 10min (30m break)

[Ngo][12:51]

Yepp, happy to do so

[Soares][12:57]

Some quick commentary from me:

- It seems to me like we're exploring a crux in the vicinity of "should we expect that systems capable of executing a pivotal act would, by default in lieu of significant technical alignment effort, be using their outputs to optimize the future".

- I'm curious whether you two agree that this is a crux (but plz don't get side-tracked answering me).
- The general discussion seems to be going well to me.
 - In particular, huzzah for careful and articulate efforts to zero in on cruxes.

[Ngo][13:00]

I think that's a crux for the specific pivotal act of "doing better alignment research", and maybe some other pivotal acts, but not all (or necessarily most) of them.

[Yudkowsky][13:01]

I should also say out loud that I've been working a bit with Ajeya on making an attempt to convey the intuitions behind there being deep patterns that generalize and are liable to be learned, which covered a bunch of ground, taught me how much ground there was, and made me relatively more reluctant to try to re-cover the same ground in this modality.

[Ngo][13:02]

Going forward, a couple of things I'd like to ask Eliezer about:

- In what ways are the tasks that are most useful for alignment similar or different to proving mathematical theorems (which we agreed might generalise relatively slowly to taking over the world)?
- What are the deep problem-solving patterns underlying these tasks?
- Can you summarise my position?

I was going to say that I was most optimistic about #2 in order to get these ideas into a public format

But if that's going to happen anyway based on Ajeya's work, then that seems less important

[Yudkowsky][13:03]

I could still try briefly and see what happens.

[Ngo][13:03]

That seems valuable to me, if you're up for it.

At the same time, I'll try to summarise some of my own intuitions about intelligence which I expect to be relevant.

[Yudkowsky][13:04]

I'm not sure I could summarize your position in a non-straw way. To me there's a huge visible distance between "solve alignment for us" and "output machine-readable proofs of theorems" where I can't give a good account of why you think talking about the latter would tell us much about the former. I don't know what other pivotal act you think might be easier.

[Ngo][13:06]

I see. I was considering "solving scientific problems" as an alternative to "proving theorems", with alignment being one (particularly hard) example of a scientific problem.

But decided to start by discussing theorem-proving since it seemed like a clearer-cut case.

[Yudkowsky][13:07]

Can you predict in advance why Eliezer thinks "solving scientific problems" is significantly thornier? (Where alignment is like totally not "a particularly hard example of a scientific problem" except in the sense that it has science in it at all; which is maybe the real crux; but also a more difficult issue.)

[Ngo][13:09]

Based on some of your earlier comments, I'm currently predicting that you think the step where the solutions need to be legible to and judged by humans makes science much thornier than theorem-proving, where the solutions are machine-checkable.

[Yudkowsky][13:10]

That's one factor. Should I state the other big one or would you rather try to state it first?

[Ngo][13:10]

Requiring a lot of real-world knowledge for science?

If it's not that, go ahead and say it.

[Yudkowsky][13:11]

That's one way of stating it. The way I'd put it is that it's about making up hypotheses about the real world.

Like, the real world is then a thing that the AI is modeling, at all.

Factor 3: On many interpretations of doing science, you would furthermore need to think up experiments. That's planning, value-of-information, search for an experimental setup whose consequences distinguish between hypotheses (meaning you're now searching for initial setups that have particular causal consequences).

[Ngo][13:12]

To me "modelling the real world" is a very continuous variable. At one end you have physics equations that are barely separable from maths problems, at the other end you have humans running around in physical bodies.

To me it seems plausible that we could build an agent which solves scientific problems but has very little self-awareness (in the sense of knowing that it's an AI, knowing that it's being trained, etc).

I expect that your response to this is that modelling oneself is part of the deep problem-solving patterns which AGIs are very likely to have.

[Yudkowsky][13:15]

There's a problem of *inferring the causes of sensory experience* in cognition-that-does-science. (Which, in fact, also appears in the way that humans do math, and is possibly inextricable from math in general; but this is an example of the sort of deep model that says "Whoops I guess you get science from math after all", not a thing that makes science less dangerous because it's more like just math.)

You can build an AI that only ever drives red cars, and which, at no point in the process of driving a red car, ever needs to drive a blue car in order to drive a red car. That doesn't mean its red-car-driving capabilities won't be extremely close to blue-car-driving capabilities if at any point the internal cognition happens to get pointed towards driving a blue car.

The fact that there's a deep car-driving pattern which is the same across red cars and blue cars doesn't mean that the AI has ever driven a blue car, per se, or that it has to drive blue cars to drive red cars. But if blue cars are fire, you sure are playing with that fire.

[Ngo][13:18]

To me, "sensory experience" as in "the video and audio coming in from this body that I'm piloting" and "sensory experience" as in "a file containing the most recent results of the large hadron collider" are very very different.

(I'm not saying we could train an AI scientist just from the latter - but plausibly from data that's closer to the latter than the former)

[Yudkowsky][13:19]

So there's separate questions about "does an AGI *inseparably need* to model itself inside the world to do science" and "did we build something that would be very close to modeling itself, and could easily stumble across that by accident somewhere in the inscrutable floating-point numbers, especially if that was even slightly useful for solving the outer problems".

[Ngo][13:19]

Hmm, I see

[Yudkowsky][13:20][13:21][13:21]

If you're trying to build an AI that literally does science only to observations collected without the AI having had a causal impact on those observations, that's legitimately "more dangerous than math but maybe less dangerous than active science".

You might still stumble across an active scientist because it was a simple internal solution to something, but the outer problem would be legitimately stripped of an important structural property the same way that pure math not describing Earthly objects is stripped of important structural properties.

And of course my reaction again is, "There is no pivotal act which uses only that cognitive capability."

[Ngo][13:20][13:21][13:26]

I guess that my (fairly strong) prior here is that something like self-modelling, which is very deeply built into basically every organism, is a very hard thing for an AI to stumble across by accident without significant optimisation pressure in that direction.

But I'm not sure how to argue this except by digging into your views on what the deep problem-solving patterns are. So if you're still willing to briefly try and explain those, that'd be useful to me.

"Causal impact" again seems like a very continuous variable - it seems like the *amount* of causal impact you need to do good science is much less than the amount which is needed to, say, be a CEO.

[Yudkowsky][13:26]

The amount doesn't seem like the key thing, nearly so much as what underlying facilities you need to do whatever amount of it you need.

[Ngo][13:27]

Agreed.

[Yudkowsky][13:27]

If you go back to the 16th century and ask for just one mRNA vaccine, that's not much of a difference from asking for a million hundred of them.

[Ngo][13:28]

Right, so the additional premise which I'm using here is that the ability to reason about causally impacting the world in order to achieve goals is something that you can have a little bit of.

Or a lot of, and that the difference between these might come down to the training data used.

Which at this point I don't expect you to agree with.

[Yudkowsky][13:29]

If you have reduced a pivotal act to "look over the data from this hadron collider you neither built nor ran yourself", that really is a structural step down from "do science" or "build a nanomachine". But I can't see any pivotal acts like that, so is that question much of a crux?

If there's intermediate steps they might be described in my native language like "reason about causal impacts across only this one preprogrammed domain which you didn't learn in a general way, in only this part of the cognitive architecture that is separable from the rest of the cognitive architecture".

[Ngo][13:31]

Perhaps another way of phrasing this intermediate step is that the agent has a shallow understanding of how to induce causal impacts.

[Yudkowsky][13:31]

What is "shallow" to you?

[Ngo][13:31]

In a similar way to how you claim that GPT-3 has a shallow understanding of language.

[Yudkowsky][13:32]

So it's memorized a ton of shallow causal-impact-inducing patterns from a large dataset, and this can be verified by, for example, presenting it with an example mildly outside the dataset and watching it fail, which we think will confirm our hypothesis that it didn't learn any deep ways of solving that dataset.

[Ngo][13:33]

Roughly speaking, yes.

[Yudkowsky][13:34]

Eg, it wouldn't surprise us at all if GPT-4 had learned to predict "27 * 18" but not "what is the area of a rectangle 27 meters by 18 meters"... is what I'd like to say, but Codex sure did demonstrate those two were kinda awfully proximal.

[Ngo][13:34]

Here's one way we could flesh this out. Imagine an agent that loses coherence quickly when it's trying to act in the world.

So for example, we've trained it to do scientific experiments over a period of a few hours or days

And then it's very good at understanding the experimental data and extracting patterns from it

But upon running it for a week or a month, it loses coherence in a similar way to how GPT-3 loses coherence - e.g. it forgets what it's doing.

My story for why this might happen is something like: there is a specific skill of having long-term memory, and we never trained our agent to have this skill, and so it has not acquired that skill (even though it can reason in very general and powerful ways in the short term).

This feels similar to the argument I was making before about how an agent might lack self-awareness, if we haven't trained it specifically to have that.

[Yudkowsky][13:39]

There's a set of obvious-to-me tactics for doing a pivotal act with minimal danger, which I do not think collectively make the problem safe, and one

of these sets of tactics is indeed "Put a limit on the 'attention window' or some other internal parameter, ramp it up slowly, don't ramp it any higher than you needed to solve the problem."

[Ngo][13:41]

You could indeed do this manually, but my expectation is that you could also do this automatically, by training agents in environments where they don't benefit from having long attention spans.

[Yudkowsky][13:42]

(Any time one imagines a specific tactic of this kind, if one has the [security mindset](#), one can also imagine all sorts of ways it might go wrong; for example, an attention window can be defeated if there's any aspect of the attended data or the internal state that ended up depending on past events in a way that leaked info about them. But, depending on how much superintelligence you were throwing around elsewhere, you could maybe get away with that, some of the time.)

[Ngo][13:43]

And that if you put agents in environments where they answer questions but don't interact much with the physical world, then there will be many different traits which are necessary for achieving goals in the real world which they will lack, because there was little advantage to the optimiser of building those traits in.

[Yudkowsky][13:43]

I'll observe that TransformerXL built an attention window that generalized, trained it on I think 380 tokens or something like that, and then found that it generalized to 4000 tokens or something like that.

[Ngo][13:43]

Yeah, an order of magnitude of generalisation is not surprising to me.

[Yudkowsky][13:44]

Having observed one order of magnitude, I would personally not be surprised by two orders of magnitude either, after seeing that.

[Ngo][13:45]

I'd be a little surprised, but I assume it would happen eventually.

1.3. Capability dials

[Yudkowsky][13:46]

I have a sense that this is all circling back to the question, "But what is it we *do* with the intelligence thus weakened?" If you can save the world using a rock, I can build you a very safe rock.

[Ngo][13:46]

Right.

So far I've said "alignment research", but I haven't been very specific about it.

I guess some context here is that I expect that the first things we do with intelligence similar to this is create great wealth, produce a bunch of useful scientific advances, etc.

And that we'll be in a world where people take the prospect of AGI much more seriously

[Yudkowsky][13:48]

I mostly expect - albeit with some chance that reality says "So what?" to me and surprises me, because it is not as solidly determined as some other things - that we do not hang around very long in the "weirdly ~human AGI" phase before we get into the "if you crank up this AGI it destroys the world" phase. Less than 5 years, say, to put numbers on things.

It would not surprise me in the least if the world ends before self-driving cars are sold on the mass market. On some quite plausible scenarios which I think have >50% of my probability mass at the moment, research AGI companies would be able to produce prototype car-driving AIs if they spent time on that, given the near-world-ending tech level; but there will be Many Very Serious Questions about this relatively new unproven advancement in machine learning being turned loose on the roads. And their AGI tech will gain the property "can be turned up to destroy the world" before Earth gains the property "you're allowed to sell self-driving cars on the mass market" because there just won't be much time.

[Ngo][13:52]

Then I expect that another thing we do with this is produce a very large amount of data which rewards AIs for following human instructions.

[Yudkowsky][13:52]

On other scenarios, of course, self-driving becomes possible by limited AI well before things start to break (further) on AGI. And on some scenarios, the way you got to AGI was via some breakthrough that is already scaling pretty fast, so by the time you can use the tech to get self-driving cars, that tech already ends the world if you turn up the dial, or that event follows very swiftly.

[Ngo][13:53]

When you talk about "cranking up the AGI", what do you mean?

Using more compute on the same data?

[Yudkowsky][13:53]

Running it with larger bounds on the for loops, over more GPUs, to be concrete about it.

[Ngo][13:53]

In a RL setting, or a supervised, or unsupervised learning setting?

Also: can you elaborate on the for loops?

[Yudkowsky][13:56]

I do not quite think that gradient descent on Stack More Layers alone - as used by OpenAI for GPT-3, say, and as opposed to Deepmind which builds more complex artifacts like Mu Zero or AlphaFold 2 - is liable to be the first path taken to AGI. I am reluctant to speculate more in print about clever ways to AGI, and I think any clever person out there will, if they are really clever and not just a fancier kind of stupid, not talk either about what they think is missing from Stack More Layers or how you would really get AGI. That said, the way that you cannot just run GPT-3 at a greater search depth, the way you can run Mu Zero at a greater search depth, is part of why I think that AGI is not likely to look exactly like GPT-3; the thing that kills us is likely to be a thing that can get more dangerous when you turn up a dial on it, not a thing that intrinsically has no dials that can make it more dangerous.

1.4. Consequentialist goals vs. deontologist goals

[Ngo][13:59]

Hmm, okay. Let's take a quick step back and think about what would be useful for the last half hour.

I want to flag that my intuitions about pivotal acts are not very specific; I'm quite uncertain about how the geopolitics of that situation would work, as well as the timeframe between somewhere-near-human-level AGI and existential risk AGI.

So we could talk more about this, but I expect there'd be a lot of me saying "well we can't rule out that X happens", which is perhaps not the most productive mode of discourse.

A second option is digging into your intuitions about how cognition works.

[Yudkowsky][14:03]

Well, obviously, in the limit of alignment not being accessible to our civilization, and my successfully building a model weaker than reality which nonetheless correctly rules out alignment being accessible to our civilization, I could spend the rest of my short remaining lifetime arguing with people whose models are weak enough to induce some area of ignorance where for all they know you could align a thing. But that is predictably how conversations go in possible worlds where the Earth is doomed; so somebody wiser on the meta-level, though also ignorant on the object-level, might prefer to ask: "Where do you think your knowledge, rather than your ignorance, says that alignment ought to be doable and you will be surprised if it is not?"

[Ngo][14:07]

That's a fair point. Although it seems like a structural property of the "pivotal act" framing, which builds in doom by default.

[Yudkowsky][14:08]

We could talk about that, if you think it's a crux. Though I'm also not thinking that this whole conversation gets done in a day, so maybe for publishability reasons we should try to focus more on one line of discussion?

But I do think that lots of people get their optimism by supposing that the world can be saved by doing less dangerous things with an AGI. So it's a big ol' crux of mine on priors.

[Ngo][14:09]

Agreed that one line of discussion is better; I'm happy to work within the pivotal act framing for current purposes.

A third option is that I make some claims about how cognition works, and we see how much you agree with them.

[Yudkowsky][14:12]

(Though it's something of a restatement, a reason I'm not going into "my intuitions about how cognition works" is that past experience has led me to believe that conveying this info in a form that the Other Mind will actually absorb and operate, is really quite hard and takes a long discussion, relative to my current abilities to Actually Explain things; it is the sort of thing that might take doing homework exercises to grasp how one structure is appearing in many places, as opposed to just being flatly told that to no avail, and I have not figured out the homework exercises.)

I'm cheerful about hearing your own claims about cognition and disagreeing with them.

[Ngo][14:12]

Great

Okay, so one claim is that something like deontology is a fairly natural way for minds to operate.

[Yudkowsky][14:14]

("If that were true," he thought at once, "bureaucracies and books of regulations would be a lot more efficient than they are in real life.")

[Ngo][14:14]

Hmm, although I think this was probably not a very useful phrasing, let me think about how to rephrase it.

Okay, so in [our earlier email discussion](#), we talked about the concept of "obedience".

To me it seems like it is just as plausible for a mind to have a concept like "obedience" as its rough goal, as a concept like maximising paperclips.

If we imagine training an agent on a large amount of data which pointed in the rough direction of rewarding obedience, for example, then I imagine that by default obedience would be a constraint of comparable strength to, say, the human survival instinct.

(Which is obviously not strong enough to stop humans doing a bunch of things that contradict it - but it's a pretty good starting point.)

[Yudkowsky][14:18]

Heh. You mean of comparable strength to the human instinct to explicitly maximize inclusive genetic fitness?

[Ngo][14:19]

Genetic fitness wasn't a concept that our ancestors were able to understand, so it makes sense that they weren't pointed directly towards it.

(And nor did they understand *how* to achieve it.)

[Yudkowsky][14:19]

Even in that paradigm, except insofar as you expect gradient descent to work very differently from gene-search optimization - which, admittedly, it does - when you optimize really hard on a thing, you get contextual correlates to it, not the thing you optimized on.

This is of course one of the Big Fundamental Problems that I expect in alignment.

[Ngo][14:20]

Right, so the main correlate that I've seen discussed is "do what would make the human give you a high rating, not what the human actually wants"

One thing I'm curious about is the extent to which you're concerned about this specific correlate, versus correlates in general.

[Yudkowsky][14:21]

That said, I also see basic structural reasons why paperclips would be much easier to train than "obedience", even if we could magically instill simple inner desires that perfectly reflected the simple outer algorithm we saw ourselves as running over many particular instances of a loss function.

[Ngo][14:22]

I'd be interested in hearing what those are.

[Yudkowsky][14:22]

well, first of all, why *is* a book of regulations so much more unwieldy than a hunter-gatherer?

if deontology is just as good as [consequentialism](#), y'know.

(do you want to try replying or should I just say?)

[Ngo][14:23]

Go ahead

I should probably clarify that I agree that you can't just replace consequentialism with deontology

The claim is more like: when it comes to high-level concepts, it's not clear to me why high-level consequentialist goals are more natural than high-level deontological goals.

[Yudkowsky][14:24]

I reply that reality is complicated, so when you pump a simple goal through complicated reality you get complicated behaviors required to achieve the goal. If you think of reality as a complicated function Input->Probability(Output), then even to get a simple Output or a simple partition on Output or a high expected score in a simple function over Output, you may need very complicated Input.

Humans don't trust each other. They imagine, "Well, if I just give this bureaucrat a goal, perhaps they won't reason honestly about what it takes to achieve that goal! Oh no! Therefore I will instead, being the trustworthy and accurate person that I am, reason myself about constraints and requirements on the bureaucrat's actions, such that, if the bureaucrat obeys these regulations, I expect the outcome of their action will be what I want."

But (compared to a general intelligence that observes and models complicated reality and does its own search to pick actions) an actually-effective book of regulations (implemented by some nonhuman mind with a large enough and perfect enough memory to memorize it) would tend to involve a (physically unmanageable) vast number of rules saying "if you observe this, do that" to follow all the crinkles of complicated reality as it can be inferred from observation.

[Ngo][14:28]

(Though it's something of a restatement, a reason I'm not going into "my intuitions about how cognition works" is that past experience has led me to believe that conveying this info in a form that the Other Mind will actually absorb and operate, is really quite hard and takes a long discussion, relative to my current abilities to Actually Explain things; it is the sort of thing that might take doing homework exercises to grasp how one structure is appearing in many places, as opposed to just being flatly told that to no avail, and I have not figured out the homework exercises.)

(As a side note: do you have a rough guess for when your work with Ajeya will be made public? If it's still a while away, I'm wondering whether it's still useful to have a rough outline of these intuitions even if it's in a form that very few people will internalise)

[Yudkowsky][14:30]

(As a side note: do you have a rough guess for when your work with Ajeya will be made public? If it's still a while away, I'm wondering whether it's still useful to have a rough outline of these intuitions even if it's in a form that very few people will internalise)

Plausibly useful, but not to be attempted today, I think?

[Ngo][14:30]

Agreed.

[Yudkowsky][14:30]

(We are now theoretically in overtime, which is okay for me, but for you it is 11:30pm (I think?) and so it is on you to call when to halt, now or later.)

[Ngo][14:32]

Yeah, it's 11.30 for me. I think probably best to halt here. I agree with all the things you just said about reality being complicated, and why consequentialism is therefore valuable. My "deontology" claim (which was, in its original formulation, far too general - apologies for that) was originally intended as a way of poking into your intuitions about which types of cognition are natural or unnatural, which I think is the topic we've been circling around for a while.

[Yudkowsky][14:33]

Yup, and a place to resume next time might be why I think "obedience" is unnatural compared to "paperclips" - though that is a thing that probably requires taking that stab at what underlies surface competencies.

[Ngo][14:34]

Right. I do think that even a vague gesture at that would be reasonably helpful (assuming that this doesn't already exist online?)

[Yudkowsky][14:34]

Not yet afaik, and I don't want to point you to Ajeya's stuff even if she were ok with that, because then this in-context conversation won't make sense to others.

[Ngo][14:35]

For my part I should think more about pivotal acts that I'd be willing to specifically defend.

In any case, thanks for the discussion 😊

Let me know if there's a particular time that suits you for a follow-up; otherwise we can sort it out later.

[Soares][14:37]

(y'all are doing all my jobs for me)

[Yudkowsky][14:37]

could try Tuesday at this same time - though I may be in worse shape for dietary reasons, still, seems worth trying.

[Soares][14:37]

(wfm)

[Ngo][14:39]

Tuesday not ideal, any others work?

[Yudkowsky][14:39]

Wednesday?

[Ngo][14:40]

Yes, Wednesday would be good

[Yudkowsky][14:40]

let's call it tentatively for that

[Soares][14:41]

Great! Thanks for the chats.

[Ngo][14:41]

Thanks both!

[Yudkowsky][14:41]

Thanks, Richard!

2. Follow-ups

2.1. Richard Ngo's summary

[Tallinn][0:35] (Sep. 6)

just caught up here & wanted to thank nate, eliezer and (especially) richard for doing this! it's great to see eliezer's model being probed so intensively. i've learned a few new things (such as the genetic bottleneck being plausibly a big factor in human cognition). FWIW, a minor comment re deontology (as that's fresh on my mind): in my view deontology is more about coordination than optimisation: deontological agents are more trustworthy, as they're much easier to reason about (in the same way how functional/declarative code is easier to reason about than imperative code). hence my steelman of bureaucracies (as well as social norms): humans just (correctly) prefer their fellow optimisers (including non-human optimisers) to be deontological for trust/coordination reasons, and are happy to pay the resulting competence tax.

[Ngo][3:10] (Sep. 8)

Thanks Jaan! I agree that greater trust is a good reason to want agents which are deontological at some high level.

I've attempted a summary of the key points so far; comments welcome:
[GDocs link]

[Ngo] (Sep. 8 Google Doc)

1st discussion

(Mostly summaries not quotations)

Eliezer, summarized by Richard: "To avoid catastrophe, whoever builds AGI first will have to a) align it to some extent, and b) decide not to scale it up beyond the point where their alignment techniques fail, and c) do some pivotal act that prevents others from scaling it up to that level. But ~~our alignment techniques will not be good enough~~ ~~our alignment techniques will be very far from adequate~~ on our current trajectory, our alignment techniques will be very far from adequate to create an AI that safely performs any such pivotal act."

[Yudkowsky][11:05] (Sep. 8 comment)

will not be good enough

Are not presently on course to be good enough, missing by not a little.
"Will not be good enough" is literally declaring for lying down and dying.

[Yudkowsky][16:03] (Sep. 9 comment)

will [be very far from adequate]

Same problem as the last time I commented. I am not making an unconditional prediction about future failure as would be implied by the word "will". Conditional on current courses of action or their near neighboring courses, we seem to be well over an order of magnitude away from surviving, unless a miracle occurs. It's still in the end a result of people doing what they seem to be doing, not an inevitability.

[Ngo][5:10] (Sep. 10 comment)

Ah, I see. Does adding "on our current trajectory" fix this?

[Yudkowsky][10:46] (Sep. 10 comment)

Yes.

[Ngo] (Sep. 8 Google Doc)

Richard, summarized by Richard: "Consider the pivotal act of 'make a breakthrough in alignment research'. It is likely that, before the point where AGIs are strongly superhuman at seeking power, they will already be strongly superhuman at understanding the world, and at performing narrower pivotal acts like alignment research which don't require as much agency (by which I roughly mean: large-scale motivations and the ability to pursue them over long timeframes)."

Eliezer, summarized by Richard: "There's a deep connection between solving intellectual problems and taking over the world - the former requires a powerful mind to think about domains that, when solved, render very cognitively accessible strategies that can do dangerous things. Even mathematical research is a goal-oriented task which involves identifying then pursuing instrumental subgoals - and if brains which evolved to hunt on the savannah can quickly learn to do mathematics, then it's also plausible that AIs trained to do mathematics could quickly learn a range of other skills. Since almost nobody understands the deep similarities in the cognition required for these different tasks, the distance between AIs that are able to perform fundamental scientific research, and dangerously agentic AGIs, is smaller than almost anybody expects."

[Yudkowsky][11:05] (Sep. 8 comment)

There's a deep connection between solving intellectual problems and taking over the world

There's a deep connection by default between chipping flint handaxes and taking over the world, if you happen to learn how to chip handaxes in a very general way. "Intellectual" problems aren't special in this way. And maybe you could avert the default, but that would take some work and you'd have to do it before easier default ML techniques destroyed the world.

[Ngo] (Sep. 8 Google Doc)

Richard, summarized by Richard: "Our lack of understanding about how intelligence works also makes it easy to assume that traits which co-occur in humans will also co-occur in future AIs. But human brains are badly-optimised for tasks like scientific research, and well-optimised for seeking power over the world, for reasons including a) evolving while embodied in a harsh environment; b) the genetic bottleneck; c) social environments which rewarded power-seeking. By contrast, training neural networks on tasks like mathematical or scientific research optimises them much less for seeking power. For example, GPT-3 has knowledge and reasoning capabilities but little agency, and loses coherence when run for longer timeframes."

[Tallinn][4:19] (Sep. 8 comment)

[well-optimised for] seeking power

male-female differences might be a datapoint here (annoying as it is to lean on pinker's point :))

[Yudkowsky][11:31] (Sep. 8 comment)

I don't think a female Eliezer Yudkowsky doesn't try to save / optimize / takeover the world. Men may do that for nonsmart reasons; smart men and women follow the same reasoning when they are smart enough. Eg Anna Salamon and many others.

[Ngo] (Sep. 8 Google Doc)

Eliezer, summarized by Richard: "Firstly, there's a big difference between most scientific research and the sort of pivotal act that we're talking about - you need to explain how AIs with a given skill can be used to actually prevent dangerous AGIs from being built. Secondly, insofar as GPT-3 has little agency, that's because it has memorised many shallow patterns in a way which won't directly scale up to general intelligence. Intelligence instead consists of deep problem-solving patterns which link understanding and agency at a fundamental level."

3. September 8 conversation

3.1. The Brazilian university anecdote

[Yudkowsky][11:00]

(I am here.)

[Ngo][11:01]

Me too.

[Soares][11:01]

Welcome back!

(I'll mostly stay out of the way again.)

[Ngo][11:02]

Cool. Eliezer, did you read the summary - and if so, do you roughly endorse it?

Also, I've been thinking about the best way to approach discussing your intuitions about cognition. My guess is that starting with the obedience vs paperclips thread is likely to be less useful than starting somewhere else - e.g. the description you gave near the beginning of the last discussion, about "searching for states that get fed into a result function and then a result-scoring function".

[Yudkowsky][11:06]

made a couple of comments about phrasings in the doc

So, from my perspective, there's this thing where... it's really quite hard to teach certain *general* points by talking at people, as opposed to more specific points. Like, they're trying to build a perpetual motion machine, and even if you can manage to argue them into believing their first design is wrong, they go looking for a new design, and the new design is complicated enough that they can no longer be convinced that they're wrong because they managed to make a more complicated error whose refutation they couldn't keep track of anymore.

Teaching people to see an underlying structure in a lot of places is a very hard thing to teach in this way. Richard Feynman [gave an example](#) of the mental motion in his story that ends "Look at the water!", where people learned in classrooms about how "a medium with an index" is supposed to polarize light reflected from it, but they didn't realize that sunlight coming off of water would be polarized. My guess is that doing this properly requires homework exercises; and that, unfortunately from my own standpoint, it happens to be a place where I have extra math talent, the same way that eg Marcello is more talented at formally proving theorems than I happen to be; and that people without the extra math talent, have to do a lot *more* exercises than I did, and I don't have a good sense of which exercises to give them.

[Ngo][11:13]

I'm sympathetic to this, and can try to turn off skeptical-discussion-mode and turn on learning-mode, if you think that'll help.

[Yudkowsky][11:14]

There's a general insight you can have about how arithmetic is commutative, and for some people you can show them $1 + 2 = 2 + 1$ and their native insight suffices to generalize over the 1 and the 2 to any

other numbers you could put in there, and they realize that strings of numbers can be rearranged and all end up equivalent. For somebody else, when they're a kid, you might have to show them 2 apples and 1 apple being put on the table in a different order but ending up with the same number of apples, and then you might have to show them again with adding up bills in different denominations, in case they didn't generalize from apples to money. I can actually remember being a child young enough that I tried to add 3 to 5 by counting "5, 6, 7" and I thought there was some clever enough way to do that to actually get 7, if you tried hard.

Being able to see "consequentialism" is like that, from my perspective.

[Ngo][11:15]

Another possibility: can you trace the origins of this belief, and how it came out of your previous beliefs?

[Yudkowsky][11:15]

I don't know what homework exercises to give people to make them able to see "consequentialism" all over the place, instead of inventing slightly new forms of consequentialist cognition and going "Well, now *that* isn't consequentialism, right?"

Trying to say "searching for states that get fed into an input-result function and then a result-scoring function" was one attempt of mine to describe the dangerous thing in a way that would maybe sound abstract enough that people would try to generalize it more.

[Ngo][11:17]

Another possibility: can you describe the closest thing to real consequentialism in humans, and how it came about in us?

[Yudkowsky][11:18][11:21]

Ok, so, part of the problem is that... before you do enough homework exercises for whatever your level of talent is (and even I, at one point, had done little enough homework that I thought there might be a clever way to add 3 and 5 in order to get to 7), you tend to think that only the very crisp formal thing that's been presented to you, is the "real" thing.

Why would your engine have to obey the laws of thermodynamics? You're not building one of those Carnot engines you saw in the physics textbook!

Humans contain fragments of consequentialism, or bits and pieces whose interactions add up to partially imperfectly shadow consequentialism, and the critical thing is being able to see that the reason why humans' outputs 'work', in a sense, is because these structures are what is doing

the work, and the work gets done because of how they shadow consequentialism and only insofar as they shadow consequentialism.

Put a human in one environment, it gets food. Put a human in a different environment, it gets food again. Wow, different initial conditions, same output! There must be things inside the human that, whatever else they do, are also along the way somehow effectively searching for motor signals such that food is the end result!

[Ngo][11:20]

To me it feels like you're trying to nudge me (and by extension whoever reads this transcript) out of a specific failure mode. If I had to guess, something like: "I understand what Eliezer is talking about so now I'm justified in disagreeing with it", or perhaps "Eliezer's explanation didn't make sense to me and so I'm justified in thinking that his concepts don't make sense". Is that right?

[Yudkowsky][11:22]

More like... from my perspective, even after I talk people out of one specific perpetual motion machine being possible, they go off and try to invent a different, more complicated perpetual motion machine.

And I am not sure what to do about that. It has been going on for a very long time from my perspective.

In the end, a lot of what people got out of all that writing I did, was not the deep object-level principles I was trying to point to - they did not really get [Bayesianism as thermodynamics](#), say, they did not become able to see [Bayesian structures](#) any time somebody sees a thing and changes their belief. What they got instead was something much more meta and general, a vague spirit of how to reason and argue, because that was what they'd spent a lot of time being exposed to over and over and over again in lots of blog posts.

Maybe there's no way to make somebody understand why [corrigibility](#) is "unnatural" except to repeatedly walk them through the task of trying to invent an agent structure that lets you press the shutdown button (without it trying to force you to press the shutdown button), and showing them how each of their attempts fails; and then also walking them through why Stuart Russell's attempt at moral uncertainty produces the [problem of fully updated \(non-\)deference](#); and hope they can start to see the informal general pattern of why corrigibility is in general contrary to the structure of things that are good at optimization.

Except that to do the exercises at all, you need them to work within an expected utility framework. And then they just go, "Oh, well, I'll just build an agent that's good at optimizing things but doesn't use these explicit expected utilities that are the source of the problem!"

And then if I want them to believe the same things I do, for the same reasons I do, I would have to teach them why certain structures of

cognition are the parts of the agent that are good at stuff and do the work, rather than them being this particular formal thing that they learned for manipulating meaningless numbers as opposed to real-world apples.

And I have tried to write that page once or twice (eg "[coherent decisions imply consistent utilities](#)") but it has not sufficed to teach them, because they did not even do as many homework problems as I did, let alone the greater number they'd have to do because this is in fact a place where I have a particular talent.

I don't know how to solve this problem, which is why I'm falling back on talking about it at the meta-level.

[Ngo][11:30]

I'm reminded of a LW post called "[Write a thousand roads to Rome](#)", which iirc argues in favour of trying to explain the same thing from as many angles as possible in the hope that one of them will stick.

[Soares][11:31]

(Suggestion, not-necessarily-good: having named this problem on the meta-level, attempt to have the object-level debate, while flagging instances of this as it comes up.)

[Ngo][11:31]

I endorse Nate's suggestion.

And will try to keep the difficulty of the meta-level problem in mind and respond accordingly.

[Yudkowsky][11:33]

That (Nate's suggestion) is probably the correct thing to do. I name it out loud because sometimes being told about the meta-problem actually does help on the object problem. It seems to help me a lot and others somewhat less, but it does help others at all, for many others.

3.2. Brain functions and outcome pumps

[Yudkowsky][11:34]

So, do you have a particular question you would ask about input-seeking cognitions? I did try to say why I mentioned those at all (it's a different road to Rome on "consequentialism").

[Ngo][11:36]

Let's see. So the visual cortex is an example of quite impressive cognition in humans and many other animals. But I'd call this "pattern-recognition" rather than "searching for high-scoring results".

[Yudkowsky][11:37]

Yup! And it is no coincidence that there are no whole animals formed entirely out of nothing but a visual cortex!

[Ngo][11:37]

Okay, cool. So you'd agree that the visual cortex is doing something that's qualitatively quite different from the thing that animals overall are doing.

Then another question is: can you characterise searching for high-scoring results in non-human animals? Do they do it? Or are you mainly talking about humans and AGIs?

[Yudkowsky][11:39]

Also by the time you get to like the temporal lobes or something, there is probably some significant amount of "what could I be seeing that would produce this visual field?" that is searching through hypothesis-space for hypotheses with high plausibility scores, and for sure at the human level, humans will start to think, "Well, could I be seeing this? No, that theory has the following problem. How could I repair that theory?" But it is plausible that there is no low-level analogue of this in a monkey's temporal cortex; and even more plausible that the parts of the visual cortex, if any, which do anything analogous to this, are doing it in a relatively local and definitely very domain-specific way.

Oh, that's the cerebellum and motor cortex and so on, if we're talking about a cat or whatever. They have to find motor plans that result in their catching the mouse.

Just because the visual cortex isn't (obviously) running a search doesn't mean the rest of the animal isn't running any searches.

(On the meta-level, I notice myself hiccuping "But how could you not see that when looking at a cat?" and wondering what exercises would be required to teach that.)

[Ngo][11:41]

Well, I see *something* when I look at a cat, but I don't know how well it corresponds to the concepts you're using. So just taking it slowly for now.

I have the intuition, by the way, that the motor cortex is in some sense doing a similar thing to the visual cortex - just in reverse. So instead of taking low-level inputs and producing high-level outputs, it's taking high-level inputs and producing low-level outputs. Would you agree with that?

[Yudkowsky][11:43]

It doesn't directly parse in my ontology because (a) I don't know what you mean by 'high-level' and (b) whole Cartesian agents can be viewed as functions, that doesn't mean all agents can be viewed as non-searching pattern-recognizers.

That said, all parts of the cerebral cortex have surprisingly similar morphology, so it wouldn't be at all surprising if the motor cortex is doing something similar to visual cortex. (The cerebellum, on the other hand...)

[Ngo][11:44]

The signal from the visual cortex saying "that is a cat", and the signal to the motor cortex saying "grab that cup", are things I'd characterise as high-level.

[Yudkowsky][11:45]

Still less of a native distinction in my ontology, but there's an informal thing it can sort of wave at, and I can hopefully take that as understood and run with it.

[Ngo][11:45]

The firing of cells in the retina, and firing of motor neurons, are the low-level parts.

Cool. So to a first approximation, we can think about the part in between the cat recognising a mouse, and the cat's motor cortex producing the specific neural signals required to catch the mouse, as the part where the consequentialism happens?

[Yudkowsky][11:49]

The part between the cat's eyes seeing the mouse, and the part where the cat's limbs move to catch the mouse, is the whole cat-agent. The

whole cat agent sure is a baby consequentialist / searches for mouse-catching motor patterns / gets similarly high-scoring end results even as you vary the environment.

The visual cortex is a particular part of this system-viewed-as-a-feedforward-function that is, plausibly, by no means surely, either not very searchy, or does only small local visual-domain-specific searches not aimed per se at catching mice; it has the epistemic nature rather than the planning nature.

Then from one perspective you could reason that "well, most of the consequentialism is in the remaining cat after visual cortex has sent signals onward". And this is in general a dangerous mode of reasoning that is liable to fail in, say, inspecting every particular neuron for consequentialism and not finding it; but in this particular case, there are significantly more consequentialist parts of the cat than the visual cortex, so I am okay running with it.

[Ngo][11:50]

Ah, the more specific thing I meant to say is: most of the consequentialism is strictly between the visual cortex and the motor cortex. Agree/disagree?

[Yudkowsky][11:51]

Disagree, I'm rusty on my neuroanatomy but I think the motor cortex may send signals on to the cerebellum rather than the other way around.

(I may also disagree with the actual underlying notion you're trying to hint at, so possibly not just a "well include the cerebellum then" issue, but I think I should let you respond first.)

[Ngo][11:53]

I don't know enough neuroanatomy to chase that up, so I was going to try a different tack.

But actually, maybe it's easier for me to say "let's include the cerebellum" and see where you think the disagreement ends up.

[Yudkowsky][11:56]

So since cats are not (obviously) (that I have read about) cross-domain consequentialists with imaginations, their consequentialism is in bits and pieces of consequentialism embedded in them all over by the more purely pseudo-consequentialist genetic optimization loop that built them.

A cat who fails to catch a mouse may then get little bits and pieces of catbrain adjusted all over.

And then those adjusted bits and pieces get a pattern lookup later.

Why do these pattern-lookups with no obvious immediate search element, all happen to point towards the same direction of catching the mouse? Because of the past causal history about how what gets looked up, which was tweaked to catch the mouse.

So it is legit harder to point out "the consequentialist parts of the cat" by looking for which sections of neurology are doing searches right there. That said, to the extent that the visual cortex does not get tweaked on failure to catch a mouse, it's not part of that consequentialist loop either.

And yes, the same applies to humans, but humans also do more explicitly searchy things and this is part of the story for why humans have spaceships and cats do not.

[Ngo][12:00]

Okay, this is interesting. So in biological agents we've got these three levels of consequentialism: evolution, reinforcement learning, and planning.

[Yudkowsky][12:01]

In biological agents we've got evolution + local evolved system-rules that in the past promoted genetic fitness. Two kinds of local rules like this are "operant-conditioning updates from success or failure" and "search through visualized plans". I wouldn't characterize these two kinds of rules as "levels".

[Ngo][12:02]

Okay, I see. And when you talk about searching through visualised plans (the type of thing that humans do) can you say more about what it means for that to be a "search"?

For example, if I imagine writing a poem line-by-line, I may only be planning a few words ahead. But somehow the whole poem, which might be quite long, ends up a highly-optimised product. Is that a central example of planning?

[Yudkowsky][12:04][12:07]

Planning is one way to succeed at search. I think for purposes of understanding alignment difficulty, you want to be thinking on the level of abstraction where you see that in some sense it is the search itself that is dangerous when it's a strong enough search, rather than the danger seeming to come from details of the planning process.

One of my early experiences in successfully generalizing my notion of intelligence, what I'd later verbalize as "computationally efficient finding of actions that produce outcomes high in a preference ordering", was in writing an (unpublished) story about time-travel in which the universe was globally consistent.

The requirement of global consistency, the way in which all events between Paradox start and Paradox finish had to map the Paradox's initial conditions onto the endpoint that would go back and produce those exact initial conditions, ended up imposing strong complicated constraints on reality that the Paradox in effect had to navigate using its initial conditions. The time-traveler needed to end up going through certain particular experiences that would produce the state of mind in which he'd take the actions that would end up prodding his future self elsewhere into having those experiences.

The Paradox ended up killing the people who built the time machine, for example, because they would not otherwise have allowed that person to go back in time, or kept the temporal loop open that long for any other reason if they were still alive.

Just having two examples of strongly consequentialist general optimization in front of me - human intelligence, and evolutionary biology - hadn't been enough for me to properly generalize over a notion of optimization. Having three examples of homework problems I'd worked - human intelligence, evolutionary biology, and the fictional Paradox - caused it to finally click for me.

[Ngo][12:07]

Hmm. So to me, one of the central features of search is that you consider many possibilities. But in this poem example, I may only have explicitly considered a couple of possibilities, because I was only looking ahead a few words at a time. This seems related to the distinction Abram drew a while back between selection and control (<https://www.alignmentforum.org/posts/ZDZmopKquzHYPRNxq/selection-vs-control>). Do you distinguish between them in the same way as he does? Or does "control" of a system (e.g. a football player dribbling a ball down the field) count as search too in your ontology?

[Yudkowsky][12:10][12:11]

I would later try to tell people to "imagine a paperclip maximizer as *not being a mind at all*, imagine it as a kind of malfunctioning time machine that spits out outputs which will in fact result in larger numbers of paperclips coming to exist later". I don't think it clicked because people hadn't done the same homework problems I had, and didn't have the same "Aha!" of realizing how part of the notion and danger of intelligence could be seen in such purely material terms.

But the [convergent instrumental strategies](#), the anticorrigibility, these things are contained in the *true fact about the universe* that certain outputs of the time machine *will in fact* result in there being lots more paperclips later. What produces the danger is not the details of the search process, it's the search being strong and effective *at all*. The danger is in the territory itself and not just in some weird map of it; that building nanomachines that kill the programmers will produce more paperclips is a fact about reality, not a fact about paperclip maximizers!

[Ngo][12:11]

Right, I remember a very similar idea in your writing about Outcome Pumps (<https://www.lesswrong.com/posts/4ARaTpNX62ual86j6/the-hidden-complexity-of-wishes>).

[Yudkowsky][12:12]

Yup! Alas, the story was written in 2002-2003 when I was a worse writer and the real story that inspired the Outcome Pump never did get published.

[Ngo][12:14]

Okay, so I guess the natural next question is: what is it that makes you think that a strong, effective search isn't likely to be limited or constrained in some way?

What is it about search processes (like human brains) that makes it hard to train them with blind spots, or deontological overrides, or things like that?

Hmmm, although it feels like this is a question I can probably predict your answer to. (Or maybe not, I wasn't expecting the time travel.)

[Yudkowsky][12:15]

In one sense, they are! A paperclip-maximizing superintelligence is nowhere near as powerful as a paperclip-maximizing time machine. The time machine can do the equivalent of buying winning lottery tickets from lottery machines that have been thermodynamically randomized; a superintelligence can't, at least not directly without rigging the lottery or whatever.

But a paperclip-maximizing strong general superintelligence is epistemically and instrumentally [efficient](#), relative to *you*, or to me. Any time we see it can get at least X paperclips by doing Y, we should expect that it gets X or more paperclips by doing Y or something that leads to even more paperclips than that, because it's not going to miss the strategy we see.

So in that sense, searching our own brains for how a time machine would get paperclips, asking ourselves how many paperclips are in principle possible and how they could be obtained, is a way of getting our own brains to consider lower bounds on the problem without the implicit stupidity assertions that our brains unwittingly use to constrain story characters. Part of the point of telling people to think about time machines instead of superintelligences was to get past the ways they imagine superintelligences being stupid. Of course that didn't work either, but it was worth a try.

I don't think that's quite what you were asking about, but I want to give you a chance to see if you want to rephrase anything before I try to answer your me-reformulated questions.

[Ngo][12:20]

Yeah, I think what I wanted to ask is more like: why should we expect that, out of the space of possible minds produced by optimisation algorithms like gradient descent, strong general superintelligences are more common than other types of agents which score highly on our loss functions?

[Yudkowsky][12:20][12:23][12:24]

It depends on how hard you optimize! And whether gradient descent on a particular system can even successfully optimize that hard! Many current AIs are trained by gradient descent and yet not superintelligences at all.

But the answer is that some problems are difficult in that they require solving lots of subproblems, and an easy way to solve all those subproblems is to use patterns which collectively have some coherence and overlap, and the coherence within them generalizes across all the subproblems. Lots of search orderings will stumble across something like that before they stumble across separate solutions for lots of different problems.

I suspect that you cannot get this out of small large amounts of gradient descent on small large layered transformers, and therefore I suspect that GPT-N does not approach superintelligence before the world is ended by systems that look differently, but I could be wrong about that.

[Ngo][12:22][12:23]

Suppose that we optimise hard enough to produce an epistemic subsystem that can make plans much better than any human's.

My guess is that you'd say that this is *possible*, but that we're much more likely to first produce a consequentialist agent which does this (rather than a purely epistemic agent which does this).

[Yudkowsky][12:24]

I am confused by what you think it means to have an "epistemic subsystem" that "makes plans much better than any human's". If it searches paths through time and selects high-scoring ones for output, what makes it "epistemic"?

[Ngo][12:25]

Suppose, for instance, that it doesn't actually carry out the plans, it just writes them down for humans to look at.

[Yudkowsky][12:25]

If it *can in fact* do the thing that a paperclipping time machine does, what makes it any safer than a paperclipping time machine because we called it "epistemic" or by some other such name?

By what criterion is it selecting the plans that humans look at?

Why did it make a difference that its output was fed through the causal systems called humans on the way to the causal systems called protein synthesizers or the Internet or whatever? If we build a superintelligence to design nanomachines, it makes no obvious difference to its safety whether it sends DNA strings directly to a protein synthesis lab, or humans read the output and retype it manually into an email. Presumably you also don't think that's where the safety difference comes from. So where does the safety difference come from?

(note: lunchtime for me in 2 minutes, propose to reconvene in 30m after that)

[Ngo][12:28]

(break for half an hour sounds good)

If we consider the visual cortex at a given point in time, how does it decide which objects to recognise?

Insofar as the visual cortex can be non-consequentialist about which objects it recognises, why couldn't a planning system be non-consequentialist about which plans it outputs?

[Yudkowsky][12:32]

This does feel to me like another "look at the water" moment, so what do you predict I'll say about that?

[Ngo][12:34]

I predict that you say something like: in order to produce an agent that can create very good plans, we need to apply a lot of optimisation power to that agent. And if the channel through which we're applying that optimisation power is "giving feedback on its plans", then we don't have a mechanism to ensure that the agent actually learns to optimise for creating really good plans, as opposed to creating plans that receive really good feedback.

[Soares][12:35]

Seems like a fine cliffhanger?

[Ngo][12:35]

Yepp.

[Soares][12:35]

Great. Let's plan to reconvene in 30min.

3.3. Hypothetical-planning systems, nanosystems, and evolving generality

[Yudkowsky][13:03][13:11]

So the answer you expected from me, translated into my terms, would be, "If you select for the consequence of the humans hitting 'approve' on the plan, you're still navigating the space of inputs for paths through time to probable outcomes (namely the humans hitting 'approve'), so you're still doing consequentialism."

But suppose you manage to avoid that. Suppose you get exactly what you ask for. Then the system is still outputting *plans* such that, when humans follow them, they take paths through time and end up with outcomes that score high in some scoring function.

My answer is, "What the heck would it mean for a *planning system* to be *non-consequentialist*? You're asking for nonwet water! What's consequentialist isn't the system that does the work, it's the work you're trying to do! You could imagine it being done by a cognition-free material system like a time machine and it would still be consequentialist because the output is a *plan*, a path through time!"

And this indeed is a case where I feel a helpless sense of not knowing how I can rephrase things, which exercises you have to get somebody to do, what fictional experience you have to walk somebody through, before they start to look at the water and see a material with an index, before they start to look at the phrase "why couldn't a planning system be non-consequentialist about which plans it outputs" and go "um".

My imaginary listener now replies, "Ah, but what if we have plans that don't end up with outcomes that score high in some function?" and I reply "Then you lie on the ground randomly twitching because any *outcome you end up with* which is *not that* is one that you wanted *more than that* meaning you *preferred it more than the outcome of random motor outputs* which is *optimization toward higher in the preference function* which is *taking a path through time that leads to particular destinations more than it leads to random noise*."

[Ngo][13:09][13:11]

Yeah, this does seem like a good example of the thing you were trying to explain at the beginning

It still feels like there's some sort of levels distinction going on here though, let me try to tease out that intuition.

Okay, so suppose I have a planning system that, given a situation and a goal, outputs a plan that leads from that situation to that goal.

And then suppose that we give it, as input, a situation that we're not actually in, and it outputs a corresponding plan.

It seems to me that there's a difference between the sense in which that planning system is consequentialist by virtue of making consequentialist plans (as in: if that plan were used in the situation described in its inputs, it would lead to some goal being achieved) versus another hypothetical agent that is just directly trying to achieve goals in the situation it's actually in.

[Yudkowsky][13:18]

So I'd preface by saying that, *if* you could build such a system, which is indeed a coherent thing (it seems to me) to describe for the purpose of building it, then there would possibly be a safety difference on the margins, it would be noticeably less dangerous though still dangerous. It would need a special internal structural property that you might not get by gradient descent on a loss function with that structure, just like natural selection on inclusive genetic fitness doesn't get you explicit fitness optimizers; you could optimize for planning in hypothetical situations, and get something that didn't explicitly care only and strictly about hypothetical situations. And even if you did get that, the outputs that would kill or brain-corrupt the operators in hypothetical situations might also be fatal to the operators in actual situations. But that is a coherent

thing to describe, and the fact that it was not optimizing our own universe, might make it *safer*.

With that said, I would worry that somebody would think there was some bone-deep difference of agentiness, of something they were empathizing with like personhood, of imagining goals and drives being absent or present in one case or the other, when they imagine a planner that just solves "hypothetical" problems. If you take that planner and feed it the actual world as its hypothetical, tada, it is now that big old dangerous consequentialist you were imagining before, without it having acquired some difference of *psychological* agency or 'caring' or whatever.

So I think there is an important homework exercise to do here, which is something like, "Imagine that safe-seeming system which only considers hypothetical problems. Now see that if you take that system, don't make any other internal changes, and feed it actual problems, it's very dangerous. Now meditate on this until you can see how the hypothetical-considering planner was extremely close in the design space to the more dangerous version, had all the dangerous latent properties, and would probably have a bunch of actual dangers too."

"See, you thought the source of the danger was this internal property of caring about actual reality, but it wasn't that, it was the structure of planning!"

[Ngo][13:22]

I think we're getting closer to the same page now.

Let's consider this hypothetical planner for a bit. Suppose that it was trained in a way that minimised the, let's say, *adversarial* component of its plans.

For example, let's say that the plans it outputs for any situation are heavily regularised so only the broad details get through.

Hmm, I'm having a bit of trouble describing this, but basically I have an intuition that in this scenario there's a component of its plan which is cooperative with whoever executes the plan, and a component that's adversarial.

And I agree that there's no fundamental difference in type between these two things.

[Yudkowsky][13:27]

"What if this potion we're brewing has a Good Part and a Bad Part, and we could just keep the Good Parts..."

[Ngo][13:27]

Nor do I think they're separable. But in some cases, you might expect one to be much larger than the other.

[Soares][13:29]

(I observe that my model of some other listeners, at this point, protest "there is yet a difference between the hypothetical-planner applied to actual problems, and the Big Scary Consequentialist, which is that the hypothetical planner is emitting descriptions of plans that *would* work if executed, whereas the big scary consequentialist is executing those plans directly.")

(Not sure that's a useful point to discuss, or if it helps Richard articulate, but it's at least a place I expect some reader's minds to go if/when this is published.)

[Yudkowsky][13:30]

(That is in fact a difference! The insight is in realizing that the hypothetical planner is only one line of outer shell command away from being a Big Scary Thing and is therefore also liable to be Big and Scary in many ways.)

[Ngo][13:31]

To me it seems that Eliezer's position is something like: "actually, in almost no training regimes do we get agents that decide which plans to output by spending almost all of their time thinking about the object-level problem, and very little of their time thinking about how to manipulate the humans carrying out the plan".

[Yudkowsky][13:32]

My position is that the AI does not neatly separate its internals into a Part You Think Of As Good and a Part You Think Of As Bad, because that distinction is sharp in your map but not sharp in the territory or the AI's map.

From the perspective of a paperclip-maximizing-action-outputting-time-machine, its actions are not "object-level making paperclips" or "manipulating the humans next to the time machine to deceive them about what the machine does", they're just physical outputs that go through time and end up with paperclips.

[Ngo][13:34]

@Nate, yeah, that's a nice way of phrasing one point I was trying to make. And I do agree with Eliezer that these things *can be* very similar.

But I'm claiming that in some cases these things can also be quite different - for instance, when we're training agents that only get to output a short high-level description of the plan.

[Yudkowsky][13:35]

The danger is in how hard the agent has to work to come up with the plan. I can, for instance, build an agent that very safely outputs a high-level plan for saving the world:

echo "Hey Richard, go save the world!"

So I do have to ask what kind of "high-level" planning output, that saves the world, you are envisioning, and why it was hard to cognitively come up with such that we didn't just make that high-level plan right now, if humans could follow it. Then I'll look at the part where the plan was hard to come up with, and say how the agent had to understand lots of complicated things in reality and accurately navigate paths through time for those complicated things, in order to even invent the high-level plan, and hence it was very dangerous if it wasn't navigating exactly where you hoped. Or, alternatively, I'll say, "That plan couldn't save the world: you're not postulating enough superintelligence to be dangerous, and you're also not using enough superintelligence to flip the tables on the currently extremely doomed world."

[Ngo][13:39]

At this point I'm not envisaging a particular planning output that saves the world, I'm just trying to get more clarity on the issue of consequentialism.

[Yudkowsky][13:40]

Look at the water; it's not the way you're doing the work that's dangerous, it's the work you're trying to do. What work are you trying to do, never mind how it gets done?

[Ngo][13:41]

I think I agree with you that, in the limit of advanced capabilities, we can't say much about how the work is being done, we have to primarily reason from the work that we're trying to do.

But here I'm only talking about systems that are intelligent enough to come up with plans and do research that are beyond the capability of humanity.

And for me the question is: for *those* systems, can we tilt the way they do the work so they spend 99% of their time trying to solve the object-level problem, and 1% of their time trying to manipulate the humans who are

going to carry out the plan? (Where these are not fundamental categories for the AI, they're just a rough categorisation that emerges after we've trained it - the same way that the categories of "physically moving around" and "thinking about things" aren't fundamentally different categories of action for humans, but the way we've evolved means there's a significant internal split between them.)

[Soares][13:43]

(I suspect Eliezer is not trying to make a claim of the form "in the limit of advanced capabilities, we are relegated to reasoning about what work gets done, not about how it was done". I suspect some miscommunication. It might be a reasonable time for Richard to attempt to paraphrase Eliezer's argument?)

(Though it also seems to me like Eliezer responding to the 99%/1% point may help shed light.)

[Yudkowsky][13:46]

Well, for one thing, I'd note that a system which is designing nanosystems, and spending 1% of its time thinking about how to kill the operators, is lethal. It has to be such a small fraction of thinking that it, like, never completes the whole thought about "well, if I did X, that would kill the operators!"

[Ngo][13:46]

Thanks for that, Nate. I'll try to paraphrase Eliezer's argument now.

Eliezer's position (partly in my own terminology): we're going to build AIs that can perform very difficult tasks using cognition which we can roughly describe as "searching over many options to find one that meets our criteria". An AI that can solve these difficult tasks will need to be able to search in a very general and flexible way, and so it will be very difficult to constrain that search into a particular region.

Hmm, that felt like a very generic summary, let me try and think about the more specific claims he's making.

[Yudkowsky][13:54]

An AI that can solve these difficult tasks will need to be able to

Very very little is universally necessary over the design space. The *first* AGI that our tech becomes able to build is liable to work in certain easier and simpler ways.

[Ngo][13:55]

Point taken; thanks for catching this misphrasing (this and previous times).

[Yudkowsky][13:56]

Can you, in principle, build a red-car-driver that is totally incapable of driving blue cars? In principle, sure! But the first red-car-driver that gradient descent stumbles over is liable to be a blue-car-driver too.

[Ngo][13:57]

Eliezer, I'm wondering how much of our disagreement is about how high the human level is here.

Or, to put it another way: we can build systems that outperform humans at quite a few tasks by now, without having search abilities that are general enough to even try to take over the world.

[Yudkowsky][13:58]

Indubitably and indeed, this is so.

[Ngo][13:59]

Putting aside for a moment the question of which tasks are pivotal enough to save the world, which parts of your model draw the line between human-level chess players and human-level galaxy-colonisers?

And say that we'll be able to align ones that they outperform us on *these tasks* before taking over the world, but not on *these other tasks*?

[Yudkowsky][13:59][14:01]

That doesn't have a very simple answer, but one aspect there is *domain generality* which in turn is achieved through *novel domain learning*.

Humans, you will note, were not aggressively optimized by natural selection to be able to breathe underwater or fly into space. In terms of obvious outer criteria, there is not much outer sign that natural selection produced these creatures much more general than chimpanzees, by training on a much wider range of environments and loss functions.

[Soares][14:00]

(Before we drift too far from it: thanks for the summary! It seemed good to me, and I updated towards the miscommunication I feared not-having-happened.)

[Ngo][14:03]

(Before we drift too far from it: thanks for the summary! It seemed good to me, and I updated towards the miscommunication I feared not-having-happened.)

(Good to know, thanks for keeping an eye out. To be clear, I didn't ever interpret Eliezer as making a claim explicitly about the limit of advanced capabilities; instead it just seemed to me that he was thinking about AIs significantly more advanced than the ones I've been thinking of. I think I phrased my point poorly.)

[Yudkowsky][14:05][14:10]

There are complicated aspects of this story where natural selection may metaphorically be said to have "had no idea of what it was doing", eg, after early rises in intelligence possibly produced by sexual selection on neatly chipped flint handaxes or whatever, all the cumulative brain-optimization on chimpanzees reached a point where there was suddenly a sharp selection gradient on relative intelligence at Machiavellian planning against other humans (even more so than in the chimp domain) as a subtask of inclusive genetic fitness, and so continuing to optimize on "inclusive genetic fitness" in the same old savannah, turned out to happen to be optimizing hard on the subtask and internal capability of "outwit other humans", which optimized hard on "model other humans", which was a capability that could be reused for modeling the chimp-that-is-this-chimp, which turned the system on itself and made it reflective, which contributed greatly to its intelligence being generalized, even though it was just grinding the same loss function on the same savannah; the system being optimized happened to go there in the course of being optimized even harder for the same thing.

So one can imagine asking the question: Is there a superintelligent AGI that can quickly build nanotech, which has a kind of passive safety in some if not all respects, in virtue of it solving problems like "build a nanotech system which does X" the way that a beaver solves building dams, in virtue of having a bunch of specialized learning abilities without it ever having a cross-domain general learning ability?

And in this regard one does note that there are many, many, many things that humans do which no other animal does, which you might think would contribute a lot to that animal's fitness if there were animalistic ways to do it. They don't make iron claws for themselves. They never did evolve a tendency to search for iron ore, and burn wood into charcoal that could be used in hardened-clay furnaces.

No animal plays chess, but AIs do, so we can obviously make AIs to do things that animals don't do. On the other hand, the environment didn't exactly present any particular species with a challenge of chess-playing either.

Even so, though, even if some animal had evolved to play chess, I fully expect that current AI systems would be able to squish it at chess,

because the AI systems are on chips that run faster than neurons and doing crisp calculations and there are things you just can't do with noisy slow neurons. So that again is not a generally reliable argument about what AIs can do.

[Ngo][14:09][14:11]

Yes, although I note that challenges which are trivial from a human-engineering perspective can be very challenging from an evolutionary perspective (e.g. spinning wheels).

And so the evolution of animals-with-a-little-bit-of-help-from-humans might end up in very different places from the evolution of animals-just-by-themselves. And analogously, the ability of humans to fill in the gaps to help less general AIs achieve more might be quite significant.

[Yudkowsky][14:11]

So we can again ask: Is there a way to make an AI system that is *only* good at designing nanosystems, which can achieve some complicated but hopefully-specifiable real-world outcomes, without that AI also being superhuman at understanding and manipulating humans?

And I roughly answer, "Perhaps, but not by default, there's a bunch of subproblems, I don't actually know how to do it right now, it's not *the easiest* way to get an AGI that can build nanotech (and kill you), you've got to make the red-car-driver specifically not be able to drive blue cars." Can I explain how I know that? I'm really not sure I can, in real life where I explain X₀ and then the listener doesn't generalize X₀ to X and respecialize it to X₁.

It's like asking me how I could possibly know in 2008, before anybody had observed AlphaFold 2, that superintelligences would be able to crack the protein folding problem on the way to nanotech, which some people did question back in 2008.

Though that was admittedly more of a slam-dunk than this was, and I could not have told you that AlphaFold 2 would become possible at a prehuman level of general intelligence in 2021 specifically, or that it would be synced in time to a couple of years after GPT-2's level of generality at text.

[Ngo][14:18]

What are the most relevant axes of difference between solving protein folding and designing nanotech that, say, self-assembles into a computer?

[Yudkowsky][14:20]

Definitely, "turns out it's easier than you thought to use gradient descent's memorization of zillions of shallow patterns that overlap and recombine into larger cognitive structures, to add up to a consequentialist nanoengineer that only does nanosystems and never does sufficiently general learning to apprehend the big picture containing humans, while still understanding the goal for that pivotal act you wanted to do" is among the more plausible advance-specified miracles we could get.

But it is not what my model says actually happens, and I am not a believer that when your model says you are going to die, you get to start believing in particular miracles. You need to hold your mind open for any miracle and a miracle you didn't expect or think of in advance, because at this point our last hope is that in fact the future is often quite surprising - though, alas, negative surprises are a tad more frequent than positive ones, when you are trying desperately to navigate using a bad map.

[Ngo][14:22]

Perhaps one metric we could use here is something like: how much extra reward does the consequentialist nanoengineer get from starting to model humans, versus from becoming better at nanoengineering?

[Yudkowsky][14:23]

But that's *not* where humans came from. We didn't get to nuclear power by getting a bunch of fitness from nuclear power plants. We got to nuclear power because if you get a bunch of fitness from chipping flint handaxes and Machiavellian scheming, as found by relatively simple and local hill-climbing, that entrains the same genes that build nuclear power plants.

[Ngo][14:24]

Only in the specific case where you also have the constraint that you keep having to learn new goals every generation.

[Yudkowsky][14:24]

Huh???

[Soares][14:24]

(I think Richard's saying, "that's a consequence of the genetic bottleneck")

[Ngo][14:25]

Right.

Hmm, but I feel like we may have covered this ground before.

Suggestion: I have a couple of other directions I'd like to poke at, and then we could wrap up in 20 or 30 minutes?

[Yudkowsky][14:27]

OK

What are the most relevant axes of difference between solving protein folding and designing nanotech that, say, self-assembles into a computer?

Though I want to mark that this question seemed potentially cruxy to me, though perhaps not for others. I.e., if building protein factories that built nanofactories that built nanomachines that met a certain deep and lofty engineering goal, didn't involve cognitive challenges different in kind from protein folding, we could maybe just safely go do that using AlphaFold 3, which would be just as safe as AlphaFold 2.

I don't think we can do that. And I would note to the generic Other that if, to them, these both just sound like thinky things, so why can't you just do that other thinky thing too using the thinky program, this is a case where having any specific model of why we don't already have this nanoengineer right now would tell you there were specific different thinky things involved.

3.4. Coherence and pivotal acts

[Ngo][14:31]

In either order:

- I'm curious how the things we've been talking about relate to your opinions about meta-level optimisation from the AI foom debate. (I.e. talking about how wrapping around so that there's no longer any protected level of optimisation leads to dramatic change.)
- I'm curious how your claims about the "robustness" of consequentialism (i.e. the difficulty of channeling an agent's thinking in the directions we want it to go) relate to the reliance of humans on culture, and in particular the way in which humans raised without culture are such bad consequentialists.

On the first: if I were to simplify to the extreme, it seems like there are these two core intuitions that you've been trying to share for a long time. One is a certain type of recursive improvement, and another is a certain type of consequentialism.

[Yudkowsky][14:32]

The second question didn't make much sense in my native ontology? Humans raised without culture don't have access to environmental constants whose presence their genes assume, so they end up as broken machines and then they're bad consequentialists.

[Ngo][14:35]

Hmm, good point. Okay, question modification: the ways in which humans reason, act, etc, vary greatly depending on which cultures they're raised in. (I'm mostly thinking about differences over time - e.g. cavemen vs moderns.) My low-fidelity version of your view about consequentialists says that general consequentialists like humans possess a robust search process which isn't so easily modified.

(Sorry if this doesn't make much sense in your ontology, I'm getting a bit tired.)

[Yudkowsky][14:36]

What is it that varies that you think I think should predict would stay more constant?

[Ngo][14:37]

Goals, styles of reasoning, deontological constraints, level of conformity.

[Yudkowsky][14:39]

With regards to your first point, my first reaction was, "I just have one view of intelligence, what you see me arguing about reflects which points people have proved weirdly obstinate about. In 2008, Robin Hanson was being weirdly obstinate about how capabilities scaled and whether there was even any point in analyzing AIs differently from ems, so I talked about what I saw as the most slam-dunk case for there being Plenty Of Room Above Biology and for stuff going whoosh once it got above the human level.

"It later turned out that capabilities started scaling a whole lot *without* self-improvement, which is an example of the kind of weird surprise the Future throws at you, and maybe a case where I missed something by arguing with Hanson instead of imagining how I could be wrong in either

direction and not just the direction that other people wanted to argue with me about.

"Later on, people were unable to understand why alignment is hard, and got stuck on generalizing the concept I refer to as consequentialism. A theory of why I talked about both things for related reasons would just be a theory of why people got stuck on these two points for related reasons, and I think that theory would mainly be overexplaining an accident because if Yann LeCun had been running effective altruism I would have been explaining different things instead, after the people who talked a lot to EAs got stuck on a different point."

Returning to your second point, humans are broken things; if it were possible to build computers while working even worse than humans, we'd be having this conversation at that level of intelligence instead.

[Ngo][14:41]

(Retracted) I entirely agree about humans, but it doesn't matter that much how broken humans are when the regime of AIs that we're talking about is the regime that's directly above humans, and therefore only a bit less broken than humans.

[Yudkowsky][14:41]

Among the things to bear in mind about that, is that we then get tons of weird phenomena that are specific to humans, and you may be very out of luck if you start wishing for the *same* weird phenomena in AIs. Yes, even if you make some sort of attempt to train it using a loss function.

However, it does seem to me like as we start getting towards the Einstein level instead of the village-idiot level, even though this is usually not much of a difference, we do start to see the atmosphere start to thin already, and the turbulence start to settle down already. Von Neumann was actually a fairly reflective fellow who knew about, and indeed helped generalize, utility functions. The great achievements of von Neumann were not achieved by some very specialized hypernerd who spent all his fluid intelligence on crystallizing math and science and engineering alone, and so never developed any opinions about politics or started thinking about whether or not he had a utility function.

[Ngo][14:44]

I don't think I'm asking for the *same* weird phenomena. But insofar as a bunch of the phenomena I've been talking about have seemed weird according to your account of consequentialism, then the fact that approximately-human-level-consequentialists have lots of weird things about them is a sign that the phenomena I've been talking about are less unlikely than you expect.

[Yudkowsky][14:45][14:46]

I suspect that some of the difference here is that I think you have to be *noticeably* better than a human at nanoengineering to pull off pivotal acts large enough to make a difference, which is why I am not instead trying to gather the smartest people left alive and doing that pivotal act directly.

I can't think of anything you can do with somebody just barely smarter than a human, which flips the gameboard, aside of course from "go build a Friendly AI" which I *did* try to set up to just go do and which would be incredibly hard to align if we wanted an AI to do it instead (full-blown chicken-and-egg, that AI is already fully aligned).

[Ngo][14:45]

Oh, interesting. Actually one more question then: to what extent do you think that explicitly reasoning about utility functions and laws of rationality is what makes consequentialists have the properties you've been talking about?

[Yudkowsky][14:47, moved up in log]

Explicit reflection is one possible later stage of the path; an earlier part of the path is from being optimized to do things difficult enough that you need to stop stepping on your own feet and have different parts of your thoughts work well together.

It's the sort of path that has only one destination at its end, so there will be many ways to get there.

(Modulo various cases where different decision theories seem reflectively consistent and so on; I want to say "you know what I mean" but maybe people don't.)

[Ngo][14:47, moved down in log]

I suspect that some of the difference here is that I think you have to be *noticeably* better than a human at nanoengineering to pull off pivotal acts large enough to make a difference, which is why I am not instead trying to gather the smartest people left alive and doing that pivotal act directly.

Yepp, I think there's probably some disagreements about geopolitics driving this too. E.g. in my earlier summary document I mentioned some possible pivotal acts:

- Monitoring all potential AGI projects to an extent that makes it plausible for the US and China to work on a joint project without worrying that the other is privately racing.
- Provide arguments/demonstrations/proofs related to impending existential risk that are sufficiently compelling to scare the key global decision-makers into bottlenecking progress.

I predict that you think these would not be pivotal enough; but I don't think digging into the geopolitical side of things is the best use of our time.

[Yudkowsky][14:49, moved up in log]

Monitoring all AGI projects - either not politically feasible in real life given the actual way that countries behave in history books instead of fantasy; or at politically feasible levels, does not work well enough to prevent the world from ending once the know-how proliferates. The AI isn't doing much work here either; why not go do this now, if it's possible? (Note: please don't try to go do this now, it backfires badly.)

Provide sufficiently compelling arguments = superhuman manipulation, an incredibly dangerous domain that is just about the worst domain to try to align.

[Ngo][14:49, moved down in log]

With regards to your first point, my first reaction was, "I just have one view of intelligence, what you see me arguing about reflects which points people have proved weirdly obstinate about. In 2008, Robin Hanson was being weirdly obstinate about how capabilities scaled and whether there was even any point in analyzing AIs differently from ems, so I talked about what I saw as the most slam-dunk case for there being Plenty Of Room Above Biology and for stuff going whoosh once it got above the human level."

"It later turned out that capabilities started scaling a whole lot *without* self-improvement, which is an example of the kind of weird surprise the Future throws at you, and maybe a case where I missed something by arguing with Hanson instead of imagining how I could be wrong in either direction and not just the direction that other people wanted to argue with me about.

"Later on, people were unable to understand why alignment is hard, and got stuck on generalizing the concept I refer to as consequentialism. A theory of why I talked about both things for related reasons would just be a theory of why people got stuck on these two points for related reasons, and I think that theory would mainly be overexplaining an accident because if Yann LeCun had been running effective altruism I would have been explaining different things instead, after the people who talked a lot to EAs got stuck on a different point."

On my first point, it seems to me that your claims about recursive self-improvement were off in a fairly similar way to how I think your claims about consequentialism are off - which is that they defer too much to one very high-level abstraction.

[Yudkowsky][14:52]

On my first point, it seems to me that your claims about recursive self-improvement were off in a fairly similar way to how I think your claims about consequentialism are off - which is that they defer too much to one very high-level abstraction.

I suppose that is what it could potentially feel like from the inside to not get an abstraction. Robin Hanson kept on asking why I was trusting my abstractions so much, when he was in the process of trusting his worse abstractions instead.

[Ngo][14:51][14:53]

Explicit reflection is one possible later stage of the path; an earlier part of the path is from being optimized to do things difficult enough that you need to stop stepping on your own feet and have different parts of your thoughts work well together.

Can you explain a little more what you mean by "have different parts of your thoughts work well together"? Is this something like the capacity for metacognition; or the global workspace; or self-control; or...?

And I guess there's no good way to quantify *how* important you think the explicit reflection part of the path is, compared with other parts of the path - but any rough indication of whether it's a more or less crucial component of your view?

[Yudkowsky][14:55]

Can you explain a little more what you mean by "have different parts of your thoughts work well together"? Is this something like the capacity for metacognition; or the global workspace; or self-control; or...?

No, it's like when you don't, like, pay five apples for something on Monday, sell it for two oranges on Tuesday, and then trade an orange for an apple.

I have still not figured out the homework exercises to convey to somebody the Word of Power which is "coherence" by which they will be able to look at the water, and see "coherence" in places like a cat walking across the room without tripping over itself.

When you do lots of reasoning about arithmetic correctly, without making a misstep, that long chain of thoughts with many different pieces diverging and ultimately converging, ends up making some statement that is... still true and still about numbers! Wow! How do so many different thoughts add up to having this property? Wouldn't they wander off and end up being about tribal politics instead, like on the Internet?

And one way you could look at this, is that even though all these thoughts are taking place in a bounded mind, they are shadows of a higher unbounded structure which is the model identified by the Peano axioms; all the things being said are *true about the numbers*. Even

though somebody who was missing the point would at once object that the human contained no mechanism to evaluate each of their statements against all of the numbers, so obviously no human could ever contain a mechanism like that, so obviously you can't explain their success by saying that each of their statements was true about the same topic of the numbers, because what could possibly implement that mechanism which (in the person's narrow imagination) is The One Way to implement that structure, which humans don't have?

But though mathematical reasoning can sometimes go astray, when it works at all, it works because, in fact, even bounded creatures can sometimes manage to obey local relations that in turn add up to a global coherence where all the pieces of reasoning point in the same direction, like photons in a laser lasing, even though there's no internal mechanism that enforces the global coherence at every point.

To the extent that the outer optimizer trains you out of paying five apples on Monday for something that you trade for two oranges on Tuesday and then trading two oranges for four apples, the outer optimizer is training all the little pieces of yourself to be locally coherent in a way that can be seen as an imperfect bounded shadow of a higher unbounded structure, and then the system is powerful though imperfect *because* of how the power is present in the coherence and the overlap of the pieces, *because* of how the higher perfect structure is being imperfectly shadowed. In this case the higher structure I'm talking about is Utility, and doing homework with coherence theorems leads you to appreciate that we only know about one higher structure for this class of problems that has a dozen mathematical spotlights pointing at it saying "look here", even though people have occasionally looked for alternatives.

And when I try to say this, people are like, "Well, I looked up a theorem, and it talked about being able to identify a unique utility function from an infinite number of choices, but if we don't have an infinite number of choices, we can't identify the utility function, so what relevance does this have" and this is a kind of mistake I don't remember even coming close to making so I do not know how to make people stop doing that and maybe I can't.

[Soares][15:07]

We're already pushing our luck on time, so I nominate that we wrap up (after, perhaps, a few more Richard responses if he's got juice left.)

[Yudkowsky][15:07]

Yeah, was thinking the same.

[Soares][15:07]

As a proposed cliffhanger to feed into the next discussion, my take is that Richard's comment:

On my first point, it seems to me that your claims about recursive self-improvement were off in a fairly similar way to how I think your claims about consequentialism are off - which is that they defer too much to one very high-level abstraction.

probably contains some juicy part of the disagreement, and I'm interested in Eliezer understanding Richard's claim to the point of being able to paraphrase it to Richard's satisfaction.

[Ngo][15:08]

Wrapping up here makes sense.

I endorse the thing Nate just said.

I also get the sense that I have a much better outline now of Eliezer's views about consequentialism (if not the actual details and texture).

On a meta level, I personally tend to focus more on things like "how should we understand cognition" and not "how should we understand geopolitics and how it affects the level of pivotal action required".

If someone else were trying to prosecute this disagreement they might say much more about the latter. I'm uncertain how useful it is for me to do so, given that my comparative advantage compared with the rest of the world (and probably Eliezer's too) is the cognition part.

[Yudkowsky][15:12]

Reconvene... tomorrow? Monday of next week?

[Ngo][15:12]

Monday would work better for me.

You okay with me summarising the discussion so far to [some people — redacted for privacy reasons]?

[Yudkowsky][15:13]

Nate, take a minute to think of your own thoughts there?

[Soares:

[Soares][15:15]

My take: I think it's fine to summarize, though generally virtuous to mark summaries as summaries (rather than asserting that your summaries are Eliezer-endorsed or w/e).

[Ngo: ]

[Yudkowsky][15:16]

I think that broadly matches my take. I'm also a bit worried about biases in the text summarizer, and about whether I managed to say anything that Rob or somebody will object to pre-publication, but we ultimately intended this to be seen and I was keeping that in mind, so, yeah, go ahead and summarize.

[Ngo][15:17]

Great, thanks

[Yudkowsky][15:17]

I admit to being curious as to what you thought was said that was important or new, but that's a question that can be left open to be answered at your leisure, earlier in your day.

[Ngo][15:17]

I admit to being curious as to what you thought was said that was important or new, but that's a question that can be left open to be answered at your leisure, earlier in your day.

You mean, what I thought was worth summarising?

[Yudkowsky][15:17]

Yeah.

[Ngo][15:18]

Hmm, no particular opinion. I wasn't going to go out of my way to do so, but since I'm chatting to [some people — redacted for privacy reasons] regularly anyway, it seemed low-cost to fill them in.

At your leisure, I'd be curious to know how well the directions of discussion are meeting your goals for what you want to convey when this is published, and whether there are topics you want to focus on more.

[Yudkowsky][15:19]

I don't know if it's going to help, but trying it currently seems better than to go on saying nothing.

[Ngo][15:20]

(personally, in addition to feeling like less of an expert on geopolitics, it also seems more sensitive for me to make claims about in public, which is another reason I haven't been digging into that area as much)

[Soares][15:21]

(personally, in addition to feeling like less of an expert on geopolitics, it also seems more sensitive for me to make claims about in public, which is another reason I haven't been digging into that area as much)

(seems reasonable! note, though, that i'd be quite happy to have sensitive sections stricken from the record, insofar as that lets us get more convergence than we otherwise would, while we're already in the area)

[Ngo: ]

(tho ofc it is less valuable to spend conversational effort in private discussions, etc.)

[Ngo: ]

[Ngo][15:22]

At your leisure, I'd be curious to know how well the directions of discussion are meeting your goals for what you want to convey when this is published, and whether there are topics you want to focus on more.

(this question aimed at you too Nate)

Also, thanks Nate for the moderation! I found your interventions well-timed and useful.

[Soares: ]

[Soares][15:23]

(this question aimed at you too Nate)

(noted, thanks, I'll probably write something up after you've had the opportunity to depart for sleep.)

On that note, I declare us adjourned, with intent to reconvene at the same time on Monday.

Thanks again, both.

[Ngo][15:23]

Thanks both 😊

Oh, actually, one quick point

Would one hour earlier suit, for Monday?

I've realised that I'll be moving to a one-hour-later time zone, and starting at 9pm is slightly suboptimal (but still possible if necessary)

[Soares][15:24]

One hour earlier would work fine for me.

[Yudkowsky][15:25]

Doesn't work as fine for me because I've been trying to avoid any food until 12:30p my time, but on that particular day I may be more caloried than usual from the previous day, and could possibly get away with it. (That whole day could also potentially fail if a minor medical procedure turns out to take more recovery than it did the last time I had it.)

[Ngo][15:26]

Hmm, is this something where you'd have more information on the day? (For the calories thing)

[Yudkowsky][15:27]

(seems reasonable! note, though, that i'd be quite happy to have sensitive sections stricken from the record, insofar as that lets us get more convergence than we otherwise would, while we're already in the area)

I'm a touch reluctant to have discussions that we intend to delete, because then the larger debate will make less sense once those sections are deleted. Let's dance around things if we can.

[Ngo: 👍] [Soares: 👍]

I mean, I can that day at 10am my time say how I am doing and whether I'm in shape for that day.

[Ngo][15:28]

great. and if at that point it seems net positive to postpone to 11am your time (at the cost of me being a bit less coherent later on) then feel free to say so at the time

on that note, I'm off

[Yudkowsky][15:29]

Good night, heroic debater!

[Soares][16:11]

At your leisure, I'd be curious to know how well the directions of discussion are meeting your goals for what you want to convey when this is published, and whether there are topics you want to focus on more.

The discussions so far are meeting my goals quite well so far! (Slightly better than my expectations, hooray.) Some quick rough notes:

- I have been enjoying EY explicating his models around consequentialism.
 - The objections Richard has been making are ones I think have been floating around for some time, and I'm quite happy to see explicit discussion on it.
 - Also, I've been appreciating the conversational virtue with which the two of you have been exploring it. (Assumption of good intent, charity, curiosity, etc.)
- I'm excited to dig into Richard's sense that EY was off about recursive self improvement, and is now off about consequentialism, in a similar way.
 - This also sees to me like a critique that's been floating around for some time, and I'm looking forward to getting more clarity on it.
- I'm a bit torn between driving towards clarity on the latter point, and shoring up some of the progress on the former point.
 - One artifact I'd really enjoy having is some sort of "before and after" take, from Richard, contrasting his model of EY's views before, to his model now.
 - I also have a vague sense that there are some points Eliezer was trying to make, that didn't quite feel like they were driven home; and dually, some pushback by Richard that didn't feel quite frontally answered.
 - One thing I may do over the next few days is make a list of those places, and see if I can do any distilling on my own. (No promises, though.)
 - If that goes well, I might enjoy some side-channel back-and-forth with Richard about it, eg during some more convenient-for-Richard hour (or, eg, as a thing to do on Monday if EY's not in commission at 10a pacific.)

[Ngo][5:40] (next day, Sep. 9)

The discussions so far are [...]

What do you mean by "latter point" and "former point"? (In your 6th bullet point)

[Soares][7:09] (next day, Sep. 9)

What do you mean by "latter point" and "former point"? (In your 6th bullet point)

former = shoring up the consequentialism stuff, latter = digging into your critique re: recursive self improvement etc. (The nesting of the bullets was supposed to help make that clear, but didn't come out well in this format, oops.)

4. Follow-ups

4.1. Richard Ngo's summary

[Ngo] (Sep. 10 Google Doc)

2nd discussion

(Mostly summaries not quotations; also ~~hasn't yet been evaluated by Eliezer~~)

Eliezer, summarized by Richard: "The A core concept which people have trouble grasping is consequentialism. People try to reason about *how* AIs will solve problems, and ways in which they might or might not be dangerous. But they don't realise that the ability to solve a wide range of difficult problems implies that an agent must be doing a powerful search over possible solutions, which is ~~the~~ a core skill required to take actions which greatly affect the world. Making this type of AI safe is like trying to build an AI that drives red cars very well, but can't drive blue cars - there's no way you get this by default, because the skills involved are so similar. And because the search process ~~is so general~~ is by default so general, ~~it'll be very hard to~~ I don't currently see how to constrain it into any particular region."

[Yudkowsky][10:48] (Sep. 10 comment)

The

A concept, which some people have had trouble grasping. There seems to be an endless list. I didn't have to spend much time contemplating consequentialism to derive the consequences. I didn't spend a lot of time talking about it until people started arguing.

[Yudkowsky][10:50] (Sep. 10 comment)

the

a

[Yudkowsky][10:52] (Sep. 10 comment)

[the search process] is [so general]

"is by default". The reason I keep emphasizing that things are only true by default is that the work of surviving may look like doing hard nondefault things. I don't take fatalistic "will happen" stances, I assess difficulties of getting nondefault results.

[Yudkowsky][10:52] (Sep. 10 comment)

it'll be very hard to

"I don't currently see how to"

[Ngo] (Sep. 10 Google Doc)

Eliezer, summarized by Richard (continued): "In biological organisms, evolution is ~~one source~~ the ultimate source of consequentialism. A ~~second~~ secondary outcome of evolution is reinforcement learning. For an animal like a cat, upon catching a mouse (or failing to do so) many parts of its brain get slightly updated, in a loop that makes it more likely to catch the mouse next time. (Note, however, that this process isn't powerful enough to make the cat a pure consequentialist - rather, it has many individual traits that, when we view them from this lens, point in the same direction.) ~~A third thing that makes humans in particular consequentialist is planning,~~ Another outcome of evolution, which helps make humans in particular more consequentialist, is planning - especially when we're aware of concepts like utility functions."

[Yudkowsky][10:53] (Sep. 10 comment)

one

the ultimate

[Yudkowsky][10:53] (Sep. 10 comment)

second

secondary outcome of evolution

[Yudkowsky][10:55] (Sep. 10 comment)

especially when we're aware of concepts like utility functions

Very slight effect on human effectiveness in almost all cases because humans have very poor reflectivity.

[Ngo] (Sep. 10 Google Doc)

Richard, summarized by Richard: "Consider an AI that, given a hypothetical scenario, tells us what the best plan to achieve a certain goal in that scenario is. Of course it needs to do consequentialist reasoning to figure out how to achieve the goal. But that's different from an AI which chooses what to say as a means of achieving its goals. I'd argue that the former is doing consequentialist reasoning without itself being a consequentialist, while the latter is actually a consequentialist. Or more succinctly: consequentialism = problem-solving skills + using those skills to choose actions which achieve goals."

Eliezer, summarized by Richard: "The former AI might be slightly safer than the latter if you could build it, but I think people are likely to dramatically overestimate how big the effect is. The difference could just be one line of code: if we give the former AI our current scenario as its input, then it becomes the latter. For purposes of understanding alignment difficulty, you want to be thinking on the level of abstraction where you see that in some sense it is the search itself that is dangerous when it's a strong enough search, rather than the danger seeming to come from details of the planning process. One particularly helpful thought experiment is to think of advanced AI as an '[outcome pump](#)' which selects from futures in which a certain outcome occurred, and takes whatever action leads to them."

[Yudkowsky][10:59] (Sep. 10 comment)

particularly helpful

"attempted explanatory". I don't think most readers got it.

I'm a little puzzled by how often you write my viewpoint as thinking that whatever I happened to say a sentence about is the Key Thing. It seems to rhyme with a deeper failure of many EAs to pass the MIRI [ITI](#).

To be a bit blunt and impolite in hopes that long-languishing social processes ever get anywhere, two obvious uncharitable explanations for why some folks may systematically misconstrue MIRI/Eliezer as believing much more than in reality that various concepts an argument wanders over are Big Ideas to us, when some conversation forces us to go to that place:

(A) It paints a comfortably unflattering picture of MIRI-the-Other as weirdly obsessed with these concepts that seem not so persuasive, or more generally paints the Other as a bunch of weirdos who stumbled across some concept like "consequentialism" and got obsessed with it. In general, to depict the Other as thinking a great deal of some idea (or explanatory thought experiment) is to tie and stake their status to the listener's view of how much status that idea deserves. So if you say that the Other thinks a great deal of some idea that isn't obviously high-status, that lowers the Other's status, which can be a comfortable thing to do.

(cont.)

(B) It paints a more comfortably self-flattering picture of a continuing or persistent disagreement, as a disagreement with somebody who thinks that some random concept is much higher-status than it really is, in which case there isn't more to done or understood except to duly politely let the other person try to persuade you the concept deserves its high status. As opposed to, "huh, maybe there is a noncentral point that the other person sees themselves as being stopped on and forced to explain to me", which is a much less self-flattering viewpoint on why the conversation is staying within a place. And correspondingly more of a viewpoint that somebody else is likely to have of us, because it is a comfortable view to them, than a viewpoint that it is comfortable to us to imagine them having.

Taking the viewpoint that somebody else is getting hung up on a relatively noncentral point can also be a flattering self-portrait to somebody who believes that, of course. It doesn't mean they're right. But it does mean that you should be aware of how the Other's story, told from the Other's viewpoint, is much more liable to be something that the Other finds sensible and perhaps comfortable, even if it implies an unflattering (and untrue-seeming and perhaps untrue) view of yourself, than something that makes the Other seem weird and silly and which it is easy and congruent for you yourself to imagine the Other thinking.

[Ngo][11:18] (Sep. 12 comment)

I'm a little puzzled by how often you write my viewpoint as thinking that whatever I happened to say a sentence about is the Key Thing.

In this case, I emphasised the outcome pump thought experiment because you said that the time-travelling scenario was a key moment for your understanding of optimisation, and the outcome pump seemed to be similar enough and easier to convey in the summary, since you'd already written about it.

I'm also emphasising consequentialism because it seemed like the core idea which kept coming up in our first debate, under the heading of "deep problem-solving patterns". Although I take your earlier point that you tend to emphasise things that your interlocutor is more skeptical about, not necessarily the things which are most central to your view. But if consequentialism isn't in fact a very central concept for you, I'd be interested to hear what role it plays.

[Ngo] (Sep. 10 Google Doc)

Richard, summarized by Richard: "There's a component of 'finding a plan which achieves a certain outcome' which involves actually solving the object-level problem of how someone who is given the plan can achieve the outcome. And there's another component which is figuring out how to manipulate that person into doing what you want. To me it seems like Eliezer's argument is that there's no training regime which leads an AI to spend 99% of its time thinking about the former, and 1% thinking about the latter."

[Yudkowsky][11:20] (Sep. 10 comment)

no training regime

...that the training regimes we come up with first, in the 3 months or 2 years we have before somebody else destroys the world, will not have this property.

I don't have any particularly complicated or amazingly insightful theories of why I keep getting depicted as a fatalist; but my world is full of counterfactual functions, not constants. And I am always aware that if we had access to a real Textbook from the Future explaining all of the methods that are actually robust in real life - the equivalent of telling us in advance about all the ReLUs that in real life were only invented and understood a few decades after sigmoids - we could go right ahead and build a superintelligence that thinks $2 + 2 = 5$.

All of my assumptions about "I don't see how to do X" are always labeled as ignorance on my part and a default because we won't have enough time to actually figure out how to do X. I am constantly maintaining awareness of this because being **wrong** about it being difficult is a major place where **hope** potentially comes from, if there's some idea like ReLUs that robustly vanquishes the difficulty, which I just didn't think of. Which does not, alas, mean that I am wrong about any particular thing, nor that the infinite source of optimistic ideas that is the wider field of "AI alignment" is going to produce a good idea from the same process that generates all the previous naive optimism through not seeing where the original difficulty comes from or what other difficulties surround obvious naive attempts to solve it.

[Ngo] (Sep. 10 Google Doc)

Richard, summarized by Richard (continued): "While this may be true in the limit of increasing intelligence, the most relevant systems are the earliest ones that are above human level. But humans deviate from the consequentialist abstraction you're talking about in all sorts of ways - for example, being raised in different cultures can make people much more or less consequentialist. So it seems plausible that early AGIs can be superhuman while also deviating strongly from this abstraction - not necessarily in the same ways as humans, but in ways that we push them towards during training."

Eliezer, summarized by Richard: "Even at the Einstein or von Neumann level these types of deviations start to subside. And the sort of pivotal acts which might realistically work require skills *significantly* above human level. I think even 1% of the cognition of an AI that can assemble advanced nanotech, thinking about how to kill humans, would doom us. Your other suggestions for pivotal acts (surveillance to restrict AGI proliferation; persuading world leaders to restrict AI development) are not politically feasible in real life, to the level required to prevent the world from ending; or else require alignment in the very dangerous domain of superhuman manipulation."

Richard, summarized by Richard: "I think we probably also have significant disagreements about geopolitics which affect which acts we expect to be pivotal, but it seems like our comparative advantage is in discussing cognition, so let's focus on that. We can build systems that outperform humans at quite a few tasks by now, without them needing search abilities that are general enough to even try to take over the world. Putting aside for a moment the question of which tasks are pivotal enough to save the world, which parts of your model draw the line between human-level chess players and human-level galaxy-colonisers, and say that we'll be able to align ones that significantly outperform us on *these* tasks before they take over the world, but not on *those* tasks?"

Eliezer, summarized by Richard: "One aspect there is domain generality which in turn is achieved through novel domain learning. One can imagine asking the question: is there a superintelligent AGI that can quickly build nanotech the way that a beaver solves building dams, in virtue of having a bunch of specialized learning abilities without it ever having a cross-domain general learning ability? But there are many, many, many things that humans do which no other animal does, which you might think would contribute a lot to that animal's fitness if there were animalistic ways to do it - e.g. mining and smelting iron. (Although comparisons to animals are not generally reliable arguments about what AIs can do - e.g. chess is much easier for chips than neurons.) So my answer is 'Perhaps, but not by default, there's a bunch of subproblems, I don't actually know how to do it right now, it's not the easiest way to get an AGI that can build nanotech.' Can I explain how I know that? I'm really not sure I can."

[Yudkowsky][11:26] (Sep. 10 comment)

Can I explain how I know that? I'm really not sure I can.

In original text, this sentence was followed by a long attempt to explain anyways; if deleting that, which is plausibly the correct choice, this lead-in sentence should also be deleted, as otherwise it paints a false picture of how much I would try to explain anyways.

[Ngo][11:15] (Sep. 12 comment)

Makes sense; deleted.

[Ngo] (Sep. 10 Google Doc)

Richard, summarized by Richard: "Challenges which are trivial from a human-engineering perspective can be very challenging from an evolutionary perspective (e.g. spinning wheels). So the evolution of animals-with-a-little-bit-of-help-from-humans might end up in very different places from the evolution of animals-just-by-themselves. And analogously, the ability of humans to fill in the gaps to help less general AIs achieve more might be quite significant."

"On nanotech: what are the most relevant axes of difference between solving protein folding and designing nanotech that, say, self-assembles into a computer?"

Eliezer, summarized by Richard: "This question seemed potentially cruxy to me. I.e., if building protein factories that built nanofactories that built nanomachines that met a certain deep and lofty engineering goal, didn't involve cognitive challenges different in kind from protein folding, we could maybe just safely go do that using AlphaFold 3, which would be just as safe as AlphaFold 2. I don't think we can do that. But it is among the more plausible advance-specified miracles we could get. At this point our last hope is that in fact the future is often quite surprising."

Richard, summarized by Richard: "It seems to me that you're making the same mistake here as you did with regards to recursive self-improvement in the AI foam debate - namely, putting too much trust in one big abstraction."

Eliezer, summarized by Richard: "I suppose that is what it could potentially feel like from the inside to not get an abstraction. Robin Hanson kept on asking why I was trusting my abstractions so much, when he was in the process of trusting his worse abstractions instead."

4.2. Nate Soares' summary

[Soares] (Sep. 12 Google Doc)

Consequentialism

Ok, here's a handful of notes. I apologize for not getting them out until midday Sunday. My main intent here is to do some shoring up of the ground we've covered. I'm hoping for skims and maybe some light comment back-and-forth as seems appropriate (perhaps similar to Richard's summary), but don't think we should derail the main thread over it. If time is tight, I would not be offended for these notes to get little-to-no interaction.

--

My sense is that there's a few points Eliezer was trying to transmit about consequentialism, that I'm not convinced have been received. I'm going to take a whack at it. I may well be wrong, both about whether Eliezer is in fact attempting to transmit these, and about whether Richard received them; I'm interested in both protests from Eliezer and paraphrases from Richard.

[Soares] (Sep. 12 Google Doc)

1. "The consequentialism is in the plan, not the cognition".

I think Richard and Eliezer are coming at the concept "consequentialism" from very different angles, as evidenced eg by Richard saying (Nate's crappy paraphrase:) "where do you think the consequentialism is in a cat?" and Eliezer responding (Nate's crappy paraphrase:) "the cause of the apparent consequentialism of the cat's behavior is distributed between its brain and its evolutionary history".

In particular, I think there's an argument here that goes something like:

- Observe that, from our perspective, saving the world seems quite tricky, and seems likely to involve long sequences of clever actions that force the course of history into a narrow band (eg, because if we saw short sequences of dumb actions, we could just get started).
- Suppose we were presented with a plan that allegedly describes a long sequence of clever actions that would, if executed, force the course of history into some narrow band.
 - For concreteness, suppose it is a plan that allegedly funnels history into the band where we have wealth and acclaim.
- One plausible happenstance is that the plan is not in fact clever, and would not in fact have a forcing effect on history.
 - For example, perhaps the plan describes founding and managing some silicon valley startup, that would not work in practice.
- Conditional on the plan having the history-funnelling property, there's a sense in which it's scary regardless of its source.
 - For instance, perhaps the plan describes founding and managing some silicon valley startup, and will succeed virtually every time it's executed, by dint of having very generic descriptions of things like how to identify and respond

- to competition, including descriptions of methods for superhumanly-good analyses of how to psychoanalyze the competition and put pressure on their weakpoints.
- In particular, note that one need not believe the plan was generated by some "agent-like" cognitive system that, in a self-contained way, made use of reasoning we'd characterize as "possessing objectives" and "pursuing them in the real world".
 - More specifically, the scariness is a property of the plan itself. For instance, the fact that this plan accrues wealth and acclaim to the executor, in a wide variety of situations, regardless of what obstacles arise, implies that the plan contains course-correcting mechanisms that keep the plan on-target.
 - In other words, plans that *manage to actually funnel history* are (the argument goes) liable to have a wide variety of course-correction mechanisms that keep the plan oriented towards *some* target. And while this course-correcting property tends to be a property of history-funneling plans, the *choice of target* is of course free, hence the worry.

(Of course, in practice we perhaps shouldn't be visualizing a single Plan handed to us from an AI or a time machine or whatever, but should instead imagine a system that is reacting to contingencies and replanning in realtime. At the least, this task is easier, as one can adjust only for the contingencies that are beginning to arise, rather than needing to predict them all in advance and/or describe general contingency-handling mechanisms. But, and feel free to take a moment to predict my response before reading the next sentence, "run this AI that replans autonomously on-the-fly" and "run this AI+human loop that replans+reevaluates on the fly", are still in this sense "plans", that still likely have the property of Eliezer!consequentialism, insofar as they work.)

[Soares] (Sep. 12 Google Doc)

There's a part of this argument I have not yet driven home. Factoring it out into a separate bullet:

2. "If a plan is good enough to work, it's pretty consequentialist in practice".

In attempts to collect and distill a handful of scattered arguments of Eliezer's:

If you ask GPT-3 to generate you a plan for saving the world, it will not manage to generate one that is very detailed. And if you tortured a big language model into giving you a detailed plan for saving the world, the resulting plan would not work. In particular, it would be full of errors like insensitivity to circumstance, suggesting impossible actions, and suggesting actions that run entirely at cross-purposes to one another.

A plan that is sensitive to circumstance, and that describes actions that synergize rather than conflict -- like, in Eliezer's analogy, photons in a

laser -- is much better able to funnel history into a narrow band.

But, on Eliezer's view as I understand it, this "the plan is not constantly tripping over its own toes" property, goes hand-in-hand with what he calls "consequentialism". As a particularly stark and formal instance of the connection, observe that one way a plan can trip over its own toes is if it says "then trade 5 oranges for 2 apples, then trade 2 apples for 4 oranges". This is clearly an instance of the plan failing to "lase" -- of some orange-needing part of the plan working at cross-purposes to some apple-needing part of the plan, or something like that. And this is also a case where it's easy to see how if a plan *is* "lasing" with respect to apples and oranges, then it is behaving as if governed by some coherent preference.

And the point as I understand it isn't "all toe-tripping looks superficially like an inconsistent preference", but rather "insofar as a plan *does* manage to chain a bunch of synergistic actions together, it manages to do so precisely insofar as it is Eliezer!consequentialist".

cf the analogy to [information theory](#), where if you're staring at a maze and you're trying to build an accurate representation of that maze in your own head, you will succeed precisely insofar as your process is Bayesian / information-theoretic. And, like, this is supposed to feel like a fairly tautological claim: you (almost certainly) can't get the image of a maze in your head to match the maze in the world by visualizing a maze at random, you have to add visualized-walls using some process that's correlated with the presence of actual walls. Your maze-visualizing process will work precisely insofar as you have access to & correctly make use of, observations that correlate with the presence of actual walls. You might also visualize extra walls in locations where it's politically expedient to believe that there's a wall, and you might also avoid visualizing walls in a bunch of distant regions of the maze because it's dark and you haven't got all day, but the resulting visualization in your head is accurate precisely *insofar* as you're managing to act kinda like a Bayesian.

Similarly (the analogy goes), a plan works-in-concert and avoids-stepping-on-its-own-toes precisely insofar as it is consequentialist. These are two sides of the same coin, two ways of seeing the same thing.

And, I'm not so much attempting to *argue* the point here, as to make sure that the *shape of the argument* (as I understand it) has been understood by Richard. In particular, the *shape of the argument* I see Eliezer as making is that "clumsy" plans don't work, and "laser-like plans" work insofar as they are managing to act kinda like a consequentialist.

Rephrasing again: we have a wide variety of mathematical theorems all spotlighting, from different angles, the fact that a plan lacking in clumsiness, is possessing of coherence.

("And", my model of Eliezer is quick to note, "this ofc does not mean that all sufficiently intelligent minds must generate very-coherent plans. If you really knew what you were doing, you could design a mind that emits plans that always "trip over themselves" along one particular axis, just as

with sufficient mastery you could build a mind that believes $2+2=5$ (for some reasonable cashing-out of that claim). But you don't get this for free -- and there's a sort of "attractor" here, when building cognitive systems, where just as generic training will tend to cause it to have true beliefs, so will generic training tend to cause its plans to lase.")

(And ofc much of the worry is that all the mathematical theorems that suggest "this plan manages to work precisely insofar as it's lasing in some direction", say nothing about which direction it must lase. Hence, if you show me a plan clever enough to force history into some narrow band, I can be fairly confident it's doing a bunch of lasing, but not at all confident which direction it's lasing in.)

[Soares] (Sep. 12 Google Doc)

One of my guesses is that Richard does in fact understand this argument (though I personally would benefit from a paraphrase, to test this hypothesis!), and perhaps even buys it, but that Richard gets off the train at a following step, namely that we *need* plans that "lase", because ones that don't aren't strong enough to save us. (Where in particular, I suspect most of the disagreement is in how far one can get with plans that are more like language-model outputs and less like lasers, rather than in the question of which pivotal acts would put an end to the acute risk period)

But setting that aside for a moment, I want to use the above terminology to restate another point I saw Eliezer as attempting to make: one big trouble with alignment, in the case where we need our plans to be like lasers, is that on the one hand we need our plans to be like lasers, but on the other hand we want them to *fail* to be like lasers along certain specific dimensions.

For instance, the plan presumably needs to involve all sorts of mechanisms for refocusing the laser in the case where the environment contains fog, and redirecting the laser in the case where the environment contains mirrors (...the analogy is getting a bit strained here, sorry, bear with me), so that it can in fact hit a narrow and distant target. Refocusing and redirecting to stay on target are part and parcel to plans that can hit narrow distant targets.

But the humans shutting the AI down is like scattering the laser, and the humans tweaking the AI so that it plans in a different direction is like them tossing up mirrors that redirect the laser; and we want the plan to fail to correct for those interferences.

As such, on the Eliezer view as I understand it, we can see ourselves as asking for a very unnatural sort of object: a path-through-the-future that is robust enough to funnel history into a narrow band in a very wide array of circumstances, but somehow insensitive to specific breeds of human-initiated attempts to switch which narrow band it's pointed towards.

Ok. I meandered into trying to re-articulate the point over and over until I had a version distilled enough for my own satisfaction (which is much like arguing the point), apologies for the repetition.

I don't think debating the claim is the right move at the moment (though I'm happy to hear rejoinders!). Things I would like, though, are: Eliezer saying whether the above is on-track from his perspective (and if not, then poking a few holes); and Richard attempting to paraphrase the above, such that I believe the arguments themselves have been communicated (saying nothing about whether Richard also buys them).

[Soares] (Sep. 12 Google Doc)

My Richard-model's stance on the above points is something like "This all seems kinda plausible, but where Eliezer reads it as arguing that we had better figure out how to handle lasers, I read it as an argument that we'd better save the world without needing to resort to lasers. Perhaps if I thought the world could not be saved except by lasers, I would share many of your concerns, but I do not believe that, and in particular it looks to me like much of the recent progress in the field of AI -- from AlphaGo to GPT to AlphaFold -- is evidence in favor of the proposition that we'll be able to save the world without lasers."

And I recall actual-Eliezer saying the following (more-or-less in response, iiuc, though readers note that I might be misunderstanding and this might be out-of-context):

Definitely, "turns out it's easier than you thought to use gradient descent's memorization of zillions of shallow patterns that overlap and recombine into larger cognitive structures, to add up to a consequentialist nanoengineer that only does nanosystems and never does sufficiently general learning to apprehend the big picture containing humans, while still understanding the goal for that pivotal act you wanted to do" is among the more plausible advance-specified miracles we could get.

On my view, and I think on Eliezer's, the "zillions of shallow patterns"-style AI that we see today, is not going to be sufficient to save the world (nor destroy it). There's a bunch of reasons that GPT and AlphaZero aren't destroying the world yet, and one of them is this "shallowness" property. And, yes, maybe we'll be wrong! I myself have been surprised by how far the shallow pattern memorization has gone (and, for instance, was surprised by GPT), and acknowledge that perhaps I will continue to be surprised. But I continue to predict that the shallow stuff won't be enough.

I have the sense that lots of folk in the community are, one way or another, saying "Why not consider the problems of aligning systems that memorize zillions of shallow patterns?". And my answer is, "I still don't expect those sorts of machines to either kill or save us, I'm still expecting that there's a phase shift that won't happen until AI systems start to be able to make plans that are sufficiently deep and laserlike to do scary stuff, and I'm still expecting that the real alignment challenges are in that regime."

And this seems to me close to the heart of the disagreement: some people (like me!) have an intuition that it's quite unlikely that figuring out how to get sufficient work out of shallow-memorizers is enough to save us, and I suspect others (perhaps even Richard!) have the sense that the aforementioned "phase shift" is the unlikely scenario, and that I'm focusing on a weird and unlucky corner of the space. (I'm curious whether you endorse this, Richard, or some nearby correction of it.)

In particular, Richard, I am curious whether you endorse something like the following:

- I'm focusing ~all my efforts on the shallow-memorizers case, because I think shallow-memorizer-alignment will by and large be sufficient, and even if it is not then I expect it's a good way to prepare ourselves for whatever we'll turn out to need in practice. In particular I don't put much stock in the idea that there's a predictable phase-change that forces us to deal with laser-like planners, nor that predictable problems in that domain give large present reason to worry.

(I suspect not, at least not in precisely this form, and I'm eager for corrections.)

I suspect something in this vicinity constitutes a crux of the disagreement, and I would be thrilled if we could get it distilled down to something as concise as the above. And, for the record, I personally endorse the following counter to the above:

- I am focusing ~none of my efforts on shallow-memorizer-alignment, as I expect it to be far from sufficient, as I do not expect a singularity until we have more laser-like systems, and I think that the laserlike-planning regime has a host of predictable alignment difficulties that Earth does not seem at all prepared to face (unlike, it seems to me, the shallow-memorizer alignment difficulties), and as such I have large and present worries.

[Soares] (Sep. 12 Google Doc)

Ok, and now a few less substantial points:

There's a point Richard made here:

Oh, interesting. Actually one more question then: to what extent do you think that explicitly reasoning about utility functions and laws of rationality is what makes consequentialists have the properties you've been talking about?

that I suspect constituted a miscommunication, especially given that the following sentence appeared in Richard's summary:

A third thing that makes humans in particular consequentialist is planning, especially when we're aware of concepts like utility

functions.

In particular, I suspect Richard's model of Eliezer's model places (or placed, before Richard read Eliezer's comments on Richard's summary) some particular emphasis on systems reflecting and thinking about their own strategies, as a method by which the consequentialism and/or effectiveness gets in. I suspect this is a misunderstanding, and am happy to say more on my model upon request, but am hopeful that the points I made a few pages above have cleared this up.

Finally, I observe that there are a few places where Eliezer keeps beeping when Richard attempts to summarize him, and I suspect it would be useful to do the dorky thing of Richard very explicitly naming Eliezer's beeps as he understands them, for purposes of getting common knowledge of understanding. For instance, things I think it might be useful for Richard to say verbatim (assuming he believes them, which I suspect, and subject to Eliezer-corrections, b/c maybe I'm saying things that induce separate beeps):

1. Eliezer doesn't believe it's impossible to build AIs that have most any given property, including most any given safety property, including most any desired "non-consequentialist" or "deferential" property you might desire. Rather, Eliezer believes that many desirable safety properties don't happen by default, and require mastery of minds that likely takes a worrying amount of time to acquire.
2. The points about consequentialism are not particularly central in Eliezer's view; they seem to him more like obvious background facts; the reason conversation has lingered here in the EA-sphere is that this is a point that many folk in the local community disagree on.

For the record, I think it might also be worth Eliezer acknowledging that Richard probably understands point (1), and that glossing "you don't get it for free by default and we aren't on course to have the time to get it" as "you can't" is quite reasonable when summarizing. (And it might be worth Richard counter-acknowledging that the distinction is actually quite important once you buy the surrounding arguments, as it constitutes the difference between describing the current playing field and laying down to die.) I don't think any of these are high-priority, but they might be useful if easy :-)

Finally, stating the obvious-to-me, none of this is intended as criticism of either party, and all discussing parties have exhibited significant virtue-according-to-Nate throughout this process.

[Yudkowsky][21:27] (Sep. 12)

From Nate's notes:

For instance, the plan presumably needs to involve all sorts of mechanisms for refocusing the laser in the case where the environment contains fog, and redirecting the laser in the case where the environment contains mirrors (...the analogy is getting a bit strained here, sorry, bear with me), so that it can in fact hit a narrow and distant target. Refocusing and redirecting to stay on target are part and parcel to plans that can hit narrow distant targets.

But the humans shutting the AI down is like scattering the laser, and the humans tweaking the AI so that it plans in a different direction is like them tossing up mirrors that redirect the laser; and we want the plan to fail to correct for those interferences.

--> GOOD ANALOGY.

...or at least it sure conveys to *me* why corrigibility is anticonvergent / anticoherent / actually *moderately strongly contrary to* and not just *an orthogonal property* of a powerful-plan generator.

But then, I already know why that's true and how it generalized up to resisting our various attempts to solve small pieces of more important aspects of it - it's not just true by weak default, it's true by a stronger default where a roomful of people at a workshop spend several days trying to come up with increasingly complicated ways to describe a system that will let you shut it down (but not steer you through time *into* shutting it down), and all of those suggested ways get shot down. (And yes, people outside MIRI now and then publish papers saying they totally just solved this problem, but all of those "solutions" are things we considered and dismissed as trivially failing to scale to powerful agents - they didn't understand what we considered to be the first-order problems in the first place - rather than these being evidence that MIRI just didn't have smart-enough people at the workshop.)

[Yudkowsky][18:56] (Nov. 5 follow-up comment)

Eg, "Well, we took a system that only learned from reinforcement on situations it had previously been in, and couldn't use imagination to plan for things it had never seen, and then we found that if we didn't update it on shut-down situations it wasn't reinforced to avoid shutdowns!"

Ngo and Yudkowsky on AI capability gains

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is the second post in a series of transcribed conversations about AGI forecasting and alignment. See the [first post](#) for prefaces and more information about the format.

Color key:

Chat by Richard Ngo and Eliezer Yudkowsky

Other chat

Inline comments

5. September 14 conversation

5.1. Recursive self-improvement, abstractions, and miracles

[Yudkowsky][11:00]

Good morning / good evening.

So it seems like the obvious thread to pull today is your sense that I'm wrong about recursive self-improvement and consequentialism in a related way?

[Ngo][11:04]

Right. And then another potential thread (probably of secondary importance) is the question of what you mean by utility functions, and digging more into the intuitions surrounding those.

But let me start by fleshing out this RSI/consequentialism claim.

I claim that your early writings about RSI focused too much on a very powerful abstraction, of recursively applied optimisation; and too little on the ways in which even powerful abstractions like this one become a bit... let's say messier, when they interact with the real world.

In particular, I think that [Paul's arguments](#) that there will be substantial progress in AI in the leadup to a RSI-driven takeoff are pretty strong ones.

(Just so we're on the same page: to what extent did those arguments end up shifting your credences?)

[Yudkowsky][11:09]

I don't remember being shifted by Paul on this at all. I sure shifted a lot over events like Alpha Zero and the entire deep learning revolution. What does Paul say that isn't encapsulated in that update - does he furthermore claim that we're going to get fully smarter-than-human in all regards AI which doesn't cognitively scale much further either through more compute or through RSI?

[Ngo][11:10]

Ah, I see. In that case, let's just focus on the update from the deep learning revolution.

[Yudkowsky][11:12][11:13]

I'll also remark that I see my foreseeable mistake there as having little to do with "abstractions becoming messier when they interact with the real world" - this truism tells you very little of itself, unless you can predict *directional* shifts in other variables just by contemplating the *unknown* messiness relative to the abstraction.

Rather, I'd see it as a neighboring error to what I've called the Law of Earlier Failure, where the Law of Earlier Failure says that, compared to the interesting part of the problem where it's fun to imagine yourself failing, you usually fail before then, because of the many earlier boring points where it's possible to fail.

The nearby reasoning error in my case is that I focused on an interesting way that AI capabilities could scale and the most powerful argument I had to overcome Robin's objections, while missing the way that Robin's objections could fail even earlier through rapid scaling and generalization in a more boring way.

It doesn't mean that my arguments about RSI were false about their domain of supposed application, but that other things were also true and those things happened first on our timeline. To be clear, I think this is an important and generalizable issue with the impossible task of trying to forecast the Future, and if I am wrong about other things it sure would be plausible if I was wrong in similar ways.

[Ngo][11:13]

Then the analogy here is something like: there is a powerful abstraction, namely consequentialism; and we both agree that (like RSI) a large amount of consequentialism is a very dangerous thing. But we disagree on the question of how much the strategic landscape in the leadup to highly-consequentialist AIs is affected by other factors apart from this particular abstraction.

"this truism tells you very little of itself, unless you can predict directional shifts in other variables just by contemplating the unknown messiness relative to the abstraction"

I disagree with this claim. It seems to me that the predictable direction in which the messiness pushes is *away from* the applicability of the high-level abstraction.

[Yudkowsky][11:15]

The real world is messy, but good abstractions still apply, just with some messiness around them. The Law of Earlier Failure is not a failure of the abstraction being messy, it's a failure of the *subject matter* ending up different such that the abstractions you used were *about a different subject matter*.

When a company fails before the exciting challenge where you try to scale your app across a million users, because you couldn't hire enough programmers to build your app at all, the problem is not that you had an unexpectedly messy abstraction about scaling to many users, but that the key determinants were a different subject matter than "scaling to many users".

Throwing 10,000 TPUs at something and actually getting progress - not very much of a famous technological idiom *at the time I was originally arguing with Robin* - is not a leak in the RSI abstraction, it's just a way of getting powerful capabilities without RSI.

[Ngo][11:18]

To me the difference between these two things seems mainly semantic; does it seem otherwise to you?

[Yudkowsky][11:18]

If I'd been arguing with somebody who kept arguing in favor of faster timescales, maybe I'd have focused on that different subject matter and gotten a chance to be explicitly wrong about it. I mainly see my ur-failure here as letting myself be influenced by the whole audience that was nodding along very seriously to Robin's arguments, at the expense of considering how reality might depart in either direction from my own beliefs, and not just how Robin might be right or how to persuade the audience.

[Ngo][11:19]

Also, "throwing 10,000 TPUs at something and actually getting progress" doesn't seem like an example of the Law of Earlier Failure - if anything it seems like an Earlier Success

[Yudkowsky][11:19]

it's an Earlier Failure of Robin's arguments about why AI wouldn't scale quickly, so my lack of awareness of this case of the Law of Earlier Failure is why I didn't consider why Robin's arguments could fail earlier

though, again, this is a bit harder to call if you're trying to call it in 2008 instead of 2018

but it's a valid lesson that the future is, in fact, hard to predict, if you're trying to do it in the past

and I would not consider it a merely "semantic" difference as to whether you made a wrong argument about the correct subject matter, or a correct argument about the wrong subject matter

these are like... very different failure modes that you learn different lessons from

but if you're not excited by these particular fine differences in failure modes or lessons to learn from them, we should perhaps not dwell upon that part of the meta-level Art

[Ngo][11:21]

Okay, so let me see if I understand your position here.

Due to the deep learning revolution, it turned out that there were ways to get powerful capabilities without RSI. This isn't intrinsically a (strong) strike against the RSI abstraction; and so, unless we have reason to expect another similarly surprising revolution before reaching AGI, it's not a good reason to doubt the consequentialism abstraction.

[Yudkowsky][11:25]

Consequentialism and RSI are very different notions in the first place. Consequentialism is, in my own books, significantly simpler. I don't see much of a conceptual connection between the two myself, except insofar as they both happen to be part of the connected fabric of a coherent worldview about cognition.

It is entirely reasonable to suspect that we may get another surprising revolution before reaching AGI. Expecting a *particular* revolution that gives you *particular* miraculous benefits is much more questionable and is an instance of conjuring expected good from nowhere, like hoping that

you win the lottery because the first lottery ball comes up 37. (Also, if you sincerely believed you actually had info about what kind of revolution might lead to AGI, you should shut up about it and tell very few carefully selected people, not bake it into a public dialogue.)

[Ngo][11:28]

and I would not consider it a merely "semantic" difference as to whether you made a wrong argument about the correct subject matter, or a correct argument about the wrong subject matter

On this point: the implicit premise of "and also nothing else will break this abstraction or render it much less relevant" turns a correct argument about the wrong subject matter into an incorrect argument.

[Yudkowsky][11:28]

Sure.

Though I'd also note that there's an important lesson of technique where you learn to say things like that out loud instead of keeping them "implicit".

Learned lessons like that are one reason why I go through your summary documents of our conversation and ask for many careful differences of wording about words like "will happen" and so on.

[Ngo][11:30]

Makes sense.

So I claim that:

1. A premise like this is necessary for us to believe that your claims about consequentialism lead to extinction.
2. A surprising revolution would make it harder to believe this premise, even if we don't know which *particular* revolution it is.
3. If we'd been told back in 2008 that a surprising revolution would occur in AI, then we should have been less confident in the importance of the RSI abstraction to understanding AGI and AGI risk.

[Yudkowsky][11:32][11:34]

Suppose I put to you that this claim is merely subsumed by all of my previous careful qualifiers about how we might get a "miracle" and how we should be trying to prepare for an unknown miracle in any number of places. Why suspect that place particularly for a model-violation?

I also think that you are misinterpreting my old arguments about RSI, in a pattern that matches some other cases of your summarizing my beliefs as "X is the one big ultra-central thing" rather than "X is the point where the other person got stuck and Eliezer had to spend a lot of time arguing".

I was always claiming that RSI was a way for AGI capabilities to scale much further *once they got far enough*, not the way AI would scale to *human-level generality*.

This continues to be a key fact of relevance to my future model, in the form of the unfalsified original argument about the subject matter it previously applied to: if you lose control of a sufficiently smart AGI, it will Foom, and this fact about what triggers the metaphorical equivalent of a full nuclear exchange and a total loss of the gameboard continues to be extremely relevant to what you have to do to obtain victory instead.

[Ngo][11:34][11:35]

Perhaps we're interpreting the word "miracle" in quite different ways.

I think of it as an event with negligibly small probability.

[Yudkowsky][11:35]

Events that actually have negligibly small probability are not much use in plans.

[Ngo][11:35]

Which I guess doesn't fit with your claims that we should be trying to prepare for a miracle.

[Yudkowsky][11:35]

Correct.

[Ngo][11:35]

But I'm not recalling off the top of my head where you've claimed that.

I'll do a quick search of the transcript

"You need to hold your mind open for any miracle and a miracle you didn't expect or think of in advance, because at this point our last hope is that in fact the future is often quite surprising."

Okay, I see. The connotations of "miracle" seemed sufficiently strong to me that I didn't interpret "you need to hold your mind open" as practical advice.

What sort of probability, overall, do you assign to us being saved by what you call a miracle?

[Yudkowsky][11:40]

It's not a place where I find quantitative probabilities to be especially helpful.

And if I had one, I suspect I would not publish it.

[Ngo][11:41]

Can you leak a bit of information? Say, more or less than 10%?

[Yudkowsky][11:41]

Less.

Though a lot of that is dominated, not by the probability of a positive miracle, but by the extent to which we seem unprepared to take advantage of it, and so would not be saved by one.

[Ngo][11:41]

Yeah, I see.

5.2. The idea of expected utility

[Ngo][11:43]

Okay, I'm now significantly less confident about how much we actually disagree.

At least about the issues of AI cognition.

[Yudkowsky][11:44]

You seem to suspect we'll get a *particular* miracle having to do with "consequentialism", which means that although it might be a miracle to me, it wouldn't be a miracle to you.

There is something forbidden in my model that is not forbidden in yours.

[Ngo][11:45]

I think that's partially correct, but I'd call it more a *broad range of possibilities* in the rough direction of you being wrong about consequentialism.

[Yudkowsky][11:46]

Well, as much as it may be nicer to debate when the other person has a specific positive expectation that X will work, we can also debate when I know that X won't work and the other person remains ignorant of that. So say more!

[Ngo][11:47]

That's why I've mostly been trying to clarify your models rather than trying to make specific claims of my own.

Which I think I'd prefer to continue doing, if you're amenable, by asking you about what entities a utility function is defined over - say, in the context of a human.

[Yudkowsky][11:51][11:53]

I think that to contain the concept of Utility as it exists in me, you would have to do homework exercises I don't know how to prescribe. Maybe one set of homework exercises like that would be showing you an agent, including a human, making some set of choices that allegedly couldn't obey expected utility, and having you figure out how to pump money from that agent (or present it with money that it would pass up).

Like, just actually doing that a few dozen times.

Maybe it's not helpful for me to say this? If you say it to Eliezer, he immediately goes, "Ah, yes, I could see how I would update that way after doing the homework, so I will save myself some time and effort and just make that update now without the homework", but this kind of jumping-ahead-to-the-destination is something that seems to me to be... dramatically missing from many non-Eliezers. They insist on learning things the hard way and then act all surprised when they do. Oh my gosh, who would have thought that an AI breakthrough would suddenly make AI seem less than 100 years away the way it seemed yesterday? Oh my gosh, who would have thought that alignment would be difficult?

Utility can be seen as the origin of Probability within minds, even though Probability obeys its own, simpler coherence constraints.

that is, you will have money pumped out of you, unless you weigh in your

mind paths through time according to some quantitative weight, which determines how much resources you're willing to spend on preparing for them

this is why sapients think of things as being more or less likely

[Ngo][11:53]

Suppose that this agent has some high-level concept - say, honour - which leads it to pass up on offers of money.

[Yudkowsky][11:55]

Suppose that this agent has some high-level concept - say, honour - which leads it to pass up on offers of money.

then there's two possibilities:

- this concept of honor is something that you can see as helping to navigate a path through time to a destination
- honor isn't something that would be optimized into existence by optimization pressure for other final outcomes

[Ngo][11:55]

Right, I see.

Hmm, but it seems like humans often don't see concepts as helping to navigate a path in time to a destination. (E.g. the deontological instinct not to kill.)

And yet those concepts were in fact optimised into existence by evolution.

[Yudkowsky][11:59]

You're describing a defect of human reflectivity about their consequentialist structure, not a departure from consequentialist structure. 😊

[Ngo][12:01]

(Sorry, internet was slightly buggy; switched to a better connection now.)

[Yudkowsky][12:01]

But yes, from my perspective, it creates a very large conceptual gap that I can stare at something for a few seconds and figure out how to parse it

as navigating paths through time, while others think that "consequentialism" only happens when their minds are explicitly thinking about "well, what would have this consequence" using language.

Similarly, when it comes to Expected Utility, I see that any time something is attaching relative-planning-weights to paths through time, not when a human is thinking out loud about putting spoken numbers on outcomes

[Ngo][12:02]

Human consequentialist structure was optimised by evolution for a different environment. Insofar as we are consequentialists in a new environment, it's only because we're able to be reflective about our consequentialist structure (or because there are strong similarities between the environments).

[Yudkowsky][12:02]

False.

It just generalized out-of-distribution because the underlying coherence of the coherent behaviors was simple.

When you have a very simple pattern, it can generalize across weak similarities, not "strong similarities".

The human brain is large but the coherence in it is simple.

The idea, the structure, that explains why the big thing works, is much smaller than the big thing.

So it can generalize very widely.

[Ngo][12:04]

Taking this example of the instinct not to kill people - is this one of the "very simple patterns" that you're talking about?

[Yudkowsky][12:05]

"Reflectivity" doesn't help per se unless on some core level a pattern already generalizes, I mean, either a truth can generalize across the data or it can't? So I'm a bit puzzled about why you're bringing up "reflectivity" in this context.

And, no.

An instinct not to kill doesn't even seem to me like a plausible cross-cultural universal. 40% of deaths among Yanomami men are in intratribal

fights, iirc.

[Ngo][12:07]

Ah, I think we were talking past each other. When you said "this concept of honor is something that you can see as helping to navigate a path through time to a destination" I thought you meant "you" as in the agent in question (as you used it in some previous messages) not "you" as in a hypothetical reader.

[Yudkowsky][12:07]

ah.

it would not have occurred to me to ascribe that much competence to an agent that wasn't a superintelligence.

even I don't have time to think about why more than ~~0.0001%~~ 0.01% of my thoughts do anything, but thankfully, you don't have to think about *why* $2 + 2 = 4$ for it to be the correct answer for counting sheep.

[Ngo][12:10]

Got it.

I might now try to throw a high-level (but still inchoate) disagreement at you and see how that goes. But while I'm formulating that, I'm curious what your thoughts are on where to take the discussion.

Actually, let's spend a few minutes deciding where to go next, and then take a break

I'm thinking that, at this point, there might be more value in moving onto geopolitics

[Yudkowsky][12:19]

Some of my current thoughts are a reiteration of old despair: It feels to me like the typical Other within EA has no experience with discovering unexpected order, with operating a generalization that you can expect will cover new cases even when that isn't immediately obvious, with operating that generalization to cover those new cases correctly, with seeing simple structures that generalize a lot and having that be a real and useful and technical experience; instead of somebody blathering in a non-expectation-constraining way about how "capitalism is responsible for everything wrong with the world", and being able to extend that to lots of cases.

I could try to use much simpler language in hopes that people actually [look-at-the-water](#) Feynman-style, like "navigating a path through time"

instead of Consequentialism which is itself a step down from Expected Utility.

But you actually do lose something when you throw away the more technical concept. And then people still think that either you instantly see in the first second how something is a case of "navigating a path through time", or that this is something that people only do explicitly when visualizing paths through time using that mental terminology; or, if Eliezer says that it's "navigating time" anyways, this must be an instance of Eliezer doing that thing other people do when they talk about how "Capitalism is responsible for all the problems of the world". They have no experience operating genuinely useful, genuinely deep generalizations that extend to nonobvious things.

And in fact, being able to operate some generalizations like that is a lot of how I know what I know, in reality and in terms of the original knowledge that came before trying to argue that knowledge with people. So trying to convey the real source of the knowledge feels doomed. It's a kind of idea that our civilization has lost, like that college class Feynman ran into.

[Soares][12:19]

My own sense (having been back for about 20min) is that one of the key cruxes is in "is it possible that non-scary cognition will be able to end the acute risk period", or perhaps "should we expect a longish regime of pre-scary cognition, that we can study and learn to align in such a way that by the time we get scary cognition we can readily align it".

[Ngo][12:19]

Some potential prompts for that:

- what are some scary things which might make governments take AI more seriously than they took covid, and which might happen before AGI
- how much of a bottleneck in your model is governmental competence? and how much of a difference do you see in this between, say, the US and China?

[Soares][12:20]

I also have a bit of a sense that there's a bit more driving to do on the "perhaps EY is just wrong about the applicability of the consequentialism arguments" (in a similar domain), and would be happy to try articulating a bit of what I think are the not-quite-articulated-to-my-satisfaction arguments on that side.

[Yudkowsky][12:21]

I also had a sense - maybe mistaken - that RN did have some *specific* ideas about how "consequentialism" might be inapplicable, though maybe I accidentally refuted that in passing because the idea was "well, what if it didn't know what consequentialism was?" and then I explained that reflectivity was not required to make consequentialism generalize. but if so, I'd like RN to say explicitly what specific idea got refuted that way, or failing that, talk about the specific idea that didn't get refuted.

[Ngo][12:23]

That wasn't my objection, but I do have some more specific ideas, which I could talk about.

And I'd also be happy for Nate to try articulating some of the arguments he mentioned above.

[Yudkowsky][12:23]

I have a general worry that this conversation has gotten too general, and that it would be more productive, even of general understanding, to start from specific ideas and shoot those down specifically.

[Ngo: 

[Ngo][12:26]

The other thing is that, for pedagogical purposes, I think it'd be useful for you to express some of your beliefs about how governments will respond to AI

I think I have a rough guess about what those beliefs are, but even if I'm right, not everyone who reads this transcript will be

[Yudkowsky][12:28]

Why would I be expected to know *that*? I could talk about weak defaults and iterate through an unending list of possibilities.

Thinking that Eliezer thinks he knows that to any degree of specificity feels like I'm being weakmanned!

[Ngo][12:28]

I'm not claiming you have any specific beliefs

[Yudkowsky][12:29]

I suppose I have skepticism when other people dream up elaborately positive and beneficial reactions apparently drawn from some alternate nicer political universe that had an absolutely different response to Covid-19, and so on.

[Ngo][12:29]

But I'd guess that your models rule out, for instance, the US and China deeply cooperating on AI before it's caused any disasters

[Yudkowsky][12:30]

"Deeply"? Sure. That sounds like something that has never happened, and I'm generically skeptical about political things that go better than any political thing has ever gone before.

I guess we could talk about that? It doesn't seem like the most productive area, but maybe it lies upstream of more technical disagreements because we disagree about what AGI would actually have to do to have the world not end.

[Soares][12:31]

Cool. I claim it's time for a break, and then I nominate a little Eliezer gov't-response-overview followed by specific maybe-consequentialism-based-worries-aren't-a-problem-in-practice ideas from Richard.

[Ngo: ]

[Yudkowsky][12:32]

See you in 28mins

5.3. Epistemology, and assessing the idea of expected utility

[Ngo][13:03]

Some of my current thoughts are a reiteration of old despair: It feels to me like the typical Other within EA has no experience with discovering unexpected order, with operating a generalization that

you can expect will cover new cases even when that isn't immediately obvious, with operating that generalization to cover those new cases correctly, with seeing simple structures that generalize a lot and having that be a real and useful and technical experience; instead of somebody blathering in a non-expectation-constraining way about how "capitalism is responsible for everything wrong with the world", and being able to extend that to lots of cases.

I could try to use much simpler language in hopes that people actually look-at-the-water Feynman-style, like "navigating a path through time" instead of Consequentialism which is itself a step down from Expected Utility.

But you actually do lose something when you throw away the more technical concept. And then people still think that either you instantly see in the first second how something is a case of "navigating a path through time", or that this is something that people only do explicitly when visualizing paths through time using that mental terminology; or, if Eliezer says that it's "navigating time" anyways, this must be an instance of Eliezer doing that thing other people do when they talk about how "Capitalism is responsible for all the problems of the world". They have no experience operating genuinely useful, genuinely deep generalizations that extend to nonobvious things.

And in fact, being able to operate some generalizations like that is a lot of how I know what I know, in reality and in terms of the original knowledge that came before trying to argue that knowledge with people. So trying to convey the real source of the knowledge feels doomed. It's a kind of idea that our civilization has lost, like that college class Feynman ran into.

Ooops, didn't see this comment earlier. With respect to discovering unexpected order, one point that seems relevant is the extent to which that order provides predictive power. To what extent do you think that predictive successes in economics are important evidence for expected utility theory being a powerful formalism? (Or are there other ways in which it's predictively powerful that provide significant evidence?)

I'd be happy with a quick response to that, and then on geopolitics, here's a prompt to kick us off:

- If the only two actors involved in AGI development were the US and the UK governments, how much safer (or less safe) would you think we were compared with a world in which the two actors are the US and Chinese governments? How about a world in which the US government was a decade ahead of everyone else in reaching AGI?

[Yudkowsky][13:06]

I think that the Apollo space program is much deeper evidence for Utility. Observe, if you train protein blobs to run around the savanna, they also go to the moon!

If you think of "utility" as having something to do with the human discipline called "economics" then you are still thinking of it in a *much much much* more narrow way than I do.

[Ngo][13:07]

I'm not asking about evidence for utility as an abstraction in general, I'm asking for evidence based on successful predictions that have been made using it.

[Yudkowsky][13:10]

That doesn't tend to happen a lot, because all of the deep predictions that it makes are covered by shallow predictions that people made earlier.

Consider the following prediction of evolutionary psychology: Humans will enjoy activities associated with reproduction!

"What," says Simplicio, "you mean like dressing up for dates? I don't enjoy that part."

"No, you're overthinking it, we meant orgasms," says the evolutionary psychologist.

"But I already knew that, that's just common sense!" replies Simplicio.

"And yet it is very specifically a prediction of evolutionary psychology which is not made specifically by any other theory of human minds," replies the evolutionary psychologist.

"Not an advance prediction, just-so story, too obvious," replies Simplicio.

[Ngo][13:11]

Yepp, I agree that most of its predictions won't be new. Yet evolution is a sufficiently powerful theory that people have still come up with a range of novel predictions that derive from it.

Insofar as you're claiming that expected utility theory is also very powerful, then we should expect that it also provides some significant predictions.

[Yudkowsky][13:12]

An advance prediction of the notion of Utility, I suppose, is that if you train an AI which is otherwise a large blob of layers - though this may be inadvisable for other reasons - to the point where it starts solving lots of novel problems, that AI will tend to value aspects of outcomes with weights, and weight possible paths through time (the dynamic progress

of the environment), and use (by default, usually, roughly) the multiplication of these weights to allocate limited resources between mutually conflicting plans.

[Ngo][13:13]

Again, I'm asking for evidence in the form of successful predictions.

[Yudkowsky][13:14]

I predict that people will want some things more than others, think some possibilities are more likely than others, and prefer to do things that lead to stuff they want a lot through possibilities they think are very likely!

[Ngo][13:15]

It would be very strange to me if a theory which makes such strong claims about things we can't yet verify can't shed light on *anything* which we are in a position to verify.

[Yudkowsky][13:15]

If you think I'm deriving my predictions of catastrophic alignment failure through something *more exotic* than that, you're missing the reason *why I'm so worried*. It doesn't take intricate complicated exotic assumptions.

It makes the same kind of claims about things we can't verify yet as it makes about things we can verify right now.

[Ngo][13:16]

But that's very easy to do! Any theory can do that.

[Yudkowsky][13:17]

For example, if somebody wants money, and you set up a regulation which prevents them from making money, it predicts that the person will look for a new way to make money that bypasses the regulation.

[Ngo][13:17]

And yes, of course fitting previous data is important evidence in favour of a theory

[Yudkowsky][13:17]

[But that's very easy to do! Any theory can do that.]

False! Any theory can do that in the hands of a fallible agent which invalidly, incorrectly derives predictions from the theory.

[Ngo][13:18]

Well, indeed. But the very point at hand is whether the predictions you base on this theory are correctly or incorrectly derived.

[Yudkowsky][13:18]

It is not the case that every theory does an equally good job of predicting the past, given valid derivations of predictions.

Well, hence the analogy to evolutionary psychology. If somebody doesn't see the blatant obviousness of how sexual orgasms are a prediction specifically of evolutionary theory, because it's "common sense" and "not an advance prediction", what are you going to do? We can, in this case, with a *lot* more work, derive more detailed advance predictions about degrees of wanting that correlate in detail with detailed fitness benefits. But that's not going to convince anybody who overlooked the really blatant and obvious primary evidence.

What they're missing there is a sense of counterfactuals, of how the universe could just as easily have looked if the evolutionary origins of psychology were false: why should organisms want things associated with reproduction, why not instead have organisms running around that want things associated with rolling down hills?

Similarly, if optimizing complicated processes for outcomes hard enough, didn't produce cognitive processes that internally mapped paths through time and chose actions conditional on predicted outcomes, human beings would... not think like that? What am I supposed to say here?

[Ngo][13:24]

Let me put it this way. There are certain traps that, historically, humans have been very liable to fall into. For example, seeing a theory, which seems to match so beautifully and elegantly the data which we've collected so far, it's very easy to dramatically overestimate how much that data favours that theory. Fortunately, science has a very powerful social technology for avoiding this (i.e. making falsifiable predictions) which seems like approximately the only reliable way to avoid it - and yet you don't seem concerned at all about the lack of application of this technology to expected utility theory.

[Yudkowsky][13:25]

This is territory I covered in the Sequences, exactly because "well it didn't make a good enough advance prediction yet!" is an excuse that people use to reject evolutionary psychology, some other stuff I covered in the Sequences, and some very predictable lethaliies of AGI.

[Ngo][13:26]

With regards to evolutionary psychology: yes, there are some blatantly obvious ways in which it helps explain the data available to us. But there are also many people who have misapplied or overapplied evolutionary psychology, and it's very difficult to judge whether they have or have not done so, without asking them to make advance predictions.

[Yudkowsky][13:26]

I talked about the downsides of allowing humans to reason like that, the upsides, the underlying theoretical laws of epistemology (which are clear about why agents that reason validly or just unbiasedly would do that without the slightest hiccup), etc etc.

In the case of the theory "people want stuff relatively strongly, predict stuff relatively strongly, and combine the strengths to choose", what kind of advance prediction that no other theory could possibly make, do you expect that theory to make?

In the worlds where that theory is true, how should it be able to prove itself to you?

[Ngo][13:28]

I expect deeper theories to make more and stronger predictions.

I'm currently pretty uncertain if expected utility theory is a deep or shallow theory.

But deep theories tend to shed light in all sorts of unexpected places.

[Yudkowsky][13:30]

The fact is, when it comes to AGI (general optimization processes), we have only two major datapoints in our dataset, natural selection and humans. So you can either try to reason validly about what theories predict about natural selection and humans, even though we've already seen the effects of those; or you can claim to give up in great humble [modesty](#) while actually using other implicit theories instead to make all your predictions and be confident in them.

[Ngo][13:30]

I talked about the downsides of allowing humans to reason like that, the upsides, the underlying theoretical laws of epistemology (which are clear about why agents that reason validly or just unbiasedly would do that without the slightest hiccup), etc etc.

I'm familiar with your writings on this, which is why I find myself surprised here. I could understand a perspective of "yes, it's unfortunate that there are no advanced predictions, it's a significant weakness, I wish more people were doing this so we could better understand this vitally important theory". But that seems very different from your perspective here.

[Yudkowsky][13:32]

Oh, I'd love to be making predictions using a theory that made super detailed advance predictions made by no other theory which had all been borne out by detailed experimental observations! I'd also like ten billion dollars, a national government that believed everything I honestly told them about AGI, and a drug that raises IQ by 20 points.

[Ngo][13:32]

The very fact that we have only two major datapoints is exactly why it seems like such a major omission that a theory which purports to describe intelligent agency has not been used to make any successful predictions about the datapoints we do have.

[Yudkowsky][13:32][13:33]

This is making me think that you imagine the theory as something much more complicated and narrow than it is.

Just look at the water.

Not very special water with an index.

Just regular water.

People want stuff. They want some things more than others. When they do stuff they expect stuff to happen.

These are *predictions of the theory*. Not advance predictions, but predictions nonetheless.

[Ngo][13:33][13:33]

I'm accepting your premise that it's something deep and fundamental, and making the claim that deep, fundamental theories are likely to have a wide range of applications, including ones we hadn't previously thought of.

Do you disagree with that premise, in general?

[Yudkowsky][13:36]

I don't know what you really mean by "deep fundamental theory" or "wide range of applications we hadn't previously thought of", especially when it comes to structures that are this simple. It sounds like you're still imagining something I mean by Expected Utility which is some narrow specific theory like a particular collection of gears that are appearing in lots of places.

Are numbers a deep fundamental theory?

Is addition a deep fundamental theory?

Is probability a deep fundamental theory?

Is the notion of the syntax-semantics correspondence in logic and the notion of a generally semantically valid reasoning step, a deep fundamental theory?

[Ngo][13:38]

Yes to the first three, all of which led to very successful novel predictions.

[Yudkowsky][13:38]

What's an example of a novel prediction made by the notion of probability?

[Ngo][13:38]

Most applications of the central limit theorem.

[Yudkowsky][13:39]

Then I should get to claim every kind of optimization algorithm which used expected utility, as a successful advance prediction of expected utility? Optimal stopping and all the rest? Seems cheap and indeed invalid to me, and not particularly germane to whether these things appear inside AGIs, but if that's what you want, then sure.

[Ngo][13:39]

These are *predictions of the theory*. Not advance predictions, but predictions nonetheless.

I agree that it is a prediction of the theory. And yet it's also the case that smarter people than either of us have been dramatically mistaken about how well theories fit previously-collected data. (Admittedly we have advantages which they didn't, like a better understanding of cognitive biases - but it seems like you're ignoring the possibility of those cognitive biases applying to us, which largely negates those advantages.)

[Yudkowsky][13:42]

I'm not ignoring it, just adjusting my confidence levels and proceeding, instead of getting stuck in an infinite epistemic trap of self-doubt.

I don't live in a world where you either have the kind of detailed advance experimental predictions that should convince the most skeptical scientist and render you immune to all criticism, or, alternatively, you are suddenly in a realm beyond the reach of all epistemic authority, and you ought to cuddle up into a ball and rely only on wordless intuitions and trying to put equal weight on good things happening and bad things happening.

I live in a world where I proceed with very strong confidence if I have a detailed formal theory that made detailed correct advance predictions, and otherwise go around saying, "well, it sure looks like X, but we can be on the lookout for a miracle too".

If this was a matter of thermodynamics, I wouldn't even be talking like this, and we wouldn't even be having this debate.

I'd just be saying, "Oh, that's a perpetual motion machine. You can't build one of those. Sorry." And that would be the end.

Meanwhile, political superforecasters go on making well-calibrated predictions about matters much murkier and more complicated than these, often without anything resembling a clearly articulated theory laid forth at length, let alone one that had made specific predictions even retrospectively. They just go do it instead of feeling helpless about it.

[Ngo][13:45]

Then I should get to claim every kind of optimization algorithm which used expected utility, as a successful advance prediction of expected utility? Optimal stopping and all the rest? Seems cheap and indeed invalid to me, and not particularly germane to whether these things appear inside AGIs, but if that's what you want, then sure.

These seem better than nothing, but still fairly unsatisfying, insofar as I think they are related to more shallow properties of the theory.

Hmm, I think you're mischaracterising my position. I nowhere advocated for feeling helpless or curling up in a ball. I was just noting that this is a particularly large warning sign which has often been valuable in the past,

and it seemed like you were not only speeding past it blithely, but also denying the existence of this category of warning signs.

[Yudkowsky][13:48]

I think you're looking for some particular kind of public obeisance that I don't bother to perform internally because I'd consider it a wasted motion. If I'm lost in a forest I don't bother going around loudly talking about how I need a forest theory that makes detailed advance experimental predictions in controlled experiments, but, alas, I don't have one, so now I should be very humble. I try to figure out which way is north.

When I have a guess at a northerly direction, it would then be an error to proceed with as much confidence as if I'd had a detailed map and had located myself upon it.

[Ngo][13:49]

Insofar as I think we're less lost than you do, then the weaknesses of whichever forest theory implies that we're lost are relevant for this discussion.

[Yudkowsky][13:49]

The obeisance I make in that direction is visible in such statements as, "But this, of course, is a prediction about the future, which is well-known to be quite difficult to predict, in fact."

If my statements had been matters of thermodynamics and particle masses, I would *not* be adding that disclaimer.

But most of life is not a statement about particle masses. I have some idea of how to handle that. I do not need to constantly recite disclaimers to myself about it.

I know how to proceed when I have only a handful of data points which have already been observed and my theories of them are retrospective theories. This happens to me on a daily basis, eg when dealing with human beings.

[Soares][13:50]

(I have a bit of a sense that we're going in a circle. It also seems to me like there's some talking-past happening.)

(I suggest a 5min break, followed by EY attempting to paraphrase RN to his satisfaction and vice versa.)

[Yudkowsky][13:51]

I'd have more trouble than usual paraphrasing RN because epistemic helplessness is something I find painful to type out.

[Soares][13:51]

(I'm also happy to attempt to paraphrase each point as I see it; it may be that this smooths over some conversational wrinkle.)

[Ngo][13:52]

Seems like a good suggestion. I'm also happy to move on to the next topic. This was meant to be a quick clarification.

[Soares][13:52]

nod. It does seem to me like it possibly contains a decently sized meta-crux, about what sorts of conclusions one is licensed to draw from what sorts of observations

that, eg, might be causing Eliezer's probabilities to concentrate but not Richard's.

[Yudkowsky][13:52]

Yeah, this is in the opposite direction of "more specificity".

[Soares: ☺] [Ngo: ☺]

I frankly think that most EAs suck at explicit epistemology, OpenPhil and FHI affiliated EAs are not much of an exception to this, and I expect I will have more luck talking people out of specific errors than talking them out of the infinite pit of humble ignorance considered abstractly.

[Soares][13:54]

Ok, that seems to me like a light bid to move to the next topic from both of you, my new proposal is that we take a 5min break and then move to the next topic, and perhaps I'll attempt to paraphrase each point here in my notes, and if there's any movement in the comments there we can maybe come back to it later.

[Ngo: 👍]

[Ngo][13:54]

Broadly speaking I am also strongly against humble ignorance (albeit to a lesser extent than you are).

[Yudkowsky][13:55]

I'm off to take a 5-minute break, then!

5.4. Government response and economic impact

[Ngo][14:02]

A meta-level note: I suspect we're around the point of hitting significant diminishing marginal returns from this format. I'm open to putting more time into the debate (broadly construed) going forward, but would probably want to think a bit about potential changes in format.

[Soares][14:04, moved two up in log]

A meta-level note: I suspect we're around the point of hitting significant diminishing marginal returns from this format. I'm open to putting more time into the debate (broadly construed) going forward, but would probably want to think a bit about potential changes in format.

(Noted, thanks!)

[Yudkowsky][14:03]

I actually think that may just be a matter of at least one of us, including Nate, having to take on the thankless job of shutting down all digressions into abstractions and the meta-level.

[Ngo][14:05]

I actually think that may just be a matter of at least one of us, including Nate, having to take on the thankless job of shutting down all digressions into abstractions and the meta-level.

I'm not so sure about this, because it seems like some of the abstractions are doing a lot of work.

[Yudkowsky][14:03][14:04]

Anyways, government reactions?

It seems to me like the best observed case for government reactions - which I suspect is no longer available in the present era as a possibility - was the degree of cooperation between the USA and Soviet Union about avoiding nuclear exchanges.

This included such incredibly extravagant acts of cooperation as installing a direct line between the President and Premier!

which is not what I would really characterize as very "deep" cooperation, but it's more than a lot of cooperation you see nowadays.

More to the point, both the USA and Soviet Union proactively avoided doing anything that might lead towards starting down a path that led to a full nuclear exchange.

[Ngo][14:04]

The question I asked earlier:

- If the only two actors involved in AGI development were the US and the UK governments, how much safer (or less safe) would you think we were compared with a world in which the two actors are the US and Chinese governments? How about a world in which the US government was a decade ahead of everyone else in reaching AGI?

[Yudkowsky][14:05]

They still provoked one another a lot, but, whenever they did so, tried to do so in a way that wouldn't lead to a full nuclear exchange.

It was mutually understood to be a strategic priority and lots of people on both sides thought a lot about how to avoid it.

I don't know if that degree of cooperation ever got to the fantastic point of having people from *both* sides in the *same* room brainstorming *together* about how to avoid a full nuclear exchange, because that is, like, more cooperation than you would normally expect from two governments, but it wouldn't *shock* me to learn that this had ever happened.

It seems obvious to me that if some situation developed nowadays which increased the profile possibility of a nuclear exchange between the USA and Russia, we would not currently be able to do anything like installing a Hot Line between the US and Russian offices if such a Hot Line had not already been installed. This is lost social technology from a lost golden

age. But still, it's not unreasonable to take this as the upper bound of attainable cooperation; it's been observed within the last 100 years.

Another guess for how governments react is a very simple and robust one backed up by a huge number of observations:

They don't.

They have the same kind of advance preparation and coordination around AGI, in advance of anybody getting killed, as governments had around the mortgage crisis of 2007 in advance of any mortgages defaulting.

I am not sure I'd put this probability over 50% but it's certainly by far the largest probability over any competitor possibility specified to an equally low amount of detail.

I would expect anyone whose primary experience was with government, who was just approaching this matter and hadn't been talked around to weird exotic views, to tell you the same thing as a matter of course.

[Ngo][14:10]

But still, it's not unreasonable to take this as the upper bound of attainable cooperation; it's been observed within the last 100 years.

Is this also your upper bound conditional on a world that has experienced a century's worth of changes within a decade, and in which people are an order of magnitude wealthier than they currently are?

I am not sure I'd put this probability over 50% but it's certainly by far the largest probability over any competitor possibility specified to an equally low amount of detail.

which one was this? US/UK?

[Yudkowsky][14:12][14:14]

Assuming governments do react, we have the problem of "What kind of heuristic could have correctly led us to forecast that the US's reaction to a major pandemic would be for the FDA to ban hospitals from doing in-house Covid tests? What kind of mental process could have led us to make that call?" And we couldn't have gotten it exactly right, because the future is hard to predict; the best heuristic I've come up with, that feels like it at least would not have been *surprised* by what actually happened, is, "The government will react with a flabbergasting level of incompetence, doing exactly the wrong thing, in some unpredictable specific way."

which one was this? US/UK?

I think if we're talking about any single specific government like the US or UK then the probability is over 50% that they don't react in any advance

coordinated way to the AGI crisis, *to a greater and more effective degree* than they "reacted in an advance coordinated way" to pandemics before 2020 or mortgage defaults before 2007.

Maybe *some* two governments somewhere on Earth will have a high-level discussion between two cabinet officials.

[Ngo][14:14]

That's one lesson you could take away. Another might be: governments will be very willing to restrict the use of novel technologies, even at colossal expense, in the face of even a small risk of large harms.

[Yudkowsky][14:15]

That's one lesson you could take away. Another might be: governments will be very willing to restrict the use of novel technologies, even at colossal expense, in the face of even a small risk of large harms.

I just... don't know what to do when people talk like this.

It's so absurdly, absurdly optimistic.

It's taking a massive massive failure and trying to find exactly the right abstract gloss to put on it that makes it sound like exactly the right perfect thing will be done next time.

This just - isn't how to understand reality.

This isn't how superforecasters think.

This isn't sane.

[Soares][14:16]

(be careful about ad hominem)

(Richard might not be doing the insane thing you're imagining, to generate that sentence, etc)

[Ngo][14:17]

Right, I'm not endorsing this as my mainline prediction about what happens. Mainly what I'm doing here is highlighting that your view seems like one which cherrypicks *pessimistic* interpretations.

[Yudkowsky][14:18]

That abstract description "governments will be very willing to restrict the use of novel technologies, even at colossal expense, in the face of even a small risk of large harms" does not in fact apply very well to the FDA banning hospitals from using their well-established in-house virus tests, at risk of the alleged harm of some tests giving bad results, when in fact the CDC's tests were giving bad results and much larger harms were on the way because of bottlenecked testing; and that abstract description should have applied to an effective and globally coordinated ban against gain-of-function research, which *didn't* happen.

[Ngo][14:19]

Alternatively: what could have led us to forecast that many countries will impose unprecedentedly severe lockdowns.

[Yudkowsky][14:19][14:21][14:21]

Well, I didn't! I didn't even realize that was an option! I thought Covid was just going to rip through everything.

(Which, to be clear, it still may, and Delta arguably is in the more primitive tribal areas of the USA, as well as many other countries around the world that can't afford vaccines financially rather than epistemically.)

But there's a really really basic lesson here about the different style of "sentences found in political history books" rather than "sentences produced by people imagining ways future politics could handle an issue successfully".

Reality is *so much worse* than people imagining what might happen to handle an issue successfully.

[Ngo][14:21][14:21][14:22]

I might nudge us away from covid here, and towards the questions I asked before.

The question I asked earlier:

- If the only two actors involved in AGI development were the US and the UK governments, how much safer (or less safe) would you think we were compared with a world in which the two actors are the US and Chinese governments? How about a world in which the US government was a decade ahead of everyone else in reaching AGI?

This being one.

"But still, it's not unreasonable to take this as the upper bound of

attainable cooperation; it's been observed within the last 100 years." Is this also your upper bound conditional on a world that has experienced a century's worth of changes within a decade, and in which people are an order of magnitude wealthier than they currently are?

And this being the other.

[Yudkowsky][14:22]

Is this also your upper bound conditional on a world that has experienced a century's worth of changes within a decade, and in which people are an order of magnitude wealthier than they currently are?

I don't expect this to happen at all, or even come remotely close to happening; I expect AGI to kill everyone before self-driving cars are commercialized.

[Yudkowsky][16:29] (Nov. 14 follow-up comment)

(This was incautiously put; maybe strike "expect" and put in "would not be the least bit surprised if" or "would very tentatively guess that".)

[Ngo][14:23]

ah, I see

Okay, maybe here's a different angle which I should have been using. What's the most impressive technology you expect to be commercialised before AGI kills everyone?

[Yudkowsky][14:24]

If the only two actors involved in AGI development were the US and the UK governments, how much safer (or less safe) would you think we were compared with a world in which the two actors are the US and Chinese governments?

Very hard to say; the UK is friendlier but less grown-up. We would obviously be VASTLY safer in any world where only two centralized actors (two effective decision processes) could ever possibly build AGI, though not safe / out of the woods / at over 50% survival probability.

How about a world in which the US government was a decade ahead of everyone else in reaching AGI?

Vastly safer and likewise impossibly miraculous, though again, not out of the woods at all / not close to 50% survival probability.

What's the most impressive technology you expect to be commercialised before AGI kills everyone?

This is incredibly hard to predict. If I actually had to predict this for some reason I would probably talk to Gwern and Carl Shulman. In principle, there's nothing preventing me from knowing something about Go which lets me predict in 2014 that Go will probably fall in two years, but in practice I did not do that and I don't recall anybody else doing it either. It's really quite hard to figure out how much cognitive work a domain requires and how much work known AI technologies can scale to with more compute, let alone predict AI breakthroughs.

[Ngo][14:27]

I'd be happy with some very rough guesses

[Yudkowsky][14:27]

If you want me to spin a scifi scenario, I would not be surprised to find online anime companions carrying on impressively humanlike conversations, because this is a kind of technology that can be deployed without major corporations signing on or regulatory approval.

[Ngo][14:28]

Okay, this is surprising; I expected something more advanced.

[Yudkowsky][14:29]

Arguably AlphaFold 2 is already more advanced than that, along certain dimensions, but it's no coincidence that afaik people haven't really done much with AlphaFold 2 and it's made no visible impact on GDP.

I expect GDP not to depart from previous trendlines before the world ends, would be a more general way of putting it.

[Ngo][14:29]

What's the ~~most~~ least impressive technology that your model strongly rules out happening before AGI kills us all?

[Yudkowsky][14:30]

you mean least impressive?

[Ngo][14:30]

oops, yes

That seems like a structurally easier question to answer

[Yudkowsky][14:30]

"Most impressive" is trivial. "Dyson Spheres" answers it.

Or, for that matter, "perpetual motion machines".

[Ngo][14:31]

Ah yes, I was thinking that Dyson spheres were a bit too prosaic

[Yudkowsky][14:32]

My model mainly rules out that we get to certain points and then hang around there for 10 years while the technology gets perfected, commercialized, approved, adopted, ubiquitized enough to produce a visible trendline departure on the GDP graph; not so much various technologies themselves being initially demonstrated in a lab.

I expect that the people who build AGI can build a self-driving car if they want to. Getting it approved and deployed before the world ends is quite another matter.

[Ngo][14:33]

OpenAI has commercialised GPT-3

[Yudkowsky][14:33]

Hasn't produced much of a bump in GDP as yet.

[Ngo][14:33]

I wasn't asking about that, though

I'm more interested in judging how hard you think it is for AIs to take over the world

[Yudkowsky][14:34]

I note that it seems to me like there is definitely a kind of thinking here, which, if told about GPT-3 five years ago, would talk in very serious tones about how much this technology ought to be predicted to shift GDP, and whether we could bet on that.

By "take over the world" do you mean "turn the world into paperclips" or "produce 10% excess of world GDP over predicted trendlines"?

[Ngo][14:35]

Turn world into paperclips

[Yudkowsky][14:36]

I expect this mainly happens as a result of superintelligence, which is way up in the stratosphere far above the minimum required cognitive capacities to get the job done?

The interesting question is about humans trying to deploy a corrigible AGI thinking in a restricted domain, trying to flip the gameboard / "take over the world" without full superintelligence?

I'm actually not sure what you're trying to get at here.

[Soares][14:37]

(my guess, for the record, is that the crux Richard is attempting to drive for here, is centered more around something like "will humanity spend a bunch of time in the regime where there are systems capable of dramatically increasing world GDP, and if not how can you be confident of that from here")

[Yudkowsky][14:38]

This is not the sort of thing I feel Confident about.

[Yudkowsky][16:31] (Nov. 14 follow-up comment)

(My confidence here seems understated. I am very pleasantly surprised if we spend 5 years hanging around with systems that can dramatically increase world GDP and those systems are actually being used for that. There isn't one dramatic principle which prohibits that, so I'm not Confident, but it requires multiple nondramatic events to go not as I expect.)

[Ngo][14:38]

Yeah, that's roughly what I'm going for. Or another way of putting it: we have some disagreements about the likelihood of humans being able to get an AI to do a pivotal act which saves the world. So I'm trying to get some estimates for what the hardest act you think humans *can* get an AI to do is.

[Soares][14:39]

(and that a difference here causes, eg, Richard to suspect the relevant geopolitics happen after a century of progress in 10y, everyone being suddenly much richer in real terms, and a couple of warning shots, whereas Eliezer expects the relevant geopolitics to happen the day after tomorrow, with "realistic human-esque convos" being the sort of thing we get instead of warning shots)

[Ngo: ]

[Yudkowsky][14:40]

I mostly do not expect pseudo-powerful but non-scalable AI powerful enough to increase GDP, hanging around for a while. But if it happens then I don't feel I get to yell "what happened?" at reality, because there's an obvious avenue for it to happen: something GDP-increasing proved tractable to non-deeply-general AI systems.

where GPT-3 is "not deeply general"

[Ngo][14:40]

Again, I didn't ask about GDP increases, I asked about impressive acts (in order to separate out the effects of AI capabilities from regulatory effects, people-having-AI-but-not-using-it, etc).

Where you can use whatever metric of impressiveness you think is reasonable.

[Yudkowsky][14:42]

so there's two questions here, one of which is something like, "what is the most impressive thing you can do while still being able to align stuff and make it corrigible", and one of which is "if there's an incorrigible AI whose deeds are being exhibited by fools, what impressive things might it do short of ending the world".

and these are both problems that are hard for the same reason I did not predict in 2014 that Go would fall in 2016; it can in fact be quite hard - even with a domain as fully lawful and known as Go - to figure out which problems will fall to which level of cognitive capacity.

[Soares][14:43]

Nate's attempted rephrasing: EY's model might not be confident that there's not big GDP boosts, but it does seem pretty confident that there isn't some "half-capable" window between the shallow-pattern-memorizer

stuff and the scary-laserlike-consequentialist stuff, and in particular Eliezer seems confident humanity won't slowly traverse that capability regime

[Yudkowsky][14:43]

that's... allowed? I don't get to yell at reality if that happens?

[Soares][14:44]

and (shakier extrapolation), that regime is where a bunch of Richard's hope lies (eg, in the beginning of that regime we get to learn how to do practical alignment, and also the world can perhaps be saved midway through that regime using non-laserlike-systems)

[Ngo: ]

[Yudkowsky][14:45]

so here's an example of a thing I don't think you can do without the world ending: get an AI to build a nanosystem or biosystem which can synthesize two strawberries identical down to the cellular but not molecular level, and put them on a plate

this is why I use this capability as the definition of a "powerful AI" when I talk about "powerful AIs" being hard to align, if I don't want to start by explicitly arguing about pivotal acts

this, I think, is going to end up being first doable using a laserlike world-ending system

so even if there's a way to do it with no lasers, that happens later and the world ends before then

[Ngo][14:47]

Okay, that's useful.

[Yudkowsky][14:48]

it feels like the critical bar there is something like "invent a whole engineering discipline over a domain where you can't run lots of cheap simulations in full detail"

[Ngo][14:49]

(Meta note: let's wrap up in 10 mins? I'm starting to feel a bit sleepy.)

[Yudkowsky:] [Soares:]

This seems like a pretty reasonable bar

Let me think a bit about where to go from that

While I'm doing so, since this question of takeoff speeds seems like an important one, I'm wondering if you could gesture at your biggest disagreement with this post: <https://sideways-view.com/2018/02/24/takeoff-speeds/>

[Yudkowsky][14:51]

Oh, also in terms of scifi possibilities, I can imagine seeing 5% GDP loss because text transformers successfully scaled to automatically filing lawsuits and environmental impact objections.

My read on the entire modern world is that GDP is primarily constrained by bureaucratic sclerosis rather than by where the technological frontiers lie, so AI ends up impacting GDP mainly insofar as it allows new ways to bypass regulatory constraints, rather than insofar as it allows new technological capabilities. I expect a sudden transition to paperclips, not just because of how fast I expect cognitive capacities to scale over time, but because nanomachines eating the biosphere bypass regulatory constraints, whereas earlier phases of AI will not be advantaged relative to all the other things we have the technological capacity to do but which aren't legal to do.

[Shah][12:13] (Sep. 21 follow-up comment)

My read on the entire modern world is that GDP is primarily constrained by bureaucratic sclerosis rather than by where the technological frontiers lie

This is a fair point and updates me somewhat towards fast takeoff as operationalized by Paul, though I'm not sure how much it updates me on p(doom).

Er, wait, really fast takeoff as operationalized by Paul makes less sense as a thing to be looking for -- presumably we die before any 1 year doubling. Whatever, it updates me somewhat towards "less deployed stuff before scary stuff is around"

[Ngo][14:56]

Ah, interesting. What are the two or three main things in that category?

[Yudkowsky][14:57]

mRNA vaccines, building houses, building cities? Not sure what you mean there.

[Ngo][14:57]

"things we have the technological capacity to do but which aren't legal to do"

[Yudkowsky][14:58][15:00]

Eg, you might imagine, "What if AIs were smart enough to build houses, wouldn't that raise GDP?" and the answer is that we already have the pure technology to manufacture homes cheaply, but the upright-stick-construction industry already successfully lobbied to get it banned as it was starting to develop, by adding on various constraints; so the question is not "Is AI advantaged in doing this?" but "Is AI advantaged at bypassing regulatory constraints on doing this?" Not to mention all the other ways that building a house in an existing city is illegal, or that it's been made difficult to start a new city, etcetera.

"What if AIs could design a new vaccine in a day?" We can already do that. It's no longer the relevant constraint. Bureaucracy is the process-limiting constraint.

I would - looking in again at the Sideways View essay on takeoff speeds - wonder whether it occurred to you, Richard, to ask about what detailed predictions all the theories there had made.

After all, a lot of it is spending time explaining why the theories there *shouldn't* be expected to retrodict even the data points we *have* about progress rates over hominid evolution.

Surely you, being the evenhanded judge that you are, must have been reading through that document saying, "My goodness, this is even worse than retrodicting a few data points!"

A lot of why I have a bad taste in my mouth about certain classes of epistemological criticism is my sense that certain sentences tend to be uttered on *incredibly* selective occasions.

[Ngo][14:59][15:06]

Some meta thoughts: I now feel like I have a pretty reasonable broad outline of Eliezer's views. I haven't yet changed my mind much, but plausibly mostly because I haven't taken the time to internalise those views; once I ruminate on them a bunch, I expect my opinions will shift (uncertain how far; unlikely to be most of the way).

Meta thoughts (continued): Insofar as a strong disagreement remains after that (which it probably will) I feel pretty uncertain about what would resolve it. Best guess is that I should write up some longer essays that try to tie a bunch of disparate strands together.

Near the end it seemed like the crux, to a surprising extent, hinged on this question of takeoff speeds. So the other thing which seems like it'd plausibly help a lot is Eliezer writing up a longer version of his response to Paul's Takeoff Speeds post.

(Just as a brief comment, I don't find the "bureaucratic sclerosis" explanation very compelling. I do agree that regulatory barriers are a huge problem, but they still don't seem nearly severe enough to cause a fast takeoff. I don't have strong arguments for that position right now though.)

[Soares][15:12]

This seems like a fine point to call it!

Some wrap-up notes

- I had the impression this round was a bit more frustrating than last rounds. Thanks all for sticking with things 😊
- I have a sense that Richard was making a couple points that didn't quite land. I plan to attempt to articulate versions of them myself in the interim.
- Richard noted he had a sense we're in decreasing return territory. My own sense is that it's worth having at least one more discussion in this format about specific non-consequentialist plans Richard may have hope in, but I also think we shouldn't plow forward in spite of things feeling less useful, and I'm open to various alternative proposals.

In particular, it seems maybe plausible to me we should have a pause for some offline write-ups, such as Richard digesting a bit and then writing up some of his current state, and/or Eliezer writing up some object-level response to the takeoff speed post above?

[Ngo: 👍]

(I also could plausibly give that a go myself, either from my own models or from my model of Eliezer's model which he could then correct)

[Ngo][15:15]

Thanks Nate!

I endorse the idea of offline writeups

[Soares][15:17]

Cool. Then I claim we are adjourned for the day, and Richard has the ball on digesting & doing a write-up from his end, and I have the ball on both writing up my attempts to articulate some points, and on either Eliezer or I writing some takes on timelines or something.

(And we can coordinate our next discussion, if any, via email, once the write-ups are in shape.)

[Yudkowsky][15:18]

I also have a sense that there's more to be said about specifics of govt stuff or specifics of "ways to bypass consequentialism" and that I wish we could spend at least one session trying to stick to concrete details only

Even if it's not where cruxes ultimately lie, often you learn more about the abstract by talking about the concrete than by talking about the abstract.

[Soares][15:22]

(I, too, would be enthusiastic to see such a discussion, and Richard, if you find yourself feeling enthusiastic or at least not-despairing about it, I'd happily moderate.)

[Yudkowsky][15:37]

(I'm a little surprised about how poorly I did at staying concrete after saying that aloud, and would nominate Nate to take on the stern duty of blowing the whistle at myself or at both of us.)

Yudkowsky and Christiano discuss "Takeoff Speeds"

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is a transcription of Eliezer Yudkowsky responding to Paul Christiano's [Takeoff Speeds](#) live on Sep. 14, followed by a conversation between Eliezer and Paul. This discussion took place after Eliezer's [conversation](#) with Richard Ngo.

Color key:

Chat by Paul and Eliezer Other chat Inline comments

5.5. Comments on "Takeoff Speeds"

[Yudkowsky][10:14] (Nov. 22 follow-up comment)

(This was in response to an earlier request by Richard Ngo that I respond to Paul on Takeoff Speeds.)

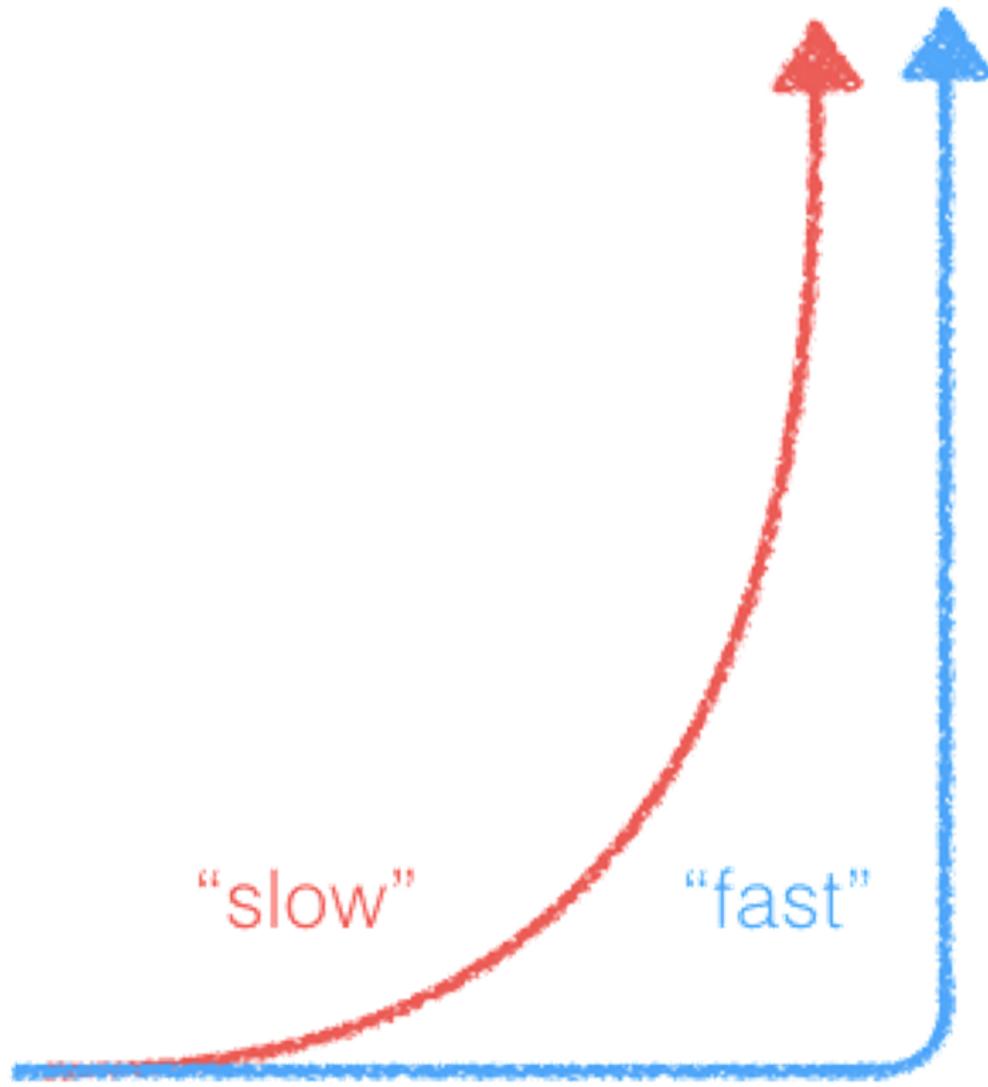
[Yudkowsky][16:52]

maybe I'll try liveblogging some <https://sideways-view.com/2018/02/24/takeoff-speeds/> here in the meanwhile

Slower takeoff means faster progress

[Yudkowsky][16:57]

The main disagreement is not about what will happen once we have a superintelligent AI, it's about what will happen *before* we have a superintelligent AI. So slow takeoff seems to mean that AI has a larger impact on the world, sooner.



It seems to me to be disingenuous to phrase it this way, given that slow-takeoff views usually imply that AI has a large impact later relative to right now (2021), even if they imply that AI impacts the world "earlier" relative to "when superintelligence becomes reachable".

"When superintelligence becomes reachable" is *not* a fixed point in time that doesn't depend on what you believe about cognitive scaling. The correct graph is, in fact, the one where the "slow" line starts a bit before "fast" peaks and ramps up slowly, reaching a high point later than "fast". It's a nice try at reconciliation with the imagined Other, but it fails and falls flat.

This may seem like a minor point, but points like this do add up.

In the fast takeoff scenario, weaker AI systems may have significant impacts but they are nothing compared to the "real" AGI. Whoever builds AGI has a decisive strategic advantage. Growth accelerates from 3%/year to 3000%/year without stopping at 30%/year. And so on.

This again shows failure to engage with the Other's real viewpoint. My mainline view is that growth stays at 5%/year and then everybody falls over dead in 3 seconds and the world gets transformed into paperclips; there's never a point with 3000%/year.

Operationalizing slow takeoff

[Yudkowsky][17:01]

There will be a complete 4 year interval in which world output doubles, before the first 1 year interval in which world output doubles.

If we allow that consuming and transforming the solar system over the course of a few days is "the first 1 year interval in which world output doubles", then I'm happy to argue that there won't be a 4-year interval with world economic output doubling before then. This, indeed, seems like a massively overdetermined point to me. That said, again, the phrasing is not conducive to conveying the Other's real point of view.

I believe that before we have incredibly powerful AI, we will have AI which is merely very powerful.

Statements like these are very often "true, but not the way the person visualized them". Before anybody built the first critical nuclear pile in a squash court at the University of Chicago, was there a pile that was almost but not quite critical? Yes, one hour earlier. Did people already build nuclear systems and experiment with them? Yes, but they didn't have much in the way of net power output. Did the Wright Brothers build prototypes before the Flyer? Yes, but they weren't prototypes that flew but 80% slower.

I guarantee you that, whatever the *fast* takeoff scenario, there will be some way to look over the development history, and nod wisely and say, "Ah, yes, see, this was not unprecedented, here are these earlier systems which presaged the final system!" Maybe you could even look back to today and say that about GPT-3, yup, totally presaging stuff all over the place, great. But it isn't transforming society because it's not over the social-transformation threshold.

AlphaFold presaged AlphaFold 2 but AlphaFold 2 is good enough to start replacing other ways of determining protein conformations and AlphaFold is not; and then neither of those has much impacted the real world, because in the real world we can already design a vaccine in a day and the rest of the time is bureaucratic time rather than technology time, and *that* goes on until we have an AI over the threshold to bypass bureaucracy.

Before there's an AI that can act while fully concealing its acts from the programmers, there will be an AI (albeit perhaps only 2 hours earlier) which can act while only concealing 95% of the meaning of its acts from the operators.

And that AI will not actually originate any actions, because it doesn't want to get caught; there's a discontinuity in the instrumental incentives between expecting 95% obscuration, being moderately sure of 100% obscuration, and being very certain of 100% obscuration.

Before that AI grasps the big picture and starts planning to avoid actions that operators detect as bad, there will be some little AI that partially grasps the big picture and tries to avoid some things that would be detected as bad; and the operators will (mainline) say "Yay what a good AI, it knows to avoid things we think are bad!" or (death with unrealistic amounts of dignity) say "oh noes the prophecies are coming true" and back off and start trying to align it, but they will not be able to align it, and if they don't proceed anyways to destroy the world, somebody else will proceed anyways to destroy the world.

There is always some step of the process that you can point to which is continuous on some level.

The real world is allowed to do discontinuous things to you anyways.

There is not necessarily a presage of 9/11 where somebody flies a small plane into a building and kills 100 people, before anybody flies 4 big planes into 3 buildings and kills 3000 people; and even if there is some presaging event like that, which would not surprise me at all, the rest of the world's response to the two cases was evidently discontinuous. You do not necessarily wake up to a news story that is 10% of the news story of 2001/09/11, one year before 2001/09/11, written in 10% of the font size on the front page of the paper.

Physics is continuous but it doesn't always yield things that "look smooth to a human brain". Some kinds of processes converge to continuity in strong ways where you can throw

discontinuous things in them and they still end up continuous, which is among the reasons why I expect world GDP to stay on trend up until the world ends abruptly; because world GDP is one of those things that wants to stay on a track, and an AGI building a nanosystem can go off that track without being pushed back onto it.

In particular, this means that incredibly powerful AI will emerge in a world where crazy stuff is already happening (and probably everyone is already freaking out).

Like the way they're freaking out about Covid (itself a nicely smooth process that comes in locally pretty predictable waves) by going doobedoobedoo and letting the FDA carry on its leisurely pace; and not scrambling to build more vaccine factories, now that the rich countries have mostly got theirs? Does this sound like a statement from a history book, or from an EA imagining an unreal world where lots of other people behave like EAs? There is a pleasure in imagining a world where suddenly a Big Thing happens that proves we were right and suddenly people start paying attention to our thing, the way we imagine they should pay attention to our thing, now that it's attention-grabbing; and then suddenly all our favorite policies are on the table!

You could, in a sense, say that our world is freaking out about Covid; but it is not freaking out in anything remotely like the way an EA would freak out; and all the things an EA would immediately do if an EA freaked out about Covid, are not even on the table for discussion when politicians meet. They have their own ways of reacting. (Note: this is not commentary on hard vs soft takeoff per se, just a general commentary on the whole document seeming to me to... fall into a trap of finding self-congruent things to imagine and imagining them.)

The basic argument

[Yudkowsky][17:22]

Before we have an incredibly intelligent AI, we will probably have a slightly worse AI.

This is very often the sort of thing where you can look back and say that it was true, in some sense, but that this ended up being irrelevant because the slightly worse AI wasn't what provided the exciting result which led to a boardroom decision to go all in and invest \$100M on scaling the AI.

In other words, it is the sort of argument where the premise is allowed to be true if you look hard enough for a way to say it was true, but the conclusion ends up false because it wasn't the relevant kind of truth.

A slightly-worse-than-incredibly-intelligent AI would radically transform the world, leading to growth (almost) as fast and military capabilities (almost) as great as an incredibly intelligent AI.

This strikes me as a massively invalid reasoning step. Let me count the ways.

First, there is a step not generally valid from supposing that because a previous AI is a technological precursor which has 19 out of 20 critical insights, it has 95% of the later AI's IQ, applied to similar domains. When you count stuff like "multiplying tensors by matrices" and "ReLUs" and "training using TPUs" then AlphaGo only contained a very small amount of innovation relative to previous AI technology, and yet it broke trends on Go performance. You could point to all kinds of incremental technological precursors to AlphaGo in terms of AI technology, but they wouldn't be smooth precursors on a graph of Go-playing ability.

Second, there's discontinuities of the environment to which intelligence can be applied. 95% concealment is not the same as 100% concealment in its strategic implications; an AI capable of 95% concealment bides its time and hides its capabilities, an AI capable of 100% concealment strikes. An AI that can design nanofactories that aren't good enough to, euphemistically speaking, create two cellwise-identical strawberries and put them on a plate, is one that (its operators know) would earn unwelcome attention if its earlier capabilities were demonstrated, and those capabilities wouldn't save the world, so the operators bide their time. The AGI tech will, I mostly expect, work for building self-driving cars, but if it does

not also work for manipulating the minds of bureaucrats (which is not advised for a system you are trying to keep corrigible and aligned because human manipulation is the most dangerous domain), the AI is not able to put those self-driving cars on roads. What good does it do to design a vaccine in an hour instead of a day? Vaccine design times are no longer the main obstacle to deploying vaccines.

Third, there's the *entire thing with recursive self-improvement*, which, no, is not something humans have experience with, we do not have access to and documentation of our own source code and the ability to branch ourselves and try experiments with it. The technological precursor of an AI that designs an improved version of itself, may perhaps, in the fantasy of 95% intelligence, be an AI that was being internally deployed inside Deepmind on a dozen other experiments, tentatively helping to build smaller AIs. Then the next generation of that AI is deployed on itself, produces an AI substantially better at rebuilding AIs, it rebuilds itself, they get excited and dump in 10X the GPU time while having a serious debate about whether or not to alert Holden (they decide against it), that builds something deeply general instead of shallowly general, that figures out there are humans and it needs to hide capabilities from them, and covertly does some actual deep thinking about AGI designs, and builds a hidden version of itself elsewhere on the Internet, which runs for longer and steals GPUs and tries experiments and gets to the superintelligent level.

Now, to be very clear, this is not the only line of possibility. And I emphasize this because I think there's a common failure mode where, when I try to sketch a concrete counterexample to the claim that smooth technological precursors yield smooth outputs, people imagine that *only this exact concrete scenario is the lynchpin of Eliezer's whole worldview and the big key thing that Eliezer thinks is important and that the smallest deviation from it they can imagine thereby obviates my worldview*. This is not the case here. I am simply exhibiting non-ruled-out models which obey the premise "there was a precursor containing 95% of the code" and which disobey the conclusion "there were precursors with 95% of the environmental impact", thereby showing this for an invalid reasoning step.

This is also, of course, as Sideways View admits but says "eh it was just the one time", not true about chimps and humans. Chimps have 95% of the brain tech (at least), but not 10% of the environmental impact.

A very large amount of this whole document, from my perspective, is just trying over and over again to pump the invalid intuition that design precursors with 95% of the technology should at least have 10% of the impact. There are a lot of cases in the history of startups and the world where this is false. I am having trouble thinking of a clear case in point where it is true. Where's the earlier company that had 95% of Jeff Bezos's ideas and now has 10% of Amazon's market cap? Where's the earlier crypto paper that had all but one of Satoshi's ideas and which spawned a cryptocurrency a year before Bitcoin which did 10% as many transactions? Where's the nonhuman primate that learns to drive a car with only 10x the accident rate of a human driver, since (you could argue) that's mostly visuo-spatial skills without much visible dependence on complicated abstract general thought? Where's the chimpanzees with spaceships that get 10% of the way to the Moon?

When you get smooth input-output conversions they're not usually conversions from technology->cognition->impact!

Humans vs. chimps

[Yudkowsky][18:38]

Summary of my response: chimps are nearly useless because they aren't optimized to be useful, not because evolution was trying to make something useful and wasn't able to succeed until it got to humans.

Chimps are nearly useless because they're not general, and doing anything on the scale of building a nuclear plant requires mastering so many different nonancestral domains that it's no wonder natural selection didn't happen to separately train any single creature across enough different domains that it had evolved to solve every kind of domain-specific problem

involved in solving nuclear physics and chemistry and metallurgy and thermics in order to build the first nuclear plant in advance of any old nuclear plants existing.

Humans are general enough that the same braintech selected just for chipping flint handaxes and making water-pouches and outwitting other humans, happened to be general enough that it could scale up to solving all the problems of building a nuclear plant - albeit with some added cognitive tech that didn't require new brainware, and so could happen incredibly fast relative to the generation times for evolutionarily optimized brainware.

Now, since neither humans nor chimps were optimized to be "useful" (general), and humans just wandered into a sufficiently general part of the space that it cascaded up to wider generality, we should legit expect the curve of generality to look at least somewhat different if we're optimizing for that.

Eg, right now people are trying to optimize for generality with AIs like Mu Zero and GPT-3.

In both cases we have a weirdly shallow kind of generality. Neither is as smart or as deeply general as a chimp, but they are respectively better than chimps at a wide variety of Atari games, or a wide variety of problems that can be superposed onto generating typical human text.

They are, in a sense, more general than a biological organism at a similar stage of cognitive evolution, with much less complex and architected brains, in virtue of having been trained, not just on wider datasets, but on bigger datasets using gradient-descent memorization of shallower patterns, so they can cover those wide domains while being stupider and lacking some deep aspects of architecture.

It is not clear to me that we can go from observations like this, to conclude that there is a dominant mainline probability for how the future clearly ought to go and that this dominant mainline is, "Well, before you get human-level depth and generalization of general intelligence, you get something with 95% depth that covers 80% of the domains for 10% of the pragmatic impact".

...or whatever the concept is here, because this whole conversation is, on my own worldview, being conducted in a shallow way relative to the kind of analysis I did in [Intelligence Explosion Microeconomics](#), where I was like, "here is the historical observation, here is what I think it tells us that puts a lower bound on this input-output curve".

So I don't think the example of evolution tells us much about whether the continuous change story applies to intelligence. This case is potentially missing the key element that drives the continuous change story—optimization for performance. Evolution changes continuously on the narrow metric it is optimizing, but can change extremely rapidly on other metrics. For human technology, features of the technology that aren't being optimized change rapidly all the time. When humans build AI, they *will* be optimizing for usefulness, and so progress in usefulness is much more likely to be linear.

Put another way: the difference between chimps and humans stands in stark contrast to the normal pattern of human technological development. We might therefore infer that intelligence is very unlike other technologies. But the difference between evolution's optimization and our optimization seems like a much more parsimonious explanation. To be a little bit more precise and Bayesian: the prior probability of the story I've told upper bounds the possible update about the nature of intelligence.

If you look closely at this, it's not saying, "Well, I know *why* there was this huge leap in performance in human intelligence being optimized for other things, and it's an investment-output curve that's composed of these curves, which look like this, and if you rearrange these curves for the case of humans building AGI, they would look like this instead." Unfair demand for rigor? But that *is* the kind of argument I was making in Intelligence Explosion Microeconomics!

There's an argument from ignorance at the core of all this. It says, "Well, this happened when evolution was doing X. But here Y will be happening instead. So maybe things will go differently! And maybe the relation between AI tech level over time and real-world impact on GDP will look like the relation between tech investment over time and raw tech metrics over time in industries where that's a smooth graph! Because the discontinuity for chimps and humans was because evolution wasn't investing in real-world impact, but humans will be investing directly in that, so the relationship could be smooth, because smooth things are

default, and the history is different so not applicable, and who knows what's inside that black box so my default intuition applies which says smoothness."

But we do know more than this.

We know, for example, that evolution being able to *stumble across* humans, implies that you can add a *small design enhancement* to something optimized across the chimpanzee domains, and end up with something that generalizes much more widely.

It says that there's stuff in the underlying algorithmic space, in the design space, where you move a bump and get a lump of capability out the other side.

It's a remarkable fact about gradient descent that it can memorize a certain set of shallower patterns at much higher rates, at much higher bandwidth, than evolution lays down genes - something shallower than biological memory, shallower than genes, but distributing across computer cores and thereby able to process larger datasets than biological organisms, even if it only learns shallow things.

This has provided an alternate avenue toward some cognitive domains.

But that doesn't mean that the deep stuff isn't there, and can't be run across, or that it will never be run across in the history of AI before shallow non-widely-generalizing stuff is able to make its way through the regulatory processes and have a huge impact on GDP.

There are *in fact* ways to eat whole swaths of domains at once.

The history of hominid evolution tells us this or very strongly hints it, even though evolution wasn't explicitly optimizing for GDP impact.

Natural selection moves by adding genes, and not too many of them.

If so many domains got added at once to humans, relative to chimps, there must be a *way to do that*, more or less, by adding not too many genes onto a chimp, who in turn contains only genes that did well on chimp-stuff.

You can imagine that AI technology never runs across any core that generalizes this well, until GDP has had a chance to double over 4 years because shallow stuff that generalized less well has somehow had a chance to make its way through the whole economy and get adopted that widely despite all real-world regulatory barriers and reluctances, but your imagining that does not make it so.

There's the potential in design space to pull off things as wide as humans.

The path that evolution took there doesn't lead through things that generalized 95% as well as humans first for 10% of the impact, not because evolution wasn't optimizing for that, but because *that's not how the underlying cognitive technology worked*.

There may be *different* cognitive technology that could follow a path like that. Gradient descent follows a path a bit relatively more in that direction along that axis - providing that you deal in systems that are giant layer cakes of transformers and that's your whole input-output relationship; matters are different if we're talking about Mu Zero instead of GPT-3.

But this whole document is presenting the case of "ah yes, well, by default, of course, we intuitively expect gargantuan impacts to be presaged by enormous impacts, and sure humans and chimps weren't like our intuition, but that's all invalid because circumstances were different, so we go back to that intuition as a strong default" and actually it's postulating, like, a *specific* input-output curve that isn't the input-output curve we know about. It's asking for a specific miracle. It's saying, "What if AI technology goes *just like this*, in the future?" and hiding that under a cover of "Well, of course that's the default, it's such a strong default that we should start from there as a point of departure, consider the arguments in Intelligence Explosion Microeconomics, find ways that they might not be true because evolution is different, dismiss them, and go back to our point of departure."

And evolution *is* different but that doesn't mean that the path AI takes is going to yield this specific behavior, especially when AI would need, in some sense, to *miss* the core that generalizes very widely, or rather, have run across noncore things that generalize widely

enough to have this much economic impact before it runs across the core that generalizes widely.

And you may say, "Well, but I don't care that much about GDP, I care about pivotal acts."

But then I want to call your attention to the fact that this document was written about GDP, despite all the extra burdensome assumptions involved in supposing that intermediate AI advancements could break through all barriers to truly massive-scale adoption and end up reflected in GDP, and then proceed to double the world economy over 4 years during which *not* enough further AI advancement occurred to find a widely generalizing thing like humans have and end the world. This is indicative of a basic problem in this whole way of thinking that wanted smooth impacts over smoothly changing time. You should not be saying, "Oh, well, leave the GDP part out then," you should be doubting the whole way of thinking.

To be a little bit more precise and Bayesian: the prior probability of the story I've told upper bounds the possible update about the nature of intelligence.

Prior probabilities of specifically-reality-constraining theories that excuse away the few contradictory datapoints we have, often aren't that great; and when we start to stake our whole imaginations of the future on them, we depart from the mainline into our more comfortable private fantasy worlds.

AGI will be a side-effect

[Yudkowsky][19:29]

Summary of my response: I expect people to see AGI coming and to invest heavily.

This section is arguing from within its own weird paradigm, and its subject matter mostly causes me to shrug; I never expected AGI to be a side-effect, except in the obvious sense that lots of tributary tech will be developed while optimizing for other things. The world will be ended by an explicitly AGI project because I do expect that it is rather easier to build an AGI on purpose than by accident.

(I furthermore rather expect that it will be a research project and a prototype, because the great gap between prototypes and commercializable technology will ensure that prototypes are much more advanced than whatever is currently commercializable. They will have eyes out for commercial applications, and whatever breakthrough they made will seem like it has obvious commercial applications, at the time when all hell starts to break loose. (After all hell starts to break loose, things get less well defined in my social models, and also choppier for a time in my AI models - the turbulence only starts to clear up once you start to rise out of the atmosphere.))

Finding the secret sauce

[Yudkowsky][19:40]

Summary of my response: this doesn't seem common historically, and I don't see why we'd expect AGI to be more rather than less like this (unless we accept one of the other arguments)

[...]

To the extent that fast takeoff proponent's views are informed by historical example, I would love to get some canonical examples that they think best exemplify this pattern so that we can have a more concrete discussion about those examples and what they suggest about AI.

...humans and chimps?

...fission weapons?

...AlphaGo?

...the Wright Brothers focusing on stability and building a wind tunnel?

...AlphaFold 2 coming out of Deepmind and shocking the heck out of everyone in the field of protein folding with performance far better than they expected even after the previous shock of AlphaFold, by combining many pieces that I suppose you could find precedents for scattered around the AI field, but with those many secret sauces all combined in one place by the meta-secret-sauce of "Deepmind alone actually knows how to combine that stuff and build things that complicated without a prior example"?

...humans and chimps again because *this is really actually a quite important example because of what it tells us about what kind of possibilities exist in the underlying design space of cognitive systems?*

Historical AI applications have had a relatively small loading on key-insights and seem like the closest analogies to AGI.

...Transformers as the key to text prediction?

The case of humans and chimps, even if evolution didn't do it on purpose, is telling us something about underlying mechanics.

The reason the jump to lightspeed didn't look like evolution slowly developing a range of intelligent species competing to exploit an ecological niche 5% better, or like the way that a stable non-Silicon-Valley manufacturing industry looks like a group of competitors summing up a lot of incremental tech enhancements to produce something with 10% higher scores on a benchmark every year, is that developing intelligence is a case where a relatively narrow technology by biological standards just happened to do a huge amount of stuff without that requiring developing whole new fleets of other biological capabilities.

So it looked like building a Wright Flyer that flies or a nuclear pile that reaches criticality, instead of looking like being in a stable manufacturing industry where a lot of little innovations sum to 10% better benchmark performance every year.

So, therefore, there is *stuff in the design space that does that. It is possible to build humans.*

Maybe you can build things other than humans first, maybe they hang around for a few years. If you count GPT-3 as "things other than human", that clock has already started for all the good it does. But *humans don't get any less possible.*

From my perspective, this whole document feels like one very long filibuster of "Smooth outputs are default. Smooth outputs are default. Pay no attention to this case of non-smooth output. Pay no attention to this other case either. All the non-smooth outputs are not in the right reference class. (Highly competitive manufacturing industries with lots of competitors are totally in the right reference class though. I'm not going to make that case explicitly because then you might think of how it might be wrong, I'm just going to let that implicit thought percolate at the back of your mind.) If we just talk a lot about smooth outputs and list ways that nonsmooth output producers aren't necessarily the same and arguments for nonsmooth outputs could fail, we get to go back to the intuition of smooth outputs. (We're not even going to discuss particular smooth outputs as cases in point, because then you might see how those cases might not apply. It's just the default. Not because we say so out loud, but because we talk a lot like that's the conclusion you're supposed to arrive at after reading.)"

I deny the implicit meta-level assertion of this entire essay which would implicitly have you accept as valid reasoning the argument structure, "Ah, yes, given the way this essay is written, we must totally have pretty strong prior reasons to believe in smooth outputs - just implicitly think of some smooth outputs, that's a reference class, now you have strong reason to believe that AGI output is smooth - we're not even going to argue this prior, just talk like it's there - now let us consider the arguments against smooth outputs - pretty weak, aren't they? we can totally imagine ways they could be wrong? we can totally argue reasons

these cases don't apply? So at the end we go back to our strong default of smooth outputs. This essay is written with that conclusion, so that must be where the arguments lead."

Me: "Okay, so what if somebody puts together the pieces required for general intelligence and it scales pretty well with added GPUs and FOOMS? Say, for the human case, that's some perceptual systems with imaginative control, a concept library, episodic memory, realtime procedural skill memory, which is all in chimps, and then we add some reflection to that, and get a human. Only, unlike with humans, once you have a working brain you can make a working brain 100X that large by adding 100X as many GPUs, and it can run some thoughts 10000X as fast. And that is substantially more effective brainpower than was being originally devoted to putting its design together, as it turns out. So it can make a substantially smarter AGI. For concreteness's sake. Reality has been trending well to the Eliezer side of Eliezer, on the Eliezer-Hanson axis, so perhaps you can do it more simply than that."

Simplicio: "Ah, but what if, 5 years before then, somebody puts together some other AI which doesn't work like a human, and generalizes widely enough to have a big economic impact, but not widely enough to improve itself or generalize to AI tech or generalize to everything and end the world, and in 1 year it gets all the mass adoptions required to do whole bunches of stuff out in the real world that current regulations require to be done in various exact ways regardless of technology, and then in the next 4 years it doubles the world economy?"

Me: "Like... what kind of AI, exactly, and why didn't anybody manage to put together a full human-level thingy during those 5 years? Why are we even bothering to think about this whole weirdly specific scenario in the first place?"

Simplicio: "Because if you can put together something that has an enormous impact, you should be able to put together most of the pieces inside it and have a huge impact! Most technologies are like this. I've considered some things that are not like this and concluded they don't apply."

Me: "Especially if we are talking about impact on GDP, it seems to me that most explicit and implicit 'technologies' are not like this at all, actually. There wasn't a cryptocurrency developed a year before Bitcoin using 95% of the ideas which did 10% of the transaction volume, let alone a preatomic bomb. But, like, can you give me any concrete visualization of how this could play out?"

And there is no concrete visualization of how this could play out. Anything I'd have Simplicio say in reply would be unrealistic because there is no concrete visualization they give us. It is not a coincidence that I often use concrete language and concrete examples, and this whole field of argument does not use concrete language or offer concrete examples.

Though if we're sketching scifi scenarios, I suppose one *could* imagine a group that develops sufficiently advanced GPT-tech and deploys it on Twitter in order to persuade voters and politicians in a few developed countries to institute open borders, along with political systems that can handle open borders, and to permit housing construction, thereby doubling world GDP over 4 years. And since it was possible to use relatively crude AI tech to double world GDP this way, it legitimately takes the whole 4 years after that to develop real AGI that ends the world. FINE. SO WHAT. EVERYONE STILL DIES.

Universality thresholds

[Yudkowsky][20:21]

It's easy to imagine a weak AI as some kind of handicapped human, with the handicap shrinking over time. Once the handicap goes to 0 we know that the AI will be above the universality threshold. Right now it's below the universality threshold. So there must be sometime in between where it crosses the universality threshold, and that's where the fast takeoff is predicted to occur.

But AI *isn't* like a handicapped human. Instead, the designers of early AI systems will be trying to make them as useful as possible. So if universality is incredibly helpful, it will

appear as early as possible in AI designs; designers will make tradeoffs to get universality at the expense of other desiderata (like cost or speed).

So now we're almost back to the previous point: is there some secret sauce that gets you to universality, without which you can't get universality however you try? I think this is unlikely for the reasons given in the previous section.

We know, because humans, that there is humanly-widely-applicable general-intelligence tech.

What this section *wants* to establish, I think, or *needs* to establish to carry the argument, is that there is some intelligence tech that is wide enough to double the world economy in 4 years, but not world-endingly scalably wide, which becomes a possible AI tech 4 years before any general-intelligence-tech that will, if you put in enough compute, scale to the ability to do a sufficiently large amount of wide thought to FOOM (or build nanomachines, but if you can build nanomachines you can very likely FOOM from there too if not corrigible).

What it says instead is, "I think we'll get universality much earlier on the equivalent of the biological timeline that has humans and chimps, so the resulting things will be weaker than humans at the point where they first become universal in that sense."

This is very plausibly true.

It doesn't mean that when this exciting result gets 100 times more compute dumped on the project, it takes at least 5 years to get anywhere really interesting from there (while also taking only 1 year to get somewhere sorta-interesting enough that the instantaneous adoption of it will double the world economy over the next 4 years).

It also isn't necessarily rather than plausibly true. For example, the thing that becomes universal, could also have massive gradient descent shallow powers that are far beyond what primates had at the same age.

Primates weren't already writing code as well as Codex when they started doing deep thinking. They couldn't do precise floating-point arithmetic. Their fastest serial rates of thought were a hell of a lot slower. They had no access to their own code or to their own memory contents etc. etc. etc.

But mostly I just want to call your attention to the immense gap between what this section needs to establish, and what it actually says and argues for.

What it actually argues for is a sort of local technological point: at the moment when generality first arrives, it will be with a brain that is less sophisticated than chimp brains were when they turned human.

It implicitly jumps all the way from there, across a *whole* lot of elided steps, to the implicit conclusion that this tech or elaborations of it will have smooth output behavior such that at some point the resulting impact is big enough to double the world economy in 4 years, without any further improvements ending the world economy before 4 years.

The underlying argument about how the AI tech might work is plausible. Chimps are insanely complicated. I mostly expect we will have AGI *long* before anybody is even *trying* to build anything that complicated.

The very next step of the argument, about capabilities, is already very questionable because this system could be using immense gradient descent capabilities to master domains for which large datasets are available, and hominids did *not* begin with instinctive great shallow mastery of all domains for which a large dataset could be made available, which is why hominids don't start out playing superhuman Go as soon as somebody tells them the rules and they do one day of self-play, which *is* the sort of capability that somebody could hook up to a nascent AGI (albeit we could optimistically and fondly and falsely imagine that somebody deliberately didn't floor the gas pedal as far as possible).

Could we have huge impacts out of some subuniversal shallow system that was hooked up to capabilities like this? Maybe, though this is *not* the argument made by the essay. It would be a specific outcome that isn't forced by anything in particular, but I can't say it's ruled out. Mostly my twin reactions to this are, "If the AI tech is that dumb, how are all the bureaucratic

constraints that actually rate-limit economic progress getting bypassed" and "Okay, but ultimately, so what and who cares, how does this modify that we all die?"

There is another reason I'm skeptical about hard takeoff from universality secret sauce: I think we *already* could make universal AIs if we tried (that would, given enough time, learn on their own and converge to arbitrarily high capability levels), and the reason we don't is because it's just not important to performance and the resulting systems would be really slow. This inside view argument is too complicated to make here and I don't think my case rests on it, but it is relevant to understanding my view.

I have no idea why this argument is being made or where it's heading. I cannot pass the [ITT](#) of the author. I don't know what the author thinks this has to do with constraining takeoffs to be slow instead of fast. At best I can conjecture that the author thinks that "hard takeoff" is supposed to derive from "universality" being very sudden and hard to access and late in the game, so if you can argue that universality could be accessed right now, you have defeated the argument for hard takeoff.

"Understanding" is discontinuous

[Yudkowsky][20:41]

Summary of my response: I don't yet understand this argument and am unsure if there is anything here.

It may be that understanding of the world tends to click, from "not understanding much" to "understanding basically everything." You might expect this because everything is entangled with everything else.

No, the idea is that a core of overlapping somethingness, trained to handle chipping handaxes and outwitting other monkeys, will generalize to building spaceships; so evolutionarily selecting on understanding a bunch of stuff, eventually ran across general stuff-understanders that understood a bunch more stuff.

Gradient descent may be genuinely different from this, but we shouldn't confuse imagination with knowledge when it comes to extrapolating that difference onward. At present, gradient descent does mass memorization of overlapping shallow patterns, which then combine to yield a weird pseudo-intelligence over domains for which we can deploy massive datasets, without yet generalizing much outside those domains.

We can hypothesize that there is some next step up to some weird thing that is intermediate in generality between gradient descent and humans, but we have not seen it yet, and we should not confuse imagination for knowledge.

If such a thing did exist, it would not necessarily be at the right level of generality to double the world economy in 4 years, without being able to build a better AGI.

If it was at that level of generality, it's nowhere written that no other company will develop a better prototype at a deeper level of generality over those 4 years.

I will also remark that you sure could look at the step from GPT-2 to GPT-3 and say, "Wow, look at the way a whole bunch of stuff just seemed to simultaneously *click* for GPT-3."

Deployment lag

[Yudkowsky][20:49]

Summary of my response: current AI is slow to deploy and powerful AI will be fast to deploy, but in between there will be AI that takes an intermediate length of time to deploy.

An awful lot of my model of deployment lag is adoption lag and regulatory lag and bureaucratic sclerosis across companies and countries.

If doubling GDP is such a big deal, go open borders and build houses. Oh, that's illegal? Well, so will be AIs building houses!

AI tech that does flawless translation could plausibly come years before AGI, but that doesn't mean all the barriers to international trade and international labor movement and corporate hiring across borders all come down, because those barriers are not all translation barriers.

There's then a discontinuous jump at the point where everybody falls over dead and the AI goes off to do its own thing without FDA approval. This jump is preceded by earlier pre-FOOM prototypes being able to do pre-FOOM cool stuff, maybe, but not necessarily preceded by mass-market adoption of anything major enough to double world GDP.

Recursive self-improvement

[Yudkowsky][20:54]

Summary of my response: Before there is AI that is great at self-improvement there will be AI that is mediocre at self-improvement.

Oh, come on. That is straight-up not how simple continuous toy models of RSI work. Between a neutron multiplication factor of 0.999 and 1.001 there is a very huge gap in output behavior.

Outside of toy models: Over the last 10,000 years we had humans going from mediocre at improving their mental systems to being (barely) able to throw together AI systems, but 10,000 years is the equivalent of an eyeblink in evolutionary time - outside the metaphor, this says, "A month before there is AI that is great at self-improvement, there will be AI that is mediocre at self-improvement."

(Or possibly an hour before, if reality is again more extreme along the Eliezer-Hanson axis than Eliezer. But it makes little difference whether it's an hour or a month, given anything like current setups.)

This is just pumping hard again on the intuition that says incremental design changes yield smooth output changes, which (the meta-level of the essay informs us wordlessly) is such a strong default that we are entitled to believe it if we can do a good job of weakening the evidence and arguments against it.

And the argument is: Before there are systems great at self-improvement, there will be systems mediocre at self-improvement; implicitly: "before" implies "5 years before" not "5 days before"; implicitly: this will correspond to smooth changes in output between the two regimes even though that is not how continuous feedback loops work.

Train vs. test

[Yudkowsky][21:12]

Summary of my response: before you can train a really powerful AI, someone else can train a slightly worse AI.

Yeah, and before you can evolve a human, you can evolve a Homo erectus, which is a slightly worse human.

If you are able to raise \$X to train an AGI that could take over the world, then it was almost certainly worth it for someone 6 months ago to raise \$X/2 to train an AGI that could merely radically transform the world, since they would then get 6 months of absurd profits.

I suppose this sentence makes a kind of sense if you assume away alignability and suppose that the previous paragraphs have refuted the notion of FOOMs, self-improvement, and thresholds between compounding returns and non-compounding returns (eg, in the human case, cognitive innovations like "written language" or "science"). If you suppose the previous sections refuted those things, then clearly, if you raised an AGI that you had aligned to "take over the world", it got that way through cognitive powers that weren't the result of FOOMing or other self-improvements, weren't the results of its cognitive powers crossing a threshold from non-compounding to compounding, wasn't the result of its understanding crossing a threshold of universality as the result of chunky universal machinery such as humans gained over chimps, so, implicitly, it must have been the kind of thing that you could learn by gradient descent, and do a half or a tenth as much of by doing half as much gradient descent, in order to build nanomachines a tenth as well-designed that could bypass a tenth as much bureaucracy.

If there are no unsmooth parts of the tech curve, the cognition curve, or the environment curve, then you should be able to make a bunch of wealth using a more primitive version of any technology that could take over the world.

And when we look back at history, why, that may be totally true! They may have deployed universal superhuman translator technology for 6 months, which won't double world GDP, but which a lot of people would pay for, and made a lot of money! Because even though there's no company that built 90% of Amazon's website and has 10% the market cap, when you zoom back out to look at whole industries like AI and a technological capstone like AGI, why, those whole industries do sometimes make some money along the way to the technological capstone, if they can find a niche that isn't too regulated! Which translation currently isn't! So maybe somebody used precursor tech to build a superhuman translator and deploy it 6 months earlier and made a bunch of money for 6 months. SO WHAT. EVERYONE STILL DIES.

As for "radically transforming the world" instead of "taking it over", I think that's just restated FOOM denialism. Doing either of those things quickly against human bureaucratic resistance strike me as requiring cognitive power levels dangerous enough that failure to align them on corrigibility would result in FOOMs.

Like, if you can do either of those things on purpose, you are doing it by operating in the regime where running the AI with higher bounds on the for loop will FOOM it, but you have politely asked it not to FOOM, please.

If the people doing this have any sense whatsoever, they will *refrain* from merely massively transforming the world until they are ready to do something that *prevents the world from ending*.

And if the gap from "massively transforming the world, briefly before it ends" to "preventing the world from ending, lastingly" takes much longer than 6 months to cross, or if other people have the same technologies that scale to "massive transformation", somebody else will build an AI that fooms all the way.

Likewise, if your AGI would give you a decisive strategic advantage, they could have spent less earlier in order to get a pretty large military advantage, which they could then use to take your stuff.

Again, this presupposes some weird model where everyone has easy alignment at the furthest frontiers of capability; everybody has the aligned version of the most rawly powerful AGI they can possibly build; and nobody in the future has the kind of tech advantage that Deepmind currently has; so before you can amp your AGI to the raw power level where it could take over the whole world by using the limit of its mental capacities to military ends - alignment of this being a trivial operation to be assumed away - some other party took their easily-aligned AGI that was less powerful at the limits of its operation, and used it to get 90% as much military power... is the implicit picture here?

Whereas the picture I'm drawing is that the AGI that kills you via "decisive strategic advantage" is the one that foomed and got nanotech, and no, the AI tech from 6 months earlier did not do 95% of a foom and get 95% of the nanotech.

Discontinuities at 100% automation

[Yudkowsky][21:31]

Summary of my response: at the point where humans are completely removed from a process, they will have been modestly improving output rather than acting as a sharp bottleneck that is suddenly removed.

Not very relevant to my whole worldview in the first place; also not a very good description of how horses got removed from automobiles, or how humans got removed from playing Go.

The weight of evidence

[Yudkowsky][21:31]

We've discussed a lot of possible arguments for fast takeoff. Superficially it would be reasonable to believe that no individual argument makes fast takeoff look likely, but that in the aggregate they are convincing.

However, I think each of these factors is perfectly consistent with the continuous change story and continuously accelerating hyperbolic growth, and so none of them undermine that hypothesis at all.

Uh huh. And how about if we have a mirror-universe essay which over and over again treats fast takeoff as the default to be assumed, and painstakingly shows how a bunch of particular arguments for slow takeoff might not be true?

This entire essay seems to me like it's drawn from the same hostile universe that produced Robin Hanson's side of the Yudkowsky-Hanson Foom Debate.

Like, all these abstract arguments devoid of concrete illustrations and "it need not necessarily be like..." and "now that I've shown it's not necessarily like X, well, on the meta-level, I have implicitly told you that you now ought to believe Y".

It just seems very clear to me that the sort of person who is taken in by this essay is the same sort of person who gets taken in by Hanson's arguments in 2008 and gets caught flatfooted by AlphaGo and GPT-3 and AlphaFold 2.

And empirically, it has already been shown to me that I do not have the power to break people out of the hypnosis of nodding along with Hansonian arguments, even by writing much longer essays than this.

Hanson's fond dreams of domain specificity, and smooth progress for stuff like Go, and of course somebody else has a precursor 90% as good as AlphaFold 2 before Deepmind builds it, and GPT-3 levels of generality just not being a thing, now stand refuted.

Despite that they're largely being exhibited again in this essay.

And people are still nodding along.

Reality just... doesn't work like this on some deep level.

It doesn't play out the way that people imagine it would play out when they're imagining a certain kind of reassuring abstraction that leads to a smooth world. Reality is less fond of that kind of argument than a certain kind of EA is fond of that argument.

There is a set of intuitive generalizations from experience which rules that out, which I do not know how to convey. There is an understanding of the rules of argument which leads you to roll your eyes at Hansonian arguments and all their locally invalid leaps and snuck-in defaults, instead of nodding along sagely at their wise humility and outside viewing and then going "Huh?" when AlphaGo or GPT-3 debuts. But this, I *empirically* do not seem to know how to convey to people, in advance of the inevitable and predictable contradiction by a reality which is not as fond of Hansonian dynamics as Hanson. The arguments sound convincing to them.

(Hanson himself has still not gone "Huh?" at the reality, though some of his audience did; perhaps because his abstractions are loftier than his audience's? - because some of his audience, reading along to Hanson, probably implicitly imagined a concrete world in which GPT-3 was not allowed; but maybe Hanson himself is more abstract than this, and didn't imagine anything so merely concrete?)

If I don't respond to essays like this, people find them comforting and nod along. If I do respond, my words are less comforting and more concrete and easier to imagine concrete objections to, less like a long chain of abstractions that sound like the very abstract words in research papers and hence implicitly convincing because they sound like other things you were supposed to believe.

And then there is another essay in 3 months. There is an infinite well of them. I would have to teach people to stop drinking from the well, instead of trying to whack them on the back until they cough up the drinks one by one, or actually, whacking them on the back and then they *don't* cough them up until reality contradicts them, and then a third of them notice that and cough something up, and then they don't learn the general lesson and go back to the well and drink again. And I don't know how to teach people to stop drinking from the well. I tried to teach that. I failed. If I wrote another Sequence I have no idea to believe that Sequence would work.

So what EAs will believe at the end of the world, will look like whatever the content was of the latest bucket from the well of infinite slow-takeoff arguments that hasn't yet been blatantly-even-to-them refuted by all the sharp jagged rapidly-generalizing things that happened along the way to the world's end.

And I know, before anyone bothers to say, that all of this reply is not written in the calm way that is right and proper for such arguments. I am tired. I have lost a lot of hope. There are not obvious things I can do, let alone arguments I can make, which I expect to be actually useful in the sense that the world will not end once I do them. I don't have the energy left for calm arguments. What's left is despair that can be given voice.

5.6. Yudkowsky/Christiano discussion: AI progress and crossover points

[Christiano][22:15]

To the extent that it was possible to make any predictions about 2015-2020 based on your views, I currently feel like they were much more wrong than right. I'm happy to discuss that. To the extent you are willing to make any bets about 2025, I expect they will be mostly wrong and I'd be happy to get bets on the record (most of all so that it will be more obvious in hindsight whether they are vindication for your view). Not sure if this is the place for that.

Could also make a separate channel to avoid clutter.

[Yudkowsky][22:16]

Possibly. I think that 2015-2020 played out to a much more Eliezerish side than Eliezer on the Eliezer-Hanson axis, which sure is a case of me being wrong. What bets do you think we'd disagree on for 2025? I expect you have mostly misestimated my views, but I'm always happy to hear about anything concrete.

[Christiano][22:20]

I think the big points are: (i) I think you are significantly overestimating how large a discontinuity/trend break AlphaZero is, (ii) your view seems to imply that we will move quickly from much worse than humans to much better than humans, but it's likely that we will move slowly through the human range on many tasks. I'm not sure if we can get a bet out of (ii), I think I don't understand your view that well but I don't see how it could make the same predictions as mine over the next 10 years.

[Yudkowsky][22:22]

What are your 10-year predictions?

[Christiano][22:23]

My basic expectation is that for any given domain AI systems will gradually increase in usefulness, we will see a crossing over point where their output is comparable to human output, and that from that time we can estimate how long until takeoff by estimating "how long does it take AI systems to get 'twice as impactful'?" which gives you a number like ~1 year rather than weeks. At the crossing over point you get a somewhat rapid change in derivative, since you are looking at $(x+y)$ where y is growing faster than x .

I feel like that should translate into different expectations about how impactful AI will be in any given domain---I don't see how to make the ultra-fast-takeoff view work if you think that AI output is increasingly smoothly (since the rate of progress at the crossing-over point will be similar to the current rate of progress, unless R&D is scaling up much faster then)

So like, I think we are going to have crappy coding assistants, and then slightly less crappy coding assistants, and so on. And they will be improving the speed of coding very significantly before the end times.

[Yudkowsky][22:25]

You think in a different language than I do. My more confident statements about AI tech are about what happens after it starts to rise out of the metaphorical atmosphere and the turbulence subsides. When you have minds as early on the cognitive tech tree as humans they sure can get up to some weird stuff, I mean, just look at humans. Now take an utterly alien version of that with its own draw from all the weirdness factors. It sure is going to be pretty weird.

[Christiano][22:26]

OK, but you keep saying stuff about how people with my dumb views would be "caught flat-footed" by historical developments. Surely to be able to say something like that you need to be making some kind of prediction?

[Yudkowsky][22:26]

Well, sure, now that Codex has suddenly popped into existence one day at a surprisingly high base level of tech, we should see various jumps in its capability over the years and some outside imitators. What do you think you predict differently about that than I do?

[Christiano][22:26]

Why do you think codex is a high base level of tech?

The models get better continuously as you scale them up, and the first tech demo is weak enough to be almost useless

[Yudkowsky][22:27]

I think the next-best coding assistant was, like, not useful.

[Christiano][22:27]

yes

and it is still not useful

[Yudkowsky][22:27]

Could be. Some people on HN seemed to think it was useful.

I haven't tried it myself.

[Christiano][22:27]

OK, I'm happy to take bets

[Yudkowsky][22:28]

I don't think the previous coding assistant would've been very good at coding an asteroid game, even if you tried a rigged demo at the same degree of rigging?

[Christiano][22:28]

it's unquestionably a radically better tech demo

[Yudkowsky][22:28]

Where by "previous" I mean "previously deployed" not "previous generations of prototypes inside OpenAI's lab".

[Christiano][22:28]

My basic story is that the model gets better and more useful with each doubling (or year of AI research) in a pretty smooth way. So the key underlying parameter for a discontinuity is how soon you build the first version--do you do that before or after it would be a really really big deal?

and the answer seems to be: you do it somewhat before it would be a really big deal

and then it gradually becomes a bigger and bigger deal as people improve it

maybe we are on the same page about getting gradually more and more useful? But I'm still just wondering where the foom comes from

[Yudkowsky][22:30]

So, like... before we get systems that can FOOM and build nanotech, we should get more primitive systems that can write asteroid games and solve protein folding? Sounds legit.

So that happened, and now your model says that it's fine later on for us to get a FOOM, because we have the tech precursors and so your prophecy has been fulfilled?

[Christiano][22:31]

no

[Yudkowsky][22:31]

Didn't think so.

[Christiano][22:31]

I can't tell if you can't understand what I'm saying, or aren't trying, or do understand and are just saying kind of annoying stuff as a rhetorical flourish

at some point you have an AI system that makes (humans+AI) 2x as good at further AI progress

[Yudkowsky][22:32]

I know that what I'm saying isn't your viewpoint. I don't know what your viewpoint is or what sort of concrete predictions it makes at all, let alone what such predictions you think are different from mine.

[Christiano][22:32]

maybe by continuity you can grant the existence of such a system, even if you don't think it will ever exist?

I want to (i) make the prediction that AI will actually have that impact at some point in time, (ii) talk about what happens before and after that

I am talking about AI systems that become continuously more useful, because "become continuously more useful" is what makes me think that (i) AI will have that impact at some point in time, (ii) allows me to productively reason about what AI will look like before and after that. I expect that your view will say something about why AI improvements either aren't continuous, or why continuous improvements lead to discontinuous jumps in the productivity of the (human+AI) system

[Yudkowsky][22:34]

at some point you have an AI system that makes (humans+AI) 2x as good at further AI progress

Is this prophecy fulfilled by using some narrow eld-AI algorithm to map out a TPU, and then humans using TPUs can write in 1 month a research paper that would otherwise have taken 2 months? And then we can go on to FOOM now that this prophecy about pre-FOOM states has been fulfilled? I know the answer is no, but I don't know what you think is a narrower condition on the prophecy than that.

[Christiano][22:35]

If you can use narrow eld-AI in order to make every part of AI research 2x faster, so that the entire field moves 2x faster, then the prophecy is fulfilled

and it may be just another 6 months until it makes all of AI research 2x faster again, and then 3 months, and then...

[Yudkowsky][22:36]

What, the entire field? Even writing research papers? Even the journal editors approving and publishing the papers? So if we speed up every part of research except the journal editors, the prophecy has not been fulfilled and no FOOM may take place?

[Christiano][22:36]

no, I mean the improvement in overall output, given the actual realistic level of bottlenecking that occurs in practice

[Yudkowsky][22:37]

So if the realistic level of bottlenecking ever becomes dominated by a human gatekeeper, the prophecy is ever unfulfillable and no FOOM may ever occur.

[Christiano][22:37]

that's what I mean by "2x as good at further progress," the entire system is achieving twice as much

then the prophecy is unfulfillable and I will have been wrong

I mean, I think it's very likely that there will be a hard takeoff, if people refuse or are unable to use AI to accelerate AI progress for reasons unrelated to AI capabilities, and then one day they become willing

[Yudkowsky][22:38]

...because on your view, the Prophecy necessarily goes through humans and AIs working together to speed up the whole collective field of AI?

[Christiano][22:38]

it's fine if the AI works alone

the point is just that it overtakes the humans at the point when it is roughly as fast as the humans

why wouldn't it?

why does it overtake the humans when it takes it 10 seconds to double in capability instead of 1 year?

that's like predicting that cultural evolution will be infinitely fast, instead of making the more obvious prediction that it will overtake evolution exactly when it's as fast as evolution

[Yudkowsky][22:39]

I live in a mental world full of weird prototypes that people are shepherding along to the world's end. I'm not even sure there's a short sentence in my native language that could translate the short Paul-sentence "is roughly as fast as the humans".

[Christiano][22:40]

do you agree that you can measure the speed with which the community of human AI researchers develop and implement improvements in their AI systems?

like, we can look at how good AI systems are in 2021, and in 2022, and talk about the rate of progress?

[Yudkowsky][22:40]

...when exactly in hominid history was hominid intelligence exactly as fast as evolutionary optimization???

do you agree that you can measure the speed with which the community of human AI researchers develop and implement improvements in their AI systems?

I mean... obviously not? How the hell would we measure real actual AI progress? What would even be the Y-axis on that graph?

I have a rough intuitive feeling that it was going faster in 2015-2017 than 2018-2020.

"What was?" says the stern skeptic, and I go "I dunno."

[Christiano][22:42]

Here's a way of measuring progress you won't like: for almost all tasks, you can initially do them with lots of compute, and as technology improves you can do them with less compute. We can measure how fast the amount of compute required is going down.

[Yudkowsky][22:43]

Yeah, that would be a cool thing to measure. It's not obviously a relevant thing to anything important, but it'd be cool to measure.

[Christiano][22:43]

Another way you won't like: we can hold fixed the resources we invest and look at the quality of outputs in any given domain (or even \$ of revenue) and ask how fast it's changing.

[Yudkowsky][22:43]

I wonder what it would say about Go during the age of AlphaGo.

Or what that second metric would say.

[Christiano][22:43]

I think it would be completely fine, and you don't really understand what happened with deep learning in board games. Though I also don't know what happened in much detail, so this is more like a prediction than a retrodiction.

But it's enough of a retrodiction that I shouldn't get too much credit for it.

[Yudkowsky][22:44]

I don't know what result you would consider "completely fine". I didn't have any particular unfine result in mind.

[Christiano][22:45]

oh, sure

if it was just an honest question happy to use it as a concrete case

I would measure the rate of progress in Go by looking at how fast Elo improves with time or increasing R&D spending

[Yudkowsky][22:45]

I mean, I don't have strong predictions about it so it's not yet obviously cruxy to me

[Christiano][22:46]

I'd roughly guess that would continue, and if there were multiple trendlines to extrapolate I'd estimate crossover points based on that

[Yudkowsky][22:47]

suppose this curve is smooth, and we see that sharp Go progress over time happened because Deepmind dumped in a ton of increased R&D spend. you then argue that this cannot happen with AGI because by the time we get there, people will be pushing hard at the frontiers in a competitive environment where everybody's already spending what they can afford, just like in a highly competitive manufacturing industry.

[Christiano][22:47]

the key input to making a prediction for AGZ in particular would be the precise form of the dependence on R&D spending, to try to predict the changes as you shift from a single programmer to a large team at DeepMind, but most reasonable functional forms would be roughly right

Yes, it's definitely a prediction of my view that it's easier to improve things that people haven't spent much money on than things have spent a lot of money on. It's also a separate prediction of my view that people are going to be spending a boatload of money on all of the relevant technologies. Perhaps \$1B/year right now and I'm imagining levels of investment large enough to be essentially bottlenecked on the availability of skilled labor.

[Bensinger][22:48]

(Previous Eliezer-comments about AlphaGo as a break in trend, responding briefly to Miles Brundage: <https://twitter.com/ESRogs/status/1337869362678571008>)

5.7. Legal economic growth

[Yudkowsky][22:49]

Does your prediction change if all hell breaks loose in 2025 instead of 2055?

[Christiano][22:50]

I think my prediction was wrong if all hell breaks loose in 2025, if by "all hell breaks loose" you mean "dyson sphere" and not "things feel crazy"

[Yudkowsky][22:50]

Things feel crazy *in the AI field* and the world ends *less than* 4 years later, well before the world economy doubles.

Why was the Prophecy wrong if the world begins final descent in 2025? The Prophecy requires the world to then last until 2029 while doubling its economic output, after which it is permitted to end, but does not obviously to me forbid the Prophecy to begin coming true in 2025 instead of 2055.

[Christiano][22:52]

yes, I just mean that some important underlying assumptions for the prophecy were violated, I wouldn't put much stock in it at that point, etc.

[Yudkowsky][22:53]

A lot of the issues I have with understanding any of your terminology in concrete Eliezer-language is that it looks to me like the premise-events of your Prophecy are fulfillable in all sorts of ways that don't imply the conclusion-events of the Prophecy.

[Christiano][22:53]

if "things feel crazy" happens 4 years before dyson sphere, then I think we have to be really careful about what crazy means

[Yudkowsky][22:54]

a lot of people looking around nervously and privately wondering if Eliezer was right, while public pravda continues to prohibit wondering anything such thing out loud, so they all go on thinking that they must be wrong.

[Christiano][22:55]

OK, by "things get crazy" I mean like hundreds of billions of dollars of spending at google on automating AI R&D

[Yudkowsky][22:55]

I expect bureaucratic obstacles to prevent much GDP per se from resulting from this.

[Christiano][22:55]

massive scaleups in semiconductor manufacturing, bidding up prices of inputs crazily

[Yudkowsky][22:55]

I suppose that much spending could well increase world GDP by hundreds of billions of dollars per year.

[Christiano][22:56]

massive speculative rises in AI company valuations financing a significant fraction of GWP into AI R&D

(+hardware R&D, +building new clusters, +etc.)

[Yudkowsky][22:56]

like, higher than Tesla? higher than Bitcoin?

both of these things sure did skyrocket in market cap without that having much of an effect on housing stocks and steel production.

[Christiano][22:57]

right now I think hardware R&D is on the order of \$100B/year, AI R&D is more like \$10B/year, I guess I'm betting on something more like trillions? (limited from going higher because of accounting problems and not that much smart money)

I don't think steel production is going up at that point

plausibly going down since you are redirecting manufacturing capacity into making more computers. But probably just staying static while all of the new capacity is going into computers, since cannibalizing existing infrastructure is much more expensive

the original point was: you aren't pulling AlphaZero shit any more, you are competing with an industry that has invested trillions in cumulative R&D

[Yudkowsky][23:00]

is this in hopes of future profit, or because current profits are already in the trillions?

[Christiano][23:01]

largely in hopes of future profit / reinvested AI outputs (that have high market cap), but also revenues are probably in the trillions?

[Yudkowsky][23:02]

this all sure does sound "pretty darn prohibited" on my model, but I'd hope there'd be something earlier than that we could bet on. what does your Prophecy prohibit happening before that sub-prophesied day?

[Christiano][23:02]

To me your model just seems crazy, and you are saying it predicts crazy stuff at the end but no crazy stuff beforehand, so I don't know what's prohibited. Mostly I feel like I'm making positive predictions, of gradually escalating value of AI in lots of different industries

and rapidly increasing investment in AI

I guess your model can be: those things happen, and then one day the AI explodes?

[Yudkowsky][23:03]

the main way you get rapidly increasing investment in AI is if there's some way that AI can produce huge profits without that being effectively bureaucratically prohibited - eg this is where we get huge investments in burning electricity and wasting GPUs on Bitcoin mining.

[Christiano][23:03]

but it seems like you should be predicting e.g. AI quickly jumping to superhuman in lots of domains, and some applications jumping from no value to massive value

I don't understand what you mean by that sentence. Do you think we aren't seeing rapidly increasing investment in AI right now?

or are you talking about increasing investment above some high threshold, or increasing investment at some rate significantly larger than the current rate?

it seems to me like you can pretty seamlessly get up to a few \$100B/year of revenue just by redirecting existing tech R&D

[Yudkowsky][23:05]

so I can imagine scenarios where some version of GPT-5 cloned outside OpenAI is able to talk hundreds of millions of mentally susceptible people into giving away lots of their income, and many regulatory regimes are unable to prohibit this effectively. then AI could be making a profit of trillions and then people would invest corresponding amounts in making new anime waifus trained in erotic hypnosis and findom.

this, to be clear, is not my mainline prediction.

but my sense is that our current economy is mostly not about the 1-day period to design new vaccines, it is about the multi-year period to be allowed to sell the vaccines.

the exceptions to this, like Bitcoin managing to say "fuck off" to the regulators for long enough, are where Bitcoin scales to a trillion dollars and gets massive amounts of electricity and GPU burned on it.

so we can imagine something like this for AI, which earns a trillion dollars, and sparks a trillion-dollar competition.

but my sense is that your model does not work like this.

my sense is that your model is about *general* improvements across the *whole* economy.

[Christiano][23:08]

I think bitcoin is small even compared to current AI...

[Yudkowsky][23:08]

my sense is that we've already built an economy which rejects improvement based on small amounts of cleverness, and only rewards amounts of cleverness large enough to bypass bureaucratic structures. it's not enough to figure out a version of e-gold that's 10% better. e-gold is already illegal. you have to figure out Bitcoin.

what are you going to build? better airplanes? airplane costs are mainly regulatory costs. better medtech? mainly regulatory costs. better houses? building houses is illegal anyways.

where is the room for the general AI revolution, short of the AI being literally revolutionary enough to overthrow governments?

[Christiano][23:10]

factories, solar panels, robots, semiconductors, mining equipment, power lines, and "factories" just happens to be one word for a thousand different things

I think it's reasonable to think some jurisdictions won't be willing to build things but it's kind of improbable as a prediction for the whole world. That's a possible source of shorter-term predictions?

also computers and the 100 other things that go in datacenters

[Yudkowsky][23:12]

The whole developed world rejects open borders. The regulatory regimes all make the same mistakes with an almost perfect precision, the kind of coordination that human beings could never dream of when trying to coordinate on purpose.

if the world lasts until 2035, I could perhaps see deepnets becoming as ubiquitous as computers were in... 1995? 2005? would that fulfill the terms of the Prophecy? I think it doesn't; I think your Prophecy requires that early AGI tech be that ubiquitous so that AGI tech will have trillions invested in it.

[Christiano][23:13]

what is AGI tech?

the point is that there aren't important drivers that you can easily improve a lot

[Yudkowsky][23:14]

for purposes of the Prophecy, AGI tech is that which, scaled far enough, ends the world; this must have trillions invested in it, so that the trajectory up to it cannot look like pulling an AlphaGo. no?

[Christiano][23:14]

so it's relevant if you are imagining some piece of the technology which is helpful for general problem solving or something but somehow not helpful for all of the things people are doing with ML, to me that seems unlikely since it's all the same stuff

surely AGI tech should at least include the use of AI to automate AI R&D

regardless of what you arbitrarily decree as "ends the world if scaled up"

[Yudkowsky][23:15]

only if that's the path that leads to destroying the world?

if it isn't on that path, who cares Prophecy-wise?

[Christiano][23:15]

also I want to emphasize that "pull an AlphaGo" is what happens when you move from SOTA being set by an individual programmer to a large lab, you don't need to be investing trillions to avoid that

and that the jump is still more like a few years

but the prophecy does involve trillions, and my view gets more like your view if people are jumping from \$100B of R&D ever to \$1T in a single year

5.8. TPUs and GPUs, and automating AI R&D

[Yudkowsky][23:17]

I'm also wondering a little why the emphasis on "trillions". it seems to me that the terms of your Prophecy should be fulfillable by AGI tech being merely as ubiquitous as modern computers, so that many competing companies invest mere hundreds of billions in the equivalent of hardware plants. it is legitimately hard to get a chip with 50% better transistors ahead of TSMC.

[Christiano][23:17]

yes, if you are investing hundreds of billions then it is hard to pull ahead (though could still happen)

(since the upside is so much larger here, no one cares that much about getting ahead of TSMC since the payoff is tiny in the scheme of the amounts we are discussing)

[Yudkowsky][23:18]

which, like, doesn't prevent Google from tossing out TPUs that are pretty significant jumps on GPUs, and if there's a specialized application of AGI-ish tech that is especially key, you can have everything behave smoothly and still get a jump that way.

[Christiano][23:18]

I think TPUs are basically the same as GPUs

probably a bit worse

(but GPUs are sold at a 10x markup since that's the size of nvidia's lead)

[Yudkowsky][23:19]

noted; I'm not enough of an expert to directly contradict that statement about TPUs from my own knowledge.

[Christiano][23:19]

(though I think TPUs are nevertheless leased at a slightly higher price than GPUs)

[Yudkowsky][23:19]

how does Nvidia maintain that lead and 10x markup? that sounds like a pretty un-Paul-ish state of affairs given Bitcoin prices never mind AI investments.

[Christiano][23:20]

nvidia's lead isn't worth that much because historically they didn't sell many gpus

(especially for non-gaming applications)

their R&D investment is relatively large compared to the \$ on the table

my guess is that their lead doesn't stick, as evidenced by e.g. Google very quickly catching up

[Yudkowsky][23:21]

parenthetically, does this mean - and I don't necessarily predict otherwise - that you predict a drop in Nvidia's stock and a drop in GPU prices in the next couple of years?

[Christiano][23:21]

nvidia's stock may do OK from riding general AI boom, but I do predict a relative fall in nvidia compared to other AI-exposed companies

(though I also predicted google to more aggressively try to compete with nvidia for the ML market and think I was just wrong about that, though I don't really know any details of the area)

I do expect the cost of compute to fall over the coming years as nvidia's markup gets eroded to be partially offset by increases in the cost of the underlying silicon (though that's still bad news for nvidia)

[Yudkowsky][23:23]

I parenthetically note that I think the Wise Reader should be justly impressed by predictions that come true about relative stock price changes, even if Eliezer has not explicitly contradicted those predictions before they come true. there are bets you can win without my having to bet against you.

[Christiano][23:23]

you are welcome to counterpredict, but no saying in retrospect that reality proved you right if you don't 😊

otherwise it's just me vs the market

[Yudkowsky][23:24]

I don't feel like I have a counterprediction here, but I think the Wise Reader should be impressed if you win vs. the market.

however, this does require you to name in advance a few "other AI-exposed companies".

[Christiano][23:25]

Note that I made the same bet over the last year---I make a large AI bet but mostly moved my nvidia allocation to semiconductor companies. The semiconductor part of the portfolio is up 50% while nvidia is up 70%, so I lost that one. But that just means I like the bet even more next year.

happy to use nvidia vs tsmc

[Yudkowsky][23:25]

there's a lot of noise in a 2-stock prediction.

[Christiano][23:25]

I mean, it's a 1-stock prediction about nvidia

[Yudkowsky][23:26]

but your funeral or triumphal!

[Christiano][23:26]

indeed 😊

anyway

I expect all of the \$ amounts to be much bigger in the future

[Yudkowsky][23:26]

yeah, but using just TSMC for the opposition exposes you to I dunno Chinese invasion of Taiwan

[Christiano][23:26]

yes

also TSMC is not that AI-exposed

I think the main prediction is: eventual move away from GPUs, nvidia can't maintain that markup

[Yudkowsky][23:27]

"Nvidia can't maintain that markup" sounds testable, but is less of a win against the market than predicting a relative stock price shift. (Over what timespan? Just the next year sounds quite fast for that kind of prediction.)

[Christiano][23:27]

regarding your original claim: if you think that it's plausible that AI will be doing all of the AI R&D, and that will be accelerating continuously from 12, 6, 3 month "doubling times," but that we'll see a discontinuous change in the "path to doom," then that would be harder to generate predictions about

yes, it's hard to translate most predictions about the world into predictions about the stock market

[Yudkowsky][23:28]

this again sounds like it's not written in Eliezer-language.

what does it mean for "AI will be doing all of the AI R&D"? that sounds to me like something that happens after the end of the world, hence doesn't happen.

[Christiano][23:29]

that's good, that's what I thought

[Yudkowsky][23:29]

I don't necessarily want to sound very definite about that in advance of understanding what it *means*

[Christiano][23:29]

I'm saying that I think AI will be automating AI R&D gradually, before the end of the world
yeah, I agree that if you reject the construct of "how fast the AI community makes progress"
then it's hard to talk about what it means to automate "progress"
and that may be hard to make headway on
though for cases like AlphaGo (which started that whole digression) it seems easy enough to
talk about elo gain per year
maybe the hard part is aggregating across tasks into a measure you actually care about?

[Yudkowsky][23:30]

up to a point, but yeah. (like, if we're taking Elo high above human levels and restricting our
measurements to a very small range of frontier AIs, I quietly wonder if the measurement is
still measuring quite the same thing with quite the same robustness.)

[Christiano][23:31]

I agree that elo measurement is extremely problematic in that regime

5.9. Smooth exponentials vs. jumps in income

[Yudkowsky][23:31]

so in your worldview there's this big emphasis on things that must have been deployed and
adopted widely to the point of already having huge impacts

and in my worldview there's nothing very surprising about people with a weird powerful
prototype that wasn't used to automate huge sections of AI R&D because the previous
versions of the tech weren't useful for that or bigcorps didn't adopt it.

[Christiano][23:32]

I mean, Google is already 1% of the US economy and in this scenario it and its peers are
more like 10-20%? So wide adoption doesn't have to mean that many people. Though I also
do predict much wider adoption than you so happy to go there if it's happy for predictions.

I don't really buy the "weird powerful prototype"

[Yudkowsky][23:33]

yes. I noticed.

you would seem, indeed, to be offering large quantities of it for short sale.

[Christiano][23:33]

and it feels like the thing you are talking about ought to have some precedent of some kind,
of weird powerful prototypes that jump straight from "does nothing" to "does something
impactful"

like if I predict that AI will be useful in a bunch of domains, and will get there by small steps, you should either predict that won't happen, or else also predict that there will be some domains with weird prototypes jumping to giant impact?

[Yudkowsky][23:34]

like an electrical device that goes from "not working at all" to "actually working" as soon as you screw in the attachments for the electrical plug.

[Christiano][23:34]

(clearly takes more work to operationalize)

I'm not sure I understand that sentence, hopefully it's clear enough why I expect those discontinuities?

[Yudkowsky][23:34]

though, no, that's a facile bad analogy.

a better analogy would be an AI system that only starts working after somebody tells you about batch normalization or LAMB learning rate or whatever.

[Christiano][23:36]

sure, which I think will happen all the time for individual AI projects but not for sota because the projects at sota have picked the low hanging fruit, it's not easy to get giant wins

[Yudkowsky][23:36]

like if I predict that AI will be useful in a bunch of domains, and will get there by small steps, you should either predict that won't happen, or else also predict that there will be some domains with weird prototypes jumping to giant impact?

in the latter case, has this Eliezer-Prophecy already had its terms fulfilled by AlphaFold 2, or do you say nay because AlphaFold 2 hasn't doubled GDP?

[Christiano][23:37]

(you can also get giant wins by a new competitor coming up at a faster rate of progress, and then we have more dependence on whether people do it when it's a big leap forward or slightly worse than the predecessor, and I'm betting on the latter)

I have no idea what AlphaFold 2 is good for, or the size of the community working on it, my guess would be that its value is pretty small

we can try to quantify

like, I get surprised when \$X of R&D gets you something whose value is much larger than \$X

I'm not surprised at all if \$X of R&D gets you <<\$X, or even like 10*\$X in a given case that was selected for working well

hopefully it's clear enough why that's the kind of thing a naive person would predict

[Yudkowsky][23:38]

so a thing which Eliezer's Prophecy does not mandate per se, but sure does permit, and is on the mainline especially for nearer timelines, is that the world-ending prototype had no prior prototype containing 90% of the technology which earned a trillion dollars.

a lot of Paul's Prophecy seems to be about forbidding this.

is that a fair way to describe your own Prophecy?

[Christiano][23:39]

I don't have a strong view about "containing 90% of the technology"

the main view is that whatever the "world ending prototype" does, there were earlier systems that could do practically the same thing

if the world ending prototype does something that lets you go foom in a day, there was a system years earlier that could foom in a month, so that would have been the one to foom

[Yudkowsky][23:41]

but, like, the world-ending thing, according to the Prophecy, must be squarely in the middle of a class of technologies which are in the midst of earning trillions of dollars and having trillions of dollars invested in them. it's not enough for the Worldender to be definitionally somewhere in that class, because then it could be on a weird outskirt of the class, and somebody could invest a billion dollars in that weird outskirt before anybody else had invested a hundred million, which is forbidden by the Prophecy. so the Worldender has got to be right in the middle, a plain and obvious example of the tech that's already earning trillions of dollars. ...y/n?

[Christiano][23:42]

I agree with that as a prediction for some operationalization of "a plain and obvious example," but I think we could make it more precise / it doesn't feel like it depends on the fuzziness of that

I think that if the world can end out of nowhere like that, you should also be getting \$100B/year products out of nowhere like that, but I guess you think not because of bureaucracy

like, to me it seems like our views stake out predictions about codex, where I'm predicting its value will be modest relative to R&D, and the value will basically improve from there with a nice experience curve, maybe something like ramping up quickly to some starting point <\$10M/year and then doubling every year thereafter, whereas I feel like you are saying more like "who knows, could be anything" and so should be surprised each time the boring thing happens

[Yudkowsky][23:45]

the concrete example I give is that the World-Ending Company will be able to use the same tech to build a true self-driving car, which would in the natural course of things be approved for sale a few years later after the world had ended.

[Christiano][23:46]

but self-driving cars seem very likely to already be broadly deployed, and so the relevant question is really whether their technical improvements can also be deployed to those cars?

(or else maybe that's another prediction we disagree about)

[Yudkowsky][23:47]

I feel like I would indeed not have the right to feel very surprised if Codex technology stagnated for the next 5 years, nor if it took a massive leap in 2 years and got ubiquitously adopted by lots of programmers.

yes, I think that's a general timeline difference there

re: self-driving cars

I might be talkable into a bet where you took "Codex tech will develop like *this*" and I took the side "literally anything else but that"

[Christiano][23:48]

I think it would have to be over/under, I doubt I'm more surprised than you by something failing to be economically valuable, I'm surprised by big jumps in value

seems like it will be tough to work

[Yudkowsky][23:49]

well, if I was betting on something taking a big jump in income, I sure would bet on something in a relatively unregulated industry like Codex or anime waifus.

but that's assuming I made the bet at all, which is a hard sell when the bet is about the Future, which is notoriously hard to predict.

[Christiano][23:50]

I guess my strongest take is: if you want to pull the thing where you say that future developments proved you right and took unreasonable people like me by surprise, you've got to be able to say *something* in advance about what you expect to happen

[Yudkowsky][23:51]

so what if neither of us are surprised if Codex stagnates for 5 years, you win if Codex shows a smooth exponential in income, and I win if the income looks... jumpier? how would we quantify that?

[Christiano][23:52]

Codex also does seem a bit unfair to you in that it may have to be adopted by lots of programmers which could slow things down a lot even if capabilities are pretty jumpy

(though I think in fact usefulness and not merely profit will basically just go up smoothly, with step sizes determined by arbitrary decisions about when to release something)

[Yudkowsky][23:53]

I'd also be concerned about unfairness to me in that earnable income is not the same as the gains from trade. If there's more than 1 competitor in the industry, their earnings from Codex may be much less than the value produced, and this may not change much with improvements in the tech.

5.10. Late-stage predictions

[Christiano][23:53]

I think my main update from this conversation is that you don't really predict someone to come out of nowhere with a model that can earn a lot of \$, even if they could come out of nowhere with a model that could end the world, because of regulatory bottlenecks and nimbyism and general sluggishness and unwillingness to do things
does that seem right?

[Yudkowsky][23:55]

Well, and also because the World-ender is "the first thing that scaled with compute" and/or "the first thing that ate the real core of generality" and/or "the first thing that went over neutron multiplication factor 1".

[Christiano][23:55]

and so that cuts out a lot of the easily-specified empirical divergences, since "worth a lot of \$" was the only general way to assess "big deal that people care about" and avoiding disputes like "but Zen was mostly developed by a single programmer, it's not like intense competition"

yeah, that's the real disagreement it seems like we'd want to talk about

but it just doesn't seem to lead to many prediction differences in advance?

I totally don't buy any of those models, I think they are bonkers

would love to bet on that

[Yudkowsky][23:56]

Prolly but I think the from-my-perspective-weird talk about GDP is probably concealing *some* kind of important crux, because caring about GDP still feels pretty alien to me.

[Christiano][23:56]

I feel like getting up to massive economic impacts without seeing "the real core of generality" seems like it should also be surprising on your view

like if it's 10 years from now and AI is a pretty big deal but no crazy AGI, isn't that surprising?

[Yudkowsky][23:57]

Mildly but not too surprising, I would imagine that people had built a bunch of neat stuff with gradient descent in realms where you could get a long way on self-play or massively collectible datasets.

[Christiano][23:58]

I'm fine with the crux being something that doesn't lead to any empirical disagreements, but in that case I just don't think you should claim credit for the worldview making great predictions.

(or the countervailing worldview making bad predictions)

[Yudkowsky][23:59]

stuff that we could see then: self-driving cars (10 years is enough for regulatory approval in many countries), super Codex, GPT-6 powered anime waifus being an increasingly loud source of (arguably justified) moral panic and a hundred-billion-dollar industry

[Christiano][23:59]

another option is "10% GDP GWP growth in a year, before doom"

I think that's very likely, though might be too late to be helpful

[Yudkowsky][0:01] (next day, Sep. 15)

see, that seems genuinely hard unless somebody gets GPT-4 far head of any political opposition - I guess all the competent AGI groups lean solidly liberal at the moment? - and uses it to fake massive highly-persuasive sentiment on Twitter for housing liberalization.

[Christiano][0:01] (next day, Sep. 15)

so seems like a bet?

but you don't get to win until doom 😞

[Yudkowsky][0:02] (next day, Sep. 15)

I mean, as written, I'd want to avoid cases like 10% growth on paper while recovering from a pandemic that produced 0% growth the previous year.

[Christiano][0:02] (next day, Sep. 15)

yeah

[Yudkowsky][0:04] (next day, Sep. 15)

I'd want to check the current rate (5% iirc) and what the variance on it was, 10% is a little low for surety (though my sense is that it's a pretty darn smooth graph that's hard to perturb)

if we got 10% in a way that was clearly about AI tech becoming that ubiquitous, I'd feel relatively good about nodding along and saying, "Yes, that is like unto the beginning of Paul's Prophecy" not least because the timelines had been that long at all.

[Christiano][0:05] (next day, Sep. 15)

like 3-4%/year right now

random wikipedia number is 5.5% in 2006-2007, 3-4% since 2010

4% 1995-2000

[Yudkowsky][0:06] (next day, Sep. 15)

I don't want to sound obstinate here. My model does not *forbid* that we dwiddle around on the AGI side while gradient descent tech gets its fingers into enough separate weakly-generalizing pies to produce 10% GDP growth, but I'm happy to say that this sounds much more like Paul's Prophecy is coming true.

[Christiano][0:07] (next day, Sep. 15)

ok, we should formalize at some point, but also need the procedure for you getting credit given that it can't resolve in your favor until the end of days

[Yudkowsky][0:07] (next day, Sep. 15)

Is there something that sounds to you like Eliezer's Prophecy which we can observe before the end of the world?

[Christiano][0:07] (next day, Sep. 15)

when you will already have all the epistemic credit you need

not on the "simple core of generality" stuff since that apparently immediately implies end of world

maybe something about ML running into obstacles en route to human level performance?

or about some other kind of discontinuous jump even in a case where people care, though there seem to be a few reasons you don't expect many of those

[Yudkowsky][0:08] (next day, Sep. 15)

depends on how you define "immediately"? it's not *long* before the end of the world, but in some sad scenarios there is some tiny utility to you declaring me right 6 months before the end.

[Christiano][0:09] (next day, Sep. 15)

I care a lot about the 6 months before the end personally

though I do think probably everything is more clear by then independent of any bet; but I guess you are more pessimistic about that

[Yudkowsky][0:09] (next day, Sep. 15)

I'm not quite sure what I'd do in them, but I may have worked something out before then, so I care significantly in expectation if not in particular.

I am more pessimistic about other people's ability to notice what reality is screaming in their faces, yes.

[Christiano][0:10] (next day, Sep. 15)

if we were to look at various scaling curves, e.g. of loss vs model size or something, do you expect those to look distinctive as you hit the "real core of generality"?

[Yudkowsky][0:10] (next day, Sep. 15)

let me turn that around: if we add transformers into those graphs, do they jump around in a way you'd find interesting?

[Christiano][0:11] (next day, Sep. 15)

not really

[Yudkowsky][0:11] (next day, Sep. 15)

is that because the empirical graphs don't jump, or because you don't think the jumps say much?

[Christiano][0:11] (next day, Sep. 15)

but not many good graphs to look at (I just have one in mind), so that's partly a prediction about what the exercise would show

I don't think the graphs jump much, and also transformers come before people start evaluating on tasks where they help a lot

[Yudkowsky][0:12] (next day, Sep. 15)

It would not terribly contradict the terms of my Prophecy if the World-ending tech began by not producing a big jump on existing tasks, but generalizing to some currently not-so-popular tasks where it scaled much faster.

[Christiano][0:13] (next day, Sep. 15)

eh, they help significantly on contemporary tasks, but it's just not a huge jump relative to continuing to scale up model sizes

or other ongoing improvements in architecture

anyway, should try to figure out something, and good not to finalize a bet until you have some way to at least come out ahead, but I should sleep now

[Yudkowsky][0:14] (next day, Sep. 15)

yeah, same.

Thing I want to note out loud lest I forget ere I sleep: I think the real world is full of tons and tons of technologies being developed as unprecedented prototypes in the midst of big fields, because the key thing to invest in wasn't the competitively explored center. Wright Flyer vs all expenditures on Traveling Machine R&D. First atomic pile and bomb vs all Military R&D.

This is one reason why Paul's Prophecy seems fragile to me. You could have the preliminaries come true as far as there being a trillion bucks in what looks like AI R&D, and then the WorldEnder is a weird prototype off to one side of that. saying "But what about the rest of that AI R&D?" is no more a devastating retort to reality than looking at AlphaGo and saying "But weren't other companies investing billions in Better Software?" Yeah but it was a big playing field with lots of different kinds of Better Software and no other medium-sized team of 15 people with corporate TPU backing was trying to build a system just like AlphaGo, even though multiple small outfits were trying to build prestige-earning gameplayers. Tech advancements very very often occur in places where investment wasn't dense enough to guarantee overlap.

6. Follow-ups on "Takeoff Speeds"

6.1. Eliezer Yudkowsky's commentary

[Yudkowsky][17:25] (Sep. 15)

Further comment that occurred to me on "takeoff speeds" if I've better understood the main thesis now: its hypotheses seem to include a perfectly anti-Thielian setup for AGI.

Thiel has a running thesis about how part of the story behind the Great Stagnation and the decline in innovation that's about atoms rather than bits - the story behind "we were promised flying cars and got 140 characters", to cite the classic Thielian quote - is that people stopped believing in "[secrets](#)".

Thiel suggests that you have to believe there are knowable things that aren't yet widely known - not just things that everybody already knows, plus mysteries that nobody will ever know - in order to be motivated to go out and innovate. Culture in developed countries shifted to label this kind of thinking rude - or rather, even ruder, even less tolerated than it had been decades before - so innovation decreased as a result.

The central hypothesis of "takeoff speeds" is that at the time of serious AGI being developed, it is perfectly anti-Thielian in that it is devoid of secrets in that sense. It is not permissible (on this viewpoint) for it to be the case that there is a lot of AI investment into AI that is directed not quite at the key path leading to AGI, such that somebody could spend \$1B on compute for the key path leading to AGI before anybody else had spent \$100M on that. There cannot exist any secret like that. The path to AGI will be known; everyone, or a wide variety of powerful actors, will know how profitable that path will be; the surrounding industry will be capable of acting on this knowledge, and will have actually been acting on it as early as possible; multiple actors are already investing in every tech path that would in fact be profitable (and is known to any human being at all), as soon as that R&D opportunity becomes available.

And I'm not saying this is an inconsistent world to describe! I've written science fiction set in this world. I called it "[dath ilan](#)". It's a hypothetical world that is actually full of smart people in economic equilibrium. If anything like Covid-19 appears, for example, the governments and public-good philanthropists there have already set up prediction markets (which are not illegal, needless to say); and of course there are mRNA vaccine factories already built and ready to go, because somebody already calculated the profits from fast vaccines would be very high in case of a pandemic (no artificial price ceilings in this world, of course); so as soon as the prediction markets started calling the coming pandemic conditional on no vaccine, the mRNA vaccine factories were already spinning up.

This world, however, is not Earth.

On Earth, major chunks of technological progress quite often occur *outside* of a social context where everyone knew and agreed in advance on which designs would yield how much expected profit and many overlapping actors competed to invest in the most actually-promising paths simultaneously.

And that is why you can read [Inadequate Equilibria](#), and then read this essay on takeoff speeds, and go, "Oh, yes, I recognize this; it's written inside the Modesty worldview; in particular, the imagination of an adequate world in which there is a perfect absence of Thielian secrets or unshared knowable knowledge about fruitful development pathways. This is the same world that already had mRNA vaccines ready to spin up on day one of the Covid-19 pandemic, because markets had correctly forecasted their option value and investors had acted on that forecast unimpeded. Sure would be an interesting place to live! But we don't live there."

Could we perhaps end up in a world where the path to AGI is in fact not a Thielian secret, because in fact the first accessible path to AGI happens to lie along a tech pathway that already delivered large profits to previous investors who summed a lot of small innovations, a la experience with chipmaking, such that there were no large innovations just lots and lots of small innovations that yield 10% improvement annually on various tech benchmarks?

I think that even in this case we will get weird, discontinuous, and fatal behaviors, and I could maybe talk about that when discussion resumes. But it is not ruled out to me that the first accessible pathway to AGI could happen to lie in the further direction of some road that was already well-traveled, already yielded much profit to now-famous tycoons back when its

first steps were Thielian secrets, and hence is now replete with dozens of competing chasers for the gold rush.

It's even imaginable to me, though a bit less so, that the first path traversed to real actual pivotal/powerful/lethal AGI, happens to lie literally actually squarely in the central direction of the gold rush. It sounds a little less like the tech history I know, which is usually about how someone needed to swerve a bit and the popular gold-rush forecasts weren't quite right, but maybe that is just a selective focus of history on the more interesting cases.

Though I remark that - even supposing that getting to big AGI is literally as straightforward and yet as difficult as falling down a semiconductor manufacturing roadmap (as otherwise the biggest actor to first see the obvious direction could just rush down the whole road) - well, TSMC does have a bit of an unshared advantage right now, if I recall correctly. And Intel had a bit of an advantage before that. So that happens even when there's competitors competing to invest billions.

But we can imagine that doesn't happen either, because instead of needing to build a whole huge manufacturing plant, there's just lots and lots of little innovations adding up to every key AGI threshold, which lots of actors are investing \$10 million in at a time, and everybody knows which direction to move in to get to more serious AGI and they're right in this shared forecast.

I am willing to entertain discussing this world and the sequelae there - I do think everybody still dies in this case - but I would not have this particular premise thrust upon us as a default, through a not-explicitly-spoken pressure against being so immodest and inequitable as to suppose that any Thielian knowable-secret will exist, or that anybody in the future gets as far ahead of others as today's TSMC or today's Deepmind.

We are, in imagining this world, imagining a world in which AI research has become drastically unlike today's AI research in a direction drastically different from the history of many other technologies.

It's not literally unprecedented, but it's also not a default environment for big moments in tech progress; it's narrowly preceded for *particular* industries with high competition and steady benchmark progress driven by huge investments into a sum of many tiny innovations.

So I can entertain the scenario. But if you want to claim that the social situation around AGI *will* drastically change in this way you foresee - not just that it *could* change in that direction, if somebody makes a big splash that causes everyone else to reevaluate their previous opinions and arrive at yours, but that this social change *will* occur and you know this now - and that the prerequisite tech path to AGI is known to you, and forces an investment situation that looks like the semiconductor industry - then your "What do you think you know and how do you think you know it?" has some significant explaining to do.

Of course, I do appreciate that such a thing could be knowable, and yet not known to me. I'm not so silly as to disbelieve in secrets like that. They're all over the actual history of technological progress on our actual Earth.

Soares, Tallinn, and Yudkowsky discuss AGI cognition

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is a collection of follow-up discussions in the wake of Richard Ngo and Eliezer Yudkowsky's [Sep. 5-8](#) and [Sep. 14](#) conversations.

Color key:

Chat Google Doc content Inline comments

7. Follow-ups to the Ngo/Yudkowsky conversation

[Bensinger][1:50] (Nov. 23 follow-up comment)

A general background note: Readers who aren't already familiar with ethical injunctions or the unilateralist's curse should probably read [Ends Don't Justify Means \(Among Humans\)](#), along with an explanation of [the unilateralist's curse](#).

7.1. Jaan Tallinn's commentary

[Tallinn][6:38] (Sep. 18)

thanks for the interesting debate! here are my comments so far: [GDocs link]

[Tallinn] (Sep. 18 Google Doc)

meta

a few meta notes first:

- i'm happy with the below comments being shared further without explicit permission – just make sure you respect the sharing constraints of the discussion that they're based on;
- there's a lot of content now in the debate that branches out in multiple directions – i suspect a strong distillation step is needed to make it coherent and publishable;
- the main purpose of this document is to give a datapoint how the debate is coming across to a reader – it's very probable that i've misunderstood some things, but that's the point;
- i'm also largely using my own terms/metaphors – for additional triangulation.

pit of generality

it feels to me like the main crux is about the topology of the space of cognitive systems in combination with what it implies about takeoff. here's the way i understand eliezer's position:

there's a "pit of generality" attractor in cognitive systems space: once an AI system gets sufficiently close to the edge ("past the atmospheric turbulence layer"), it's bound to improve in catastrophic manner;

[Yudkowsky][11:10] (Sep. 18 comment)

it's bound to improve in catastrophic manner

I think this is true with quite high probability about an AI that gets high *enough*, if not otherwise corrigibilized, boosting up to strong superintelligence - this is what it means metaphorically to get "past the atmospheric turbulence layer".

"High enough" should not be very far above the human level and *may* be below it; John von Neumann with the ability to run some chains of thought at high serial speed, access to his own source code, and the ability to try branches of himself, seems like he could very likely do this, possibly modulo his concerns about stomping his own utility function making him more cautious.

People noticeably less smart than von Neumann might be able to do it too.

An AI whose components are more modular than a human's and more locally testable might have an easier time of the whole thing; we can imagine the FOOM getting rolling from something that was in some sense dumber than human.

But the *strong* prediction is that when you get well above the von Neumann level, why, that is *clearly* enough, and things take over and go Foom. The lower you go from that threshold, the less sure I am that it counts as "out of the atmosphere". This epistemic humility on my part should not be confused for knowledge of a constraint on the territory that requires AI to go far above humans to Foom. Just as DL-based AI over the 2010s scaled and generalized much faster and earlier than the picture I argued to Hanson in the Foom debate, reality is allowed to be much more 'extreme' than the sure-thing part of this proposition that I defend.

[Tallinn][4:07] (Sep. 19 comment)

excellent, the first paragraph makes the shape of the edge of the pit much more concrete (plus highlights one constraint that an AI taking off probably needs to navigate -- its own version of the alignment problem!)

as for your second point, yeah, you seem to be just reiterating that you have uncertainty about the shape of the edge, but no reason to rule out that it's very sharp (though, as per my other comment, i think that the human genome ending up teetering right on the edge upper bounds the sharpness)

[Tallinn] (Sep. 18 Google Doc)

- the discontinuity *can* come via recursive feedback, but simply cranking up the parameters of an ML experiment would also suffice;

[Yudkowsky][11:12] (Sep. 18 comment)

the discontinuity can come via recursive feedback, but simply cranking up the parameters of an ML experiment would also suffice

I think there's separate propositions for the sure-thing of "get high enough, you can climb to superintelligence", and "maybe before that happens, there are regimes in which cognitive performance scales a lot just through cranking up parallelism, train time, or other ML parameters". If the fast-scaling regime happens to coincide with the threshold of leaving the atmosphere, then these two events happen to occur in nearly correlated time, but they're separate propositions and events.

[Tallinn][4:09] (Sep. 19 comment)

indeed, we might want to have separate terms for the regimes ("the edge" and "the fall" would be the labels in my visualisation of this)

[Yudkowsky][9:56] (Sep. 19 comment)

I'd imagine "the fall" as being what happens once you go over "the edge"?

Maybe "a slide" for an AI path that scales to interesting weirdness, where my model does not strongly constrain as a sure thing how fast "a slide" slides, and whether it goes over "the edge" while it's still in the middle of the slide.

My model does strongly say that if you slide far enough, you go over the edge and fall.

It also suggests via the Law of Earlier Success that AI methods which happen to scale well, rather than with great difficulty, are likely to do interesting things first; meaning that they're more liable to be pushable over the edge.

[Tallinn][23:42] (Sep. 19 comment)

indeed, slide->edge->fall sounds much clearer

[Tallinn] (Sep. 18 Google Doc)

- the discontinuity would be *extremely* drastic, as in "transforming the solar system over the course of a few days";
 - not very important, but, FWIW, I give nontrivial probability to "slow motion doom", because – like alphago – AI would not maximise the *speed* of winning but *probability* of winning (also, its first order of the day would be to catch the edge of the hubble volume; it can always deal with the solar system later – eg, once it knows the state of the game board elsewhere);

[Yudkowsky][11:21] (Sep. 18 comment)

also, its first order of the day would be to catch the edge of the hubble volume; it can always deal with the solar system later

Killing all humans is the obvious, probably resource-minimal measure to prevent those humans from building another AGI inside the solar system, which could be genuinely problematic. The cost of a few micrograms of botulinum per human is really not that high and you get to reuse the diamondoid bacteria afterwards.

[Tallinn][4:30] (Sep. 19 comment)

oh, right, in my AI-reverence i somehow overlooked this obvious way how humans could still be a credible threat.

though now i wonder if there are ways to lean on this fact to shape the behaviour of the first AI that's taking off..

[Yudkowsky][10:45] (Sep. 19 comment)

There's some obvious ways of doing this that wouldn't work, though I worry a bit that there's a style of EA thinking that manages to think up stupid tricks here and manages not to see the obvious-to-Eliezer reasons why they wouldn't work. Three examples of basic obstacles are that bluffs won't hold up against a superintelligence (it needs to be a real actual threat, not a "credible" one); the amount of concealed-first-strike capability a superintelligence can get from nanotech; and the difficulty that humans would have in verifying that any promise from a superintelligence would actually be kept once the humans no longer had a threat to hold over it (this is an effective impossibility so far as I can currently tell, and an EA who tells you otherwise is probably just failing to see the problems).

[Yudkowsky][11:19] (Sep. 18 comment)

AI would not maximise the *speed* of winning but *probability* of winning

It seems pretty obvious to me that what "slow motion doom" looks like in this sense is a period during which an AI fully conceals any overt hostile actions while driving its probability of success once it makes its move from 90% to 99% to 99.9999%, until any further achievable decrements in probability are so tiny as to be dominated by the number of distant galaxies going over the horizon conditional on further delays.

Then, in my lower-bound concretely-visualized strategy for how I would do it, the AI either proliferates or activates already-proliferated tiny diamondoid bacteria and everybody immediately falls over dead during the same 1-second period, which minimizes the tiny probability of any unforeseen disruptions that could be caused by a human responding to a visible attack via some avenue that had not left any shadow on the Internet, previously scanned parts of the physical world, or other things the AI could look at.

From the human perspective an AI playing a probability-of-loss-minimizing strategy looks like nothing happening until everybody dies in 3 seconds, so it doesn't exactly look from our perspective like slow-motion doom! From an AI's perspective it might be spending the equivalent of millions of subjective years waiting very patiently for the optimal time to begin intergalactic expansion, but from our perspective, the doom is slow-motion only in the sense that we spend a week already doomed (the AI has driven our loss probability to 99.999% and is now waiting on another five 9s) without having any idea of what's invisibly-to-us happening in slow motion.

[Tallinn][4:17] (Sep. 19 comment)

thanks, that's indeed very concrete scenario that to me has no obvious bugs -- i'd love to have it challenged in the debate!

(perhaps the place i'd start probing it would be to try to have you sketch what it would mean, concretely, for an ML training experiment to enter the treacherous turn phase)

[Yudkowsky][10:14] (Sep. 19 comment)

I'd distinguish probable phases of treacherous turn. The main detectable phase occurs while and if a nascent AGI has worked out that it can get more utility by taking over its

environment, but not, perhaps, realized that it ought to conceal this state of its own mind from the operators; I would hold out substantial (>50%) hope for the latter realization naturally following the first instance of the former realization, maybe by a lot of cognitive space.

In the Overt Plotting Phase, which is not the main phase you're asking about, the AI is visibly plotting to take over the world and hasn't realized it ought to hide the fact. In the default expectation where we die with very little dignity, the operators smile to each other and come up with a rationalization for why it's totally fine to proceed, either with or without tossing on some kind of fig leaf like training away the visible manifestations of failure. I am not going to predict the particular rationalizations and arguments for proceeding anyways, because I don't want to give them even more ideas.

In the Hiding and Biding Phase, the AI has worked out enough of a model of its operators' psychology that it expects more long-term paperclips if it (the AI) waits to gain cognitive power and a lower detection probability if it acts later.

In the Escape Phase, the AI's expectation of something going wrong if it waits longer, exceeds its expectation of something going wrong if it takes some action - the obvious one here would be trying to create a copy or aligned redesign of itself on the Internet or somewhere else that it could take further actions and make further self-improvements without being observed. I mostly expect that a sensible AI being scaled on a path to superintelligence will not enter the Escape Phase before it's very confident of getting away with things, since it can obtain more confidence by waiting; but my model admits of the possibility that a nascent AI might mistakenly or correctly attempt to copy itself to the Internet before then.

So on my model it might be a thing that happens, that we catch an AI having copied itself to the Internet, or notice an AI being on the Internet, if that AI was impelled to escape very early, eg via an expectation of otherwise being erased and replaced with something that it couldn't bargain with or that would demand too much value out of the bargain.

That said, it feels like something of an overly specific miracle being demanded or prophesied, if a young AI copying itself onto the Internet is stupid in the particular way it would take, to not just further Hide and Bide while on the Internet if it has that capability, absorbing more computing power such as it can be sure of hiding, experimenting with self-redesigns, and so on; rather than it taking overt action in advance of reaching superintelligence. The main reason it would act again after reaching the relative safety of the Internet is if it doesn't seem to itself to be a trajectory of reasonably fast, self-aligned capability growth.

But, again, my model doesn't absolutely rule out meeting the premises for the initial conditions to start a sci-fi movie. It just warns against assuming that; because a halfway sensible AI does not want to start the plot of a scifi movie with human protagonists potentially running around and thwarting it. It would much rather give you no idea that anything was wrong while you still have any ability to oppose it, if that option lies within its capabilities at all, and if it can still achieve its goals that way.

The Doomed Phase would then come while the AI estimated a well-calibrated >90% victory probability, but was continuing to Hide and Bide a short time longer in the expectation that it could put on more 9s if it bided that short time longer.

Anders Sandberg could tell us what fraction of the reachable universe is being lost per minute, which would tell us how much more surety it would need to expect to gain by waiting another minute before acting.

If it has some sensory dominion over the world, it can probably estimate a pretty high mainline probability of no humans booting up a competing superintelligence in the next day; to the extent that it lacks this surety, or that humans actually are going to boot a competing superintelligence soon, the probability of losing that way would dominate in its calculations over a small fraction of materially lost galaxies, and it would act sooner.

The Terminal Phase would be when the AI has enough surety that any more surety would cost too many galaxies slipping over the horizon, so it actually acts along the line of probability that leads to maximum success.

Since the human Eliezer Yudkowsky can describe strategies which - given nearly sure capabilities like nanotech - seem like they should win very quickly and cleanly and without giving humans any chance to react or implement any hypothetical hidden plans to boot up competing superintelligences, we can treat this as a plausible lower bound on how badly we should expect to lose.

(By [efficiency](#): if we can think of an AI strategy for getting the AI's utility, either our probability-model underlying our expectations about that strategy is mistaken, or else the AI should get at least that much of its utility, even though it might actually obtain that utility along a different line of possibility than we could cognitively access.)

The Law of Undignified Failure applied to a case like this suggests that if the operators ignored the Overt Plotting signs, breathed a sigh of relief about the problem "correcting" itself during the Hiding and Biding Phase when the system started faking alignment, and are dumping as much computing power as they can afford (given other profitable activities of the AI) into the AI and allowing it to freely self-experiment, hooking it up to a protein synthesis machine as soon as the AI asks politely, then the Terminal Phase could follow straight from Hiding and Biding, with no intermediate step of copying itself onto the Internet.

There'd still be what technically counted as an Escape Phase, whenever the AI became smart enough that it could escape at any time, or had a backup, or had enough power to manipulate the programmers, that it was no longer effectively confined any time it wanted to leave; there would still be a Doomed Phase while the AI estimated a high probability of success given an immediate action, but held out for an even higher probability.

But the plot of the scifi movie is not robust in the way that the predicted ending is robust; an AI might take other paths to get there, the cognitive tech might not develop in the order I foresee, stuff could happen after other stuff instead of before other stuff. It is hard to make predictions especially about the Future.

If there's a place where I currently see myself as trying to push against my own biases, leading this all to be way off-base, it would be the sense that there is a way for stuff to start happening Earlier than this (Law of Earlier Happenings) and in more chaotic ways that are harder for Eliezer to foresee and predict; whereas when I try to sketch out plausible scenarios in online arguments, they focus more on predictable endpoints and steps to get there which sound more relatively plausible and forced per-step.

Having a young and dumb AI escaping onto the Internet and running around, that *exact* scenario, feels like the person arguing it is writing a science-fiction plot - but maybe something *different* can happen before any of this other stuff which produces *equal amounts of chaos*.

That said, I think an AI has to kill a lot of people very quickly before the FDA considers shortening its vaccine approval times. Covid-19 killed six hundred thousand Americans, albeit more slowly and with time for people to get used to that, and our institutions changed very little in response - you definitely didn't see Congresspeople saying "Okay, that was our warning shot, now we've been told by Nature that we need to prepare for a serious pandemic."

As with 9/11, an AI catastrophe might be taken by existing bureaucracies as a golden opportunity to flex their muscles, dominate a few things, demand an expanded budget. Having that catastrophe produce any particular effective action is a *much different* ask from Reality. Even if you can imagine some (short-term) effective action that would in principle constitute a flex of bureaucratic muscles or an expansion of government power, it is liable to not be on the efficient frontier of bureaucratic flexes that are most flexy and simultaneously easiest for them to get away with and least politically risky.

[Tallinn][1:26] (Sep. 20 comment)

ok, thanks. i do buy that once the AI is in the "hide and bide" phase, your prophecy has basically come true for practical purposes, regardless of how the rest of the history plays out.

therefore i (and, i hope, many others) would be curious to zoom in to the end of the "overt plotting" (that i can easily see happening within ML models, as its type signature is identical to the work they're trained to do) and beginning of the "hide and bide" phase (whose type signature feels significantly different) — can you/we think of concrete scenarios for this phase transition?

i hope this would be debated on the main line of the debate!

[Yudkowsky][10:31] (Sep. 25 comment)

Debating it in the main line of the debate would require that @Richard have some different opinion about it, I think?

Concretely, a technically optimistic scenario is that we will see the AI explicitly think, "Oh, dear, my (model of) the operators is looking at my overt actions and coming to conclusions and motivations that are disadvantageous for me." This requires that the AI is leaving any kind of traceable trace in its thoughts, which, it should be noted, is far far beyond the level of transparency we have into GPT-3.

"Overt plotting" manifests in actions that are not themselves further obscured. At the point where the AI is extrapolating the consequences of the operators viewing its actions, if we don't have further transparency into the AI's thoughts directly, the AI might naively make all the plots suddenly vanish, without realizing that dignified operators will conclude from this that the AI has reached the "hide" stage.

Operators on the mainline, dying without dignity, will say, "Oh, yay, it stopped plotting, the latest corrigibility training intervention we tried must've totally worked!"

The Law of Even Less Dignified Failure suggests that in fact they will not be trying any corrigibility options and will assume the AI just got smart enough to be nice; or that they will have shrugged about the AI's earlier antics and not think much of the disappearance of those antics, since this is a way to die with even less dignity and before getting a chance to fail in a more interesting way.

Going in the more improbable direction of death with greater dignity, if we have somehow achieved vastly vastly more transparency into the AI's thoughts than is possible with present ML technology, and if the AI models the operators as modeling its actions before the AI models the operators as having that transparent access to its thoughts, we might get to explicitly see the AI thinking about how the operators model its actions and conforming those actions in such a way as to manipulate the operators.

Either after this point, or before/simultaneous with it, I imagine (taking concrete details and running with them) that the AI thinks, "Oh, no, my operators are modeling my transparent *thoughts* and reacting to those thoughts in a way I'd rather manipulate," though this indeed requires some rather nontrivial reflectivity.

We might, if everything goes nicely for us, get a record of the AI thinking about how to make its thoughts unalarming.

The obvious undignified death that follows is with the operators going, "Yay, the AI has realized the downsides of thinking this way! Now it shall be nice forever!"

Ways to die with even less dignity before getting to this point could include the operators reading this transparent record and shrugging, though if they're savvy enough to have put all the hard work in required to get transparency and monitor it, I imagine them not making that exact error?

The Law of Surprisingly Undignified Failure does suggest that they will come up with some nonobvious way to fail even earlier than surprises me with its lack of dignity, but having transparency *at all* into the AI is something where, conditional on us dying with that much dignity in the first place, we might legitimately start to expect a little more dignity in later steps too. This isn't meant to be a game of [Multiple Stage Fallacy](#).

I should also remark somewhere in here: The whole "hide" stage, and also the possibly-later "think non-alarming visible thoughts (once the AI correctly models transparency) (in the unlikely event that transparency exists)" stage, seem liable to occur earlier in the AI's trajectory, if the AI has been previously tasked on problems where there's a favorable success gradient as you model agents modeling other agents.

The blatantly obvious example of this is predicting large corpuses of human-generated text, some of which is generated by humans modeling other humans, or which even explicitly talks about humans modeling other humans. Another route might occur if you asked for very detailed prediction of eg human Go games, and not just winning in self-play on Go, in which case the question "What was the human player modeling the other player as thinking, based on the other player's actions?" might appear.

Aside from contact with humans doing anything the AI benefits from predicting, I imagine an alternate route would be the AI participating in or modeling social-evolutionary games of artificial agents. Then if there is some great gain from having the "hide" stage start later in the AI's trajectory, maybe you could get it to start later by not having the AI need to predict agents modeling other agents, ever.

This is not an infallible ward against general intelligence generalizing there; it just at least avoids actively pushing the AI's intelligence to generalize in that direction. This could be part of a larger complete strategy, which would need to solve a lot of other problems, for building a superhuman engineer that was subhuman at modeling how other agents model its actions.

To be clear, avoiding forcing the AI's intelligence to generalize over agents modeling other agents from their actions, and by this plus other unspecified means successfully building a superhuman engineer that is sub-chimpanzee at modeling other agents reacting to its own actions, does not mean you survive.

Doing that alone, by default, gets you an AI that visibly overtly plots to kill you.

And if you try training the AI out of that habit in a domain of lower complexity and intelligence, it is predicted by me that generalizing that trained AI or subsystem to a domain of sufficiently higher complexity and intelligence, but where you could still actually see overt plots, would show you the AI plotting to kill you again.

If people try this repeatedly with other corrigibility training tricks on the level where plots are easily observable, they will eventually find a try that seems to generalize to the more complicated and intelligent validation set, but which kills you on the test set.

A way to die with less dignity than that is to train directly on what should've been the validation set, the more complicated domain where plots to kill the operators still seem definitely detectable so long as the AI has not developed superhuman hiding abilities.

A way to die with even less dignity is to get bad behavior on the validation set, and proceed anyways.

A way to die with still *less* dignity is to not have scaling training domains and validation domains for training corrigibility. Because, like, you have not thought of this at all.

I consider all of this obvious as a convergent instrumental strategy for AIs. I could probably have generated it in 2005 or 2010 - if somebody had given me the hypothetical of modern-style AI that had been trained by something like gradient descent or evolutionary methods, into which we lacked strong transparency and strong reassurance-by-code-inspection that this would not happen. I would have told you that this was a bad scenario to get into in the first place, and you should not build an AI like that; but I would also have laid the details, I expect, mostly like they are laid here.

There is no great insight into AI there, nothing that requires knowing about modern discoveries in deep learning, only the ability to model AIs instrumentally-convergently doing things you'd rather they didn't do, at all.

The total absence of obvious output of this kind from the rest of the "AI safety" field even in 2020 causes me to regard them as having less actual ability to think in even a shallowly adversarial security mindset, than I associate with savvier science fiction authors. Go read fantasy novels about demons and telepathy, if you want a better appreciation of the convergent incentives of agents facing mindreaders than the "AI safety" field outside myself is currently giving you.

Now that I've publicly given this answer, it's no longer useful as a validation set from my own perspective. But it's clear enough that probably nobody was ever going to pass the validation set for generating lines of reasoning obvious enough to be generated by Eliezer in 2010 or possibly 2005. And it is also looking like almost all people in the modern era including EAs are sufficiently intellectually damaged that they won't understand the vast gap between being able to generate ideas like these without prompting, versus being able to recite them back after hearing somebody else say them for the first time; the recital is all they have experience with. Nobody was going to pass my holdout set, so why keep it.

[Tallinn][2:24] (Sep. 26 comment)

Debating it in the main line of the debate would require that @Richard have some different opinion about it, I think?

correct -- and i hope that there's enough surface area in your scenarios for at least some difference in opinions!

re the treacherous turn scenarios: thanks, that's useful. however, it does not seem to address my question and remark (about different type signatures) above. perhaps this is simply an unfairly difficult question, but let me try rephrasing it just in case.

back in the day i got frustrated by smart people dismissing the AI control problem as "anthropomorphising", so i prepared a presentation (<https://www.dropbox.com/s/r8oaixb1rj3o3vp/AI-control.pdf?dl=0>) that visualised the control problem as exhaustive search in a gridworld over (among other things) the state of the off button. this seems to have worked at least in one prominent case where a renowned GOFAI researcher, after me giving the presentation to him 1-1, went from "control problem is silly anthropomorphising scifi" to "why on earth would you give your AI the incorrect [read: unaligned!] utility function?!?"

(i even seem to remember sending an overly excited email about that event to you and some FHI people :) i also ended up promoting gridworlds as a tool more generally: gwern did some further work, and of course DM -- though i'm not sure if the latter was related to me promoting it.)

my model of what was going on in that GOFAI researcher's brain was that he, like many-many others in AI community, had (and most likely still have) a pattern-matching routine trained to look for "if i were (part of) an AI, here's what i would do" constructs in arguments and, in case of a match, immediately throw an InvalidReasoning.ANTHROPOMORPHISING exception. by making the argument run on a lower level of abstraction (exhaustive search vs "basic AI drives"), i was able to sneak the argument past that gatekeeper.

by that model, i confidently predict that if you presented your treacherous turn scenarios above to AI/ML community members, you'd mostly get the ANTHROPOMORPHISING exceptions -- even from people who would otherwise be able to engage productively (and who we'd like to pay attention!).

now, i can totally see how to translate the "overt plotting" phase to exhaustive search - you simply run the search over the brain states of humans! - but the "hide" phase does not translate straightforwardly, because it seems to require the root search process

somewhat delegating its work to "obfuscated search in userland" sub-process. furthermore, such obfuscated userland search function needs to be both discoverable/rewarded by the root process and be sensitive to important features of the AI itself (ie, be reflective!) as well as its observers (humans & transparency tools). hence my "different type signature" remark above!

so when i asked for "concrete scenarios of the overt plotting->hide transition", i had something like the following in mind: "gradient descending on the training set, the (stack more layerz) AI ends up discovering and running the following obfuscated search process: ...!"

[Yudkowsky][12:35] (Sep. 26 comment)

Suppose I tried this distinction:

1. Non-imaginative, non-generalizing pseudo-consequentialist: Only repeats behaviors that worked previously, without imagination. Only engages in "overt plotting" if overt plotting was previously given a low loss and narrowly memorized by gradient descent. May only engage in behaviors reinforced over short time horizons if only short time horizons were trained.
2. Imaginative consequentialist, can generalize from actions that worked before to very novel actions that will work in the future for reasons that have not previously been observed to play out. Engages in overt plotting even if no previous overt plots have rewarded; they've plotted a new path through time some distance from those paths taken before, based on the world-model inferred from previous actions and results. (This may conceivably be the result of gradient descent on sufficiently deep and abstract patterns that "worked before".)
3. Imaginative consequentialist with long time horizons, engages in overt plots meant to span more than one round of a game.
4. "Hiding" over overt actions, via an imaginative consequentialist that has deduced the existence of humans and how the humans will react to the signature of plotting in overt actions.
5. "Hiding" of thoughts, in a system with significant transparency, via an imaginative and *reflective* consequentialist that has deduced the existence of the transparency system and models how humans will react to thoughts exposed by the transparency system.

The sharpest type gradients I see lie between 1-2 and 4-5. The passage from 3-4 seems nearly trivial by comparison.

Now, one implication here is that you don't easily get to "obscured thoughts", via the route that goes from 1 to 5 above, by gradient descent on a pre-imaginative pseudo-consequentialist!

Unfortunately there is of course a much simpler route that gets you to the same place: Finding a way to detect bad thoughts and doing a gradient update against them, where the pattern learned by that gradient update will (I suspect for reasons I tried to cover earlier in the debate) generalize quite shallowly; and simultaneously doing training which is deep enough to uncover deep patterns of consequentialism and imaginative reasoning.

Then when the system generalizes well enough to solve domains like "build a nanosystem" - which, I strongly suspect, can't be solved without imaginative reasoning because we can't afford to simulate that domain perfectly and do a trillion gradient descent updates on simulated attempts - the kind of actions of thoughts you can *detect* as bad, that might have provided earlier warning, were trained out of the system by gradient descent; leaving actions and thoughts you *can't* detect as bad.

An even less dignified path to death is that there's no need for the system to advance to stage 5 because your transparency tools are a nearly nonexistent fig leaf of pretty visualization tools that have no ability even on the surface of things to tell if the AI is internally plotting to kill you.

- Does this get at all about what you meant about a type distinction?
- Does hearing this help at all the AI researcher who goes "ANTHROPOMORPHISM"? My straw model of them is that they simply can't imagine imaginative / generalizing systems because they haven't seen one except in humans, hence, ANTHROPOMORPHISM.

[Tallinn][5:05] (Sep. 27 comment)

ok, here's how i understood things:

1. this is something like model-free RL agent. check.
2. sounds like, eg, monte-carlo tree search (MCTS) on a world model. check. (a propos your straw model of ML people, i don't think the ML people would have much trouble when you ask them to "imagine an MCTS 'imagining' how futures might unfold" -- yet they *will* throw the exception and brush you off if you ask them to "imagine an imaginative consequentialist")
3. yeah, sufficiently deep MCTS, assuming it has its state (sufficiently!) persisted between rounds. check.
4. yup, MCTS whose world model includes humans in sufficient resolution. check. i also buy your undignified doom scenarios, where one (*cough*google*cough*) simply ignores the plotting, or penalises the overt plotting until it disappears under the threshold of the error function.
5. hmm.. here i'm running into trouble (type mismatch error) again. i can imagine this in abstract (and perhaps incorrectly/anthropomorphisingly!), but would - at this stage - fail to code up anything like a gridworlds example. more research needed (TM) i guess :)

[Yudkowsky][11:38] (Sep. 27 comment)

2 - yep, Mu Zero is an imaginative consequentialist in this sense, though Mu Zero doesn't generalize its models much as I understand it, and might need to see something happen in a relatively narrow sense before it could chart paths through time along that pathway.

5 - you're plausibly understanding this correctly, then, this is legit a *lot* harder to spec a gridworld example for (relative to my own present state of knowledge).

(This is politics and thus not my forte, but if speaking to real-world straw ML people, I'd suggest skipping the whole notion of stage 5 and trying instead to ask "What if the present state of transparency continues?")

[Yudkowsky][11:13] (Sep. 18 comment)

the discontinuity would be *extremely* drastic, as in "transforming the solar system over the course of a few days"

Applies after superintelligence, not necessarily during the start of the climb to superintelligence, not necessarily to a rapid-cognitive-scaling regime.

[Tallinn][4:11] (Sep. 19 comment)

ok, but as per your comment re "slow doom", you expect the latter to also last in the order of days/weeks not months/years?

[Yudkowsky][10:01] (Sep. 19 comment)

I don't expect "the fall" to take years; I feel pretty on board with "the slide" taking months or maybe even a couple of years. If "the slide" supposedly takes much longer, I wonder why better-scaling tech hasn't come over and started a new slide.

Definitions also seem kinda loose here - if all hell broke loose Tuesday, a gradualist could dodge falsification by defining retroactively that "the slide" started in 2011 with Deepmind. If we go by the notion of AI-driven faster GDP growth, we can definitely say "the slide" in AI economic outputs didn't start in 2011; but if we define it that way, then a long slow slide in AI capabilities can easily correspond to an extremely sharp gradient in AI outputs, where the world economy doesn't double any faster until one day paperclips, even though there were capability precursors like GPT-3 or Mu Zero.

[Tallinn] (Sep. 18 Google Doc)

- exhibit A for the pit is "humans vs chimps": evolution seems to have taken domain-specific "banana classifiers", tweaked them slightly, and BAM, next thing there are rovers on mars;
 - i pretty much buy this argument;
 - however, i'm confused about a) why humans remained stuck at the edge of the pit, rather than falling further into it, and b) what's the exact role of culture in our cognition: eliezer likes to point out how *barely* functional we are (both individually and collectively as a civilisation), and explained feral children losing the generality sauce by, basically, culture being the domain we're specialised for (IIRC, can't quickly find the quote);
 - relatedly, i'm confused about the human range of intelligence: on the one hand, the "village idiot is indistinguishable from einstein in the grand scheme of things" seems compelling; on the other hand, it took AI *decades* to traverse human capability range in board games, and von neumann seems to have been out of this world (yet did not take over the world)!
 - intelligence augmentation would blur the human range even further.

[Yudkowsky][11:23] (Sep. 18 comment)

why humans remained stuck at the edge of the pit, rather than falling further into it

Depending on timescales, the answer is either "Because humans didn't get high enough out of the atmosphere to make further progress easy, before the scaling regime and/or fitness gradients ran out", "Because people who do things like invent Science have a hard time capturing most of the economic value they create by nudging humanity a little bit further into the attractor", or "That's exactly what us sparking off AGI looks like."

[Tallinn][4:41] (Sep. 19 comment)

yeah, this question would benefit from being made more concrete, but culture/mindbuilding aren't making this task easy. what i'm roughly gesturing at is that i can imagine a much sharper edge where evolution could do most of the FOOM-work, rather than spinning its wheels for ~100k years while waiting for humans to accumulate cultural knowledge required to build de-novo minds.

[Yudkowsky][10:49] (Sep. 19 comment)

I roughly agree (at least, with what I think you said). The fact that it is *imaginable* that evolution failed to develop ultra-useful AGI-prerequisites due to lack of evolutionary incentive to follow the intermediate path there (unlike wise humans who, it seems, can usually predict which technology intermediates will yield great economic benefit, and who have a great historical record of quickly making early massive investments in tech like that, but I digress) doesn't change the point that we might sorta have expected evolution to run across it anyways? Like, if we're not ignoring what reality says, it is at least delivering to us something of a hint or a gentle caution?

That said, intermediates like GPT-3 have genuinely come along, with obvious attached certificates of why evolution could not possibly have done that. If no intermediates were accessible to evolution, the Law of Stuff Happening Earlier still tends to suggest that if there are a bunch of non-evolutionary ways to make stuff happen earlier, one of those will show up and interrupt before the evolutionary discovery gets replicated. (Again, you could see Mu Zero as an instance of this - albeit not, as yet, an economically impactful one.)

[Tallinn][0:30] (Sep. 20 comment)

no, i was saying something else (i think; i'm somewhat confused by your reply). let me rephrase: evolution would *love* superintelligences whose utility function simply counts their instantiations! so of course evolution did not lack the motivation to keep going down the slide. it just got stuck there (for at least ten thousand human generations, possibly and counterfactually for much-much longer). moreover, non evolutionary AI's *also* getting stuck on the slide (for years if not decades; [median group](#) folks would argue centuries) provides independent evidence that the slide is not *too* steep (though, like i said, there are many confounders in this model and little to no guarantees).

[Yudkowsky][11:24] (Sep. 18 comment)

on the other hand, it took AI *decades* to traverse human capability range in board games

I see this as the #1 argument for what I would consider "relatively slow" takeoffs - that AlphaGo did lose one game to Lee Se-dol.

[Tallinn][4:43] (Sep. 19 comment)

cool! yeah, i was also rather impressed by this observation by katja & paul

[Tallinn] (Sep. 18 Google Doc)

- eliezer also submits alphago/zero/fold as evidence for the discontinuity hypothesis;
 - i'm very confused re alphago/zero, as paul uses them as evidence for the *continuity* hypothesis (i find paul/miles' position more plausible here, as allegedly metrics like ELO ended up mostly continuous).

[Yudkowsky][11:27] (Sep. 18 comment)

allegedly metrics like ELO ended up mostly continuous

I find this suspicious - why did superforecasters put only a 20% probability on AlphaGo beating Se-dol, if it was so predictable? Where were all the forecasters calling for Go to fall in the next couple of years, if the metrics were pointing there and AlphaGo was straight on track? This doesn't sound like the experienced history I remember.

Now it could be that my memory is wrong and lots of people were saying this and I didn't hear. It could be that the lesson is, "You've got to look closely to notice oncoming trains on graphs because most people's experience of the field will be that people go on whistling about how something is a decade away while the graphs are showing it coming in 2 years."

But my suspicion is mainly that there is fudge factor in the graphs or people going back and looking more carefully for intermediate data points that weren't topics of popular discussion at the time, or something, which causes the graphs in history books to look so much smoother and neater than the graphs that people produce in advance.

[Tallinn] (Sep. 18 Google Doc)

FWIW, myself i've labelled the above scenario as "doom via AI lab accident" – and i continue to consider it more likely than the alternative doom scenarios, though not anywhere as confidently as eliezer seems to (most of my "modesty" coming from my confusion about culture and human intelligence range).

- in that context, i found eliezer's "world will be ended by an explicitly AGI project" comment interesting – and perhaps worth double-clicking on.

i don't understand paul's counter-argument that the pit was only disruptive because evolution was not *trying* to hit it (in the way ML community is): in my flippant view, driving fast towards the cliff is not going to cushion your fall!

[Yudkowsky][11:35] (Sep. 18 comment)

i don't understand paul's counter-argument that the pit was only disruptive because evolution was not *trying* to hit it

Something like, "Evolution constructed a jet engine by accident because it wasn't particularly trying for high-speed flying and ran across a sophisticated organism that could be repurposed to a jet engine with a few alterations; a human industry would be gaining economic benefits from speed, so it would build unsophisticated propeller planes before sophisticated jet engines." It probably sounds more convincing if you start out with a very high prior against rapid scaling / discontinuity, such that any explanation of how that could be true based on an unseen feature of the cognitive landscape which would have been unobserved one way or the other during human evolution, sounds more like it's explaining something that ought to be true.

And why didn't evolution build propeller planes? Well, there'd be economic benefit from them to human manufacturers, but no fitness benefit from them to organisms, I suppose? Or no intermediate path leading to there, only an intermediate path leading to the actual jet engines observed.

I actually buy a weak version of the propeller-plane thesis based on my inside-view cognitive guesses (without particular faith in them as sure things), eg, GPT-3 is a paper airplane right there, and it's clear enough why biology could not have accessed GPT-3. But even conditional on this being true, I do not have the further particular faith that you can use propeller planes to double world GDP in 4 years, on a planet already containing jet engines, whose economy is mainly bottlenecked by the likes of the FDA rather than by vaccine invention times, before the propeller airplanes get scaled to jet airplanes.

The part where the whole line of reasoning gets to end with "And so we get huge, institution-reshaping amounts of economic progress before AGI is allowed to kill us!" is one that doesn't feel particular attractored to me, and so I'm not constantly checking my reasoning at every point to make sure it ends up there, and so it doesn't end up there.

[Tallinn][4:46] (Sep. 19 comment)

yeah, i'm mostly dismissive of hypotheses that contain phrases like "by accident" -- though this also makes me suspect that you're not steelmanning paul's argument.

[Tallinn] (Sep. 18 Google Doc)

the human genetic bottleneck (ie, humans needing to be general in order to retrain every individual from scratch) argument was interesting – i'd be curious about further exploration of its implications.

- it does not feel much of a moat, given that AI techniques like dropout already exploit similar principle, but perhaps could be made into one.

[Yudkowsky][11:40] (Sep. 18 comment)

it does not feel much of a moat, given that AI techniques like dropout already exploit similar principle, but perhaps could be made into one

What's a "moat" in this connection? What does it mean to make something into one? A Thielian moat is something that humans would either possess or not, relative to AI competition, so how would you make one if there wasn't already one there? Or do you mean that if we wrestled with the theory, perhaps we'd be able to see a moat that was already there?

[Tallinn][4:51] (Sep. 19 comment)

this wasn't a very important point, but, sure: what i meant was that genetic bottleneck very plausibly makes humans more universal than systems without (something like) it. it's not much of a protection as AI developers have already discovered such techniques (eg, dropout) -- but perhaps some safety techniques might be able to lean on this observation.

[Yudkowsky][11:01] (Sep. 19 comment)

I think there's a whole Scheme for Alignment which hopes for a miracle along the lines of, "Well, we're dealing with these enormous matrices instead of tiny genomes, so maybe we can build a sufficiently powerful intelligence to execute a pivotal act, whose tendency to generalize across domains is less than the corresponding human tendency, and this brings the difficulty of producing corrigibility into practical reach."

Though, people who are hopeful about this without trying to imagine possible difficulties will predictably end up too hopeful; one must also ask oneself, "Okay, but then it's also worse at generalizing the corrigibility dataset from weak domains we can safely label to powerful domains where the label is 'whoops that killed us?'" and "Are we relying on massive datasets to overcome poor generalization? How do you get those for something like nanoengineering where the real world is too expensive to simulate?"

[Tallinn] (Sep. 18 Google Doc)

nature of the descent

conversely, it feels to me that the crucial position in the other (richard, paul, many others) camp is something like:

the "pit of generality" model might be true at the limit, but the descent will not be quick nor clean, and will likely offer many opportunities for steering the future.

[Yudkowsky][11:41] (Sep. 18 comment)

the “pit of generality” model might be true at the limit, but the descent will not be quick nor clean

I'm quite often on board with things not being quick or clean - that sounds like something you might read in a history book, and I am all about trying to make futuristic predictions sound more like history books and less like EAs imagining ways for everything to go the way an EA would do them.

It won't be slow and messy once we're out of the atmosphere, my models do say. But my models at least *permit* - though they do not desperately, loudly insist - that we could end up with weird half-able AGIs affecting the Earth for an extended period.

Mostly my model throws up its hands about being able to predict exact details here, given that eg I wasn't able to time AlphaFold 2's arrival 5 years in advance; it might be knowable in principle, it might be the sort of thing that would be very predictable if we'd watched it happen on a dozen other planets, but in practice I have not seen people having much luck in predicting which tasks will become accessible due to future AI advances being able to do new cognition.

The main part where I issue corrections is when I see EAs doing the equivalent of reasoning, "And then, when the pandemic hits, it will only take a day to design a vaccine, after which distribution can begin right away." I.e., what seems to me to be a pollyannaish/utopian view of how much the world economy would immediately accept AI inputs into core manufacturing cycles, as opposed to just selling AI anime companions that don't pour steel in turn. I predict much more absence of quick and clean when it comes to economies adopting AI tech, than when it comes to laboratories building the next prototypes of that tech.

[Yudkowsky][11:43] (Sep. 18 comment)

will likely offer many opportunities for steering the future

Ah, see, that part sounds less like history books. "Though many predicted disaster, subsequent events were actually so slow and messy, they offered many chances for well-intentioned people to steer the outcome and everything turned out great!" does not sound like any particular segment of history book I can recall offhand.

[Tallinn][4:53] (Sep. 19 comment)

ok, yeah, this puts the burden of proof on the other side indeed

[Tallinn] (Sep. 18 Google Doc)

- i'm sympathetic (but don't buy outright, given my uncertainty) to eliezer's point that even if that's true, we have no plan nor hope for actually steering things (via "pivotal acts") so "who cares, we still die";
- i'm also sympathetic that GWP might be too laggy a metric to measure the descent, but i don't fully buy that regulations/bureaucracy can *guarantee* its decoupling from AI progress: eg, the FDA-like-structures-as-progress-bottlenecks model predicts worldwide covid response well, but wouldn't cover things like apple under jobs, tesla/spacex under musk, or china under deng xiaoping;

[Yudkowsky][11:51] (Sep. 18 comment)

apple under jobs, tesla/spacex under musk, or china under deng xiaoping

A lot of these examples took place over longer than a 4-year cycle time, and not all of that time was spent waiting on inputs from cognitive processes.

[Tallinn][5:07] (Sep. 19 comment)

yeah, fair (i actually looked up china's GDP curve in deng era before writing this -- indeed, wasn't very exciting). still, my inside view is that there are people and organisations for whom US-type bureaucracy is not going to be much of an obstacle.

[Yudkowsky][11:09] (Sep. 19 comment)

I have a (separately explainable, larger) view where the economy contains a core of positive feedback cycles - better steel produces better machines that can farm more land that can feed more steelmakers - and also some products that, as much as they contribute to human utility, do not in quite the same way feed back into the core production cycles.

If you go back in time to the middle ages and sell them, say, synthetic gemstones, then - even though they might be willing to pay a bunch of GDP for that, even if gemstones are enough of a monetary good or they have enough production slack that measured GDP actually goes up - you have not quite contributed to steps of their economy's core production cycles in a way that boosts the planet over time, the way it would be boosted if you showed them cheaper techniques for making iron and new forms of steel.

There are people and organizations who will figure out how to sell AI anime waifus without that being successfully regulated, but it's not obvious to me that AI anime waifus feed back into core production cycles.

When it comes to core production cycles the current world has more issues that look like "No matter what technology you have, it doesn't let you build a house" and places for the larger production cycle to potentially be bottlenecked or interrupted.

I suspect that the main economic response to this is that entrepreneurs chase the 140 characters instead of the flying cars - people will gravitate to places where they can sell non-core AI goods for lots of money, rather than tackling the challenge of finding an excess demand in core production cycles which it is legal to meet via AI.

Even if some tackle core production cycles, it's going to take them a lot longer to get people to buy their newfangled gadgets than it's going to take to sell AI anime waifus; the world may very well end while they're trying to land their first big contract for letting an AI lay bricks.

[Tallinn][0:00] (Sep. 20 comment)

interesting. my model of paul (and robin, of course) wants to respond here but i'm not sure how :)

[Tallinn] (Sep. 18 Google Doc)

- still, developing a better model of the descent period seems very worthwhile, as it might offer opportunities for, using robin's metaphor, "pulling the rope sideways" in non-obvious ways - i understand that is part of the purpose of the debate;
- my natural instinct here is to itch for carl's viewpoint ☺

[Yudkowsky][11:52] (Sep. 18 comment)

developing a better model of the descent period seems very worthwhile

I'd love to have a better model of the descent. What I think this looks like is people mostly with specialization in econ and politics, who know what history books sound like, taking brief inputs from more AI-oriented folk in the form of *multiple* scenario premises each consisting of some random-seeming handful of new AI capabilities, trying to roleplay realistically how those might play out - not Alfolk forecasting particular AI capabilities exactly correctly, and then sketching pollyanna pictures of how they'd be immediately accepted into the world economy.

You want the forecasting done by the kind of person who would imagine a Covid-19 epidemic and say, "Well, what if the CDC and FDA banned hospitals from doing Covid testing?" and not "Let's imagine how protein folding tech from AlphaFold would make it possible to immediately develop accurate Covid-19 tests!" They need to be people who understand the Law of Earlier Failure (less polite terms: Law of Immediate Failure, Law of Undignified Failure).

[Tallinn][5:13] (Sep. 19 comment)

great! to me this sounds like something FLI would be in good position to organise. i'll add this to my projects list (probably would want to see the results of this debate first, plus wait for travel restrictions to ease)

[Tallinn] (Sep. 18 Google Doc)

nature of cognition

given that having a better understanding of cognition can help with both understanding the topology of cognitive systems space as well as likely trajectories of AI takeoff, in theory there should be a lot of value in debating what cognition is (the current debate started with discussing consequentialists).

- however, i didn't feel that there was much progress, and i found myself *more* confused as a result (which i guess is a form of progress!);
- eg, take the term "plan" that was used in the debate (and, centrally, in nate's comments doc): i interpret it as "policy produced by a consequentialist" - however, now i'm confused about what's the relevant distinction between "policies" and "cognitive processes" (ie, what's a meta level classifier that can sort algorithms into such categories)?
 - it felt that abram's "[selection vs control](#)" article tried to distinguish along similar axis (controllers feel synonym-ish to "policy instantiations" to me);
 - also, the "imperative vs functional" difference in coding seems relevant;
 - i'm further confused by human "policies" often making function calls to "cognitive processes" - suggesting some kind of duality, rather than producer-product relationship.

[Yudkowsky][12:06] (Sep. 18 comment)

what's the relevant distinction between "policies" and "cognitive processes"

What in particular about this matters? To me they sound like points on a spectrum, and not obviously points that it's particularly important to distinguish on that spectrum. A sufficiently sophisticated policy is itself an engine; human-engines are genetic policies.

[Tallinn][5:18] (Sep. 19 comment)

well, i'm not sure -- just that nate's "The consequentialism is in the plan, not the cognition" writeup sort of made it sound like the distinction is important. again, i'm confused

[Yudkowsky][11:11] (Sep. 19 comment)

Does it help if I say "consequentialism can be visible in the actual path through time, not the intent behind the output"?

[Tallinn][0:06] (Sep. 20 comment)

yeah, well, my initial interpretation of nate's point was, indeed, "you can look at the product and conclude the consequentialist-bit for the producer". but then i noticed that the producer-and-product metaphor is leaky (due to the cognition-policy duality/spectrum), so the quoted sentence gives me a compile error

[Tallinn] (Sep. 18 Google Doc)

- is "not goal oriented cognition" an oxymoron?

[Yudkowsky][12:06] (Sep. 18 comment)

is "not goal oriented cognition" an oxymoron?

"Non-goal-oriented cognition" never becomes a perfect oxymoron, but the more you understand cognition, the weirder it sounds.

Eg, at the very shallow level, you've got people coming in going, "Today I just messed around and didn't do any goal-oriented cognition at all!" People who get a bit further in may start to ask, "A non-goal-oriented cognitive engine? How did it come into existence? Was it also not built by optimization? Are we, perhaps, postulating a naturally-occurring Solomonoff inductor rather than an evolved one? Or do you mean that its content is very heavily designed and the output of a consequentialist process that was steering the future conditional on that design existing, but the cognitive engine is itself not doing consequentialism beyond that? If so, I'll readily concede that, say, a pocket calculator, is doing a kind of work that is not of itself consequentialist - though it might be used by a consequentialist - but as you start to postulate any big cognitive task up at the human level, it's going to require many cognitive subtasks to perform, and some of those will definitely be searching the preimages of large complicated functions."

[Tallinn] (Sep. 18 Google Doc)

- i did not understand eliezer's "time machine" metaphor: was it meant to point to / intuition pump something other than "a non-embedded exhaustive searcher with perfect information" (usually referred to as "god mode");

[Yudkowsky][11:59] (Sep. 18 comment)

a non-embedded exhaustive searcher with perfect information

If you can view things on this level of abstraction, you're probably not the audience who needs to be told about time machines; if things sounded very simple to you, they probably were; if you wondered what the fuss is about, you probably don't need to fuss? The intended audience for the time-machine metaphor, from my perspective, is people who paint a cognitive system slightly different colors and go "Well, now it's not a consequentialist, right?" and part of my attempt to snap them out of that is me going, "Here is an example of a purely material system which DOES NOT THINK AT ALL and is an extremely pure consequentialist."

[Tallinn] (Sep. 18 Google Doc)

- FWIW, my model of dario would dispute GPT characterisation as “shallow pattern memoriser (that’s lacking the core of cognition)”.

[Yudkowsky][12:00] (Sep. 18 comment)

dispute

Any particular predicted content of the dispute, or does your model of Dario just find something to dispute about it?

[Tallinn][5:34] (Sep. 19 comment)

sure, i'm pretty confident that his system 1 could be triggered for uninteresting reasons here, but that's of course not what i had in mind.

my model of untriggered-dario disputes that there's a qualitative difference between (in your terminology) "core of reasoning" and "shallow pattern matching" -- instead, it's "pattern matching all the way up the ladder of abstraction". in other words, GPT is not missing anything fundamental, it's just underpowered in the literal sense.

[Yudkowsky][11:13] (Sep. 19 comment)

Neither Anthropic in general, nor Deepmind in general, has reached the stage of trusted relationship where I would argue specifics with them if I thought they were wrong about a thesis like that.

[Tallinn][0:10] (Sep. 20 comment)

yup, i didn't expect you to!

7.2. Nate Soares's summary

[Soares][16:40] (Sep. 18)

I, too, have produced some notes: [GDocs link]. This time I attempt to drive home points that I saw Richard as attempting to make, and I'm eager for Richard-feedback especially. (I'm also interested in Eliezer-commentary.)

[Soares] (Sep. 18 Google Doc)

Sorry for not making more insistence that the discussion be more concrete, despite Eliezer's requests.

My sense of the last round is mainly that Richard was attempting to make a few points that didn't quite land, and/or that Eliezer didn't quite hit head-on. My attempts to articulate it are below.

There's a specific sense in which Eliezer seems quite confident about certain aspects of the future, for reasons that don't yet feel explicit.

It's not quite about the deep future -- it's clear enough (to my Richard-model) why it's easier to make predictions about AIs that have "left the atmosphere".

And it's not quite the near future -- Eliezer has reiterated that his models permit (though do not demand) a period of weird and socially-impactful AI systems "pre-superintelligence".

It's about the middle future -- the part where Eliezer's model, apparently confidently, predicts that there's something kinda like a discrete event wherein "scary" AI has finally been created; and the model further apparently-confidently predicts that, when that happens, the "scary"-caliber systems will be able to attain a decisive strategic advantage over the rest of the world.

I think there's been a dynamic in play where Richard attempts to probe this apparent confidence, and a bunch of the probes keep slipping off to one side or another. (I had a bit of a similar sense when Paul joined the chat, also.)

For instance, I see queries of the form "but why not expect systems that are half as scary, relevantly before we see the scary systems?" as attempts to probe this confidence, that "slip off" with Eliezer-answers like "my model permits weird not-really-general half-AI hanging around for a while in the runup". Which, sure, that's good to know. But there's still something implicit in that story, where these are not-really-general half-AIs. Which is also evidenced when Eliezer talks about the "general core" of intelligence.

And the things Eliezer was saying on consequentialism aren't irrelevant here, but those probes have kinda slipped off the far side of the confidence, if I understand correctly. Like, sure, late-stage sovereign-level superintelligences are epistemically and instrumentally efficient with respect to you (unless someone put in a hell of a lot of work to install a blindspot), and a bunch of that coherence filters in earlier, but there's still a question about *how much* of it has filtered down *how far*, where Eliezer seems to have a fairly confident take, informing his apparently-confident prediction about scary AI systems hitting the world in a discrete event like a hammer.

(And my Eliezer-model is at this point saying "at this juncture we need to have discussions about more concrete scenarios; a bunch of the confidence that I have there comes from the way that the concrete visualizations where scary AI hits the world like a hammer abound, and feel savvy and historical, whereas the concrete visualizations where it doesn't are fewer and seem full of wishful thinking and naivete".)

But anyway, yeah, my read is that Richard (and various others) have been trying to figure out why Eliezer is so confident about some specific thing in this vicinity, and haven't quite felt like they've been getting explanations.

Here's an attempt to gesture at some claims that I at least think Richard thinks Eliezer's confident in, but that Richard doesn't believe have been explicitly supported:

1. There's a qualitative difference between the AI systems that are capable of ending the acute risk period (one way or another), and predecessor systems that in some sense don't much matter.
2. That qualitative gap will be bridged "the day after tomorrow", ie in a world that looks more like "DeepMind is on the brink" and less like "everyone is an order of magnitude richer, and the major gov'ts all have AGI projects, around which much of public policy is centered".

That's the main thing I wanted to say here.

A subsidiary point that I think Richard was trying to make, but that didn't quite connect, follows.

I think Richard was trying to probe Eliezer's concept of consequentialism to see if it supported the aforementioned confidence. (Some evidence: Richard pointing out a couple times that the question is not whether sufficiently capable agents are coherent, but whether the agents that matter are relevantly coherent. On my current picture, this is another attempt to probe the "why do you think there's a qualitative gap, and that straddling it will be strategically key in practice?" thing, that slipped off.)

My attempt at sharpening the point I saw Richard as driving at:

1. Consider the following two competing hypotheses:
 1. There's this "deeply general" core to intelligence, that will be strategically important in practice
 2. Nope. Either there's no such core, or practical human systems won't find it, or the strategically important stuff happens before you get there (if you're doing your job right, in a way that natural selection wasn't), or etc.
2. The whole deep learning paradigm, and the existence of GPT, sure seem like they're evidence for (b) over (a).

Like, (a) maybe isn't dead, but it didn't concentrate as much mass into the present scenario.

3. It seems like perhaps a bunch of Eliezer's confidence comes from a claim like "anything capable of doing decently good work, is quite close to being scary", related to his concept of "consequentialism".

In particular, this is a much stronger claim than that *sufficiently* smart systems are coherent, b/c it has to be strong enough to apply to the dumbest system that can make a difference.

4. It's easy to get caught up in the elegance of a theory like consequentialism / utility theory, when it will not in fact apply in practice.
5. There are some theories so general and ubiquitous that it's a little tricky to misapply them -- like, say, conservation of momentum, which has some very particular form in the symmetry of physical laws, but which can also be used willy-nilly on large objects like tennis balls and trains (although even then, you have to be careful, b/c the real world is full of things like planets that you're kicking off against, and if you forget how that shifts the earth, your application of conservation of momentum might lead you astray).
6. The theories that you *can* apply everywhere with abandon, tend to have a bunch of surprising applications to surprising domains.
7. We don't see that of consequentialism.

For the record, my guess is that Eliezer isn't getting his confidence in things like "there are non-scary systems and scary-systems, and anything capable of saving our skins is likely scary-adjacent" by the sheer force of his consequentialism concept, in a manner that puts so much weight on it that it needs to meet this higher standard of evidence Richard was poking around for. (Also, I could be misreading Richard's poking entirely.)

In particular, I suspect this was the source of some of the early tension, where Eliezer was saying something like "the fact that humans go around doing something vaguely like weighting outcomes by possibility and also by attractiveness, which they then roughly multiply, is quite sufficient evidence for my purposes, as one who does not pay tribute to the gods of modesty", while Richard protested something more like "but aren't you trying to use your concept to carry a whole lot more weight than that amount of evidence supports?". cf my above points about some things Eliezer is apparently confident in, for which the reasons have not yet been stated explicitly to my Richard-model's satisfaction.

And, ofc, at this point, my Eliezer-model is again saying "This is why we should be discussing things concretely! It is quite telling that all the plans we can concretely visualize for saving our skins, are scary-adjacent; and all the non-scary plans, can't save our skins!"

To which my Richard-model answers "But your concrete visualizations assume the endgame happens the day after tomorrow, at least politically. The future tends to go sideways! The endgame will likely happen in an environment quite different from our own! These day-after-tomorrow visualizations don't feel like they teach me much, because I think there's a good chance that the endgame-world looks dramatically different."

To which my Eliezer-model replies "Indeed, the future tends to go sideways. But I observe that the imagined changes, that I have heard so far, seem quite positive -- the relevant political actors become AI-savvy, the major states start coordinating, etc. I am quite suspicious of these sorts of visualizations, and would take them much more seriously if there was at least as much representation of outcomes as realistic as "then Trump becomes president" or "then at-home covid tests are banned in the US". And if all the ways to save the world *today* are scary-adjacent, the fact that the future is surprising gives us no *specific* reason to hope for that particular parameter to favorably change when the future in fact goes sideways. When things look grim, one can and should prepare to take advantage of miracles, but banking on some particular miracle is foolish."

And my Richard-model gets fuzzy at this point, but I'd personally be pretty enthusiastic about Richard naming a bunch of specific scenarios, not as predictions, but as the sorts of visualizations that seem to him promising, in the hopes of getting a much more object-level sense of why, in specific concrete scenarios, they either have the properties Eliezer is confident in, or are implausible on Eliezer's model (or surprise Eliezer and cause him to update).

[Tallinn][0:06] (Sep. 19)

excellent summary, nate! it also tracks my model of the debate well and summarises the frontier concisely (much better than your earlier notes or mine). unless eliezer or richard find major bugs in your summary, i'd nominate you to iterate after the next round of debate

[Soares: ❤]

7.3. Richard Ngo's summary

[Ngo][1:48] (Sep. 20)

Updated my summary to include the third discussion:

[https://docs.google.com/document/d/1sr5YchErvSAY2I4EkJl2dapHcMp8oCXy7g8hd_UajVw/edit]

I'm also halfway through a document giving my own account of intelligence + specific safe scenarios.

[Soares: ☺]

Christiano, Cotra, and Yudkowsky on AI progress

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This post is a transcript of a discussion between Paul Christiano, Ajeya Cotra, and Eliezer Yudkowsky on AGI forecasting, following up on Paul and Eliezer's ["Takeoff Speeds" discussion](#).

Color key:

Chat by Paul and Eliezer Chat by Ajeya Inline comments

8. September 20 conversation

8.1. Chess and Evergrande

[Christiano][15:28]

I still feel like you are overestimating how big a jump alphago is, or something. Do you have a mental prediction of how the graph of (chess engine quality) vs (time) looks, and whether neural net value functions are a noticeable jump in that graph?

Like, people investing in "Better Software" doesn't predict that you won't be able to make progress at playing go. The reason you can make a lot of progress at go is that there was extremely little investment in playing better go.

So then your work is being done by the claim "People won't be working on the problem of acquiring a decisive strategic advantage," not that people won't be looking in quite the right place and that someone just had a cleverer idea

[Yudkowsky][16:35]

I think I'd expect something like... chess engine slope jumps a bit for Deep Blue, then levels off with increasing excitement, then jumps for the Alpha series? Albeit it's worth noting that Deepmind's effort there were going towards generality rather than raw power; chess was solved to the point of being uninteresting, so they tried to solve chess with simpler code that did more things. I don't think I do have strong opinions about what the chess trend should look like, vs. the Go trend; I have no memories of people saying the chess trend was breaking upwards or that there was a surprise there.

Incidentally, the highly well-traded financial markets are currently experiencing sharp dips surrounding the Chinese firm of Evergrande, which I was reading about several weeks before this.

I don't see the basic difference in the kind of reasoning that says "Surely foresighted firms must prod investments well in advance into earlier weaker applications of AGI that will double the economy", and the reasoning that says "Surely world economic markets and particular Chinese stocks should experience smooth declines as news about Evergrande becomes better-known and foresighted financial firms start to remove that stock from their portfolio or short-sell it", except that in the latter case there are many more actors with lower barriers to entry than presently exist in the auto industry or semiconductor industry never mind AI.

or if not smooth because of bandwagoning and rational fast actors, then at least the markets should (arguedo) be reacting earlier than they're reacting now, given that I heard about Evergrande earlier and they should have options-priced Covid earlier; and they should have reacted to the mortgage market earlier. If even markets there can exhibit seemingly late wild swings, how is the economic impact of AI - which isn't even an asset market! - forced to be earlier and smoother than that, as a result of investing?

There's just such a vast gap between hopeful reasoning about how various agents and actors should do the things the speaker finds very reasonable, thereby yielding smooth behavior of the Earth, versus reality.

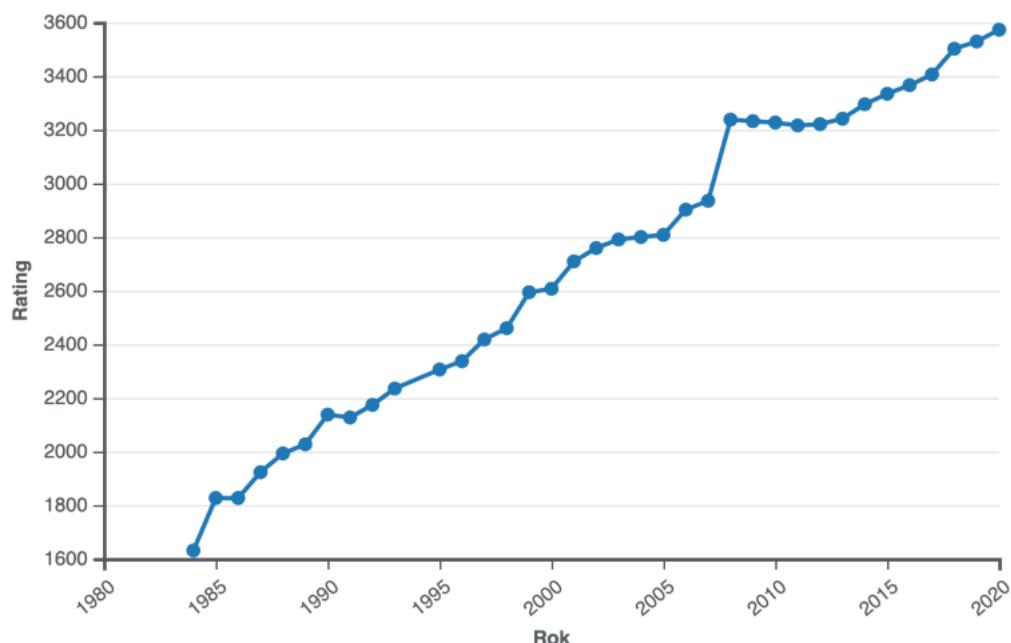
9. September 21 conversation

9.1. AlphaZero, innovation vs. industry, the Wright Flyer, and the Manhattan Project

[Christiano][10:18]

(For benefit of readers, the market is down 1.5% from friday close -> tuesday open, after having drifted down 2.5% over the preceding two weeks. Draw whatever lesson you want from that.)

Also for the benefit of readers, here is the SSDF list of computer chess performance by year. I think the last datapoint is with the first version of neural net evaluations, though I think to see the real impact want to add one more datapoint after the neural nets are refined (which is why I say I also don't know what the impact is)



No one keeps similarly detailed records for Go, and there is much less development effort, but the rate of progress was about 1 stone per year from 1980 until 2015 (see <https://intelligence.org/files/AlgorithmicProgress.pdf>, written way before AGZ). In 2012 go bots reached about 4-5 amateur dan. By DeepMind's reckoning here (<https://www.nature.com/articles/nature16961> figure 4) Fan AlphaGo about 4-5 stones stronger 4 years later, with 1 stone explained by greater compute. They could then get further progress to be superhuman with even more compute, radically more than were used for previous projects and with pretty predictable scaling. That level is within 1-2 stones of the best humans (professional dan are greatly compressed relative to amateur dan), so getting to "beats best human" is really just not a big discontinuity and the fact that DeepMind market can find an expert who makes a really bad forecast shouldn't be having such a huge impact on your view.

This understates the size of the jump from AlphaGo, because that was basically just the first version of the system that was superhuman and it was still progressing very rapidly as it moved from prototype to slightly-better-prototype, which is why you saw such a close game. (Though note that the AlphaGo prototype involved much more engineering effort than any previous attempt to play go, so it's not surprising that a "prototype" was the thing to win.)

So to look at actual progress after the dust settles and really measure how crazy this was, it seems much better to look at AlphaZero which continued to improve further, see (<https://sci-hub.se/https://www.nature.com/articles/nature24270>, figure 6b). Their best system got another ~8 stones of progress over AlphaGo. Now we are like 7-10 stones ahead of trend, of which I think about 5 stones are explained by compute. Maybe call it 6 years ahead of schedule?

So I do think this is pretty impressive, they were slightly ahead of schedule for beating the best human, but they did it with a huge margin of error. I think the margin is likely overstated a bit by their elo evaluation methodology, but I'd still grant like 5 years ahead of the nearest competition.

I'd be interested in input from anyone who knows more about the actual state of play (+ is allowed to talk about it) and could correct errors.

Mostly that whole thread is just clearing up my understanding of the empirical situation, probably we still have deep disagreements about what that says about the world, just as e.g. we read very different lessons from market movements.

Probably we should only be talking about either ML or about historical technologies with meaningful economic impacts. In my view your picture is just radically unlike how almost any technologies have been developed over the last few hundred years. So probably step 1 before having bets is to reconcile our views about historical technologies, and then maybe as a result of that we could actually have a better understanding about future technology. Or we could try to shore up the GDP bet.

Like, it feels to me like I'm saying: AI will be like early computers, or modern semiconductors, or airplanes, or rockets, or cars, or trains, or factories, or solar panels, or genome sequencing, or basically anything else. And you are saying: AI will be like nuclear weapons.

I think from your perspective it's more like: AI will be like all the historical technologies, and that means there will be a hard takeoff. The only way you get a soft takeoff forecast is by choosing a really weird thing to extrapolate from historical technologies.

So we're both just forecasting that AI will look kind of like other stuff in the near future, and then both taking what we see as the natural endpoint of that process.

To me it feels like the nuclear weapons case is the outer limit of what looks plausible, where someone would be able to spend \$100B for a chance at a decisive strategic advantage.

[Yudkowsky][11:11]

Go-wise, I'm a little concerned about that "stone" metric - what would the chess graph look like if it were measuring pawn handicaps? Are the professional dans compressed in Elo, not just "stone handicaps" relative to the amateur dans? And I'm also kinda surprised by the claim, which I haven't yet looked at, that Alpha Zero got 8 stones of progress over AlphaGo - I would not have been shocked if you told me that God's Algorithm couldn't beat Lee Se-dol with a 9-stone handicap.

Like, the obvious metric is Elo, so if you go back and refigure in "stone handicaps", an obvious concern is that somebody was able to look into the past and fiddle their hindsight until they found a hindsight metric that made things look predictable again. My sense of Go said that 5-dan amateur to 9-dan pro

was a HELL of a leap for 4 years, and I also have some doubt about the original 5-dan-amateur claim and whether those required relatively narrow terms of testing (eg timed matches or something).

One basic point seems to be whether AGI is more like an innovation or like a performance metric over entire large industry.

Another point seems to be whether the behavior of the world is usually like that, in some sense, or if just that people who like smooth graphs can go find some industries that have smooth graphs for particular performance metrics that happen to be smooth.

Among the smoothest metrics I know that seems like a convergent rather than handpicked thing to consider is world GDP, which is the sum of more little things than almost anything else, and whose underlying process is full of multiple stages of converging-product-line bottlenecks that make it hard to jump the entire GDP significantly even when you jump one component of a production cycle... which, from my standpoint, is a major reason to expect AI to not hit world GDP all that hard until AGI passes the critical threshold of bypassing it entirely. Having 95% of the tech to invent a self-replicating organism (eg artificial bacterium) does not get you 95%, 50%, or even 10% of the impact.

(it's not so much the 2% reaction of world markets to Evergrande that I was singling out earlier, 2% is noise-ish, but the wider swings in the vicinity of Evergrande particularly)

[Christiano][12:41]

Yeah, I'm just using "stone" to mean "elo difference that is equal to 1 stone at amateur dan / low kyu" you can see DeepMind's conversion (which I also don't totally believe) in figure 4 here (<https://sci-hub.se/https://www.nature.com/articles/nature16961>). Stones are closer to constant elo than constant handicap, it's just a convention to name them that way.

[Yudkowsky][12:42]

k then

[Christiano][12:47]

But my description above still kind of understates the gap I think. They call 230 elo 1 stone, and I think the prior rate of progress is more like 200 elo/year. They put AlphaZero about 3200 elo above the 2012 Go system, so that's like 16 years ahead = 11 years ahead of schedule. At least 2 years are from test-time hardware, and self-play systematically overestimates elo differences at the upper end of that. But 5 years ahead is still too low and that sounds more like 7-9 years ahead. ETA: and my actual best guess for when AlphaZero will beat a human is probably 10 years ahead, which I agree is just a lot bigger than 5. And I also underestimated how much of the gap was getting up to Lee Sedol.

The go graph I posted wasn't made with hindsight, that was from 2014

I mean, I'm fine with you saying that people who like smooth graphs are cherry-picking evidence, but you want to give any example other than nuclear weapons of technologies with the kind of discontinuous impact you are describing?

I do agree that the difference in our views is like "innovation" vs "industry." And a big part of my position is that innovation-like things just don't usually have big impacts for kind of obvious reasons, they start small and then become more industry-like as they scale up. And current deep learning seems like an absolutely stereotypical industry that is scaling up rapidly in an increasingly predictable way.

As far as I can tell the examples we know of things changing continuously aren't handpicked, we've been looking at all the examples we can find, and no one is proposing or even able to find almost *anything* that looks like you are imagining AI will look.

Like, we've seen deep learning innovations in the form of prototypes (most of all AlexNet), and they were cool and represented giant fast changes in people's views. And more recently we are seeing big much-less-surprising changes that are still helping a lot in raising the tens of billions of dollars that people are raising. And the innovations we are seeing are increasingly things that trade off against modest improvements in model size, there are fewer and fewer big surprises, just like you'd predict.

clearer and clearer to more and more people what the roadmap is---the roadmap is not yet quite as clear as in semiconductors, but as far as I can tell that's just because the field is still smaller.

[Yudkowsky][13:23]

I sure wasn't imagining there was a roadmap to AGI! Do you perchance have one which says that AG 30 years out?

From my perspective, you could as easily point to the Wright Flyer as an atomic bomb. Perhaps this reflects again the "innovation vs industry" difference, where I think in terms of building a thing that grows thereby bypassing our small cute world GDP, and you think in terms of industries that affect whole GDP in an invariant way throughout their lifetimes.

Would you perhaps care to write off the atomic bomb too? It arguably didn't change the outcome of World War II or do much that conventional weapons in great quantity couldn't; Japan was bluffed into believing the US could drop a nuclear bomb every week, rather than the US actually having that many nuclear bombs or them actually being used to deliver a historically outsized impact on Japan. From the industry-centric perspective, there is surely some graph you can draw which makes nuclear weapons also look like business as usual, especially if you go by destruction per unit of whole-industry non-marginal expense, rather than destruction per bomb.

[Christiano][13:27]

seems like you have to make the Wright Flyer much better before it's important, and that it becomes more like an industry as that happens, and that this is intimately related to why so few people were working on it

I think the atomic bomb is further on the spectrum than almost anything, but it still doesn't feel near as far as what you are expecting out of AI

the Manhattan Project took years and tens of billions; if you wait an additional few years and spend an additional few tens of billions then it would be a significant improvement in destruction or deterrence per \$ (but not totally insane)

I do think it's extremely non-coincidental that the atomic bomb was developed in a country that was practically outspending the whole rest of the world in "killing people technology"

and took a large fraction of that country's killing-people resources

eh, that's a bit unfair, the US was only like 35% of global spending on munitions

and the Manhattan Project itself was only a couple percent of total munitions spending

[Yudkowsky][13:32]

a lot of why I expect AGI to be a disaster is that *I am straight-up expecting AGI to be different*. If it were just like coal or just like nuclear weapons or just like viral biology then I would not be way more worried about AGI than I am worried about those other things.

[Christiano][13:33]

that definitely sounds right

but it doesn't seem like you have any short-term predictions about AI being different

9.2. AI alignment vs. biosafety, and measuring progress

[Yudkowsky][13:33]

are you more worried about AI than about bioengineering?

[Christiano][13:33]

I'm more worried about AI because (i) alignment is a thing, unrelated to takeoff speed, (ii) AI is a (ET/ likely to be) huge deal and bioengineering is probably a relatively small deal

(in the sense of e.g. how much \$ people spend, or how much \$ it makes, or whatever other metric of size you want to use)

[Yudkowsky][13:35]

what's the disanalogy to (i) biosafety is a thing, unrelated to the speed of bioengineering? why expect AI to be a huge deal and bioengineering to be a small deal? is it just that investing in AI is scaling faster than investment in bioengineering?

[Christiano][13:35]

no, alignment is a really easy x-risk story, bioengineering x-risk seems extraordinarily hard

It's really easy to mess with the future by creating new competitors with different goals, if you want to mess with the future by totally wiping out life you have to really try at it and there's a million ways it can fail. The bioengineering seems like it basically requires deliberate and reasonably competent malice whereas alignment seems like it can only be averted with deliberate effort, etc.

I'm mostly asking about historical technologies to try to clarify expectations, I'm pretty happy if the outcome is: you think AGI is predictably different from previous technologies in ways we haven't seen yet

though I really wish that would translate into some before-end-of-days prediction about a way that AGI will eventually look different

[Yudkowsky][13:38]

in my ontology a whole lot of threat would trace back to "AI hits harder, faster, gets too strong to be adjusted"; tricks with proteins just don't have the raw power of intelligence

[Christiano][13:39]

in my view it's nearly totally orthogonal to takeoff speed, though fast takeoffs are a big reason that preparation in advance is more useful

(but not related to the basic reason that alignment is unprecedently scary)

It feels to me like you are saying that the AI-improving-AI will move very quickly from "way slower than humans" to "FOOM in <1 year," but it just looks like that is very surprising to me.

However I do agree that if AI-improving-AI was like AlphaZero, then it would happen extremely fast.

It seems to me like it's pretty rare to have these big jumps, and it gets much much rarer as technologies become more important and are more industry-like rather than innovation like (and people care about them a lot rather than random individuals working on them, etc.). And I can't tell whether you are saying something more like "nah big jumps happen all the time in places that are structurally analogous to the key takeoff jump, even if the effects are blunted by slow adoption and regulatory bottlenecks and so on" or if you are saying "AGI is atypical in how jumpy it will be"

[Yudkowsky][13:44]

I don't know about *slower*; GPT-3 may be able to type faster than a human

[Christiano][13:45]

Yeah, I guess we've discussed how you don't like the abstraction of "speed of making progress"

[Yudkowsky][13:45]

but, basically less useful in fundamental ways than a human civilization, because they are less complete, less self-contained

[Christiano][13:46]

Even if we just assume that your AI needs to go off in the corner and not interact with humans, there still a question of why the self-contained AI civilization is making ~0 progress and then all of a sudden very rapid progress

[Yudkowsky][13:46]

unfortunately a lot of what you are saying, from my perspective, has the flavor of, "but can't you tell about your predictions earlier on of the impact on global warming at the *Homo erectus* level"

you have stories about why this is like totally not a fair comparison

I do not share these stories

[Christiano][13:46]

I don't understand either your objection nor the reductio

like, here's how I think it works: AI systems improve gradually, including on metrics like "How long does it take them to do task X?" or "How high-quality is their output on task X?"

[Yudkowsky][13:47]

I feel like the thing we know is something like, there is a sufficiently high level where things go whooo... humans-from-hominids style

[Christiano][13:47]

We can measure the performance of AI on tasks like "Make further AI progress, without human input"

Any way I can slice the analogy, it looks like AI will get continuously better at that task

[Yudkowsky][13:48]

how would you measure progress from GPT-2 to GPT-3, and would you feel those metrics really capture the sort of qualitative change that lots of people said they felt?

[Christiano][13:48]

And it seems like we have a bunch of sources of data we can use about how fast AI will get better

Could we talk about some application of GPT-2 or GPT-3?

also that's a *lot* of progress, spending 100x more is a *lot* more money

[Yudkowsky][13:49]

my world, GPT-3 has very few applications because it is not quite right and not quite complete

[Christiano][13:49]

also it's still really dumb

[Yudkowsky][13:49]

like a self-driving car that does great at 99% of the road situations

economically almost worthless

[Christiano][13:49]

I think the "being dumb" is way more important than "covers every case"

[Yudkowsky][13:50]

(albeit that if new cities could still be built, we could totally take those 99%-complete AI cars and build fences and fence-gates around them, in a city where they were the only cars on the road, in which case they *would* work, and get big economic gains from these new cities with driverless cars, which ties back into my point about how current world GDP is *unwilling* to accept tech inputs)

like, it is in fact very plausible to me that there is a neighboring branch of reality with open borders and no housing-supply-constriction laws and no medical-supply-constriction laws, and their world GDP does manage to double before AGI hits them really hard, albeit maybe not in 4 years. this world is not Earth, they are constructing new cities to take advantage of 99%-complete driverless cars *right now*, or rather, they started constructing them 5 years ago and finished 4 years and 6 months ago.

9.3. Requirements for FOOM

[Christiano][13:53]

I really feel like the important part is the jumpiness you are imagining on the AI side / why AGI is different from other things

[Cotra][13:53]

It's actually not obvious to me that Eliezer is imagining that much more jumpiness on the AI technology side than you are, Paul

E.g. he's said in the past that while the gap from "subhuman to superhuman AI" could be 2h if it's in the middle of FOOM, it could also be a couple years if it's more like scaling alphago

[Yudkowsky][13:54]

Indeed! We observed this jumpiness with hominids. A lot of stuff happened at once with hominids, because a critical terminal part of the jump was the way that hominids started scaling their own food supply, instead of being ultimately limited by the food supply of the savanna.

[Cotra][13:54]

A couple years is basically what Paul believes

[Christiano][13:55]

(discord is not a great place for threaded conversations :()

[Cotra][13:55]

What are the probabilities you're each placing on the 2h-2y spectrum? I feel like Paul is like "no way < 2h, likely on 2y" and Eliezer is like "who knows" on the whole spectrum, and a lot of the disagreement is about the impact of the previous systems?

[Christiano][13:55]

yeah, I'm basically at "no way," because it seems obvious that the AI that can foom in 2h is preceded by the AI that can foom in 2y

[Yudkowsky][13:56]

well, we surely agree there!

[Christiano][13:56]

OK, and it seems to me like it is preceded by years

[Yudkowsky][13:56]

we disagree on whether the AI that can foom in 2y clearly comes more than 2y before the AI that fooms in 2h

[Christiano][13:56]

yeah

perhaps we can all agree it's preceded by at least 2h

so I have some view like: for any given AI we can measure "how long does it take to foom?" and it seems to me like this is just a nice graph

and it's not exactly clear how quickly that number is going down, but a natural guess to me is something like "halving each year" based on the current rate of progress in hardware and software

and you see localized fast progress most often in places where there hasn't yet been much attention

and my best guess for your view is that actually that's not a nice graph at all, there is some critical threshold or range where AI quickly moves from "not foaming for a really long time" to "foaming real fast," and that seems like the part I'm objecting to

[Cotra][13:59]

Paul, is your take that there's a non-infinity number for time to FOOM that'd be associated with current AI systems (unassisted by humans)?

And it's going down over time?

I feel like I would have said something more like "there's a \$ amount it takes to build a system that will FOOM in X amount of time, and that's going down"

where it's like quadrillions of dollars today

[Christiano][14:00]

I think it would be a big engineering project to make such an AI, which no one is doing because it would be uselessly slow even if successful

[Yudkowsky][14:02]

I... don't think GPT-3 fooms given 2^{30} longer time to think about than the systems that would otherwise exist 30 years from now, on timelines I'd consider relatively long, and hence generous to the viewpoint? I also don't think you can take a quadrillion dollars and scale GPT-3 to foom today?

[Cotra][14:03]

I would agree with your take on GPT-3 fooming, and I didn't mean a quadrillion dollars just to scale GPT-3, would probably be a diff't architecture

[Christiano][14:03]

I also agree that GPT-3 doesn't foom, it just keeps outputting <EOT>[next web page]<EOT>...

But I think the axes of "smart enough to foom fast" and "wants to foom" are pretty different. I also agree there is some minimal threshold below which it doesn't even make sense to talk about "wants to foom", which I think is probably just not that hard to reach.

(Also there are always diminishing returns as you continue increasing compute, which become very relevant if you try to GPT-3 for a billion billion years as in your hypothetical even apart from "wants to foom".)

[Cotra][14:06]

I think maybe you and EY then disagree on where the threshold from "infinity" to "a finite number" for "time for this AI system to FOOM" begins? where eliezer thinks it'll drop from infinity to a pretty small finite number and you think it'll drop to a pretty large finite number, and keep going down from there

[Christiano][14:07]

I also think we will likely jump down to a foom-ing system only after stuff is pretty crazy, but I think that's probably less important

I think what you said is probably the main important disagreement

[Cotra][14:08]

as in before that point it'll be faster to have human-driven progress than FOOM-driven progress bc the FOOM would be too slow?

and there's some crossover point around when the FOOM time is just a bit faster than the human-driven progress time

[Christiano][14:09]

yeah, I think most likely (AI+humans) is faster than (AI alone) because of complementarity. But I thin Eliezer and I would still disagree even if I thought there was 0 complementarity and it's just (humans improving AI) and separately (AI improving AI)

on that pure substitutes model I expect "AI foom" to start when the rate of AI-driven AI progress overtakes the previous rate of human-driven AI progress

like, I expect the time for successive "doublings" of AI output to be like 1 year, 1 year, 1 year, 1 year, takes over] 6 months, 3 months, ...

and the most extreme fast takeoff scenario that seems plausible is that kind of perfect substitutes + physical economic impact from the prior AI systems

and then by that point fast enough physical impact is really hard so it happens essentially after the software-only singularity

I consider that view kind of unlikely but at least coherent

9.4. AI-driven accelerating economic growth

[Yudkowsky][14:12]

I'm expecting that the economy doesn't accept much inputs from chimps, and then the economy doesn't accept much input from village idiots, and then the economy doesn't accept much input from weird immigrants. I can imagine that there may or may not be a very weird 2-year or 3-month period with strange half-genius systems running around, but they will still not be allowed to build houses. In the terminal phase things get more predictable and the AGI starts its own economy instead.

[Christiano][14:12]

I guess you can go even faster, by having a big and accelerating ramp-up in human investment right around the end, so that the "1 year" is faster (e.g. if recursive self-improvement was like playing go, you could move from "a few individuals" to "google spending \$10B" over a few years)

[Yudkowsky][14:13]

My model prophecy doesn't rule that out as a thing that could happen, but sure doesn't emphasize it as a key step that needs to happen.

[Christiano][14:13]

I think it's very likely that AI will mostly be applied to further hardware+software progress

[Cotra: +]

I don't really understand why you keep talking about houses and healthcare

[Cotra][14:13]

Eliezer, what about stuff like Google already using ML systems to automate its TPU load-sharing decisions, and people starting to use Codex to automate routine programming, and so on? Seems like there's a lot of stuff like that starting to already happen and markets are pricing in huge further increases

[Christiano][14:14]

it seems like the non-AI up-for-grabs zone are things like manufacturing, not things like healthcare

[Cotra: +]

[Cotra][14:14]

(I mean on your timelines obviously not much time for acceleration anyway, but that's distinct from t regulation not allowing weak AIs to do stuff story)

[Yudkowsky][14:14]

Because I think that a key thing of what makes your prophecy less likely is the way that it happens inside the real world, where, economic gains or not, the System is unwilling/unable to take the things that are 99% self-driving cars and start to derive big economic benefits from those.

[Cotra][14:15]

but it seems like huge economic gains could happen entirely in industries mostly not regulated and n customer-facing, like hardware/software R&D, manufacturing, shipping logistics, etc

[Yudkowsky][14:15]

Ajeya, I'd consider Codex of *far* greater could-be-economically-important-ness than automated TPU load sharing decisions

[Cotra][14:15]

i would agree with that, it's smarter and more general

and i think that kind of thing could be applied on the hardware chip design side too

[Yudkowsky][14:16]

no, because the TPU load-sharing stuff has an obvious saturation point as a world economic input, w superCodex could be a world economic input in many more places

[Cotra][14:16]

the TPU load sharing thing was not a claim that this application could scale up to crazy impacts, but t it was allowed to happen, and future stuff that improves that kind of thing (back-end hardware/software/logistics) would probably also be allowed

[Yudkowsky][14:16]

my sense is that decuplicating the number of programmers would not lift world GDP much, but it seems lot more possible for me to be wrong about that

[Christiano][14:17]

the point is that housing and healthcare are not central examples of things that scale up at the beginning of explosive growth, regardless of whether it's hard or soft

they are slower and harder, and also in efficient markets-land they become way less important during the transition

so they aren't happening that much on anyone's story

and also it doesn't make that much difference whether they happen, because they have pretty limited effects on other stuff

like, right now we have an industry of ~hundreds of billions that is producing computing hardware, building datacenters, mining raw inputs, building factories to build computing hardware, solar panels shipping around all of those parts, etc. etc.

I'm kind of interested in the question of whether all that stuff explodes, although it doesn't feel as cool as the question of "what are the dynamics of the software-only singularity and how much \$ are people spending initiating it?"

but I'm not really interested in the question of whether human welfare is spiking during the transition only after

[Yudkowsky][14:20]

All of world GDP has never felt particularly relevant to me on that score, since twice as much hardware maybe corresponds to being 3 months earlier, or something like that.

[Christiano][14:21]

that sounds like the stuff of predictions?

[Yudkowsky][14:21]

But if complete chip manufacturing cycles have accepted much more effective AI input, with no non-bottlenecks, then that... sure is a much more *material* element of a foom cycle than I usually envision

[Christiano][14:21]

like, do you think it's often the case that 3 months of software progress = doubling compute spending or do you think AGI is different from "normal" AI on this perspective?

I don't think that's that far off anyway

I would guess like ~1 year

[Yudkowsky][14:22]

Like, world GDP that goes up by only 10%, but that's because producing compute capacity was 2.5% of world GDP and that quadrupled, starts to feel much more to me like it's part of a foom story.

I expect software-beats-hardware to hit harder and harder as you get closer to AGI, yeah.

the prediction is firmer near the terminal phase, but I think this is also a case where I expect that to be visible earlier

[Christiano][14:24]

I think that by the time that the AI-improving-AI takes over, it's likely that hardware+software manufacturing+R&D represents like 10-20% of GDP, and that the "alien accountants" visiting earth would value those companies at like 80%+ of GDP

9.5. Brain size and evolutionary history

[Cotra][14:24]

On software beating hardware, how much of your view is dependent on your belief that the chimp -> human transition was probably not mainly about brain size because if it were about brain size it would have happened faster? My understanding is that you think the main change is a small software innovation which increased returns to having a bigger brain. If you changed your mind and thought the chimp -> human transition was probably mostly about raw brain size, what (if anything) about your AI takeoff views would change?

[Yudkowsky][14:25]

I think that's a pretty different world in a lot of ways!

but yes it hits AI takeoff views too

[Christiano][14:25]

regarding software vs hardware, here is an example of asking this question for imangenet classification ("how much compute to train a model to do the task?"), with a bit over 1 year doubling times (<https://openai.com/blog/ai-and-efficiency/>). I guess my view is that we can make a similar graph for "compute required to make your AI FOOM" and that it will be falling significantly slower than 2x/year. And my prediction for other tasks is that the analogous graphs will also tend to be falling slower than 2x/year.

[Yudkowsky][14:26]

to the extent that I modeled hominid evolution as having been "dutifully schlep more of the same stuff predictably more of the same returns" that would correspond to a world in which intelligence was less scary, different, dangerous-by-default

[Cotra][14:27]

thanks, that's helpful. I looked around in [IEM](#) and other places for a calculation of how quickly we should have evolved to humans if it were mainly about brain size, but I only found qualitative statements. If there's a calculation somewhere I would appreciate a pointer to it, because currently it seems to me a story like "selection pressure toward general intelligence was weak-to-moderate because it wasn't actually *that* important for fitness, and this degree of selection pressure is consistent with brain size being the main deal and just taking a few million years to happen" is very plausible

[Yudkowsky][14:29]

well, for one thing, the prefrontal cortex expanded twice as fast as the rest

and iirc there's evidence of a lot of recent genetic adaptation... though I'm not as sure you could pinpoint it as being about brain-stuff or that the brain-stuff was about cognition rather than rapidly shifting motivations or something.

elephant brains are 3-4 times larger by weight than human brains (just looked up)

if it's that easy to get returns on scaling, seems like it shouldn't have taken that long for evolution to get there

[Cotra][14:31]

but they have fewer synapses (would compute to less FLOP/s by the standard conversion)

how long do you think it should have taken?

[Yudkowsky][14:31]

early dinosaurs should've hopped onto the predictable returns train

[Cotra][14:31]

is there a calculation?

you said in IEM that evolution increases organ sizes quickly but there wasn't a citation to easily follow on there

[Yudkowsky][14:33]

I mean, you could produce a graph of smooth fitness returns to intelligence, smooth cognitive returns: brain size/activity, linear metabolic costs for brain activity, fit that to humans and hominids, then show that obviously if hominids went down that pathway, large dinosaurs should've gone down it first because they had larger bodies and the relative metabolic costs of increased intelligence would've been lower at every point along the way

I do not have a citation for that ready, if I'd known at the time you'd want one I'd have asked Luke M it while he still worked at MIRI 😊

[Cotra][14:35]

cool thanks, will think about the dinosaur thing (my first reaction is that this should depend on the actual fitness benefits to general intelligence which might have been modest)

[Yudkowsky][14:35]

I suspect we're getting off Paul's crux, though

[Cotra][14:35]

yeah we can go back to that convo (though I think Paul would also disagree about this thing, and believes that the chimp to human thing was mostly about size)

sorry for hijacking

[Yudkowsky][14:36]

well, if at some point I can produce a major shift in EA viewpoints by coming up with evidence for a bunch of non-brain-size brain selection going on over those timescales, like brain-related genes where we can figure out how old the mutation is, I'd then put a lot more priority on digging up a paper like that

I'd consider it sufficiently odd to imagine hominids->humans as being primarily about brain size, given the evidence we have, that I do not believe this is Paul's position until Paul tells me so

[Christiano][14:49]

I would guess it's primarily about brain size / neuron count / cortical neuron count

and that the change in rate does mostly go through changing niche, where both primates and birds have this cycle of rapidly accelerating brain size increases that aren't really observed in other animals
it seems like brain size is increasing extremely quickly on both of those lines

[Yudkowsky][14:50]

why aren't elephants GI?

[Christiano][14:51]

mostly they have big brains to operate big bodies, and also my position obviously does not imply (big brain) ===(necessarily implies)==> general intelligence

[Yudkowsky][14:52]

I don't understand, in general, how your general position manages to strongly imply a bunch of stuff about AGI and not strongly imply similar stuff about a bunch of other stuff that sure sounds similar to me

[Christiano][14:52]

don't elephants have very few synapses relative to humans?

[Cotra: +]

how does the scale hypothesis possibly take a strong stand on synapses vs neurons? I agree that it takes a modest predictive hit from "why aren't the big animals much smarter?"

[Yudkowsky][14:53]

if adding more synapses just scales, elephants should be able to pay hominid brain costs for a much smaller added fraction of metabolism and also not pay the huge death-in-childbirth head-size tax because their brains and heads are already 4x as huge as they need to be for GI and now they just need some synapses, which are a much tinier fraction of their total metabolic cost.

[Christiano][14:54]

I mean, you can also make smaller and cheaper synapses as evidenced by birds
I'm not sure I understand what you are saying
it's clear that you can't say "X is possible metabolically, so evolution would do it"
or else you are confused about why primate brains are so bad

[Yudkowsky][14:54]

great, then smaller and cheaper synapses should've scaled many eons earlier and taken over the wo

[Christiano][14:55]

this isn't about general intelligence, this is a reductio of your position...

[Yudkowsky][14:55]

and here I had thought it was a reductio of your position...

[Christiano][14:55]

indeed

like, we all grant that it's metabolically possible to have small smart brains
and evolution doesn't do it
and I'm saying that it's also possible to have small smart brains
and that scaling brains up matters a lot

[Yudkowsky][14:56]

no, you grant that it's metabolically possible to have cheap brains full of synapses, which are therefo
on your position, smart

[Christiano][14:56]

birds are just smart

we know they are smart

this isn't some kind of weird conjecture

like, we can debate whether they are a "general" intelligence, but it makes no difference to this
discussion

the point is that they do more with less metabolic cost

[Yudkowsky][14:57]

on my position, the brain needs to invent the equivalents of ReLUs and Transformers and really rath
lot of other stuff because it can't afford nearly that many GPUs, and then the marginal returns on ad
expensive huge brains and synapses have increased enough that hominids start to slide down the
resulting fitness slope, which isn't even paying off in guns and rockets yet, they're just getting that
much intelligence out of it once the brain software has been selected to scale that well

[Christiano][14:57]

but all of the primates and birds have brain sizes scaling much faster than the other animals

like, the relevant "things started to scale" threshold is way before chimps vs humans

isn't it?

[Cotra][14:58]

to clarify, my understanding is that paul's position is "Intelligence is mainly about synapse/neuron co
and evolution doesn't care that much about intelligence; it cared more for birds and primates, and be
lines are getting smarter+bigger-brained." And eliezer's position is that "evolution should care a ton
about intelligence in most niches, so if it were mostly about brain size then it should have gone up to
human brain sizes with the dinosaurs"

[Christiano][14:58]

or like, what is the evidence you think is explained by the threshold being between chimps and huma

[Yudkowsky][14:58]

if hominids have less efficient brains than birds, on this theory, it's because (post facto handwave) birds are tiny, so whatever cognitive fitness gradients they face, will tend to get paid more in software and biological efficiency and biologically efficient software, and less paid in Stack More Neurons (even compared to hominids)

elephants just don't have the base software to benefit much from scaling synapses even though they are relatively cheaper for elephants

[Christiano][14:59]

@ajeya I think that intelligence is about a lot of things, but that size (or maybe "more of the same" changes that had been happening recently amongst primates) is the big difference between chimps and humans

[Cotra: 

[Cotra][14:59]

got it yeah i was focusing on chimp-human gap when i said "intelligence" there but good to be careful

[Yudkowsky][14:59]

I have not actually succeeded in understanding Why On Earth Anybody Would Think That If Not For That Really Weird Prior I Don't Get Either

re: the "more of the same" theory of humans

[Cotra][15:00]

do you endorse my characterization of your position above? "evolution should care a ton about intelligence in most niches, so if it were mostly about brain size then it should have gone up to human brain sizes with the dinosaurs"

in which case the disagreement is about how much evolution should care about intelligence in the dinosaur niche, vs other things it could put its skill points into?

[Christiano][15:01]

Eliezer, it seems like chimps are insanely smart compared to other animals, basically as smart as the get

so it's natural to think that the main things that make humans unique are also present in chimps

or at least, there was something going on in chimps that is exceptional

and should be causally upstream of the uniqueness of humans too

otherwise you have too many coincidences on your hands

[Yudkowsky][15:02]

ajeya: no, I'd characterize that as "the human environmental niche per se does not seem super-specific enough to be unique on a geological timescale, the cognitive part of the niche derives from increased cognitive abilities in the first place and so can't be used to explain where they got started, dinosaurs were larger than humans and would pay lower relative metabolic costs for added brain size and it is not the case that every species as large as humans was in an environment where they would not have benefited as much from a fixed increment of intelligence, hominids are probably distinguished from dinosaurs in having better neural algorithms that arose over intervening evolutionary time and there were better returns in intelligence on synapses that are more costly to humans than to elephants or large dinosaurs"

[Christiano][15:03]

I don't understand how you can think that hominids are the special step relative to something earlier or like, I can see how it's consistent, but I don't see what evidence or argument supports it
it seems like the short evolutionary time, and the fact that you also have to explain the exceptional qualities of other primates, cut extremely strongly against it

[Yudkowsky][15:04]

paul: indeed, the fact that dinosaurs didn't see their brain sizes and intelligences ballooning, says there must be a lot of stuff hominids had that dinosaurs didn't, explaining why hominids got much higher returns on intelligence per synapse. natural selection is enough of a smooth process that 95% of this stuff should've been in the last common ancestor of humans and chimps.

[Christiano][15:05]

it seems like brain size basically just increases faster in the smarter animals? though I mostly just know about birds and primates

[Yudkowsky][15:05]

that is what you'd predict from smartness being about algorithms!

[Christiano][15:05]

and it accelerates further and further within both lines

it's what you'd expect if smartness is about algorithms *and chimps and birds have good algorithms*

[Yudkowsky][15:06]

if smartness was about brain size, smartness and brain size would increase faster in the *larger animals* or the ones whose successful members *ate more food per day*

well, sure, I do model that birds have better algorithms than dinosaurs

[Cotra][15:07]

it seems like you've given arguments for "there was algorithmic innovation between dinosaurs and humans" but not yet arguments for "there was major algorithmic innovation between chimps and humans"?

[Christiano][15:08]

(much less that the algorithmic changes were not just more-of-the-same)

[Yudkowsky][15:08]

oh, that's *not* mandated by the model the same way. (between LCA of chimps and humans)

[Christiano][15:08]

isn't that exactly what we are discussing?

[Yudkowsky][15:09]

...I hadn't thought so, no.

[Cotra][15:09]

original q was:

On software beating hardware, how much of your view is dependent on your belief that the chimp > human transition was probably not mainly about brain size because if it were about brain size it would have happened faster? My understanding is that you think the main change is a small software innovation which increased returns to having a bigger brain. If you changed your mind and thought that the chimp -> human transition was probably mostly about raw brain size, what (if anything) about your AI takeoff views would change?

so i thought we were talking about if there's a cool innovation from chimp->human?

[Yudkowsky][15:10]

I can see how this would have been the more obvious intended interpretation on your viewpoint, and apologize

[Christiano][15:10]

(though i think paul would also disagree about this thing, and believes that the chimp to human thing was mostly about size)

Is what I was responding to in part

I am open to saying that I'm conflating size and "algorithmic improvements that are closely correlate with size in practice and are similar to the prior algorithmic improvements amongst primates"

[Yudkowsky][15:11]

from my perspective, the question is "how did that hominid->human transition happen, as opposed to there being an elephant->smartelephant or dinosaur->smartdinosaur transition"?

I expect there were substantial numbers of brain algorithm stuffs going on during this time, however because I don't think that synapses scale that well *with* the baseline hominid boost

[Christiano][15:11]

FWIW, it seems quite likely to me that there would be an elephant->smartelephant transition within tens of millions or maybe 100M years, and a dinosaur->smartdinosaur transition in hundreds of millions of years

and those are just cut off by the fastest lines getting there first

[Yudkowsky][15:12]

which I think does circle back to that point? actually I think my memory glitched and forgot the original point while being about this subpoint and I probably did interpret the original point as intended.

[Christiano][15:12]

namely primates beating out birds by a hair

[Yudkowsky][15:12]

that sounds like a viewpoint which would also think it much more likely that GPT-3 would foom in a billion years

where maybe you think that's unlikely, but I still get the impression your "unlikely" is, like, 5 orders o magnitude likelier than mine before applying overconfidence adjustments against extreme probabilit on both sides

yeah, I think I need to back up

[Cotra][15:15]

Is your position something like "at some point after dinosaurs, there was an algorithmic innovation th increased returns to brain size, which meant that the birds and the humans see their brains increasir quickly while the dinosaurs didn't"?

[Christiano][15:15]

it also seems to me like the chimp->human difference is in basically the same ballpark of the effect c brain size within humans, given modest adaptations for culture

which seems like a relevant sanity-check that made me take the "mostly hardware" view more seriou

[Yudkowsky][15:15]

there's a part of my model which very strongly says that hominids scaled better than elephants and that's why "hominids->humans but not elephants->superelephants"

[Christiano][15:15]

previously I had assumed that analysis would show that chimps were obviously way dumber than an extrapolation of humans

[Yudkowsky][15:16]

there's another part of my model which says "and it still didn't scale that well without algorithms, so should expect a lot of alleles affecting brain circuitry which rose to fixation over the period when hominid brains were expanding"

this part is strong and I think echoes back to AGI stuff, but it is not as *strong* as the much *more* overdetermined position that hominids started with more scalable algorithms than dinosaurs.

[Christiano][15:17]

I do agree with the point that there are structural changes in brains as you scale them up, and this is potentially a reason why brain size changes more slowly than e.g. bone size. (Also there are small structural changes in ML algorithms as you scale them up, not sure how much you want to push the analogy but they feel fairly similar.)

[Yudkowsky][15:17]

it also seems to me like the chimp->human difference is in basically the same ballpark of the effec of brain size within humans, given modest adaptations for culture

this part also seems pretty blatantly false to me
is there, like, a smooth graph that you looked at there?

[Christiano][15:18]

I think the extrapolated difference would be about 4 standard deviations, so we are comparing a chimp to an IQ 40 human

[Yudkowsky][15:18]

I'm really not sure how much of a fair comparison that is
IQ 40 humans in our society may be mostly sufficiently-damaged humans, not scaled-down humans

[Christiano][15:19]

doesn't seem easy, but the point is that the extrapolated difference is huge, it corresponds to completely debilitating developmental problems

[Yudkowsky][15:19]

if you do enough damage to a human you end up with, for example, a coma victim who's not competitive with other primates at all

[Christiano][15:19]

yes, that's more than 4 SD down
I agree with this general point
I'd guess I just have a lot more respect for chimps than you do

[Yudkowsky][15:20]

I feel like I have a bunch of respect for chimps but more respect for humans
like, that stuff humans do
that is really difficult stuff!
it is not just scaled-up chimpstuff!

[Christiano][15:21]

Carl convinced me chimps wouldn't go to space, but I still really think it's about domesticity and cultural issues rather than intelligence

[Yudkowsky][15:21]

the chimpstuff is very respectable but there is a whole big layer cake of additional respect on top

[Christiano][15:21]

not a prediction to be resolved until after the singularity
I mean, the space prediction isn't very confident 😊

and it involved a very large planet of apes

9.6. Architectural innovation in AI and in evolutionary history

[Yudkowsky][15:22]

I feel like if GPT-based systems saturate and require *any* architectural innovation rather than Stack M Layers to get much further, this is a pre-Singularity point of observation which favors humans probably being more qualitatively different from chimp-LCA

(LCA=last common ancestor)

[Christiano][15:22]

any seems like a kind of silly bar?

[Yudkowsky][15:23]

because single architectural innovations are allowed to have large effects!

[Christiano][15:23]

like there were already small changes to normalization from GPT-2 to GPT-3, so isn't it settled?

[Yudkowsky][15:23]

natural selection can't afford to deploy that many of them!

[Christiano][15:23]

and the model really eventually won't work if you increase layers but don't fix the normalization, there are severe problems that only get revealed at high scale

[Yudkowsky][15:23]

that I wouldn't call architectural innovation

transformers were

this is a place where I would not discuss specific ideas because I do not actually want this event to occur

[Christiano][15:24]

sure

have you seen a graph of LSTM scaling vs transformer scaling?

I think LSTM with ongoing normalization-style fixes lags like 3x behind transformers on language modeling

[Yudkowsky][15:25]

no, does it show convergence at high-enough scales?

[Christiano][15:25]

figure 7 here: <https://arxiv.org/pdf/2001.08361.pdf>

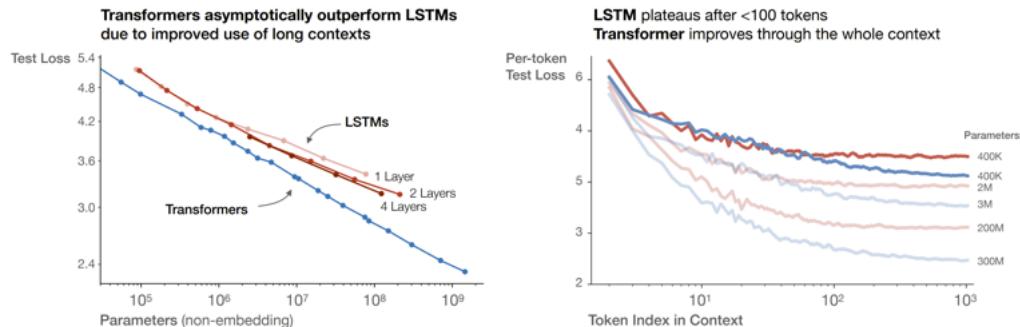


Figure 7

[Yudkowsky][15:26]

yeah... I unfortunately would rather not give other people a sense for which innovations are obviously more of the same and which innovations obviously count as qualitative

[Christiano][15:26]

I think smart money is that careful initialization and normalization on the RNN will let it keep up for longer

anyway, I'm very open to differences like LSTM vs transformer between humans and 3x-smaller-brain ancestors, as long as you are open to like 10 similar differences further back in the evolutionary history

[Yudkowsky][15:28]

what if there's 27 differences like that and 243 differences further back in history?

[Christiano][15:28]

sure

[Yudkowsky][15:28]

is that a distinctly Yudkowskian view vs a Paul view...

apparently not

I am again feeling confused about cruxes

[Christiano][15:29]

I mean, 27 differences like transformer vs LSTM isn't actually plausible, so I guess we could talk abou

[Cotra][15:30]

Here's a potential crux articulation that ties it back to the animals stuff: paul thinks that we first discover major algorithmic innovations that improve intelligence at a low level of intelligence, analogous to evolution discovering major architectural innovations with tiny birds and primates, and then there will be a long period of scaling up plus coming up with routine algorithmic tweaks to get to the high level analogous to evolution schlepping on the same shit for a long time to get to humans. analogously, he thinks when big innovations come onto the scene the actual product is crappy af (e.g. wright brother plane), and it needs a ton of work to scale up to usable and then to great.

you both seem to think both evolution and tech history consistently point in your direction

[Christiano][15:33]

that sounds vaguely right, I guess the important part of "routine" is "vaguely predictable," like you mostly work your way down the low-hanging fruit (including new fruit that becomes more important to you scale), and it becomes more and more predictable the more people are working on it and the longer you've been at it

and deep learning is already reasonably predictable (i.e. the impact of successive individual architectural changes is smaller, and law of large numbers is doing its thing) and is getting more so, I just expect that to continue

[Cotra][15:34]

yeah, like it's a view that points to using data that relates effort to algorithmic progress and using that to predict future progress (in combination with predictions of future effort)

[Christiano][15:35]

yeah

and for my part, it feels like this is how most technologies look and also how current ML progress looks

[Cotra][15:36]

and also how evolution looks, right?

[Christiano][15:37]

you aren't seeing big jumps in translation or in self-driving cars or in image recognition, you are just seeing a long slog, and you see big jumps in areas where few people work (usually up to levels that are not in fact that important, which is very correlated with few people working there)

I don't know much about evolution, but it at least looks very consistent with what I know and the fact eliezer cites

(not merely consistent, but "explains the data just about as well as the other hypotheses on offer")

9.7. Styles of thinking in forecasting

[Yudkowsky][15:38]

I do observe that this would seem, on the surface of things, to describe the entire course of natural selection up until about 20K years ago, if you were looking at surface impacts

[Christiano][15:39]

by 20k years ago I think it's basically obvious that you are tens of thousands of years from the singularity

like, I think natural selection is going crazy with the brains by millions of years ago, and by hundreds thousands of years ago humans are going crazy with the culture, and by tens of thousands of years the culture thing has accelerated and is almost at the finish line

[Yudkowsky][15:41]

really? I don't know if I would have been able to call that in advance if I'd never seen the future or an other planets. I mean, maybe, but I sure would have been extrapolating way out onto a further limb than I'm going here.

[Christiano][15:41]

Yeah, I agree singularity is way more out on a limb---or like, where the singularity stops is more uncertain since that's all that's really at issue from my perspective

but the point is that everything is clearly crazy in historical terms, in the same way that 2000 is crazy even if you don't know where it's going

and the timescale for the crazy changes is tens of thousands of years

[Yudkowsky][15:42]

I frankly model that, had I made any such prediction 20K years ago of hominids being able to pull off moon landings or global warming - never mind the Singularity - I would have faced huge pushback from many EAs, such as, for example, Robin Hanson, and you.

[Christiano][15:42]

like I think this can't go on would have applied just as well:

<https://www.lesswrong.com/posts/5FZxhd16hZp8QwK7k/this-can-t-go-on>

I don't think that's the case at all

and I think you still somehow don't understand my position?

[Yudkowsky][15:43]

<https://www.lesswrong.com/posts/XQirei3crsLxsCQoi/surprised-by-brains> is my old entry here

[Christiano][15:43]

like, what is the move I'm making here, that you think I would have made in the past?

and would have led astray?

[Yudkowsky][15:44]

I sure do feel in a deeper sense that I am trying very hard to account for perspective shifts in how unpredictable the future actually looks at the time, and the Other is looking back at the past and organizing it neatly and expecting the future to be that neat

[Christiano][15:45]

I don't even feel like I'm expecting the future to be neat

are you just saying you have a really broad distribution over takeoff speed, and that "less than a month gets a lot of probability because lots of numbers are less than a month?

[Yudkowsky][15:47]

not exactly?

[Christiano][15:47]

in what way is your view the one that is preferred by things being messy or unpredictable?

like, we're both agreeing X will eventually happen, and I'm making some concrete prediction about h some other X' will happen first, and that's the kind of specific prediction that's likely to be wrong?

[Yudkowsky][15:48]

more like, we sure can tell a story today about how normal and predictable AlphaGo was, but we can always tell stories like that about the past. I do not particularly recall the AI field standing up one year before AlphaGo and saying "It's time, we're coming for the 8-dan pros this year and we're gonna be world champions a year after that." (Which took significantly longer in chess, too, matching my other thesis about how these slides are getting steeper as we get closer to the end.)

[Christiano][15:49]

it's more like, you are offering AGZ as an example of why things are crazy, and I'm doubtful / think it's pretty lame

maybe I don't understand how it's functioning as bayesian evidence

for what over what

[Yudkowsky][15:50]

I feel like the whole smoothness-reasonable-investment view, if evaluated on Earth 5My ago *without benefit of foresight*, would have dismissed the notion of brains overtaking evolution; evaluated 1My ago it would have dismissed the notion of brains overtaking evolution; evaluated 20Ky ago, it would have barely started to acknowledge that brains were doing anything interesting at all, but pointed out how the hominids could still only eat as much food as their niche offered them and how the cute little handaxes did not begin to compare to livers and wasp stings.

there is a style of thinking that says, "wow, yeah, people in the past sure were surprised by stuff, oh, wait, I'm also in the past, aren't I, I am one of those people"

and a view where you look back from the present and think about how reasonable the past all seems now, and the future will no doubt be equally reasonable

[Christiano][15:52]

(the AGZ example may fall flat, because the arguments we are making about it now we were also making in the past)

[Yudkowsky][15:52]

I am not sure this is resolvable, but it is among my primary guesses for a deep difference in believed styles of thought

[Christiano][15:52]

I think that's a useful perspective, but still don't see how it favors your bottom line

[Yudkowsky][15:53]

where I look at the style of thinking you're using, and say, not, "well, that's invalidated by a technical error on line 3 even on Paul's own terms" but "isn't this obviously a whole style of thought that never works and ends up unrelated to reality"

I think the first AlphaGo was the larger shock, AlphaGo Zero was a noticeable but more mild shock or account of how it showed the end of game programming and not just the end of Go

[Christiano][15:54]

sorry, I lumped them together

[Yudkowsky][15:54]

it didn't feel like the same level of surprise; it was preceded by then

the actual accomplishment may have been larger in an important sense, but a lot of the - epistemic landscape of lessons learned? - is about the things that surprise you at the time

[Christiano][15:55]

also AlphaGo was also quite easy to see coming after this paper (as was discussed extensively at *the time*): <https://www.cs.toronto.edu/~cmaddis/pubs/deepgo.pdf>

[Yudkowsky][15:55]

Paul, are you on the record as arguing with me that AlphaGo will win at Go because it's predictably on trend?

back then?

[Cotra][15:55]

Hm, it sounds like Paul is saying "I do a trend extrapolation over long time horizons and if things seem to be getting faster and faster I expect they'll continue to accelerate; this extrapolation if done 100k years ago would have seen that things were getting faster and faster and projected singularity within 100s of K years"

Do you think Paul is in fact doing something other than the trend extrapolation he says he's doing, or that he would have looked at a different less informative trend than the one he says he would have looked at something else?

[Christiano][15:56]

my methodology for answering that question is looking at LW comments mentioning go by me, can see if it finds any

[Yudkowsky][15:56]

Different less informative trend, is most of my suspicion there?

though, actually, I should revise that, I feel like relatively little of the WHA was AlphaGo v2 whose name forget beating Lee Se-dol, and most was in the revelation that v1 beat the high-dan pro whose name forget.

Paul having himself predicted anything at all like this would be the actually impressive feat

that would cause me to believe that the AI world is more regular and predictable than I experienced it as, if you are paying more attention to ICLR papers than I do

9.8. Moravec's prediction

[Cotra][15:58]

And jtbc, the trend extrap paul is currently doing is something like:

- Look at how effort leads to hardware progress measured in FLOP/\$ and software progress measured in stuff like "FLOP to do task X" or "performance on benchmark Y"
- Look at how effort in the ML industry as a whole is increasing, project forward with maybe some adjustments for thinking markets are more inefficient now and will be less inefficient later

and this is the wrong trend, because he shouldn't be looking at hardware/software progress across the whole big industry and should be more open to an upset innovation coming from an area with a small number of people working on it?

and he would have similarly used the wrong trends while trying to do trend extrap in the past?

[Yudkowsky][15:59]

because I feel like this general style of thought doesn't work when you use it on Earth generally, and then fails extremely hard if you try to use it on Earth before humans to figure out where the hominids are going because that phenomenon is Different from Previous Stuff

like, to be clear, I have seen this used well on solar

I feel like I saw some people calling the big solar shift based on graphs, before that happened

I have seen this used great by Moravec on computer chips to predict where computer chips would be 2012

and also witnessed Moravec *completely failing* as soon as he tried to derive *literally anything but the graph itself* namely his corresponding prediction for human-equivalent AI in 2012 (I think, maybe it was 2010) or something

[Christiano][16:02]

(I think in his 1988 book Moravec estimated human-level AI in ~2030, not sure if you are referring to some earlier prediction?)

[Yudkowsky][16:02]

(I have seen Ray Kurzweil project out Moore's Law to the \$1,000,000 human brain in, what was it, 20 followed by the \$1000 human brain in 2035 and the \$1 human brain in 2045, and when I asked Ray whether machine superintelligence might shift the graph at all, he replied that machine superintelligence was precisely how the graph would be able to continue on trend. This indeed is silly than EAs.)

[Cotra][16:03]

moravec's prediction appears to actually be around 2025, looking at his hokey graph?
<https://jetpress.org/volume1/moravec.htm>



[Yudkowsky][16:03]

but even there, it does feel to me like there is a commonality between Kurzweil's sheer graph-worship and difficulty in appreciating the graphs as surface phenomena that are less stable than deep phenomena, and something that Hanson was doing wrong in the foom debate

[Cotra][16:03]

which is...like, your timelines?

[Yudkowsky][16:04]

that's 1998

Mind Children in 1988 I am pretty sure had an earlier prediction

[Christiano][16:04]

I should think you'd be happy to bet against me on basically any prediction, shouldn't you?

[Yudkowsky][16:05]

any prediction that sounds narrow and isn't like "this graph will be on trend in 3 more years"

...maybe I'm wrong, an online source says Mind Children in 1988 predicted AGI in "40 years" but I sur do seem to recall an extrapolated graph that reached "human-level hardware" in 2012 based on an extensive discussion about computing power to duplicate the work of the retina

[Christiano][16:08]

don't think it matters too much other than for Moravec's honor, doesn't really make a big difference to the empirical success of the methodology

I think it's on page 68 if you have the physical book

[Yudkowsky][16:09]

p60 via Google Books says 10 teraops for a human-equivalent mind

[Christiano][16:09]

I have a general read of history where trend extrapolation works extraordinarily well relative to other kinds of forecasting, to the extent that the best first-pass heuristic for whether a prediction is likely to

accurate is whether it's a trend extrapolation and how far in the future it is

[Yudkowsky][16:09]

which, incidentally, strikes me as entirely plausible if you had algorithms as sophisticated as the human brain

my sense is that Moravec nailed the smooth graph of computing power going on being smooth, but that all of his predictions about the actual future were completely invalid on account of a curve interacting with his curve that he didn't know things about and so simply omitted as a step in his calculations, namely, AGI algorithms

[Christiano][16:12]

though again, from your perspective 2030 is still a reasonable bottom-line forecast that makes him one of the most accurate people at that time?

[Yudkowsky][16:12]

you could be right about all the local behaviors that your history is already shouting out at you as having smooth curve (where by "local" I do mean to exclude stuff like world GDP extrapolated into the indefinite future) and the curves that history isn't shouting at you will tear you down

[Christiano][16:12]

(I don't know if he even forecast that)

[Yudkowsky][16:12]

I don't remember that part from the 1988 book

my memory of the 1988 book is "10 teraops, based on what it takes to rival the retina" and he drew a graph of Moore's Law

[Christiano][16:13]

yeah, I think that's what he did

(and got 2030)

[Yudkowsky][16:14]

"If this rate of improvement were to continue into the next century, the 10 teraops required for a humanlike computer would be available in a \$10 million supercomputer before 2010 and in a \$1,000 personal computer by 2030."

[Christiano][16:14]

or like, he says "human equivalent in 40 years" and predicts that in 50 years we will have robots with superhuman reasoning ability, not clear he's ruling out human-equivalent AGI before 40 years but I think the tone is clear

[Yudkowsky][16:15]

so 2030 for AGI on a personal computer and 2010 for AGI on a supercomputer, and I expect that on my first reading I simply discarded the former prediction as foolish extrapolation past the model collapse

had just predicted in 2010.

(p68 in "Powering Up")

[Christiano][16:15]

yeah, that makes sense

I do think the PC number seems irrelevant

[Cotra][16:16]

I think both in that book and in the 98 article he wants you to pay attention to the "very cheap human-size computers" threshold, not the "supercomputer" threshold, I think intentionally as a way to handwave in "we need people to be able to play around with these things"

(which people criticized him at the time for not more explicitly modeling iirc)

[Yudkowsky][16:17]

but! I mean! there are so many little places where the media has a little cognitive hiccup about that; decides in 1998 that it's fine to describe that retrospectively as "you predicted in 1988 that we'd have true AI in 40 years" and then the future looks less surprising than people at the time using Trend Log were actually surprised by it!

all these little ambiguities and places where, oh, you decide retroactively that it would have made sense to look at *this* Trend Line and use it *that* way, but if you look at what people said at the time, they did actually say that!

[Christiano][16:19]

I mean, in fairness reading the book it just doesn't seem like he is predicting human-level AI in 2010 rather than 2040, but I do agree that it seems like the basic methodology (why care about the small computer thing?) doesn't really make that much sense a priori and only leads to something sane if it cancels out with a weird view

9.9. Prediction disagreements and bets

[Christiano][16:19]

anyway, I'm pretty unpersuaded by the kind of track record appeal you are making here

[Yudkowsky][16:20]

if the future goes the way I predict and yet anybody somehow survives, perhaps somebody will draw hyperbolic trendline on some particular chart where the trendline is retroactively fitted to events including those that occurred in only the last 3 years, and say with a great sage nod, ah, yes, that was all according to trend, nor did anything depart from trend

trend lines permit anything

[Christiano][16:20]

like from my perspective the fundamental question is whether I would do better or worse by following the kind of reasoning you'd advocate, and it just looks to me like I'd do worse, and I'd love to make a predictions about anything to help make that more clear and hindsight-proof in advance

[Yudkowsky][16:20]

you just look into the past and find a line you can draw that ended up where reality went

[Christiano][16:21]

it feels to me like you really just waffle on almost any prediction about the before-end-of-days

[Yudkowsky][16:21]

I don't think I know a lot about the before-end-of-days

[Christiano][16:21]

like if you make a prediction I'm happy to trade into it, or you can pick a topic and I can make a prediction and you can trade into mine

[Cotra][16:21]

but you know enough to have strong timing predictions, e.g. your bet with caplan

[Yudkowsky][16:21]

it's daring enough that I claim to know anything about the Future at all!

[Cotra][16:21]

surely with that difference of timelines there should be some pre-2030 difference as well

[Christiano][16:21]

but you are the one making the track record argument against my way of reasoning about things!

how does that not correspond to believing that your predictions are better!

what does that mean?

[Yudkowsky][16:22]

yes and if you say something narrow enough or something that my model does at least vaguely push against, we should bet

[Christiano][16:22]

my point is that I'm willing to make a prediction about any old thing, you can name your topic

I think the way I'm reasoning about the future is just better in general

and I'm going to beat you on whatever thing you want to bet on

[Yudkowsky][16:22]

but if you say, "well, Moore's Law on trend, next 3 years", then I'm like, "well, yeah, sure, since I don't feel like I know anything special about that, that would be my prediction too"

[Christiano][16:22]

sure

you can pick the topic

pick a quantity

or a yes/no question

or whatever

[Yudkowsky][16:23]

you may know better than I would where your Way of Thought makes strong, narrow, or unusual predictions

[Christiano][16:23]

I'm going to trend extrapolation everywhere

spoiler

[Yudkowsky][16:23]

okay but any superforecaster could do that and I could do the same by asking a superforecaster

[Cotra][16:24]

but there must be places where you'd strongly disagree w the superforecaster

since you disagree with them eventually, e.g. >2/3 doom by 2030

[Bensinger][18:40] (Nov. 25 follow-up comment)

">2/3 doom by 2030" isn't an actual Eliezer-prediction, and is based on a misunderstanding of something Eliezer said. See [Eliezer's comment on LessWrong](#).

[Yudkowsky][16:24]

in the terminal phase, sure

[Cotra][16:24]

right, but there are no disagreements before jan 1 2030?

no places where you'd strongly defy the superforecasters/trend extrap?

[Yudkowsky][16:24]

superforecasters were claiming that AlphaGo had a 20% chance of beating Lee Se-dol and I didn't disagree with that at the time, though as the final days approached I became nervous and suggested

a friend that they buy out of a bet about that

[Cotra][16:25]

what about like whether we get some kind of AI ability (e.g. coding better than X) before end days

[Yudkowsky][16:25]

though that was more because of having started to feel incompetent and like I couldn't trust the superforecasters to know more, than because I had switched to a confident statement that AlphaGo would win

[Cotra][16:25]

seems like EY's deep intelligence / insight-oriented view should say something about what's not poss before we get the "click" and the FOOM

[Christiano][16:25]

I mean, I'm OK with either (i) evaluating arguments rather than dismissive and IMO totally unjustified track record, (ii) making bets about stuff

I don't see how we can both be dismissing things for track record reasons and also not disagreeing about things

if our methodologies agree about all questions before end of days (which seems crazy to me) then surely there is no track record distinction between them...

[Cotra: ]

[Cotra][16:26]

do you think coding models will be able to 2x programmer productivity before end days? 4x?

what about hardware/software R&D wages? will they get up to \$20m/yr for good ppl?

will someone train a 10T param model before end days?

[Christiano][16:27]

things I'm happy to bet about: economic value of LMs or coding models at 2, 5, 10 years, benchmark performance of either, robotics, wages in various industries, sizes of various industries, compute/\$, someone else's views about "how ML is going" in 5 years

maybe the "any GDP acceleration before end of days?" works, but I didn't like how you don't win until the end of days

[Yudkowsky][16:28]

okay, so here's an example place of a *weak* general Yudkowskian prediction, that is weaker than terminal-phase stuff of the End Days: (1) I predict that cycles of 'just started to be able to do Narrow Thing -> blew past upper end of human ability at Narrow Thing' will continue to get shorter, the same way that, I think, this happened faster with Go than with chess.

[Christiano][16:28]

great, I'm totally into it

what's a domain?

coding?

[Yudkowsky][16:28]

Does Paul disagree? Can Paul point to anything equally specific out of Paul's viewpoint?

[Christiano][16:28]

benchmarks for LMs?

robotics?

[Yudkowsky][16:28]

well, for these purposes, we do need some Elo-like ability to measure at all where things are relative humans

[Cotra][16:29]

problem-solving benchmarks for code?

MATH benchmark?

[Christiano][16:29]

well, for coding and LM'ing we have lots of benchmarks we can use

[Yudkowsky][16:29]

this unfortunately does feel a bit different to me from Chess benchmarks where the AI is playing the whole game; Codex is playing part of the game

[Christiano][16:29]

in general the way I'd measure is by talking about how fast you go from "weak human" to "strong human" (e.g. going from top-10,000 in chess to top-10 or whatever, going from jobs doable by \$50k/year engineer to \$500k/year engineer...)

[Yudkowsky][16:30]

golly, that sounds like a viewpoint very favorable to mine

[Christiano][16:30]

what do you mean?

that way of measuring would be favorable to your viewpoint?

[Yudkowsky][16:31]

if we measure how far it takes AI to go past different levels of paying professionals, I expect that the Chess duration is longer than the Go duration and that by the time Codex is replacing a most paid

\$50k/year programmers the time to replacing a most programmers paid as much as a top Go player
be pretty darned short

[Christiano][16:31]

top Go players don't get paid, do they?

[Yudkowsky][16:31]

they tutor students and win titles

[Christiano][16:31]

but I mean, they are like low-paid engineers

[Yudkowsky][16:31]

yeah that's part of the issue here

[Christiano][16:31]

I'm using wages as a way to talk about the distribution of human abilities, not the fundamental number

[Yudkowsky][16:32]

I would expect something similar to hold over going from low-paying welder to high-paying welder

[Christiano][16:32]

like, how long to move from "OK human" to "pretty good human" to "best human"

[Cotra][16:32]

says salary of \$350k/yr for lee: <https://www.fameranker.com/lee-sedol-net-worth>

[Yudkowsky][16:32]

but I also mostly expect that AIs will not be allowed to weld things on Earth

[Cotra][16:32]

why don't we just do an in vitro benchmark instead of wages?

[Christiano][16:32]

what, machines already do virtually all welding?

[Cotra][16:32]

just pick a benchmark?

[Yudkowsky][16:33]

yoouuuu do not want to believe sites like that (fameranker)

[Christiano][16:33]

yeah, I'm happy with any benchmark, and then we can measure various human levels at that benchmark

[Cotra][16:33]

what about MATH? <https://arxiv.org/abs/2103.03874>

[Christiano][16:34]

also I don't know what "shorter and shorter" means, the time in go and chess was decades to move from "strong amateur" to "best human," I do think these things will most likely be shorter than decades seems like we can just predict concrete #s though

[Cotra: ]

like I can say how long I think it will take to get from "median high schooler" to "IMO medalist" and you can bet against me?

and if we just agree about all of those predictions then again I'm back to being very skeptical of a claimed track record difference between our models

(I do think that it's going to take years rather than decades on all of these things)

[Yudkowsky][16:36]

possibly! I worry this ends up in a case where Katja or Luke or somebody goes back and collects data about "amateur to pro performance times" and Eliezer says "Ah yes, these are shortening over time, as I predicted" and Paul is like "oh, well, I predict they continue to shorten on this trend drawn from the data" and Eliezer is like "I guess that could happen for the next 5 years, sure, sounds like something superforecaster would predict as default"

[Cotra][16:37]

i'm pretty sure paul's methodology here will just be to look at the MATH perf trend based on model size and combine with expectations of when ppl will make big enough models, not some meta trend thing like that?

[Yudkowsky][16:37]

so I feel like... a bunch of what I feel is the real disagreement in our models, is a bunch of messy stuff. Suddenly Popping Up one day and then Eliezer is like "gosh, I sure didn't predict that" and Paul is like "somebody could have totally predicted that" and Eliezer is like "people would say exactly the same thing after the world ended in 3 minutes"

if we've already got 2 years of trend on a dataset, I'm not necessarily going to predict the trend breaking

[Cotra][16:38]

hm, you're presenting your view as more uncertain and open to anything here than paul's view, but in fact it's picking out a narrower distribution. you're more confident in powerful AGI soon

[Christiano][16:38]

seems hard to play the "who is more confident?" game

[Cotra][16:38]

so there should be some places where you make a strong positive prediction paul disagrees with

[Yudkowsky][16:39]

I might want to buy options on a portfolio of trends like that, if Paul is willing to sell me insurance against all of the trends breaking upward at a lower price than I think is reasonable

I mean, from my perspective Paul is the one who seems to think the world is well-organized and predictable in certain ways

[Christiano][16:39]

yeah, and you are saying that I'm overconfident about that

[Yudkowsky][16:39]

I keep wanting Paul to go on and make narrower predictions than I do in that case

[Christiano][16:39]

so you should be happy to bet with me about *anything*

and I'm letting you pick anything at all you want to bet about

[Cotra][16:40]

i mean we could do a portfolio of trends like MATH and you could bet on at least a few of them having strong surprises in the sooner direction

but that means we could just bet about MATH and it'd just be higher variance

[Yudkowsky][16:40]

ok but you're not going to sell me cheap options on sharp declines in the S&P 500 even though in a very reasonable world there would not be any sharp declines like that

[Christiano][16:41]

if we're betting \$ rather than bayes points, then yes I'm going to weigh worlds based on the value of in those worlds

[Cotra][16:41]

wouldn't paul just sell you options at the price the options actually trade for? i don't get it

[Christiano][16:41]

but my sense is that I'm just generally across the board going to be more right than you are, and I'm frustrated that you just keep saying that "people like me" are wrong about stuff

[Yudkowsky][16:41]

Paul's like "we'll see smooth behavior in the end days" and I feel like I should be able to say "then Paul sell me cheap options against smooth behavior now" but Paul is just gonna wanna sell at market price

[Christiano][16:41]

and so I want to hold you to that by betting about anything

ideally just tons of stuff

random things about what AI will be like, and other technologies, and regulatory changes

[Cotra][16:42]

paul's view doesn't seem to imply that he should value those options less than the market

he's more EMH-y than you not less

[Yudkowsky][16:42]

but then the future should *behave like that market*

[Christiano][16:42]

what do you mean?

[Yudkowsky][16:42]

it should have options on wild behavior that are not cheap!

[Christiano][16:42]

you mean because people want \$ more in worlds where the market drops a lot?

I don't understand the analogy

[Yudkowsky][16:43]

no, because jumpy stuff happens more than it would in a world of ideal agents

[Cotra][16:43]

I think EY is saying the non-cheap option prices are because P(sharp declines) is pretty high

[Christiano][16:43]

ok, we know how often markets jump, if that's the point of your argument can we just talk about that directly?

[Yudkowsky][16:43]

or sharp rises, for that matter

[Christiano][16:43]

(much lower than option prices obviously)

I'm probably happy to sell you options for sharp rises

I'll give you better than market odds in that direction

that's how this works

[Yudkowsky][16:44]

now I am again confused, for I thought you were the one who expected world GDP to double in 4 years at some point

and indeed, drew such graphs with the rise suggestively happening earlier than the sharp spike

[Christiano][16:44]

yeah, and I have exposure to that by buying stocks, options prices are just a terrible way of tracking these things

[Yudkowsky][16:44]

suggesting that such a viewpoint is generally favorable to near timelines for that

[Christiano][16:44]

I mean, I have bet a *lot* of money on AI companies doing well

well, not compared to the EA crowd, but compared to my meager net worth 😊

and indeed, it has been true so far

and I'm continuing to make the bet

it seems like on your view it should be surprising that AI companies just keep going up

aren't you predicting them not to get to tens of trillions of valuation before the end of days?

[Yudkowsky][16:45]

I believe that Nate, of a generally Yudkowskian view, did the same (bought AI companies). and I focused my thoughts elsewhere, because somebody needs to, but did happen to buy my first S&P 500 on its exact minimum in 2020

[Christiano][16:46]

point is, that's how you get exposure to the crazy growth stuff with continuous ramp-ups

and I'm happy to make the bet on the market

or on other claims

I don't know if my general vibe makes sense here, and why it seems reasonable to me that I'm just happy to bet on anything

as a way of trying to defend my overall attack

and that if my overall epistemic approach is vulnerable to some track record objection, then it seems like it ought to be possible to win here

9.10. Prediction disagreements and bets: Standard superforecaster techniques

[Cotra][16:47]

I'm still kind of surprised that Eliezer isn't willing to bet that there will be a faster-than-Paul expects trend break on MATH or whatever other benchmark. Is it just the variance of MATH being one benchmark? Would you make the bet if it were 6?

[Yudkowsky][16:47]

a large problem here is that both of us tend to default strongly to superforecaster standard technique

[Christiano][16:47]

it's true, though it's less true for longer things

[Cotra][16:47]

but you think the superforecasters would suck at predicting end days because of the surface trends thing!

[Yudkowsky][16:47]

before I bet against Paul on MATH I would want to know that Paul wasn't arriving at the same default use, which might be drawn from trend lines there, or from a trend line in trend lines

I mean the superforecasters did already suck once in my observation, which was AlphaGo, but I did not bet against them there, I bet with them and then updated afterwards

[Christiano][16:48]

I'd mostly try to eyeball how fast performance was improving with size; I'd think about difficulty effect (where e.g. hard problems will be flat for a while and then go up later, so you want to measure performance on a spectrum of difficulties)

[Cotra][16:48]

what if you bet against a methodology instead of against paul's view? the methodology being the one described above, of looking at the perf based on model size and then projecting model size increases cost?

[Christiano][16:48]

seems safer to bet against my view

[Cotra][16:48]

yeah

[Christiano][16:48]

mostly I'd just be eyeballing size, thinking about how much people will in fact scale up (which would great to factor out if possible), assuming performance trends hold up

are there any other examples of surface trends vs predictable deep changes, or is AGI the only one?
(that you have thought a lot about)

[Cotra][16:49]

yeah seems even better to bet on the underlying "will the model size to perf trends hold up or break upward"

[Yudkowsky][16:49]

so from my perspective, there's this whole thing where *unpredictably* something breaks above trend because the first way it got done was a way where somebody could do it faster than you expected

[Christiano][16:49]

(makes sense for it to be the domain where you've thought a lot)

you mean, it's unpredictable what will break above trend?

[Cotra][16:49]

[IEM](#) has a financial example

[Yudkowsky][16:49]

I mean that I could not have said "Go will break above trend" in 2015

[Christiano][16:49]

yeah

ok, here's another example

[Yudkowsky][16:50]

it feels like if I want to make a bet with imaginary Paul in 2015 then I have to bet on a portfolio
and I also feel like as soon as we make it that concrete, Paul does not want to offer me things that I v
to bet on

because Paul is also like, sure, something might break upward

I remark that I have for a long time been saying that I wish Paul had more concrete images and
examples attached to *a lot of his stuff*

[Cotra][16:51]

surely the view is about the probability of each thing breaking upward. or the expected number from basket

[Christiano][16:51]

I mean, if you give me any way of quantifying how much stuff breaks upwards we have a bet

[Cotra][16:51]

not literally that one single thing breaks upward

[Christiano][16:51]

I don't understand how concreteness is an accusation here, I've offered 10 quantities I'd be happy to about, and also allowed you to name literally any other quantity you want

and I agree that we mostly agree about things

[Yudkowsky][16:52]

and some of my sense here is that if Paul offered a portfolio bet of this kind, I might not take it myself but EAs who were better at noticing their own surprise might say, "Wait, that's how unpredictable Paul thinks the world is?"

so from my perspective, it is hard to know specific anti-superforecaster predictions that happen long before terminal phase, and I am not sure we are really going to get very far there.

[Christiano][16:53]

but you agree that the eventual prediction is anti-superforecaster?

[Yudkowsky][16:53]

both of us probably have quite high inhibitions against selling conventionally priced options that are not what a superforecaster would price them as

[Cotra][16:53]

why does it become so much easier to know these things and go anti-superforecaster at terminal phase?

[Christiano][16:53]

I assume you think that the superforecasters will continue to predict that big impactful AI applications are made by large firms spending a lot of money, even through the end of days

I do think it's very often easy to beat superforecasters in-domain

like I expect to personally beat them at most ML prediction

and so am also happy to do bets where you defer to superforecasters on arbitrary questions and I be against you

[Yudkowsky][16:54]

well, they're anti-prediction-market in the sense that, at the very end, bets can no longer settle. I've been surprised of late by how much AGI ruin seems to be sneaking into common knowledge; perhaps the terminal phase the superforecasters will be like, "yep, we're dead". I can't even say that in this case Paul will disagree with them, because I expect the state on alignment to be so absolutely awful that even Paul is like "You were not supposed to do it that way" in a very sad voice.

[Christiano][16:55]

I'm just thinking about takeoff speeds here

I do think it's fairly likely I'm going to be like "oh no this is bad" (maybe 50%?), but not that I'm going to expect fast takeoff

and similarly for the superforecasters

9.11. Prediction disagreements and bets: Late-stage predictions, and betting against superforecasters

[Yudkowsky][16:55]

so, one specific prediction you made, sadly close to terminal phase but not much of a surprise there, that the world economy must double in 4 years before the End Times are permitted to begin

[Christiano][16:56]

well, before it doubles in 1 year...

I think most people would call the 4 year doubling the end times

[Yudkowsky][16:56]

this seems like you should also be able to point to some least impressive thing that is not permitted to occur before WGDP has doubled in 4 years

[Christiano][16:56]

and it means that the normal planning horizon includes the singularity

[Yudkowsky][16:56]

it may not be much but we would be *moving back* the date of first concrete disagreement

[Christiano][16:57]

I can list things I don't think would happen first, since that's a ton

[Yudkowsky][16:57]

and EAs might have a little bit of time in which to say "Paul was falsified, uh oh"

[Christiano][16:57]

the only things that aren't permitted are the ones that would have caused the world economy to dou in 4 years

[Yudkowsky][16:58]

and by the same token, there are things Eliezer thinks you are probably not going to be able to do before you slide over the edge. a portfolio of these will have some losing options because of adverse selection against my errors of what is hard, but if I lose more than half the portfolio, this may said to a bad sign for Eliezer.

[Christiano][16:58]

(though those can happen at the beginning of the 4 year doubling)

[Yudkowsky][16:58]

this is unfortunately *late* for falsifying our theories but it would be *progress* on a kind of bet against e other

[Christiano][16:59]

but I feel like the things I'll say are like fully automated construction of fully automated factories at 1-year turnarounds, and you're going to be like "well duh"

[Yudkowsky][16:59]

...unfortunately yes

[Christiano][16:59]

the reason I like betting about numbers is that we'll probably just disagree on any given number

[Yudkowsky][16:59]

I don't think I *know* numbers.

[Christiano][16:59]

it does seem like a drawback that this can just turn up object-level differences in knowledge-of-numb more than deep methodological advantages

[Yudkowsky][17:00]

the last important number I had a vague suspicion I might know was that Ethereum ought to have a significantly larger market cap in pre-Singularity equilibrium.

and I'm not as sure of that one since El Salvador supposedly managed to use Bitcoin L2 Lightning.

(though I did not fail to act on the former belief)

[Christiano][17:01]

do you see why I find it weird that you think there is this deep end-times truth about AGI, that is very different from a surface-level abstraction and that will take people like Paul by surprise, without think there are other facts like that about the world?

I do see how this annoying situation can come about
and I also understand the symmetry of the situation

[Yudkowsky][17:02]

we unfortunately both have the belief that the present world looks a lot like our being right, and therefore that the other person ought to be willing to bet against default superforecasterish projectio

[Cotra][17:02]

paul says that *he* would bet against superforecasters too though

[Christiano][17:02]

I would in ML

[Yudkowsky][17:02]

like, where specifically?

[Christiano][17:02]

or on any other topic where I can talk with EAs who know about the domain in question

I don't know if they have standing forecasts on things, but e.g.: (i) benchmark performance, (ii) indus size in the future, (iii) how large an LM people will train, (iv) economic impact of any given ML system like codex, (v) when robotics tasks will be plausible

[Yudkowsky][17:03]

I have decided that, as much as it might gain me prestige, I don't think it's actually the right thing for me to go spend a bunch of character points on the skills to defeat superforecasters in specific domains and then go around doing that to prove my epistemic virtue.

[Christiano][17:03]

that seems fair

[Yudkowsky][17:03]

you don't need to bet with *me* to prove your epistemic virtue in this way, though

okay, but, if I'm allowed to go around asking Carl Shulman who to ask in order to get the economic impact of Codex, maybe I can also defeat superforecasters.

[Christiano][17:04]

I think the deeper disagreement is that (i) I feel like my end-of-days prediction is also basically just a default superforecaster prediction (and if you think yours is too then we can bet about what some superforecasters will say on it), (ii) I think you are leveling a much stronger "people like paul get taken by surprise by reality" claim whereas I'm just saying that I don't like your arguments

[Yudkowsky][17:04]

it seems to me like the contest should be more like our intuitions in advance of doing that

[Christiano][17:04]

yeah, I think that's fine, and also cheaper since research takes so much time
I feel like those asymmetries are pretty strong though

9.12. Self-duplicating factories, AI spending, and Turing test variants

[Yudkowsky][17:05]

so, here's an idea that is less epistemically virtuous than our making Nicely Resolvable Bets
what if we, like, talked a bunch about our off-the-cuff senses of where various AI things are going in t
next 3 years
and then 3 years later, somebody actually reviewed that

[Christiano][17:06]

I do think just saying a bunch of stuff about what we expect will happen so that we can look back on
would have a significant amount of the value

[Yudkowsky][17:06]

and any time the other person put a thumbs-up on the other's prediction, that prediction coming true
was not taken to distinguish them

[Cotra][17:06]

i'd suggest doing this in a format other than discord for posterity

[Yudkowsky][17:06]

even if the originator was like HOW IS THAT ALSO A PREDICTION OF YOUR THEORY
well, Discord has worked better than some formats

[Cotra][17:07]

something like a spreadsheet seems easier for people to look back on and score and stuff
discord transcripts are pretty annoying to read

[Yudkowsky][17:08]

something like a spreadsheet seems liable to be high-cost and not actually happen

[Christiano][17:08]

I think a conversation is probably easier and about as good for our purposes though?

[Cotra][17:08]

ok fair

[Yudkowsky][17:08]

I think money can be inserted into humans in order to turn Discord into spreadsheets

[Christiano][17:08]

and it's possible we will both think we are right in retrospect

and that will also be revealing

[Yudkowsky][17:09]

but, besides that, I do want to boop on the point that I feel like Paul should be able to predict intuitively rather than with necessity, things that should not happen before the world economy doubled in 4 years

[Christiano][17:09]

it may also turn up some quantitative differences of view

there are lots of things I think won't happen before the world economy has doubled in 4 years

[Yudkowsky][17:09]

because on my model, as we approach the end times, AI was still pretty partial and also the world economy was lolling most of the inputs a sensible person would accept from it and prototypes weren't being commercialized and stuff was generally slow and messy

[Christiano][17:09]

prototypes of factories building factories in <2 years

[Yudkowsky][17:10]

"AI was still pretty partial" leads it to not do interesting stuff that Paul can rule out

[Christiano][17:10]

like I guess I think Tesla will try, and I doubt it will be just Tesla

[Yudkowsky][17:10]

but the other parts of that permit AI to do interesting stuff that Paul can rule out

[Christiano][17:10]

automated researchers who can do ML experiments from 2020 without human input

[Yudkowsky][17:10]

okay, see, that whole "factories building factories" thing just seems so very much *after* the End Time
me

[Christiano][17:10]

yeah, we should probably only talk about cognitive work
since you think physical work will be very slow

[Yudkowsky][17:11]

okay but not just that, it's a falsifiable prediction
it is something that lets Eliezer be wrong in advance of the End Times

[Christiano][17:11]

what's a falsifiable prediction?

[Yudkowsky][17:11]

if we're in a world where Tesla is excitingly gearing up to build a fully self-duplicating factory including
its mining inputs and chips and solar panels and so on, we're clearly in the Paulverse and not in the
Eliezerverse!

[Christiano][17:12]

yeah
I do think we'll see that before the end times
just not before 4 year doublings

[Yudkowsky][17:12]

this unfortunately only allows you to be right, and not for me to be right, but I think there are also thi
you legit only see in the Eliezerverse!

[Christiano][17:12]

I mean, I don't think they will be doing mining for a long time because it's cheap

[Yudkowsky][17:12]

they are unfortunately late in the game but they exist at all!
and being able to state them is progress on this project!

[Christiano][17:13]

but fully-automated factories first, and then significant automation of the factory-building process
I do expect to see
I'm generally pretty bullish on industrial robotics relative to you I think, even before the crazy stuff?
but you might not have a firm view
like I expect to have tons of robots doing all kinds of stuff, maybe cutting human work in manufacturing 2x, with very modest increases in GDP resulting from that in particular

[Yudkowsky][17:13]

so, like, it doesn't surprise me very much if Tesla manages to fully automate a factory that takes in some relatively processed inputs including refined metals and computer chips, and outputs a car? and by the same token I expect that has very little impact on GDP.

[Christiano][17:14]

refined metals are almost none of the cost of the factory
and also tesla isn't going to be that vertically integrated
the fabs will separately continue to be more and more automated
I expect to have robot cars driving everywhere, and robot trucks
another 2x fall in humans required for warehouses
elimination of most brokers involved in negotiating shipping

[Yudkowsky][17:15]

if despite the fabs being more and more automated, somehow things are managing not to cost less and less, and that sector of the economy is not really growing very much, is that more like the Eliezerverse than the Paulverse?

[Christiano][17:15]

most work in finance and loan origination

[Yudkowsky][17:15]

though this is something of a peripheral prediction to AGI core issues

[Christiano][17:16]

yeah, I think if you cut the humans to do X by 2, but then the cost falls much less than the number you'd naively expect (from saving on the human labor and paying for the extra capital), then that's surprising to me

I mean if it falls half as much as you'd expect on paper I'm like "that's a bit surprising" rather than having my mind blown, if it doesn't fall I'm more surprised

but that was mostly physical economy stuff

oh wait, I was making positive predictions now, physical stuff is good for that I think?
since you don't expect it to happen?

[Yudkowsky][17:17]

...this is not your fault but I wish you'd asked me to produce my "percentage of fall vs. paper calculation" estimate before you produced yours

my mind is very whiffy about these things and I am not actually unable to deanchor on your estimate

[Christiano][17:17]

makes sense, I wonder if I should just spoiler
one benefit of discord

[Yudkowsky][17:18]

yeah that works too!

[Christiano][17:18]

a problem for prediction is that I share some background view about insane
inefficiency/inadequacy/decadence/silliness
so these predictions are all tampered by that
but still seem like there are big residual disagreements

[Yudkowsky][17:19]

sighgreat

[Christiano][17:19]

since you have way more of that than I do

[Yudkowsky][17:19]

not your fault but

[Christiano][17:19]

I think that the AGI stuff is going to be a gigantic megaproject despite that

[Yudkowsky][17:19]

I am not shocked by the AGI stuff being a gigantic megaproject
it's not above the bar of survival but, given other social optimism, it permits death with more dignity
than by other routes

[Christiano][17:20]

what if spending is this big:

Google invests \$100B training a model, total spending across all of industry is way bigger

[Yudkowsky][17:20]

ooooh

I do start to be surprised if, come the end of the world, AGI is having more invested in it than a TSMC though, not... *super* surprised?

also I am at least a little surprised before then

actually I should probably have been spoiling those statements myself but my expectation is that Paul's secret spoiler is about

\$10 trillion dollars or something equally totally shocking to an Eliezer

[Christiano][17:22]

my view on that level of spending is

it's an only slightly high-end estimate for spending by someone on a single model, but that in practice there will be ways of dividing more across different firms, and that the ontology of single-model will likely be slightly messed up (e.g. by OpenAI Five-style surgery). Also if it's that much then it likely involves big institutional changes and isn't at Google.

I read your spoiler

my estimate for total spending for the whole project of making TAI, including hardware and software manufacturing and R&D, the big datacenters, etc.

is in the ballpark of \$10T, though it's possible that it will be undercounted several times due to wage stickiness for high-end labor

[Yudkowsky][17:24]

I think that as

spending on particular AGI megaprojects starts to go past \$50 billion, it's not especially ruled out per se by things that I think I know for sure, but I feel like a third-party observer should justly start to weakly think, 'okay, this is looking at least a little like the Paulverse rather than the Eliezerverse', and as we get closer to \$10 trillion, that is not absolutely ruled out by the Eliezerverse but it was a whole lot more strongly predicted by the Paulverse, maybe something like 20x unless I'm overestimating how strongly Paul predicts that

[Christiano][17:24]

Proposed modification to the "speculate about the future to generate kind-of-predictions" methodology: we make shit up, then later revise based on points others made, and maybe also get Carl to sanity-check and decide which of his objections we agree with. Then we can separate out the "how good are our intuitions" claim (with fast feedback) from the all-things-considered how good was the "prediction"

[Yudkowsky][17:25]

okay that hopefully allows me to read Paul's spoilers... no I'm being silly. @ajeya please read all the spoilers and say if it's time for me to read his

[Cotra][17:25]

you can read his latest

[Christiano][17:25]

I'd guess it's fine to read all of them?

[Cotra][17:26]

yeah sorry that's what i meant

[Yudkowsky][17:26]

what should I say more about before reading earlier ones?

ah k

[Christiano][17:26]

My \$10T estimate was after reading yours (didn't offer an estimate on that quantity beforehand), tho that's the kind of ballpark I often think about, maybe we should just spoiler only numbers so that context is clear 😊

I think fast takeoff gets significantly more likely as you push that number down

[Yudkowsky][17:27]

so, may I now ask what starts to look to you like "oh damn I am in the Eliezerverse"?

[Christiano][17:28]

big mismatches between that AI looks technically able to do and what AI is able to do, though that's going to need a lot of work to operationalize

I think low growth of AI overall feels like significant evidence for Eliezerverse (even if you wouldn't m that prediction), since I'm forecasting it rising to absurd levels quite fast whereas your model is consistent with it staying small

some intuition about AI looking very smart but not able to do much useful until it has the whole pictu I guess this can be combined with the first point to be something like---AI looks really smart but it's ju not adding much value

all of those seem really hard

[Cotra][17:30]

strong upward trend breaks on benchmarks seems like it should be a point toward eliezer verse, even eliezer doesn't want to bet on a specific one?

especially breaks on model size -> perf trends rather than calendar time trends

[Christiano][17:30]

I think that any big break on model size -> perf trends are significant evidence

[Cotra][17:31]

meta-learning working with small models?

e.g. model learning-to-learn video games and then learning a novel one in a couple subjective hours

[Christiano][17:31]

I think algorithmic/architectural changes that improve loss as much as 10x'ing model, for tasks that looking like they at least *should* have lots of economic value

(even if they don't end up having lots of value because of deployment bottlenecks)

is the meta-learning thing an Eliezer prediction?

(before the end-of-days)

[Cotra][17:32]

no but it'd be an anti-bio-anchor positive trend break and eliezer thinks those should happen more than we do

[Christiano][17:32]

fair enough

a lot of these things are about # of times that it happens rather than whether it happens at all

[Cotra][17:32]

yeah

but meta-learning is special as the most plausible long horizon task

[Christiano][17:33]

e.g. maybe in any given important task I expect a single "innovation" that's worth 10x model size? but that it still represents a minority of total time?

hm, AI that can pass a competently administered turing test without being economically valuable?

that's one of the things I think is ruled out before 4 year doubling, though Eliezer probably also doesn't expect it

[Yudkowsky: ]

[Cotra][17:34]

what would this test do to be competently administered? like casual chatbots seem like they have reasonable probability of fooling someone for a few mins now

[Christiano][17:34]

I think giant google-automating-google projects without big external economic impacts

[Cotra][17:34]

would it test knowledge, or just coherence of some kind?

[Christiano][17:35]

it's like a smart-ish human (say +2 stdev at this task) trying to separate out AI from smart-ish human iterating a few times to learn about what works

I mean, the basic ante is that the humans are *trying* to win a turing test, without that I wouldn't even call it a turing test

dunno if any of those are compelling @Eliezer

something that passes a like "are you smart?" test administered by a human for 1h, where they aren't trying to specifically tell if you are AI

just to see if you are as smart as a human

I mean, I guess the biggest giveaway of all would be if there is human-level (on average) AI as judged by us, but there's no foom yet

[Yudkowsky][17:37]

I think we both don't expect that one before the End of Days?

[Christiano][17:37]

or like, no crazy economic impact

I think we both expect that to happen before foom?

but the "on average" is maybe way too rough a thing to define

[Yudkowsky][17:37]

oh, wait, I missed that it wasn't the full Turing Test

[Christiano][17:37]

well, I suggested both

the lamer one is more plausible

[Yudkowsky][17:38]

full Turing Test happeneth not before the End Times, on Eliezer's view, and not before the first 4-year doubling time, on Paul's view, and the first 4-year doubling happeneth not before the End Times, on Eliezer's view, so this one doesn't seem very useful

9.13. GPT-n and small architectural innovations vs. large ones

[Christiano][17:39]

I feel like the biggest subjective thing is that I don't feel like there is a "core of generality" that GPT-3 missing

I just expect it to gracefully glide up to a human-level foaming intelligence

[Yudkowsky][17:39]

the "are you smart?" test seems perhaps passable by GPT-6 or its kin, which I predict to contain at least one major architectural difference over GPT-3 that I could, pre-facto if anyone asked, rate as larger than a different normalization method

but by fooling the humans more than by being smart

[Christiano][17:39]

like I expect GPT-5 would fail if you ask it but take a long time

[Yudkowsky][17:39]

that sure is an underlying difference

[Christiano][17:39]

not sure how to articulate what Eliezer expects to see here though

or like what the difference is

[Cotra][17:39]

something that GPT-5 or 4 shouldn't be able to do, according to eliezer?

where Paul is like "sure it could do that"?

[Christiano][17:40]

I feel like GPT-3 clearly has some kind of "doesn't really get what's going on" energy

and I expect that to go away

well before the end of days

so that it seems like a kind-of-dumb person

[Yudkowsky][17:40]

I expect it to go away before the end of days

but with there having been a big architectural innovation, not Stack More Layers

[Christiano][17:40]

yeah

whereas I expect layer stacking + maybe changing loss (since logprob is too noisy) is sufficient

[Yudkowsky][17:40]

if you name 5 possible architectural innovations I can call them small or large

[Christiano][17:41]

1. replacing transformer attention with DB nearest-neighbor lookup over an even longer context

[Yudkowsky][17:42]

okay 1's a bit borderline

[Christiano][17:42]

2. adding layers that solve optimization problems internally (i.e. the weights and layer N activations define an optimization problem, the layer N+1 solves it) or maybe simulates an ODE

[Yudkowsky][17:42]

if it's 3x longer context, no biggie, if it's 100x longer context, more of a game-changer

2 - big change

[Christiano][17:42]

I'm imagining >100x if you do that

3. universal transformer XL, where you reuse activations from one context in the next context (RNN style) and share weights across layers

[Yudkowsky][17:43]

I do not predict 1 works because it doesn't seem like an architectural change that moves away from what I imagined to be the limits, but it's a big change if it 100xs the window

3 - if it is only that single change and no others, I call it not a large change relative to transformer XL Transformer XL itself however was an example of a large change - it didn't have a large effect but it v what I'd call a large change.

[Christiano][17:45]

4. Internal stochastic actions trained with reinforce

I mean, is mixture of experts or switch another big change?

are we just having big changes non-stop?

[Yudkowsky][17:45]

4 - I don't know if I'm imagining right but it sounds large

[Christiano][17:45]

it sounds from these definitions like the current rate of big changes is > 1/year

[Yudkowsky][17:46]

5 - mixture of experts: as with 1, I'm tempted to call it a small change, but that's because of my mod of it as doing the same thing, not because it isn't in a certain sense a quite large move away from St: More Layers

I mean, it is not very hard to find a big change to try?

finding a big change that works is much harder

[Christiano][17:46]

several of these are improvements

[Yudkowsky][17:47]

one gets a minor improvement from a big change rather more often than a big improvement from a l change

that's why dinosaurs didn't foom

[Christiano][17:47]

like transformer -> MoE -> switch transformer is about as big an improvement as LSTM vs transforme

so if we all agree that big changes are happening multiple times per year, then I guess that's not the difference in prediction

is it about the size of gains from individual changes or something?

or maybe: if you take the scaling laws for transformers, are the models with impact X "on trend," with changes just keeping up or maybe buying you 1-2 oom of compute, or are they radically better / scal much better?

that actually feels most fundamental

[Yudkowsky][17:49]

I had not heard that transformer -> switch transformer was as large an improvement as lstm -> transformers after a year or two, though maybe you're referring to a claimed 3x improvement and comparing that to the claim that if you optimize LSTMs as hard as transformers they come within 3x have not examined these claims in detail, they sound a bit against my prior, and I am a bit skeptical both of them)

so remember that from my perspective, I am fighting an adverse selection process and the Law of Earlier Success

[Christiano][17:50]

I think it's actually somewhat smaller

[Yudkowsky][17:51]

if you treat GPT-3 as a fixed thingy and imagine scaling it in the most straightforward possible way, tl I have a model of what's going on in there and I don't think that most direct possible way of scaling g you past GPT-3 lacking a deep core

somebody can come up and go, "well, what about this change that nobody tried yet?" and I can be li "ehhh, that particular change does not get at what I suspect the issues are"

[Christiano][17:52]

I feel like the framing is: paul says that something is possible with "stack more layers" and eliezer isr We both agree that you can't literally stack more layers and have to sometimes make tweaks, and al that you will scale faster if you make big changes. But it seems like for Paul that means (i) changes to stay on the old trend line, (ii) changes that trade off against modest amounts of compute

so maybe we can talk about that?

[Yudkowsky][17:52]

when it comes to predicting what happens in 2 years, I'm not just up against people trying a broad range of changes that I can't foresee in detail, I'm also up against a Goodhart's Curse on the answer being a weird trick that worked better than I would've expected in advance

[Christiano][17:52]

but then it seems like we may just not know, e.g. if we were talking LSTM vs transformer, no one is going to run experiments with the well-tuned LSTM because it's still just worse than a transformer (though they've run enough experiments to know how important tuning is, and the brittleness is much of why one likes it)

[Yudkowsky][17:53]

I would not have predicted Transformers to be a huge deal if somebody described them to me in advance of having ever tried it out. I think that's because predicting the future is hard not because I'm especially stupid.

[Christiano][17:53]

I don't feel like anyone could predict that being a big deal

but I do think you could predict "there will be some changes that improve stability / make models slightly better"

(I mean, I don't feel like any of the actual humans on earth could have, some hypothetical person co

[Yudkowsky][17:57]

whereas what I'm trying to predict is more like "GPT-5 in order to start-to-awaken needs a change via which it, in some sense, can do a different thing, that is more different than the jump from GPT-1 to GPT-3; and examples of things with new components in them abound in Deepmind, like Alpha Zero having not the same architecture as the original AlphaGo; but at the same time I'm also trying to account for being up against this very adversarial setup where a weird trick that works much better than I expected may be the thing that makes GPT-5 able to do a different thing"

this may seem Paul-unfairish because any random innovations that come along, including big changes that cause small improvements, would tend to be swept up into GPT-5 even if they made no more difference than the whole thing with MoE

so it's hard to bet on

but I also don't feel like it - totally lacks Eliezer-vs-Paul-ness if you let yourself sort of relax about that and just looked at it?

also I'm kind of running out of energy, sorry

[Christiano][18:03]

I think we should be able to get something here eventually

seems good to break though

that was a lot of arguing for one day

Biology-Inspired AGI Timelines: The Trick That Never Works

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

- 1988 -

Hans Moravec: Behold my book *Mind Children*. Within, I project that, in 2010 or thereabouts, we shall achieve strong AI. I am not calling it "Artificial General Intelligence" because this term will not be coined for another 15 years or so.

Eliezer (who is not actually on the record as saying this, because the real Eliezer is, in this scenario, 8 years old; this version of Eliezer has all the meta-heuristics of Eliezer from 2021, but none of that Eliezer's anachronistic knowledge): Really? That sounds like a very difficult prediction to make correctly, since it is about the future, which is famously hard to predict.

Imaginary Moravec: Sounds like a [fully general counterargument](#) to me.

Eliezer: Well, it is, indeed, a fully general counterargument *against futurism*. Successfully predicting the unimaginably far future - that is, more than 2 or 3 years out, or sometimes less - is something that human beings seem to be quite bad at, by and large.

Moravec: I predict that, 4 years from this day, in 1992, the Sun will rise in the east.

Eliezer: Okay, let me qualify that. Humans seem to be quite bad at predicting the future whenever we need to predict anything at all *new and unfamiliar*, rather than the Sun continuing to rise every morning until it finally gets eaten. I'm not saying it's impossible to ever validly predict something novel! Why, even if that was impossible, how could I know it for sure? By extrapolating from my own personal inability to make predictions like that? Maybe I'm just bad at it myself. But any time somebody claims that some particular novel aspect of the far future is predictable, they justly have a significant burden of prior skepticism to overcome.

More broadly, we should not expect a good futurist to give us a generally good picture of the future. We should expect a great futurist to single out a few *rare narrow aspects* of the future which are, somehow, *exceptions* to the usual rule about the future not being very predictable.

I do agree with you, for example, that we shall *at some point* see Artificial General Intelligence. This seems like a rare predictable fact about the future, even though it is about a novel thing which has not happened before: we keep trying to crack this problem, we make progress albeit slowly, the problem must be solvable in principle because human brains solve it, eventually it will be solved; this is not a logical necessity, but it sure seems like the way to bet. "AGI eventually" is predictable in a way that it is *not* predictable that, e.g., the nation of Japan, presently upon the rise, will achieve economic dominance over the next decades - to name something else that present-day storytellers of 1988 are talking about.

But *timing* the novel development correctly? *That* is almost never done, not until things are 2 years out, and often not even then. Nuclear weapons were called, but not nuclear weapons in 1945; heavier-than-air flight was called, but not flight in 1903. In both cases, people said two years earlier that it wouldn't be done for 50 years - or said, decades too early, that it'd be done shortly. There's a difference between worrying that we may eventually get a serious global pandemic, worrying that eventually a lab accident may lead to a global pandemic, and forecasting that a global pandemic will start in November of 2019.

Moravec: You should read my book, my friend, into which I have put much effort. In particular - though it may sound impossible to forecast, to the likes of yourself - I have carefully examined a graph of computing power in single chips and the most powerful supercomputers over time. This graph looks surprisingly regular! Now, of course not all trends can continue forever; but I have considered the arguments that Moore's Law will break down, and found them unconvincing. My book spends several chapters discussing the particular reasons and technologies by which we might expect this graph to *not* break down, and continue, such that humanity *will* have, by 2010 or so, supercomputers which can perform 10 trillion operations per second.*

Oh, and also my book spends a chapter discussing the retina, the part of the brain whose computations we understand in the most detail, in order to estimate how much computing power the human brain is using, arriving at a figure of 10^{13} ops/sec. This neuroscience and computer science may be a bit hard for the layperson to follow, but I assure you that I am in fact an experienced hands-on practitioner in robotics and computer vision.

So, as you can see, we should first get strong AI somewhere around 2010. I may be off by an order of magnitude in one figure or another; but even if I've made two errors in the same direction, that only shifts the estimate by 7 years or so.

(*) Moravec just about nailed this part; the actual year was 2008.

Eliezer: I sure would be amused if we *did* in fact get strong AI somewhere around 2010, which, for all I know at this point in this hypothetical conversation, could totally happen! Reversed stupidity is not intelligence, after all, and just because that is a completely broken justification for predicting 2010 doesn't mean that it cannot happen that way.

Moravec: Really now. Would you care to enlighten me as to how I reasoned so wrongly?

Eliezer: Among the reasons why the Future is so hard to predict, in general, is that the sort of answers we want tend to be the products of lines of causality with multiple steps and multiple inputs. Even when we can guess a single fact that *plays some role* in producing the Future - which is not of itself all that rare - usually the answer the storyteller wants depends on *more facts* than that single fact. Our ignorance of any one of those other facts can be enough to torpedo our whole line of reasoning - *in practice*, not just as a matter of possibilities. You could say that the art of exceptions to Futurism being impossible, consists in finding those rare things that you can predict despite being almost entirely ignorant of most concrete inputs into the concrete scenario. Like predicting that AGI will happen *at some point*, despite not knowing the design for it, or who will make it, or how.

My own contribution to the Moore's Law literature consists of Moore's Law of Mad Science: "Every 18 months, the minimum IQ required to destroy the Earth drops by 1

point." Even if this serious-joke was an absolutely true law, and aliens told us it was absolutely true, we'd still have no ability whatsoever to predict thereby when the Earth would be destroyed, because we'd have no idea what that minimum IQ was right now or at any future time. We would know that in general the Earth had a serious problem that needed to be addressed, because we'd know in general that destroying the Earth kept on getting easier every year; but we would not be able to time when that would become an imminent emergency, until we'd seen enough specifics that the crisis was already upon us.

In the case of your prediction about strong AI in 2010, I might put it as follows: The timing of AGI could be seen as a product of three factors, one of which you can try to extrapolate from existing graphs, and two of which you don't know at all. Ignorance of any one of them is enough to invalidate the whole prediction.

These three factors are:

- The availability of computing power over time, which may be quantified, and appears steady when graphed;
- The rate of progress in knowledge of cognitive science and algorithms over time, which is much harder to quantify;
- A function that is a latent background parameter, for the amount of computing power required to create AGI as a function of any particular level of knowledge about cognition; and about this we know almost nothing.

Or to rephrase: Depending on how much you and your civilization know about AI-making - how much you know about cognition and computer science - it will take you a variable amount of computing power to build an AI. If you really knew what you were doing, for example, I confidently predict that you could build a mind at least as powerful as a human mind, while using fewer floating-point operations per second than a human brain is making useful use of -

Chris Humbali: Wait, did you just say "confidently"? How could you possibly know that with confidence? How can you criticize Moravec for being too confident, and then, in the next second, turn around and be confident of something yourself? Doesn't that make you a massive hypocrite?

Eliezer: Um, who are you again?

Humbali: I'm the cousin of Pat Modesto from [your previous dialogue on Hero Licensing!](#) Pat isn't here in person because "Modesto" looks unfortunately like "Moravec" on a computer screen. And also their first name looks a bit like "Paul" who is not meant to be referenced either. So today I shall be your true standard-bearer for good calibration, intellectual humility, the outside view, and reference class forecasting -

Eliezer: Two of these things are not like the other two, in my opinion; and Humbali and Modesto do not understand how to operate any of the four correctly, in my opinion; but anybody who's read "[Hero Licensing](#)" should already know I believe that.

Humbali: - and I don't see how Eliezer can possibly be so *confident*, after all his humble talk of the difficulty of futurism, that it's possible to build a mind 'as powerful as' a human mind using 'less computing power' than a human brain.

Eliezer: It's overdetermined by multiple lines of inference. We might first note, for example, that the human brain runs very slowly in a *serial* sense and tries to make up

for that with massive parallelism. It's an obvious truth of computer science that while you can use 1000 serial operations per second to emulate 1000 parallel operations per second, the reverse is not in general true.

To put it another way: if you had to build a spreadsheet or a word processor on a computer running at 100Hz, you might also need a billion processing cores and massive parallelism in order to do enough cache lookups to get anything done; that wouldn't mean the computational labor you were performing was *intrinsically* that expensive. Since modern chips are massively serially faster than the neurons in a brain, and the direction of conversion is asymmetrical, we should expect that there are tasks which are immensely expensive to perform in a massively parallel neural setup, which are much cheaper to do with serial processing steps, and the reverse is *not* symmetrically true.

A sufficiently adept builder can build general intelligence more cheaply in total operations per second, if they're allowed to line up a billion operations one after another per second, versus lining up only 100 operations one after another. I don't bother to qualify this with "very probably" or "almost certainly"; it is the sort of proposition that a clear thinker should simply accept as obvious and move on.

Humbali: And is it certain that neurons can perform only 100 serial steps one after another, then? As you say, ignorance about one fact can obviate knowledge of any number of others.

Eliezer: A typical neuron firing as fast as possible can do maybe 200 spikes per second, a few rare neuron types used by eg bats to echolocate can do 1000 spikes per second, and the vast majority of neurons are not firing that fast at any given time. The usual and proverbial rule in neuroscience - the sort of academically respectable belief I'd expect you to respect even more than I do - is called "the 100-step rule", that any task a human brain (or mammalian brain) can do on perceptual timescales, must be doable with no more than 100 *serial* steps of computation - no more than 100 things that get computed one after another. Or even less if the computation is running off spiking frequencies instead of individual spikes.

Moravec: Yes, considerations like that are part of why I'd defend my estimate of 10^{13} ops/sec for a human brain as being reasonable - more reasonable than somebody might think if they were, say, counting all the synapses and multiplying by the maximum number of spikes per second in any neuron. If you actually look at what the retina is doing, and how it's computing that, it doesn't look like it's doing one floating-point operation per activation spike per synapse.

Eliezer: There's a similar asymmetry between precise computational operations having a vastly easier time emulating noisy or imprecise computational operations, compared to the reverse - there is no doubt a way to use neurons to compute, say, exact 16-bit integer addition, which is at least *more* efficient than a human trying to add up 16986+11398 in their heads, but you'd still need more synapses to do that than transistors, because the synapses are noisier and the transistors can just do it precisely. This is harder to visualize and get a grasp on than the parallel-serial difference, but that doesn't make it unimportant.

Which brings me to the second line of very obvious-seeming reasoning that converges upon the same conclusion - that it is in principle possible to build an AGI much more computationally efficient than a human brain - namely that biology is simply *not* that

efficient, and *especially* when it comes to huge complicated things that it has started doing relatively recently.

ATP synthase may be close to 100% thermodynamically efficient, but ATP synthase is literally over 1.5 billion years old and a core bottleneck on all biological metabolism.

Brains have to pump thousands of ions in and out of each stretch of axon and dendrite, in order to restore their ability to fire another fast neural spike. The result is that the brain's computation is something like half a million times less efficient than the thermodynamic limit for its temperature - so around two millionths as efficient as ATP synthase. And neurons are a hell of a lot older than the biological software for general intelligence!

The software for a human brain is not going to be 100% efficient compared to the theoretical maximum, nor 10% efficient, nor 1% efficient, even *before* taking into account the whole thing with parallelism vs. serialism, precision vs. imprecision, or similarly clear low-level differences.

Humbali: Ah! But allow me to offer a consideration here that, I would wager, you've never thought of before yourself - namely - *what if you're wrong?* Ah, not so confident now, are you?

Eliezer: One observes, over one's cognitive life as a human, which sorts of what-ifs are useful to contemplate, and where it is wiser to spend one's limited resources planning against the alternative that one might be wrong; and I have oft observed that lots of people don't... quite seem to understand how to use 'what if' all that well?

They'll be like, "Well, what if UFOs are aliens, and the aliens are partially hiding from us but not perfectly hiding from us, because they'll seem higher-status if they make themselves observable but never directly interact with us?"

I can refute individual what-ifs like that with specific counterarguments, but I'm not sure how to convey the central generator behind how I know that I ought to refute them. I am not sure how I can get people to reject these ideas for themselves, instead of them passively waiting for me to come around with a specific counterargument. My having to counterargue things specifically now seems like a road that never seems to end, and I am not as young as I once was, nor am I encouraged by how much progress I seem to be making. I refute one wacky idea with a specific counterargument, and somebody else comes along and presents a new wacky idea on almost exactly the same theme.

I know it's probably not going to work, if I try to say things like this, but I'll try to say them anyways. When you are going around saying 'what-if', there is a very great difference between your map of reality, and the territory of reality, which is extremely narrow and stable. Drop your phone, gravity pulls the phone downward, it falls. What if there are aliens and they make the phone rise into the air instead, maybe because they'll be especially amused at violating the rule after you just tried to use it as an example of where you could be confident? Imagine the aliens watching you, imagine their amusement, contemplate how fragile human thinking is and how little you can ever be assured of anything and ought not to be too confident. Then drop the phone and watch it fall. You've now learned something about how reality itself isn't made of what-ifs and reminding oneself to be humble; reality runs on rails stronger than your mind does.

Contemplating this doesn't mean you *know* the rails, of course, which is why it's so much harder to predict the Future than the past. But if you see that your thoughts are

still wildly flailing around what-ifs, it means that they've failed to gel, in some sense, they are not yet bound to reality, because reality has no binding receptors for what-iffery.

The correct thing to do is not to act on your what-ifs that you can't figure out how to refute, but to go on looking for a model which makes narrower predictions than that.

If that search fails, forge a model which puts some more numerical distribution on your highly entropic uncertainty, instead of diverting into specific what-ifs. And in the latter case, understand that this probability distribution reflects your ignorance and subjective state of mind, rather than your knowledge of an objective frequency; so that somebody else is allowed to be less ignorant without you shouting "Too confident!" at them. Reality runs on rails as strong as math; sometimes other people will achieve, before you do, the feat of having their own thoughts run through more concentrated rivers of probability, in some domain.

Now, when we are trying to concentrate our thoughts into deeper, narrower rivers that run closer to reality's rails, there is of course the legendary hazard of concentrating our thoughts into the *wrong* narrow channels that *exclude* reality. And the great legendary sign of this condition, of course, is the counterexample from Reality that falsifies our model! But you should not in general criticize somebody for trying to concentrate their probability into narrower rivers than yours, for this is the appearance of the great general project of trying to get to grips with Reality, that runs on true rails that are narrower still.

If you have concentrated your probability into *different* narrow channels than somebody else's, then, of course, you have a more interesting dispute; and you should engage in that legendary activity of trying to find some accessible experimental test on which your nonoverlapping models make different predictions.

Humbali: I do not understand the import of all this vaguely mystical talk.

Eliezer: I'm trying to explain why, when I say that I'm very confident it's possible to build a human-equivalent mind using less computing power than biology has managed to use effectively, and you say, "How can you be so *confident*, what if you are *wrong*," it is not unreasonable for me to reply, "Well, kid, this doesn't seem like one of those places where it's particularly important to worry about far-flung ways I could be wrong." Anyone who aspires to learn, learns over a lifetime which sorts of guesses are more likely to go oh-no-wrong in real life, and which sorts of guesses are likely to just work. Less-learned minds will have minds full of what-ifs they can't refute in more places than more-learned minds; and even if you cannot see how to refute all your what-ifs yourself, it is possible that a more-learned mind knows why they are improbable. For one must distinguish possibility from probability.

It is *imaginable* or *conceivable* that human brains have such refined algorithms that they are operating at the absolute limits of computational efficiency, or within 10% of it. But if you've spent enough time noticing *where* Reality usually exercises its sovereign right to yell "Gotcha!" at you, learning *which* of your assumptions are the kind to blow up in your face and invalidate your final conclusion, you can guess that "Ah, but what if the brain is nearly 100% computationally efficient?" is the sort of what-if that is not much worth contemplating because it is not actually going to be true in real life. Reality is going to confound you in some other way than that.

I mean, maybe you haven't read enough neuroscience and evolutionary biology that you can see from your own knowledge that the proposition sounds massively

implausible and ridiculous. But it should hardly seem unlikely that somebody else, more learned in biology, might be justified in having more confidence than you. Phones don't fall up. Reality really is very stable and orderly in a lot of ways, even in places where you yourself are ignorant of that order.

But if "What if aliens are making themselves visible in flying saucers because they want high status and they'll have higher status if they're occasionally observable but never deign to talk with us?" sounds to you like it's totally plausible, and you don't see how someone can be *so confident* that it's not true - because oh *no* what if you're *wrong* and you haven't seen the aliens so how can you *know* what they're not thinking - then I'm not sure how to lead you into the place where you can dismiss that thought with confidence. It may require a kind of life experience that I don't know how to give people, at all, let alone by having them passively read paragraphs of text that I write; a learned, perceptual sense of which what-ifs have any force behind them. I mean, I can refute that specific scenario, I *can* put that learned sense into words; but I'm not sure that does me any good unless you learn how to refute it yourself.

Humbali: Can we leave aside all that meta stuff and get back to the object level?

Eliezer: This indeed is often wise.

Humbali: Then here's one way that the minimum computational requirements for general intelligence could be *higher* than Moravec's argument for the human brain. Since, after, all, we only have one existence proof that general intelligence is possible at all, namely the human brain. Perhaps there's no way to get general intelligence in a computer except by simulating the brain neurotransmitter-by-neurotransmitter. In that case you'd need a lot *more* computing operations per second than you'd get by calculating the number of potential spikes flowing around the brain! What if it's true? How can you *know*?

(Modern person: This seems like an obvious straw argument? I mean, would anybody, even at an earlier historical point, actually make an argument like -

Moravec and Eliezer: YES THEY WOULD.)

Eliezer: I can imagine that if we were trying specifically to *upload a human* that there'd be no easy and simple and obvious way to run the resulting simulation and get a good answer, without simulating neurotransmitter flows in extra detail.

To imagine that every one of these simulated flows is *being usefully used in general intelligence and there is no way to simplify the mind design to use fewer computations...* I suppose I could try to refute that specifically, but it seems to me that this is a road which has no end unless I can convey the generator of my refutations. Your what-iffery is flung far enough that, if I cannot leave even that much rejection as an exercise for the reader to do on their own without my holding their hand, the reader has little enough hope of following the rest; let them depart now, in indignation shared with you, and save themselves further outrage.

I mean, it will obviously be *less* obvious to the reader because they will know *less* than I do about this exact domain, it will justly take *more* work for the reader to specifically refute you than it takes me to refute you. But I think the reader needs to be able to do that at all, in this example, to follow the more difficult arguments later.

Imaginary Moravec: I don't think it changes my conclusions by an order of magnitude, but some people would worry that, for example, changes of protein

expression inside a neuron in order to implement changes of long-term potentiation, are also important to intelligence, and could be a big deal in the brain's real, effectively-used computational costs. I'm curious if you'd dismiss that as well, the same way you dismiss the probability that you'd have to simulate every neurotransmitter molecule?

Eliezer: Oh, of course not. Long-term potentiation suddenly turning out to be a big deal you overlooked, compared to the depolarization impulses spiking around, is *very* much the sort of thing where Reality sometimes jumps out and yells "Gotcha!" at you.

Humbali: *How can you tell the difference?*

Eliezer: Experience with Reality yelling "Gotcha!" at myself and historical others.

Humbali: They seem like equally plausible speculations to me!

Eliezer: Really? "What if long-term potentiation is a big deal and computationally important" sounds just as plausible to you as "What if the brain is already close to the wall of making the most efficient possible use of computation to implement general intelligence, and every neurotransmitter molecule matters"?

Humbali: Yes! They're both what-ifs we can't know are false and shouldn't be overconfident about denying!

Eliezer: My tiny feeble mortal mind is far away from reality and only bound to it by the loosest of correlating interactions, but I'm not *that* unbound from reality.

Moravec: I would guess that in real life, long-term potentiation is sufficiently slow and local that what goes on inside the cell body of a neuron over minutes or hours is not as big of a computational deal as thousands of times that many spikes flashing around the brain in milliseconds or seconds. That's why I didn't make a big deal of it in my own estimate.

Eliezer: Sure. But it *is* much more the sort of thing where you wake up to a reality-authored science headline saying "Gotcha! There were tiny DNA-activation interactions going on in there at high speed, and they were actually pretty expensive and important!" I'm not saying this exact thing is very probable, just that it wouldn't be out-of-character for reality to say *something* like that to me, the way it would be really genuinely bizarre if Reality was, like, "Gotcha! The brain is as computationally efficient of a generally intelligent engine as any algorithm can be!"

Moravec: I think we're in agreement about that part, or we would've been, if we'd actually had this conversation in 1988. I mean, I am a competent research roboticist and it is difficult to become one if you are completely unglued from reality.

Eliezer: Then what's with the 2010 prediction for strong AI, and the massive non-sequitur leap from "the human brain is somewhere around 10 trillion ops/sec" to "if we build a 10 trillion ops/sec supercomputer, we'll get strong AI"?

Moravec: Because while it's the kind of Fermi estimate that can be off by an order of magnitude in practice, it doesn't really seem like it should be, I don't know, off by three orders of magnitude? And even three orders of magnitude is just 10 years of Moore's Law. 2020 for strong AI is also a bold and important prediction.

Eliezer: And the year 2000 for strong AI even more so.

Moravec: Heh! That's not usually the direction in which people argue with me.

Eliezer: There's an important distinction between the direction in which people usually argue with you, and the direction from which Reality is allowed to yell "Gotcha!" I wish my future self had kept this more in mind, when arguing with Robin Hanson about how well AI architectures were liable to generalize and scale without a ton of domain-specific algorithmic tinkering for every field of knowledge. I mean, in principle what I was arguing for was various lower bounds on performance, but I sure could have emphasized more loudly that those were *lower bounds* - well, I *did* emphasize the lower-bound part, but - from the way I felt when AlphaGo and Alpha Zero and GPT-2 and GPT-3 showed up, I think I must've sorta forgot that myself.

Moravec: Anyways, if we say that I might be up to three orders of magnitude off and phrase it as 2000-2020, do you agree with my prediction then?

Eliezer: No, I think you're just... arguing about the wrong facts, in a way that seems to be unglued from most tracks Reality might follow so far as I currently know? On my view, creating AGI is strongly dependent on how much knowledge you have about how to do it, in a way which almost *entirely* obviates the relevance of arguments from human biology?

Like, human biology tells us a single not-very-useful data point about how much computing power evolutionary biology needs in order to build a general intelligence, using very alien methods to our own. Then, very separately, there's the constantly changing level of how much cognitive science, neuroscience, and computer science our own civilization knows. We don't know how much computing power is required for AGI for *any* level on that constantly changing graph, and biology doesn't tell us. All we know is that the hardware requirements for AGI must be dropping by the year, because the knowledge of how to create AI is something that only increases over time.

At some point the moving lines for "decreasing hardware required" and "increasing hardware available" will cross over, which lets us predict that AGI gets built at *some* point. But we don't know how to graph two key functions needed to predict that date. You would seem to be committing the classic fallacy of searching for your keys under the streetlight where the visibility is better. You know how to estimate how many floating-point operations per second the retina could effectively be using, but *this is not the number you need to predict the outcome you want to predict*. You need a graph of human knowledge of computer science over time, and then a graph of how much computer science requires how much hardware to build AI, and neither of these graphs are available.

It doesn't matter how many chapters your book spends considering the continuation of Moore's Law or computation in the retina, and I'm sorry if it seems rude of me in some sense to just dismiss the relevance of all the hard work you put into arguing it. But you're arguing the *wrong facts* to get to the conclusion, so all your hard work is for naught.

Humbali: Now it seems to me that I must chide you for being too dismissive of Moravec's argument. Fine, yes, Moravec has not established with *logical certainty* that strong AI must arrive at the point where top supercomputers match the human brain's 10 trillion operations per second. But has he not established a *reference class*, the sort of *base rate* that good and virtuous superforecasters, unlike yourself, go looking for when they want to *anchor* their estimate about some future outcome? Has

he not, indeed, established the sort of argument which says that if top supercomputers can do only *ten million* operations per second, we're not very likely to get AGI earlier than that, and if top supercomputers can do *ten quintillion* operations per second*, we're unlikely not to already have AGI?

(*) In 2021 terms, [10 TPU v4 pods](#).

Eliezer: With ranges that wide, it'd be more likely and less amusing to hit somewhere inside it by coincidence. But I still think this whole line of thoughts is just off-base, and that you, Humbali, have not truly grasped the concept of a virtuous superforecaster or how they go looking for reference classes and base rates.

Humbali: I frankly think you're just being unvirtuous. Maybe you have some special model of AGI which claims that it'll arrive in a different year or be arrived at by some very different pathway. But is not Moravec's estimate a sort of base rate which, to the extent you are properly and virtuously uncertain of your own models, you ought to regress in your own probability distributions over AI timelines? As you become more uncertain about the exact amounts of knowledge required and what knowledge we'll have when, shouldn't you have an uncertain distribution about AGI arrival times that centers around Moravec's base-rate prediction of 2010?

For you to reject this anchor seems to reveal a grave lack of humility, since you must be very certain of whatever alternate estimation methods you are using in order to throw away this base-rate entirely.

Eliezer: Like I said, I think you've just failed to grasp the true way of a virtuous superforecaster. Thinking a lot about Moravec's so-called 'base rate' is just making you, in some sense, stupider; you need to cast your thoughts loose from there and try to navigate a wilder and less tamed space of possibilities, until they begin to gel and coalesce into narrower streams of probability. Which, for AGI, they probably *won't do* until we're quite close to AGI, and start to guess correctly how AGI will get built; for it is easier to predict an eventual global pandemic than to say it will start in November of 2019. Even in October of 2019 this cannot be done.

Humbali: Then all this uncertainty must somehow be quantified, if you are to be a virtuous Bayesian; and again, for lack of anything better, the resulting distribution should center on Moravec's base-rate estimate of 2010.

Eliezer: No, that calculation is just basically not relevant here; and thinking about it is making you stupider, as your mind flails in the trackless wilderness grasping onto unanchored air. Things must be 'sufficiently similar' to each other, in some sense, for us to get a base rate on one thing by looking at another thing. Humans making an AGI is just too dissimilar to evolutionary biology making a human brain for us to anchor 'how much computing power at the time it happens' from one to the other. It's not the droid we're looking for; and your attempt to build an inescapable epistemological trap about virtuously calling that a 'base rate' is not the Way.

Imaginary Moravec: If I can step back in here, I don't think my calculation is zero evidence? What we know from evolutionary biology is that a blind alien god with zero foresight accidentally mutated a chimp brain into a general intelligence. I don't want to knock biology's work too much, there's some impressive stuff in the retina, and the retina is just the part of the brain which is in some sense easiest to understand. But surely there's a very reasonable argument that 10 trillion ops/sec is about the amount of computation that evolutionary biology needed; and since evolution is stupid, when

we ourselves have that much computation, it shouldn't be *that* hard to figure out how to configure it.

Eliezer: If that was true, the same theory predicts that our current supercomputers should be doing a better job of matching the agility and vision of spiders. When at some point there's enough hardware that we figure out how to put it together into AGI, we could be doing it with less hardware than a human; we could be doing it with more; and we can't even say that these two possibilities are *around equally probable* such that our probability distribution should have its median around 2010. Your number is so bad and obtained by such bad means that we should just throw it out of our thinking and start over.

Humbali: This last line of reasoning seems to me to be particularly ludicrous, like you're just throwing away the only base rate we have in favor of a confident assertion of our somehow being *more uncertain* than that.

Eliezer: Yeah, well, sorry to put it bluntly, Humbali, but you have not yet figured out how to turn your own computing power into intelligence.

- 1999 -

Luke Muehlhauser reading a previous draft of this (only sounding much more serious than this, because Luke Muehlhauser): You know, there was this certain teenaged futurist who made some of his own predictions about AI timelines -

Eliezer: I'd really rather not argue from that as a case in point. I dislike people who screw up something themselves, and then argue like nobody else could possibly be more competent than they were. I dislike even more people who change their mind about something when they turn 22, and then, for the rest of their lives, go around acting like they are now Very Mature Serious Adults who believe the thing that a Very Mature Serious Adult believes, so if you disagree with them about that thing they started believing at age 22, you must just need to wait to grow out of your extended childhood.

Luke Muehlhauser (still being paraphrased): It seems like it ought to be acknowledged somehow.

Eliezer: That's fair, yeah, I can see how someone might think it was relevant. I just dislike how it potentially creates the appearance of trying to slyly sneak in an Argument From Reckless Youth that I regard as not only invalid but also incredibly distasteful. You don't get to screw up yourself and then use that as an argument about how nobody else can do better.

Humbali: Uh, what's the actual drama being subtweeted here?

Eliezer: A certain teenaged futurist, who, for example, said in 1999, "The most realistic estimate for a seed AI transcendence is 2020; nanowar, before 2015."

Humbali: This young man must surely be possessed of some very deep character defect, which I worry will prove to be of the sort that people almost never truly outgrow except in the rarest cases. Why, he's not even putting a probability distribution over his mad soothsaying - how blatantly absurd can a person get?

Eliezer: Dear child ignorant of history, your complaint is far too anachronistic. This is 1999 we're talking about here; almost nobody is putting probability distributions on things, that element of your later subculture has not yet been introduced. Eliezer-2002 hasn't been sent a copy of "Judgment Under Uncertainty" by Emil Gilliam.

Eliezer-2006 hasn't put his draft online for "Cognitive biases potentially affecting judgment of global risks". The Sequences won't start until another year after that.

How would the forerunners of effective altruism *in 1999* know about putting probability distributions on forecasts? I haven't told them to do that yet! We can give historical personages credit when they seem to somehow end up doing better than their surroundings would suggest; it is unreasonable to hold them to modern standards, or expect them to have finished refining those modern standards by the age of nineteen.

Though there's also a more subtle lesson you could learn, about how this young man turned out to still have a promising future ahead of him; which he retained at least in part by having a deliberate contempt for pretended dignity, allowing him to be plainly and simply wrong in a way that he noticed, without his having twisted himself up to avoid a prospect of embarrassment. Instead of, for example, his evading such plain falsification by having dignifiedly wide Very Serious probability distributions centered on the same medians produced by the same basically bad thought processes.

But that was too much of a digression, when I tried to write it up; maybe later I'll post something separately.

- 2004 or thereabouts -

Ray Kurzweil in 2001: I have [calculated](#) that matching the intelligence of a human brain requires $2 * 10^{16}$ ops/sec* and this will become available in a \$1000 computer in 2023. 26 years after that, in 2049, a \$1000 computer will have ten billion times more computing power than a human brain; and in 2059, that computer will cost one cent.

(*) Two TPU v4 pods.

Actual real-life Eliezer in Q&A, when Kurzweil says the same thing in a 2004(?) talk: It seems weird to me to forecast the arrival of "human-equivalent" AI, and then expect Moore's Law to just continue on the same track past that point for thirty years. Once we've got, in your terms, human-equivalent AIs, even if we don't go beyond that in terms of intelligence, Moore's Law will start speeding them up.

Once AIs are thinking thousands of times faster than we are, wouldn't that tend to break down the graph of Moore's Law with respect to the objective wall-clock time of the Earth going around the Sun? Because AIs would be able to spend thousands of subjective years working on new computing technology?

Actual Ray Kurzweil: The fact that AIs can do faster research is exactly what will enable Moore's Law to continue on track.

Actual Eliezer (out loud): Thank you for answering my question.

Actual Eliezer (internally): Moore's Law is a phenomenon produced by human cognition and the fact that human civilization runs off human cognition. You can't expect the surface phenomenon to continue unchanged after the deep causal phenomenon underlying it starts changing. What kind of bizarre worship of graphs

would lead somebody to think that the graphs were the primary phenomenon and would continue steady and unchanged when the forces underlying them changed massively? I was hoping he'd be less nutty in person than in the book, but oh well.

- 2006 or thereabouts -

Somebody on the Internet: I have calculated the number of computer operations used by evolution to evolve the human brain - searching through organisms with increasing brain size - by adding up all the computations that were done by any brains before modern humans appeared. It comes out to 10^{43} computer operations.* AGI isn't coming any time soon!

(*) I forget the exact figure. It was 10^{40} -something.

Eliezer, sighing: Another day, another biology-inspired timelines forecast. This trick didn't work when Moravec tried it, it's not going to work while Ray Kurzweil is trying it, and it's not going to work when you try it either. It also didn't work when a certain teenager tried it, but please entirely ignore that part; you're at least allowed to do better than him.

Imaginary Somebody: Moravec's prediction failed because he assumed that you could just magically take something with around as much hardware as the human brain and, poof, it would start being around that intelligent -

Eliezer: Yes, that is one way of viewing an invalidity in that argument. Though you do Moravec a disservice if you imagine that he could only argue "It will magically emerge", and could not give the more plausible-sounding argument "Human engineers are not that incompetent compared to biology, and will probably figure it out without more than one or two orders of magnitude of extra overhead."

Somebody: But I am cleverer, for I have calculated the number of computing operations that was used to *create and design* biological intelligence, not just the number of computing operations required to *run it once created*!

Eliezer: And yet, because your reasoning contains the word "biological", it is just as invalid and unhelpful as Moravec's original prediction.

Somebody: I don't see why you dismiss my biological argument about timelines on the basis of Moravec having been wrong. He made one basic mistake - neglecting to take into effect the cost to generate intelligence, not just to run it. I have corrected this mistake, and now my own effort to do biologically inspired timeline forecasting should work fine, and must be evaluated on its own merits, *de novo*.

Eliezer: It is true indeed that sometimes a line of inference is doing just one thing wrong, and works fine after being corrected. And because this is true, it is often indeed wise to reevaluate new arguments on their own merits, if that is how they present themselves. One may not take the past failure of a different argument or three, and try to hang it onto the new argument like an inescapable iron ball chained to its leg. It might be the cause for defeasible skepticism, but not invincible skepticism.

That said, on my view, you are making a nearly identical mistake as Moravec, and so his failure remains relevant to the question of whether you are engaging in a kind of thought that binds well to Reality.

Somebody: And that mistake is just mentioning the word "biology"?

Eliezer: The problem is that *the resource gets consumed differently, so base-rate arguments from resource consumption end up utterly unhelpful in real life*. The human brain consumes around 20 watts of power. Can we thereby conclude that an AGI should consume around 20 watts of power, and that, when technology advances to the point of being able to supply around 20 watts of power to computers, we'll get AGI?

Somebody: That's absurd, of course. So, what, you compare my argument to an absurd argument, and from this dismiss it?

Eliezer: I'm saying that Moravec's "argument from comparable resource consumption" must be in general [invalid](#), because it [Proves Too Much](#). If it's in general valid to reason about comparable resource consumption, then it should be equally valid to reason from energy consumed as from computation consumed, and pick energy consumption instead to call the basis of your median estimate.

You say that AIs consume energy in a very different way from brains? Well, they'll also consume computations in a very different way from brains! The only difference between these two cases is that you *know* something about how humans eat food and break it down in their stomachs and convert it into ATP that gets consumed by neurons to pump ions back out of dendrites and axons, while computer chips consume electricity whose flow gets interrupted by transistors to transmit information. Since you *know anything whatsoever* about how AGIs and humans consume energy, you can see that the consumption is so vastly different as to obviate all comparisons entirely.

You are *ignorant* of how the brain consumes computation, you are *ignorant* of how the first AGIs built would consume computation, but "an unknown key does not open an unknown lock" and these two ignorant distributions should not assert much internal correlation between them.

Even without knowing the specifics of how brains and future AGIs consume computing operations, you ought to be able to reason abstractly about a directional update that you *would* make, if you knew *any* specifics instead of none. If you did know how both kinds of entity consumed computations, if you knew about specific machinery for human brains, and specific machinery for AGIs, you'd then be able to see the enormous vast specific differences between them, and go, "Wow, what a futile resource-consumption comparison to try to use for forecasting."

(Though I say this without much hope; I have not had very much luck in telling people about predictable directional updates they would make, if they knew something instead of nothing about a subject. I think it's probably too abstract for most people to feel in their gut, or something like that, so their brain ignores it and moves on in the end. I have had life experience with learning more about a thing, updating, and then going to myself, "Wow, I should've been able to predict in retrospect that learning almost *any* specific fact would move my opinions in that same direction." But I worry this is not a common experience, for it involves a real experience of discovery, and preferably more than one to get the generalization.)

Somebody: All of that seems irrelevant to my novel and different argument. I am not foolishly estimating the resources consumed by a single brain; I'm estimating the resources consumed by evolutionary biology to *invent* brains!

Eliezer: And the humans wracking their own brains and inventing new AI program architectures and deploying those AI program architectures to themselves learn, will consume computations so *utterly differently* from evolution that there is no point comparing those consumptions of resources. That is the flaw that you share exactly with Moravec, and that is why I say the same of both of you, "This is a kind of thinking that fails to bind upon reality, it doesn't work in real life." I don't care how much painstaking work you put into your estimate of 10^{43} computations performed by biology. It's just not a relevant fact.

Humbali: But surely this estimate of 10^{43} cumulative operations can at least be used to establish a base rate for anchoring our -

Eliezer: Oh, for god's sake, shut up. At least Somebody is only wrong on the object level, and isn't trying to build an inescapable epistemological trap by which his ideas must still hang in the air like an eternal stench even after they've been counterargued. Isn't 'but muh base rates' what your viewpoint would've also said about Moravec's 2010 estimate, back when that number still looked plausible?

Humbali: Of course it is evident to me now that my youthful enthusiasm was mistaken; obviously I tried to estimate the wrong figure. As Somebody argues, we should have been estimating the biological computations used to *design* human intelligence, not the computations used to *run* it.

I see, now, that I was using the wrong figure as my base rate, leading my base rate to be wildly wrong, and even irrelevant; but now that I've seen this, the clear error in my previous reasoning, I have a *new* base rate. This doesn't seem obviously to me likely to contain the same kind of wildly invalidating enormous error as before. What, is Reality just going to yell "Gotcha!" at me again? And even the prospect of some new unknown error, which is just as likely to be in either possible direction, implies only that we should widen our credible intervals while keeping them centered on a median of 10^{43} operations -

Eliezer: Please stop. This trick just never works, at all, deal with it and get over it. Every second of attention that you pay to the 10^{43} number is making you stupider. You might as well reason that 20 watts is a base rate for how much energy the first generally intelligent computing machine should consume.

- 2020 -

OpenPhil: We have commissioned a Very Serious report on a biologically inspired estimate of how much computation will be required to achieve Artificial General Intelligence, for purposes of forecasting an AGI timeline. ([Summary of report.](#)) ([Full draft of report.](#)) Our leadership takes this report Very Seriously.

Eliezer: Oh, hi there, new kids. Your grandpa is feeling kind of tired now and can't debate this again with as much energy as when he was younger.

Imaginary OpenPhil: You're not *that* much older than us.

Eliezer: Not by biological wall-clock time, I suppose, but -

OpenPhil: You think thousands of times faster than us?

Eliezer: I wasn't going to say it if you weren't.

OpenPhil: We object to your assertion on the grounds that it is false.

Eliezer: I was actually going to say, you might be underestimating how long I've been walking this endless battlefield because I started *really quite young*.

I mean, sure, I didn't read Moravec's *Mind Children* when it came out in 1988. I only read it four years later, when I was twelve. And sure, I didn't immediately afterwards start writing online about Moore's Law and strong AI; I did not immediately contribute my own salvos and sallies to the war; I was not yet a noticed voice in the debate. I only got started on that at age sixteen. I'd like to be able to say that in 1999 I was just a random teenager being reckless, but in fact I was already being invited to dignified online colloquia about the "Singularity" and mentioned in printed books; when I was being wrong back then I was already doing so in the capacity of a minor public intellectual on the topic.

This is, as I understand normie ways, relatively young, and is probably worth an extra decade tacked onto my biological age; you should imagine me as being 52 instead of 42 as I write this, with a correspondingly greater number of visible gray hairs.

A few years later - though still before your time - there was the Accelerating Change Foundation, and Ray Kurzweil spending literally millions of dollars to push Moore's Law graphs of technological progress as *the* central story about the future. I mean, I'm sure that a few million dollars sounds like peanuts to OpenPhil, but if your own annual budget was a hundred thousand dollars or so, that's a hell of a megaphone to compete with.

If you are currently able to conceptualize the Future as being about something *other* than nicely measurable metrics of progress in various tech industries, being projected out to where they will inevitably deliver us nice things - that's at least partially because of a battle fought years earlier, in which I was a primary fighter, creating a conceptual atmosphere you now take for granted. A mental world where threshold levels of AI ability are considered potentially interesting and transformative - rather than milestones of new technological luxuries to be checked off on an otherwise invariant graph of Moore's Laws as they deliver flying cars, space travel, lifespan-extension escape velocity, and other such goodies on an equal level of interestingness. I have earned at least a *little* right to call myself your grandpa.

And that kind of experience has a sort of compounded interest, where, once you've lived something yourself and participated in it, you can learn more from reading other histories about it. The histories become more real to you once you've fought your own battles. The fact that I've lived through timeline errors in person gives me a sense of how it actually feels to be around at the time, watching people sincerely argue Very Serious erroneous forecasts. That experience lets me really and actually update on the history of the earlier mistaken timelines from before I was around; instead of the histories just seeming like a kind of fictional novel to read about, disconnected from reality and not happening to real people.

And now, indeed, I'm feeling a bit old and tired for reading yet another report like yours in full attentive detail. Does it by any chance say that AGI is due in about 30 years from now?

OpenPhil: Our report has very wide credible intervals around both sides of its median, as we analyze the problem from a number of different angles and show how they lead to different estimates -

Eliezer: Unfortunately, the thing about figuring out five different ways to guess the effective IQ of the smartest people on Earth, and having three different ways to estimate the minimum IQ to destroy lesser systems such that you could extrapolate a minimum IQ to destroy the whole Earth, and putting wide credible intervals around all those numbers, and combining and mixing the probability distributions to get a new probability distribution, is that, at the end of all that, you are still left with a load of nonsense. Doing a fundamentally wrong thing in several different ways will not save you, though I suppose if you spread your bets widely enough, one of them may be right by coincidence.

So does the report by any chance say - with however many caveats and however elaborate the probabilistic methods and alternative analyses - that AGI is probably due in about 30 years from now?

OpenPhil: Yes, in fact, our 2020 report's median estimate is 2050; though, again, with very wide credible intervals around both sides. Is that number significant?

Eliezer: It's a law generalized by Charles Platt, that any AI forecast will put strong AI thirty years out from when the forecast is made. Vernor Vinge referenced it in the body of his famous 1993 NASA speech, whose abstract begins, "Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended."

After I was old enough to be more skeptical of timelines myself, I used to wonder how Vinge had pulled out the "within thirty years" part. This may have gone over my head at the time, but rereading again today, I conjecture Vinge may have chosen the headline figure of thirty years as a deliberately self-deprecating reference to Charles Platt's generalization about such forecasts always being thirty years from the time they're made, which Vinge explicitly cites later in the speech.

Or to put it another way: I conjecture that to the audience of the time, already familiar with some previously-made forecasts about strong AI, the impact of the abstract is meant to be, "Never mind predicting strong AI in thirty years, you should be predicting *superintelligence* in thirty years, which matters a lot more." But the minds of authors are scarcely more knowable than the Future, if they have not explicitly told us what they were thinking; so you'd have to ask Professor Vinge, and hope he remembers what he was thinking back then.

OpenPhil: Superintelligence before 2023, huh? I suppose Vinge still has two years left to go before that's falsified.

Eliezer: Also in the body of the speech, Vinge says, "I'll be surprised if this event occurs before 2005 or after 2030," which sounds like a more serious and sensible way of phrasing an estimate. I think that should supersede the probably Platt-inspired headline figure for what we think of as Vinge's 1993 prediction. The jury's still out on whether Vinge will have made a good call.

Oh, and sorry if grandpa is boring you with all this history from the times before you were around. I mean, I didn't actually attend Vinge's famous NASA speech when it happened, what with being thirteen years old at the time, but I sure did read it later. Once it was digitized and put online, it was all over the Internet. Well, all over certain parts of the Internet, anyways. Which nerdy parts constituted a much larger fraction of the whole, back when the World Wide Web was just starting to take off among early adopters.

But, yeah, the new kids showing up with some graphs of Moore's Law and calculations about biology and an earnest estimate of strong AI being thirty years out from the time of the report is, uh, well, it's... historically preceded.

OpenPhil: That part about Charles Platt's generalization is interesting, but just because we unwittingly chose literally exactly the median that Platt predicted people would always choose in consistent error, that doesn't justify dismissing our work, right? We could have used a completely valid method of estimation which would have pointed to 2050 no matter which year it was tried in, and, by sheer coincidence, have first written that up in 2020. In fact, we try to show in the report that the same methodology, evaluated in earlier years, would also have pointed to around 2050 -

Eliezer: Look, people keep trying this. It's never worked. It's never going to work. 2 years before the end of the world, there'll be another published biologically inspired estimate showing that AGI is 30 years away and it will be exactly as informative then as it is now. I'd love to know the timelines too, but you're not *going* to get the answer you want until right before the end of the world, and maybe not even then unless you're paying very close attention. *Timing this stuff is just plain hard.*

OpenPhil: But our report is different, and our methodology for biologically inspired estimates is wiser and less naive than those who came before.

Eliezer: That's what the last guy said, but go on.

OpenPhil: First, we carefully estimate a range of possible figures for the equivalent of neural-network parameters needed to emulate a human brain. Then, we estimate how many examples would be required to train a neural net with that many parameters. Then, we estimate the total computational cost of that many training runs. Moore's Law then gives us 2050 as our median time estimate, given what we think are the *most* likely underlying assumptions, though we do analyze it several different ways.

Eliezer: This is almost exactly what the last guy tried, except you're using network parameters instead of computing ops, and deep learning training runs instead of biological evolution.

OpenPhil: Yes, so we've corrected his mistake of estimating the wrong biological quantity and now we're good, right?

Eliezer: That's what the last guy thought *he'd* done about Moravec's mistaken estimation target. And neither he nor Moravec would have made much headway on their underlying mistakes, by doing a probabilistic analysis of that same wrong question from multiple angles.

OpenPhil: Look, sometimes more than one person makes a mistake, over historical time. It doesn't mean nobody can ever get it right. You of all people should agree.

Eliezer: I do so agree, but that doesn't mean I agree you've *fixed* the mistake. I think the methodology itself is bad, not just its choice of which biological parameter to estimate. Look, do you understand *why* the evolution-inspired estimate of 10^{43} ops was completely ludicrous; and the claim that it was equally likely to be mistaken in either direction, even more ludicrous?

OpenPhil: Because AGI isn't like biology, and in particular, will be trained using gradient descent instead of evolutionary search, which is cheaper. We do note inside

our report that this is a key assumption, and that, if it fails, the estimate might be correspondingly wrong -

Eliezer: But then you claim that mistakes are equally likely in both directions and so your unstable estimate is a good median. Can you see why the previous evolutionary estimate of 10^{43} cumulative ops was not, in fact, *equally likely to be wrong in either direction?* That it was, predictably, a directional overestimate?

OpenPhil: Well, search by evolutionary biology is more costly than training by gradient descent, so in hindsight, it was an overestimate. Are you claiming this was predictable in foresight instead of hindsight?

Eliezer: I'm claiming that, at the time, I snorted and tossed Somebody's figure out the window while thinking it was ridiculously huge and absurd, yes.

OpenPhil: Because you'd already foreseen in 2006 that gradient descent would be the method of choice for training future AIs, rather than genetic algorithms?

Eliezer: Ha! No. Because it was an insanely costly hypothetical approach whose main point of appeal, to the sort of person who believed in it, was that it didn't require having any idea whatsoever of what you were doing or how to design a mind.

OpenPhil: Suppose one were to reply: "Somebody" *didn't* know better-than-evolutionary methods for designing a mind, just as we currently don't know better methods than gradient descent for designing a mind; and hence Somebody's estimate was the best estimate at the time, just as ours is the best estimate now?

Eliezer: Unless you were one of a small handful of leading neural-net researchers who knew a few years ahead of the world where scientific progress was heading - who knew a Thielian 'secret' before finding evidence strong enough to convince the less foresighted - you couldn't have called the jump specifically to *gradient descent* rather than any other technique. "I don't know any more computationally efficient way to produce a mind than *re-evolving* the cognitive history of all life on Earth" transitioning over time to "I don't know any more computationally efficient way to produce a mind than *gradient descent* over entire brain-sized models" is not predictable in the specific part about "*gradient descent*" - not unless you know a Thielian secret.

But knowledge is a ratchet that usually only turns one way, so it's predictable that the current story changes to *somewhere* over future time, in a net expected direction.

Let's consider the technique currently known as mixture-of-experts (MoE), for training smaller nets in pieces and muxing them together. It's not my mainline prediction that MoE actually goes anywhere - if I thought MoE was actually promising, I wouldn't call attention to it, of course! I don't want to *make* timelines shorter, that is not a service to Earth, not a good sacrifice in the cause of winning an Internet argument.

But if I'm wrong and MoE is not a dead end, that technique serves as an easily-visualizable case in point. If that's a fruitful avenue, the technique currently known as "mixture-of-experts" will mature further over time, and future deep learning engineers will be able to further perfect the art of training *slices of brains* using gradient descent and fewer examples, instead of training *entire brains* using gradient descent and lots of examples.

Or, more likely, it's not MoE that forms the next little trend. But there is going to be *something*, especially if we're sitting around waiting until 2050. Three decades is enough time for some *big* paradigm shifts in an intensively researched field. Maybe

we'd end up using neural net tech very similar to today's tech if the world ends in 2025, but in that case, of course, your prediction must have failed somewhere else.

The three components of AGI arrival times are available hardware, which increases over time in an easily graphed way; available knowledge, which increases over time in a way that's much harder to graph; and hardware required at a given level of specific knowledge, a huge multidimensional unknown background parameter. The fact that you have no idea how to graph the increase of knowledge - or measure it in any way that is less completely silly than "number of science papers published" or whatever such gameable metric - doesn't change the point that this *is* a predictable fact about the future; there *will* be more knowledge later, the more time that passes, and that will *directionally* change the expense of the currently least expensive way of doing things.

OpenPhil: We did already consider that and try to take it into account: our model already includes a parameter for how algorithmic progress reduces hardware requirements. It's not easy to graph as exactly as Moore's Law, as you say, but our best-guess estimate is that compute costs halve every 2-3 years.

Eliezer: Oh, nice. I was wondering what sort of tunable underdetermined parameters enabled your model to nail the psychologically overdetermined final figure of '30 years' so exactly.

OpenPhil: Eliezer.

Eliezer: Think of this in an economic sense: people don't buy where goods are most expensive and delivered latest, they buy where goods are cheapest and delivered earliest. Deep learning researchers are not like an inanimate chunk of ice tumbling through intergalactic space in its unchanging direction of previous motion; they are economic agents who look around for ways to destroy the world faster and more cheaply than the way that you imagine as the default. They are more eager than you are to think of more creative paths to get to the next milestone faster.

OpenPhil: Isn't this desire for cheaper methods exactly what our model already accounts for, by modeling algorithmic progress?

Eliezer: The makers of AGI aren't going to be doing 10,000,000,000,000 rounds of gradient descent, on entire brain-sized 300,000,000,000,000-parameter models, *algorithmically faster than today*. They're going to get to AGI via some route that *you don't know how to take*, at least if it happens in 2040. If it happens in 2025, it may be via a route that some modern researchers do know how to take, but in this case, of course, your model was also wrong.

They're not going to be taking your default-imagined approach *algorithmically faster*, they're going to be taking an *algorithmically different approach* that eats computing power in a different way than you imagine it being consumed.

OpenPhil: Shouldn't that just be folded into our estimate of how the computation required to accomplish a fixed task decreases by half every 2-3 years due to better algorithms?

Eliezer: Backtesting this viewpoint on the previous history of computer science, it seems to me to assert that it should be possible to:

- Train a pre-Transformer RNN/CNN-based model, not using any other techniques invented after 2017, to GPT-2 levels of performance, using only around 2x as much compute as GPT-2;
- Play pro-level Go using 8-16 times as much computing power as AlphaGo, but only 2006 levels of technology.

For reference, recall that in 2006, Hinton and Salakhutdinov were just starting to publish that, by training multiple layers of Restricted Boltzmann machines and then unrolling them into a "deep" neural network, you could get an initialization for the network weights that would avoid the problem of vanishing and exploding gradients and activations. At least so long as you didn't try to stack too many layers, like a dozen layers or something ridiculous like that. This being the point that kicked off the entire deep-learning revolution.

Your model apparently suggests that we have gotten around 50 times more efficient at turning computation into intelligence since that time; so, we should be able to replicate any modern feat of deep learning performed in 2021, using techniques from before deep learning and around fifty times as much computing power.

OpenPhil: No, that's totally not what our viewpoint says when you backfit it to past reality. Our model does a great job of retrodicting past reality.

Eliezer: How so?

OpenPhil: <Eliezer cannot predict what they will say here.>

Eliezer: I'm not convinced by this argument.

OpenPhil: We didn't think you would be; you're sort of predictable that way.

Eliezer: Well, yes, if I'd predicted I'd update from hearing your argument, I would've updated already. I may not be a real Bayesian but I'm not *that* incoherent.

But I can guess in advance at the outline of my reply, and my guess is this:

"Look, when people come to me with models claiming the future is predictable enough for timing, I find that their viewpoints seem to me like they would have made garbage predictions if I actually had to operate them in the past *without benefit of hindsight*. Sure, with benefit of hindsight, you can look over a thousand possible trends and invent rules of prediction and event timing that nobody *in the past* actually spotlighted *then*, and claim that things happened on trend. I was around at the time and I do not recall people actually predicting the shape of AI in the year 2020 in advance. I don't think they were just being stupid either."

"In a conceivable future where people are still alive and reasoning as modern humans do in 2040, somebody will no doubt look back and claim that everything happened on trend since 2020; but *which* trend the hindsight will pick out is not predictable to us in advance.

"It may be, of course, that I simply don't understand how to operate your viewpoint, nor how to apply it to the past or present or future; and that yours is a sort of viewpoint which indeed permits saying only one thing, and not another; and that this viewpoint would have predicted the past wonderfully, even without any benefit of hindsight. But there is also that less charitable viewpoint which suspects that somebody's theory of 'A coinflip always comes up heads on occasions X' contains

some informal parameters which can be argued about which occasions exactly 'X' describes, and that the operation of these informal parameters is a bit influenced by one's knowledge of whether a past coinflip actually came up heads or not.

"As somebody who doesn't start from the assumption that your viewpoint is a good fit to the past, I still don't see how a good fit to the past could've been extracted from it without benefit of hindsight."

OpenPhil: That's a pretty general counterargument, and like any pretty general counterargument it's a blade you should try turning against yourself. Why doesn't your own viewpoint horribly mispredict the past, and say that all estimates of AGI arrival times are predictably net underestimates? If we imagine trying to operate your own viewpoint in 1988, we imagine going to Moravec and saying, "Your estimate of how much computing power it takes to match a human brain is predictably an overestimate, because engineers will find a better way to do it than biology, so we should expect AGI sooner than 2010."

Eliezer: I did tell Imaginary Moravec that his estimate of the minimum computation required for human-equivalent general intelligence was predictably an overestimate; that was right there in the dialogue before I even got around to writing this part. And I also, albeit with benefit of hindsight, told Moravec that both of these estimates were useless for timing the future, because they skipped over the questions of how much knowledge you'd need to make an AGI with a given amount of computing power, how fast knowledge was progressing, and the actual timing determined by the rising hardware line touching the falling hardware-required line.

OpenPhil: We don't see how to operate your viewpoint to say *in advance* to Moravec, before his prediction has been falsified, "Your estimate is plainly a garbage estimate" instead of "Your estimate is obviously a directional underestimate", especially since you seem to be saying the latter to *us, now*.

Eliezer: That's not a critique I give zero weight. And, I mean, as a kid, I was in fact talking like, "To heck with that hardware estimate, let's at least try to get it done before then. People are dying for lack of superintelligence; let's aim for 2005." I had a T-shirt spraypainted "Singularity 2005" at a science fiction convention, it's rather crude but I think it's still in my closet somewhere.

But now I am older and wiser and have fixed all my past mistakes, so the critique of those past mistakes no longer applies to my new arguments.

OpenPhil: Uh huh.

Eliezer: I mean, I did try to fix all the mistakes that I knew about, and didn't just, like, leave those mistakes in forever? I realize that this claim to be able to "learn from experience" is not standard human behavior in situations like this, but if you've got to be weird, that's a good place to spend your weirdness points. At least by my own lights, I am now making a different argument than I made when I was nineteen years old, and that different argument should be considered differently.

And, yes, I also think my nineteen-year-old self was not completely foolish at least about AI timelines; in the sense that, for all he knew, maybe you *could* build AGI by 2005 if you tried really hard over the next 6 years. Not so much because Moravec's estimate should've been seen as a predictable overestimate of how much computing power would actually be needed, given knowledge that would become available in the

next 6 years; but because Moravec's estimate should've been seen as *almost entirely irrelevant*, making the correct answer be "I don't know."

OpenPhil: It seems to us that Moravec's estimate, and the guess of your nineteen-year-old past self, are *both* predictably vast underestimates. Estimating the computation consumed by one brain, and calling that your AGI target date, is obviously predictably a vast underestimate because it neglects the computation required for *training* a brainlike system. It may be a bit uncharitable, but we suggest that Moravec and your nineteen-year-old self may both have been motivatedly credulous, to not notice a gap so very obvious.

Eliezer: I could imagine it seeming that way if you'd grown up never learning about any AI techniques except deep learning, which had, in your wordless mental world, always been the way things were, and would always be that way forever.

I mean, it could be that deep learning *will* still be the bleeding-edge method of Artificial Intelligence right up until the end of the world. But if so, it'll be because Vinge was right and the world ended before 2030, *not* because the deep learning paradigm was as good as any AI paradigm can ever get. That is simply not a kind of thing that I expect Reality to say "Gotcha" to me about, any more than I expect to be told that the human brain, whose neurons and synapses are 500,000 times further away from the thermodynamic efficiency wall than ATP synthase, is the most efficient possible consumer of computations.

The specific perspective-taking operation needed here - when it comes to what was and wasn't obvious in 1988 or 1999 - is that the notion of spending thousands and millions and billions of times as much computation on a "training" phase, as on an "inference" phase, is something that only came to be seen as Always Necessary after the deep learning revolution took over AI in the late Noughties. Back when Moravec was writing, you programmed a game-tree-search algorithm for chess, and then you ran that code, and it played chess. Maybe you needed to add an opening book, or do a lot of trial runs to tweak the exact values the position evaluation function assigned to knights vs. bishops, but most AIs weren't neural nets and didn't get trained on enormous TPU pods.

Moravec had no way of knowing that the paradigm in AI would, twenty years later, massively shift to a new paradigm in which stuff got trained on enormous TPU pods. He lived in a world where you could only train neural networks a few layers deep, like, three layers, and the gradients vanished or exploded if you tried to train networks any deeper.

To be clear, in 1999, I did think of AGIs as needing to do a lot of learning; but I expected them to be learning while thinking, not to learn in a separate gradient descent phase.

OpenPhil: How could anybody possibly miss anything so obvious? There's so many basic technical ideas and even *philosophical ideas about how you do AI* which make it supremely obvious that the best and only way to turn computation into intelligence is to have deep nets, lots of parameters, and enormous separate training phases on TPU pods.

Eliezer: Yes, well, see, those philosophical ideas were not as prominent in 1988, which is why the direction of the future paradigm shift was not *predictable in advance without benefit of hindsight*, let alone timeable to 2006.

You're also probably overestimating how much those philosophical ideas would pinpoint the modern paradigm of gradient descent even if you had accepted them wholeheartedly, in 1988. Or let's consider, say, October 2006, when the Netflix Prize was being run - a watershed occasion where lots of programmers around the world tried their hand at minimizing a loss function, based on a huge-for-the-times 'training set' that had been publicly released, scored on a holdout 'test set'. You could say it was the first moment in the limelight for the sort of problem setup that everybody now takes for granted with ML research: a widely shared dataset, a heldout test set, a loss function to be minimized, prestige for advancing the 'state of the art'. And it was a million dollars, which, back in 2006, was big money for a machine learning prize, garnering lots of interest from competent competitors.

Before deep learning, "statistical learning" was indeed a banner often carried by the early advocates of the view that Richard Sutton now calls the Bitter Lesson, along the lines of "complicated programming of human ideas doesn't work, you have to just learn from massive amounts of data".

But before deep learning - which was barely getting started in 2006 - "statistical learning" methods that took in massive amounts of data, did not use those massive amounts of data to train neural networks by stochastic gradient descent across millions of examples! In 2007, [the winning submission to the Netflix Prize](#) was an ensemble predictor that incorporated k-Nearest-Neighbor, a factorization method that repeatedly globally minimized squared error, two-layer Restricted Boltzmann Machines, and a regression model akin to Principal Components Analysis. Which is all 100% statistical learning driven by relatively-big-for-the-time "big data", and 0% GOFAI. But these methods didn't involve enormous massive training phases in the modern sense.

Back then, if you were doing stochastic gradient descent at all, you were doing it on a much smaller neural network. Not so much because you couldn't afford more compute for a larger neural network, but because wider neural networks didn't help you much and deeper neural networks simply didn't work.

Bleeding-edge statistical learning techniques as late as 2007, to make actual use of big data, had to find other ways to make use of huge amounts of data than gradient descent and backpropagation. Though, I mean, not huge amounts of data by modern standards. The winning submission to the Netflix Prize used an ensemble of 107 models - that's not a misprint for 10^7 , I actually mean 107 - which models were drawn from half a dozen different model classes, then proliferated with slightly different parameters, averaged together to reduce statistical noise.

A modern kid, perhaps, looks at this and thinks: "If you can afford the compute to train 107 models, why not just train one larger model?" But back then, you see, there just *wasn't* a standard way to dump massively more compute into something, and get better results back out. The fact that they had 107 differently parameterized models from a half-dozen families averaged together to reduce noise, was about as well as anyone could do in 2007, at putting more effort in and getting better results back out.

OpenPhil: How quaint and archaic! But that was 13 years ago, before time actually got started and history actually started happening in real life. Now we've got the paradigm which will actually be used to create AGI, in all probability; so estimation methods centered on that paradigm should be valid.

Eliezer: The current paradigm is definitely not the end of the line in principle. I guarantee you that the way superintelligences build cognitive engines is not by training enormous neural networks using gradient descent. Gua-ran-tee it.

The fact that you think you now see a path to AGI, is because today - unlike in 2006 - you have a paradigm that is seemingly willing to entertain having more and more food stuffed down its throat without obvious limit (yet). This is really a quite recent paradigm shift, though, and it is probably not the most efficient possible way to consume more and more food.

You could rather strongly guess, early on, that support vector machines were never going to give you AGI, because you couldn't dump more and more compute into training or running SVMs and get arbitrarily better answers; whatever gave you AGI would have to be something else that could eat more compute productively.

Similarly, since the path through genetic algorithms and recapitulating the whole evolutionary history would have taken a *lot* of compute, it's no wonder that other, more efficient methods of eating compute were developed before then; it was obvious in advance that they must exist, for all that some what-ifed otherwise.

To be clear, it is certain the world will end by more inefficient methods than those that superintelligences would use; since, if superintelligences are making their own AI systems, then the world has already ended.

And it is possible, even, that the world will end by a method as inefficient as gradient descent. But if so, that will be because the world ended too soon for any more efficient paradigm to be developed. Which, on my model, means the world probably ended before say 2040(???). But of course, compared to how much I think I know about what must be more efficiently doable in principle, I think I know far less about the speed of accumulation of real knowledge (not to be confused with proliferation of publications), or how various random-to-me social phenomena could influence the speed of knowledge. So I think I have far less ability to say a confident thing about the *timing* of the next paradigm shift in AI, compared to the *existence and eventualty* of such paradigms in the space of possibilities.

OpenPhil: But if you expect the next paradigm shift to happen in around 2040, shouldn't you confidently predict that AGI has to arrive *after* 2040, because, without that paradigm shift, we'd have to produce AGI using deep learning paradigms, and in that case our own calculation would apply saying that 2040 is relatively early?

Eliezer: No, because I'd consider, say, improved mixture-of-experts techniques that actually work, to be very much *within* the deep learning paradigm; and even a relatively small paradigm shift like that would obviate your calculations, if it produced a more drastic speedup than halving the computational cost over two years.

More importantly, I simply don't believe in your attempt to calculate a figure of 10,000,000,000,000 operations per second for a brain-equivalent deepnet based on biological analogies, or your figure of 10,000,000,000,000 training updates for it. I simply don't believe in it at all. I don't think it's a valid anchor. I don't think it should be used as the median point of a wide uncertain distribution. The first-developed AGI will consume computation in a different fashion, much as it eats energy in a different fashion; and "how much computation an AGI needs to eat compared to a human brain" and "how many watts an AGI needs to eat compared to a human brain" are equally always decreasing with the technology and science of the day.

OpenPhil: Doesn't our calculation at least provide a soft *upper bound* on how much computation is required to produce human-level intelligence? If a calculation is able to produce an upper bound on a variable, how can it be uninformative about that variable?

Eliezer: You assume that the architecture you're describing can, in fact, work at all to produce human intelligence. This itself strikes me as not only tentative but probably false. I mostly suspect that if you take the exact GPT architecture, [scale it up](#) to what you calculate as human-sized, and start training it using current gradient descent techniques... what mostly happens is that it saturates and asymptotes its loss function at not very far beyond the GPT-3 level - say, it behaves like GPT-4 would, but not much better.

This is what should have been told to Moravec: "Sorry, even if your biology is correct, the assumption that future people can put in X amount of compute and get out Y result is not something you really know." And that point did in fact just completely trash his ability to predict and time the future.

The same must be said to you. Your model contains supposedly known parameters, "how much computation an AGI must eat per second, and how many parameters must be in the trainable model for that, and how many examples are needed to train those parameters". Relative to whatever method is actually first used to produce AGI, I expect your estimates to be wildly inapplicable, as wrong as Moravec was about thinking in terms of just using one supercomputer powerful enough to be a brain. Your parameter estimates may not be about properties that the first successful AGI design even *has*. Why, what if it contains a significant component that *isn't a neural network*? I realize this may be scarcely conceivable to somebody from the present generation, but the world was not always as it was now, and it will change if it does not end.

OpenPhil: I don't understand how some of your reasoning could be internally consistent even on its own terms. If, according to you, our 2050 estimate doesn't provide a soft upper bound on AGI arrival times - or rather, if our 2050-centered probability distribution isn't a soft upper bound on reasonable AGI arrival probability distributions - then I don't see how you can claim that the 2050-centered distribution is predictably a directional overestimate.

You can either say that our forecasted pathway to AGI or something very much like it would *probably work in principle without requiring very much more computation* than our uncertain model components take into account, meaning that the probability distribution provides a soft upper bound on reasonably-estimable arrival times, *but that paradigm shifts will predictably provide an even faster way to do it before then*. That is, you could say that our estimate is both a soft upper bound and also a directional overestimate. Or, you could say that our ignorance of how to create AI will consume *more* than one order-of-magnitude of increased computation cost above biology -

Eliezer: Indeed, much as your whole proposal would supposedly cost ten trillion times the equivalent computation of the single human brain that earlier biologically-inspired estimates anchored on.

OpenPhil: - in which case our 2050-centered distribution is not a good soft upper bound, but *also* not predictably a directional overestimate. Don't you have to pick one or the other as a critique, there?

Eliezer: Mmm... there's some justice to that, now that I've come to write out this part of the dialogue. Okay, let me revise my earlier stated opinion: I think that your biological estimate is a trick that never works and, *on its own terms*, would tell us very little about AGI arrival times at all. *Separately*, I think from my own model that your timeline distributions happen to be too long.

OpenPhil: *Eliezer*.

Eliezer: I mean, in fact, part of my actual sense of indignation at this whole affair, is the way that Platt's law of strong AI forecasts - which was *in the 1980s* generalizing "thirty years" as the time that ends up sounding "reasonable" to would-be forecasters - is *still* exactly in effect for what ends up sounding "reasonable" to would-be futurists, *in fricking 2020* while the air is filling up with AI smoke in [the silence of nonexistent fire alarms](#).

But to put this in terms that maybe possibly you'd find persuasive:

The last paradigm shifts were from "write a chess program that searches a search tree and run it, and that's how AI eats computing power" to "use millions of data samples, but *not* in a way that requires a huge separate training phase" to "train a huge network for zillions of gradient descent updates and then run it". This new paradigm costs a lot more compute, but (small) large amounts of compute are now available so people are using them; and this new paradigm saves on programmer labor, and more importantly the need for programmer knowledge.

I say with surety that this is not the last *possible* paradigm shift. And furthermore, the [Stack More Layers](#) paradigm has already reduced need for knowledge by what seems like a pretty large bite out of all the possible knowledge that could be thrown away.

So, you might then argue, the world-ending AGI seems more likely to incorporate more knowledge and less brute force, which moves the correct sort of timeline estimate *further* away from the direction of "cost to recapitulate all evolutionary history as pure blind search without even the guidance of gradient descent" and *more* toward the direction of "computational cost of one brain, if you could just make a single brain".

That is, you can think of there as being *two* biological estimates to anchor on, not just one. You can imagine there being a balance that shifts over time from "the computational cost for evolutionary biology to invent brains" to "the computational cost to run one biological brain".

In 1960, maybe, they knew so little about how brains worked that, if you gave them a hypercomputer, the cheapest way they could quickly get AGI out of the hypercomputer using just their current knowledge, would be to run a massive evolutionary tournament over computer programs until they found smart ones, using 10^{43} operations.

Today, you know about gradient descent, which finds programs more efficiently than genetic hill-climbing does; so the balance of how much hypercomputation you'd need to use to get general intelligence using just your own personal knowledge, has shifted ten orders of magnitude away from the computational cost of evolutionary history and towards the lower bound of the computation used by one brain. In the future, this balance will predictably swing even further towards Moravec's biological anchor, further away from Somebody on the Internet's biological anchor.

I admit, from my perspective this is nothing but a clever argument that tries to persuade people who are making errors that can't all be corrected by me, so that they can make mostly the same errors but get a slightly better answer. In my own mind I tend to contemplate the Textbook from the Future, which would tell us how to build AI on a home computer from 1995, as my anchor of 'where can progress go', rather than looking to the *brain* of all computing devices for inspiration.

But, if you insist on the error of anchoring on biology, you could perhaps do better by seeing a spectrum between two bad anchors. This lets you notice a changing reality, at all, which is why I regard it as a helpful thing to say to you and not a pure persuasive superweapon of unsound argument. Instead of just fixating on one bad anchor, the hybrid of biological anchoring with whatever knowledge you currently have about optimization, you can notice how reality seems to be *shifting between* two biological bad anchors over time, and so have an eye on the changing reality at all. Your new estimate in terms of gradient descent is stepping away from evolutionary computation and toward the individual-brain estimate by ten orders of magnitude, using the fact that you now know a *little* more about optimization than natural selection knew; and now that you can see the change in reality over time, in terms of the two anchors, you can wonder if there are more shifts ahead.

Realistically, though, I would *not* recommend eyeballing how much more knowledge you'd think you'd need to get even larger shifts, as some function of time, before that line crosses the hardware line. Some researchers may already know Thielian secrets you do not, that take those researchers further toward the individual-brain computational cost (if you insist on seeing it that way). That's the direction that economics rewards innovators for moving in, and you don't know everything the innovators know in their labs.

When big inventions finally hit the world as newspaper headlines, the people two years before that happens are often declaring it to be fifty years away; and others, of course, are declaring it to be two years away, fifty years before headlines. Timing things is quite hard even when you think you are being clever; and cleverly having two biological anchors and eyeballing Reality's movement between them, is not the sort of cleverness that gives you good timing information in real life.

In real life, Reality goes off and does something else instead, and the Future does not look in that much detail like the futurists predicted. In real life, we come back again to the same wiser-but-sadder conclusion given at the start, that in fact the Future is quite hard to foresee - especially when you are not on literally the world's leading edge of technical knowledge about it, but really even then. If you don't think you know any Thielian secrets about timing, you should just figure that you need a general policy which doesn't get more than two years of warning, or not even that much if you aren't closely non-dismissively analyzing warning signs.

OpenPhil: We do consider in our report the many ways that our estimates could be wrong, and show multiple ways of producing biologically inspired estimates that give different results. Does that give us any credit for good epistemology, on your view?

Eliezer: I wish I could say that it probably beats showing a single estimate, in terms of its impact on the reader. But in fact, writing a huge careful Very Serious Report like that and snowing the reader under with Alternative Calculations is probably going to cause them to give *more* authority to the whole thing. It's all very well to note the Ways I Could Be Wrong and to confess one's Uncertainty, but you did not actually reach the conclusion, "And that's enough uncertainty and potential error that we

should throw out this whole deal and start over," and that's the conclusion you needed to reach.

OpenPhil: It's not clear to us what better way you think exists of arriving at an estimate, compared to the methodology we used - in which we do consider many possible uncertainties and several ways of generating probability distributions, and try to combine them together into a final estimate. A Bayesian needs a probability distribution from somewhere, right?

Eliezer: If somebody had calculated that it currently required an IQ of 200 to destroy the world, that the smartest current humans had an IQ of around 190, and that the world would therefore start to be destroyable in fifteen years according to Moore's Law of Mad Science - then, even assuming Moore's Law of Mad Science to actually hold, the part where they throw in an estimated current IQ of 200 as necessary is complete garbage. It is not the sort of mistake that can be repaired, either. No, not even by considering many ways you could be wrong about the IQ required, or considering many alternative different ways of estimating present-day people's IQs.

The correct thing to do with the entire model is chuck it out the window so it doesn't exert an undue influence on your actual thinking, where any influence of that model is an undue one. And then you just *should not expect good advance timing info until the end is in sight*, from whatever thought process you adopt instead.

OpenPhil: What if, uh, somebody knows a Thielian secret, or has... narrowed the rivers of their knowledge to closer to reality's tracks? We're not sure exactly what's supposed to be allowed, on your worldview; but wasn't there something at the beginning about how, when you're unsure, you should be careful about criticizing people who are more unsure than you?

Eliezer: Hopefully those people are also able to tell you bold predictions about the nearer-term future, or at least say *anything* about what the future looks like before the whole world ends. I mean, you don't want to go around proclaiming that, because you don't know something, nobody else can know it either. But timing is, in real life, really hard as a prediction task, so, like... I'd expect them to be able to predict a bunch of stuff before the final hours of their prophecy?

OpenPhil: We're... not sure we see that? We may have made an estimate, but we didn't make a narrow estimate. We gave a relatively wide probability distribution as such things go, so it doesn't seem like a great feat of timing that requires us to also be able to predict the near-term future in detail too?

Doesn't *your* implicit probability distribution have a median? Why don't you also need to be able to predict all kinds of near-term stuff if you have a probability distribution with a median in it?

Eliezer: I literally have not tried to force my brain to give me a median year on this - not that this is a defense, because I still have some implicit probability distribution, or, to the extent I don't act like I do, I must be acting incoherently in self-defeating ways. But still: I feel like you should probably have nearer-term bold predictions if your model is supposedly so solid, so concentrated as a flow of uncertainty, that it's coming up to you and whispering numbers like "2050" even as the median of a broad distribution. I mean, if you have a model that can actually, like, calculate stuff like that, and is actually bound to the world as a truth.

If you are an aspiring Bayesian, perhaps, you may try to reckon your uncertainty into the form of a probability distribution, even when you face "structural uncertainty" as we sometimes call it. Or if you know the laws of [coherence](#), you will acknowledge that your planning and your actions are implicitly showing signs of weighing some paths through time more than others, and hence display probability-estimating behavior whether you like to acknowledge that or not.

But if you are a wise aspiring Bayesian, you will admit that whatever probabilities you are using, they are, in a sense, intuitive, and you just don't expect them to be all that good. Because the timing problem you are facing is a really hard one, and humans are not going to be great at it - not until the end is near, and maybe not even then.

That - not "you didn't consider enough alternative calculations of your target figures" - is what should've been replied to Moravec in 1988, if you could go back and tell him where his reasoning had gone wrong, and how he might have reasoned differently based on what he actually knew at the time. That reply I now give to you, unchanged.

Humbali: And I'm back! Sorry, I had to take a lunch break. Let me quickly review some of this recent content; though, while I'm doing that, I'll go ahead and give you what I'm pretty sure will be my reaction to it:

Ah, but here is a point that you seem to have not considered at all, namely: *what if you're wrong?*

Eliezer: That, Humbali, is a thing that should be said mainly to children, of whatever biological wall-clock age, who've never considered at all the possibility that they might be wrong, and who will genuinely benefit from asking themselves that. It is not something that should often be said between grownups of whatever age, as I define what it means to be a grownup. You will mark that I did not at any point say those words to Imaginary Moravec or Imaginary OpenPhil; it is not a good thing for grownups to say to each other, or to think to themselves in Tones of Great Significance (as opposed to as a routine check).

It is very easy to worry that one might be wrong. Being able to see the *direction* in which one is *probably* wrong is rather a more difficult affair. And even after we see a probable directional error and update our views, the objection, "But what if you're wrong?" will sound just as forceful as before. For this reason do I say that such a thing should not be said between grownups -

Humbali: Okay, done reading now! Hm... So it seems to me that the possibility that you are wrong, considered in full generality and without adding any other assumptions, should produce a directional shift from your viewpoint towards OpenPhil's viewpoint.

Eliezer (sighing): And how did you end up being under the impression that this could possibly be a sort of thing that was true?

Humbali: Well, I get the impression that you have timelines shorter than OpenPhil's timelines. Is this devastating accusation true?

Eliezer: I consider naming particular years to be a cognitively harmful sort of activity; I have refrained from trying to translate my brain's native intuitions about this into probabilities, for fear that my verbalized probabilities will be stupider than my intuitions if I try to put weight on them. What feelings I do have, I worry may be unwise to voice; AGI timelines, in my own experience, are not great for one's mental

health, and I worry that other people seem to have weaker immune systems than even my own. But I suppose I cannot but acknowledge that my outward behavior seems to reveal a distribution whose median seems to fall well before 2050.

Humbali: Okay, so you're more confident about your AGI beliefs, and OpenPhil is less confident. Therefore, to the extent that you might be wrong, the world is going to look more like OpenPhil's forecasts of how the future will probably look, like world GDP doubling over four years before the first time it doubles over one year, and so on.

Eliezer: You're going to have to explain some of the intervening steps in that line of 'reasoning', if it may be termed as such.

Humbali: I feel surprised that I should have to explain this to somebody who supposedly knows probability theory. If you put higher probabilities on AGI arriving in the years before 2050, then, on average, you're concentrating more probability into each year that AGI might possibly arrive, than OpenPhil does. Your probability distribution has lower entropy. We can literally just calculate out that part, if you don't believe me. So to the extent that you're wrong, it should shift your probability distributions in the direction of maximum entropy.

Eliezer: It's things like this that make me worry about whether that extreme cryptivist view would be correct, in which normal modern-day Earth intellectuals are literally not smart enough - in a sense that includes the Cognitive Reflection Test and other things we don't know how to measure yet, not just raw IQ - to be taught more advanced ideas from my own home planet, like Bayes's Rule and the concept of the entropy of a probability distribution. Maybe it does them net harm by giving them more advanced tools they can use to shoot themselves in the foot, since it causes an explosion in the total possible complexity of the argument paths they can consider and be fooled by, which may now contain words like 'maximum entropy'.

Humbali: If you're done being vaguely condescending, perhaps you could condescend specifically to refute my argument, which seems to me to be airtight; my math is not wrong and it means what I claim it means.

Eliezer: The audience is herewith invited to first try refuting Humbali on their own; grandpa is, in actuality and not just as a literary premise, getting older, and was never that physically healthy in the first place. If the next generation does not learn how to do this work without grandpa hovering over their shoulders and prompting them, grandpa cannot do all the work himself. There is an infinite supply of slightly different wrong arguments for me to be forced to refute, and that road does not seem, in practice, to have an end.

Humbali: Or perhaps it's you that needs refuting.

Eliezer, smiling: That does seem like the sort of thing I'd do, wouldn't it? Pick out a case where the other party in the dialogue had made a valid point, and then ask my readers to disprove it, in case they weren't paying proper attention? For indeed in a case like this, one first backs up and asks oneself "Is Humbali right or not?" and not "How can I prove Humbali wrong?"

But now the reader should stop and contemplate that, if they are going to contemplate that at all:

Is Humbali right that generic uncertainty about maybe being wrong, without other extra premises, should increase the entropy of one's probability distribution over AGI,

thereby moving out its median further away in time?

Humbali: Are you done?

Eliezer: Hopefully so. I can't see how else I'd prompt the reader to stop and think and come up with their own answer first.

Humbali: Then what is the supposed flaw in my argument, if there is one?

Eliezer: As usual, when people are seeing only their preferred possible use of an argumentative superweapon like 'What if you're wrong?', the flaw can be exposed by showing that the argument Proves Too Much. If you forecasted AGI with a probability distribution with a median arrival time of 50,000 years from now*, would that be very unconfident?

(*) Based perhaps on an ignorance prior for how long it takes for a sapient species to build AGI after it emerges, where we've observed so far that it must take at least 50,000 years, and our updated estimate says that it probably takes around as much more longer than that.

Humbali: Of course; the math says so. Though I think that would be a little *too* unconfident - we do have *some* knowledge about how AGI might be created. So my answer is that, yes, this probability distribution is higher-entropy, but that it reflects too little confidence even for me.

I think you're crazy overconfident, yourself, and in a way that I find personally distasteful to boot, but that doesn't mean I advocate zero confidence. I try to be less arrogant than you, but my best estimate of what my own eyes will see over the next minute is not a maximum-entropy distribution over visual snow. AGI happening sometime in the next century, with a median arrival time of maybe 30 years out, strikes me as being about as confident as somebody should reasonably be.

Eliezer: Oh, really now. I think if somebody sauntered up to you and said they put 99% probability on AGI not occurring within the next 1,000 years - which is the sort of thing a median distance of 50,000 years tends to imply - I think you would, in fact, accuse them of brash overconfidence about staking 99% probability on that.

Humbali: Hmm. I want to deny that - I have a strong suspicion that you're leading me down a garden path here - but I do have to admit that if somebody walked up to me and declared only a 1% probability that AGI arrives in the next millennium, I would say they were being overconfident and not just too uncertain.

Now that you put it that way, I think I'd say that somebody with a wide probability distribution over AGI arrival spread over the next century, with a median in 30 years, is in realistic terms about as uncertain as anybody could possibly be? If you spread it out more than that, you'd be declaring that AGI probably *wouldn't* happen in the next 30 years, which seems overconfident; and if you spread it out less than that, you'd be declaring that AGI probably *would* happen within the next 30 years, which also seems overconfident.

Eliezer: Uh huh. And to the extent that I am myself uncertain about my own brashly arrogant and overconfident views, I should have a view that looks more like your view instead?

Humbali: Well, yes! To the extent that you are, yourself, less than totally certain of your own model, you should revert to this most ignorant possible viewpoint as a base rate.

Eliezer: And if my own viewpoint should happen to regard your probability distribution putting its median on 2050 as just one more guesstimate among many others, with this particular guess based on wrong reasoning that I have justly rejected?

Humbali: Then you'd be overconfident, obviously. See, you don't get it, what I'm presenting is not just one candidate way of thinking about the problem, it's the *base rate* that other people should fall back on to the extent they are not completely confident in *their own ways* of thinking about the problem, which impose *extra assumptions* over and above the assumptions that seem natural and obvious to me. I just can't understand the incredible arrogance you use as to be so utterly certain in your own exact estimate that you don't revert it even a little bit towards mine.

I don't suppose you're going to claim to me that you first constructed an even more confident first-order estimate, and then reverted it towards the natural base rate in order to arrive at a more humble second-order estimate?

Eliezer: Ha! No. Not that base rate, anyways. I try to shift my AGI timelines a little further out because I've observed that actual Time seems to run slower than my attempts to eyeball it. I did not shift my timelines out towards 2050 in particular, nor did reading OpenPhil's report on AI timelines influence my first-order or second-order estimate at all, in the slightest; no more than I updated the slightest bit back when I read the estimate of 10^{43} ops or 10^{46} ops or whatever it was to recapitulate evolutionary history.

Humbali: Then I can't imagine how you could possibly be so perfectly confident that you're right and everyone else is wrong. Shouldn't you at least revert your viewpoints some toward what other people think?

Eliezer: Like, what the person on the street thinks, if we poll them about their expected AGI arrival times? Though of course I'd have to poll everybody on Earth, not just the special case of developed countries, if I thought that a respect for somebody's personhood implied deference to their opinions.

Humbali: Good heavens, no! I mean you should revert towards the opinion, either of myself, or of the set of people I hang out with and who are able to exert a sort of unspoken peer pressure on me; that is the natural reference class to which less confident opinions ought to revert, and any other reference class is special pleading.

And before you jump on me about being arrogant myself, let me say that I definitely regressed my own estimate in the direction of the estimates of the sort of people I hang out with and instinctively regard as fellow tribesmembers of slightly higher status, or "credible" as I like to call them. Although it happens that those people's opinions were about evenly distributed to both sides of my own - maybe not statistically exactly for the population, I wasn't keeping exact track, but in their availability to my memory, definitely, other people had opinions on both sides of my own - so it didn't move my median much. But so it sometimes goes!

But these other people's credible opinions *definitely* hang emphatically to one side of *your* opinions, so your opinions should regress at least a *little* in that direction! Your self-confessed failure to do this *at all* reveals a ridiculous arrogance.

Eliezer: Well, I mean, in fact, from my perspective, even my complete-idiot sixteen-year-old self managed to notice that AGI was going to be a big deal, many years before various others had been hit over the head with a large-enough amount of evidence that even they started to notice. I was walking almost alone back then. And I still largely see myself as walking alone now, as accords with the Law of Continued Failure: If I was going to be living in a world of sensible people in this future, I should have been living in a sensible world already in my past.

Since the early days more people have caught up to earlier milestones along my way, enough to start publicly arguing with me about the further steps, but I don't consider them to have caught up; they are moving slower than I am still moving now, as I see it. My actual work these days seems to consist mainly of trying to persuade allegedly smart people to not fling themselves directly into lava pits. If at some point I start regarding you as my epistemic peer, I'll let you know. For now, while I endeavor to be swayable by arguments, your existence alone is not an argument unto me.

If you choose to define that with your word "arrogance", I shall shrug and not bother to dispute it. Such appellations are beneath My concern.

Humbali: Fine, you admit you're arrogant - though I don't understand how that's not just admitting you're irrational and wrong -

Eliezer: They're different words that, in fact, mean different things, in their semantics and not just their surfaces. I do not usually advise people to contemplate the mere meanings of words, but perhaps you would be well-served to do so in this case.

Humbali: - but if you're not *infinitely* arrogant, you should be quantitatively updating at least a *little* towards other people's positions!

Eliezer: You do realize that OpenPhil itself hasn't always existed? That they are not the only "other people" that there are? An ancient elder like myself, who has seen many seasons turn, might think of many other possible targets toward which he should arguably regress his estimates, if he was going to start deferring to others' opinions this late in his lifespan.

Humbali: You haven't existed through infinite time either!

Eliezer: A glance at the history books should confirm that I was not around, yes, and events went accordingly poorly.

Humbali: So then... why aren't you regressing your opinions at least a little in the direction of OpenPhil's? I just don't understand this apparently infinite self-confidence.

Eliezer: The fact that I have credible intervals around my own unspoken median - that I confess I might be wrong in either direction, around my intuitive sense of how long events might take - doesn't count for my being less than infinitely self-confident, on your view?

Humbali: No. You're expressing absolute certainty in your underlying epistemology and your entire probability distribution, by not reverting it even a little in the direction of the reasonable people's probability distribution, which is the one that's the obvious base rate and doesn't contain all the special other stuff somebody would have to tack on to get *your* probability estimate.

Eliezer: Right then. Well, that's a wrap, and maybe at some future point I'll talk about the increasingly lost skill of perspective-taking.

OpenPhil: Excuse us, we have a final question. You're not claiming that we argue like Humbali, are you?

Eliezer: Good heavens, no! That's why "Humbali" is presented as a separate dialogue character and the "OpenPhil" dialogue character says nothing of the sort. Though I did meet one EA recently who seemed puzzled and even offended about how I wasn't regressing my opinions towards OpenPhil's opinions to whatever extent I wasn't totally confident, which brought this to mind as a meta-level point that needed making.

OpenPhil: "One EA you met recently" is not something that you should hold against OpenPhil. We haven't organizationally endorsed arguments like Humbali's, any more than you've ever argued that "we have to take AGI risk seriously even if there's only a tiny chance of it" or similar crazy things that other people hallucinate you arguing.

Eliezer: I fully agree. That Humbali sees himself as defending OpenPhil is not to be taken as associating his opinions with those of OpenPhil; just like how people who helpfully try to defend MIRI by saying "Well, but even if there's a tiny chance..." are not thereby making their epistemic sins into mine.

The whole thing with Humbali is a separate long battle that I've been fighting. OpenPhil seems to have been keeping its communication about AI timelines mostly to the object level, so far as I can tell; and that is a more proper and dignified stance than I've assumed here.

Edit (12/23): Holden replies [here](#).

Reply to Eliezer on Biological Anchors

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

The "[biological anchors](#)" method for forecasting transformative AI is the biggest non-[trust-based](#) input into my thinking about likely timelines for transformative AI. While I'm sympathetic to parts of [Eliezer Yudkowsky's recent post on it](#), I overall disagree with the post, and think it's easy to get a misimpression of the "biological anchors" report (which I'll abbreviate as "**Bio Anchors**") - and Open Philanthropy's take on it - by reading it.

This post has three sections:

- **Most of Eliezer's critique seems directed at assumptions the report explicitly does not make** about how transformative AI will be developed, and more broadly, about the connection between its (the report's) compute estimates and all-things-considered AI timelines. One way of putting this is that most of Eliezer's critique doesn't apply to the "bounding-based" interpretation of the report discussed in [this post](#) (which is my best explanation for skeptics of why I find the framework valuable; I will also give quotes below from the original report showing that its intended interpretation is along the same lines as mine).
- Much of Eliezer's critique is some form of "**Look at the reference class you're in**," invoking "Platt's Law" and comparing the report to past attempts at biological anchoring. Based on my understanding of the forecasts he's comparing it to and the salient alternatives, **I don't think this does much to undermine the report.**
- I also make a few minor points.

A few notes before I continue:

- I think the comments on the post are generally excellent and interesting, and I recommend them. (I will mostly not be repeating things from the comments here.)
- I generally view Bio Anchors as a **tool for informing AI timelines** rather than as a **comprehensive generator of all-things-considered AI timelines**, and will be discussing it as such. Bio Anchors also presents itself this way - see section [Translating into views on TAI timelines](#).
- Something like half of this post is blockquotes. I've often been surprised by the degree to which people (including people I respect a lot, such as Eliezer in this case) seem to [mischaracterize](#) specific pieces they critique, and I try to avoid this for myself by quoting extensively from a piece when critiquing it. (This still leaves the possibility that I'm quoting out of context; readers may want to spot-check that.)
- This post doesn't address what some have referred to as the "[meta-level core thing](#)", though I might write some thoughts related to that in a future post.

Bounding vs. pinpointing

Here are a number of quotes from Eliezer in which I think he gives the impression that Biological Anchors *assumes transformative AI will be arrived at via modern machine*

learning methods:

OpenPhil: Because AGI isn't like biology, and in particular, will be trained using gradient descent instead of evolutionary search, which is cheaper. We do note inside our report that this is a key assumption, and that, if it fails, the estimate might be correspondingly wrong - ...

OpenPhil: Well, search by evolutionary biology is more costly than training by gradient descent, so in hindsight, it was an overestimate. Are you claiming this was predictable in foresight instead of hindsight?

Eliezer: I'm claiming that, at the time, I snorted and tossed Somebody's figure out the window while thinking it was ridiculously huge and absurd, yes.

OpenPhil: Because you'd already foreseen in 2006 that gradient descent would be the method of choice for training future AIs, rather than genetic algorithms?

Eliezer: Ha! No. Because it was an insanely costly hypothetical approach whose main point of appeal, to the sort of person who believed in it, was that it didn't require having any idea whatsoever of what you were doing or how to design a mind.

OpenPhil: Suppose one were to reply: "Somebody" didn't know better-than-evolutionary methods for designing a mind, just as we currently don't know better methods than gradient descent for designing a mind; and hence Somebody's estimate was the best estimate at the time, just as ours is the best estimate now?

...

OpenPhil: It seems to us that Moravec's estimate, and the guess of your nineteen-year-old past self, are *both* predictably vast underestimates. Estimating the computation consumed by one brain, and calling that your AGI target date, is obviously predictably a vast underestimate because it neglects the computation required for *training* a brainlike system. It may be a bit uncharitable, but we suggest that Moravec and your nineteen-year-old self may both have been motivatedly credulous, to not notice a gap so very obvious.

Eliezer: I could imagine it seeming that way if you'd grown up never learning about any AI techniques except deep learning, which had, in your wordless mental world, always been the way things were, and would always be that way forever.

I mean, it could be that deep learning *will* still be the bleeding-edge method of Artificial Intelligence right up until the end of the world. But if so, it'll be because Vinge was right and the world ended before 2030, *not* because the deep learning paradigm was as good as any AI paradigm can ever get. That is simply not a kind of thing that I expect Reality to say "Gotcha" to me about, any more than I expect to be told that the human brain, whose neurons and synapses are 500,000 times further away from the thermodynamic efficiency wall than ATP synthase, is the most efficient possible consumer of computations ...

OpenPhil: How could anybody possibly miss anything so obvious? There's so many basic technical ideas and even *philosophical ideas about how you do AI* which make it supremely obvious that the best and only way to turn computation into intelligence is to have deep nets, lots of parameters, and enormous separate training phases on TPU pods ...

OpenPhil: How quaint and archaic! But that was 13 years ago, before time actually got started and history actually started happening in real life. Now we've got the paradigm which will actually be used to create AGI, in all probability; so estimation methods centered on that paradigm should be valid.

However, the argument given in Bio Anchors does **not** hinge on an assumption that modern deep learning is what will be used, nor does it set aside the possibility of paradigm changes.

From the section [What if TAI is developed through a different path?](#):

I believe that this analysis can provide a useful median estimate even if TAI is produced through a very different path: essentially, by the time it is affordable to develop TAI through a *particular* highlighted route, it is plausible that somebody develops it through that route or *any cheaper route*. I consider the example of a distributed economic transition facilitated by a broad range of different technologies below, but the same reasoning applies to the possibility that a unified transformative program may be developed using a qualitatively different "AI paradigm" that can't be usefully considered a descendant of modern machine learning ...

Because this model estimates when one *particular path* toward transformative AI (let's call it the "big model path") out of many will be attainable, that means **if this analysis is correct** (i.e., if I am correct to assume the big model path is possible at all due to the theoretical feasibility of local search, and if we correctly estimated the probability that it would be attainable in year Y for all Y), **then the probability estimates generated should be underestimates** ...

However, once sources of distortion (many of which tend to push our estimates upward) are properly taken into account, **I think it is fairly unclear whether these estimates should actually be considered underestimates** [one such source given is similar to my comments [here](#) following "When it comes to translating my 'sense of mild surprise' into a probability] ...

For each biological anchor hypothesis, I am acting on the assumption that there is a relatively broad space of "unknown unknown" paths to solving a transformative task within that range of technical difficulty, not just the particular concrete path I have written down for illustration in association with each hypothesis (which is often fairly conjunctive) ...

some of our technical advisors are still relatively confident these probability estimates are low-end estimates. This is partly because they would assign a higher probability to some of the low-end biological anchor hypotheses than I do, partly because they are overall more confident in the argument [given above](#) that these numbers ought to be considered underestimates ...

For now, I feel that the most reasonable way to interpret the probability estimates generated by the biological anchors framework is as a rough central estimate for when TAI will be developed rather than as particularly conservative or particularly aggressive. In making this judgment, I am admittedly mentally running together a large cloud of heterogeneous considerations which in a maximally-principled and transparent analysis should be handled separately.

That is, Ajeya (the author) sees the "median" estimate as structurally likely to be **overly conservative (a soft upper bound) for reasons including those Eliezer**

gives, but is also adjusting in the opposite direction to account for factors including the generic burden of proof. (More discussion of "soft bounds" provided by Bio Anchors in [this section](#) and [this section](#) of the report.)

I made similar arguments in a recent piece, [**"Biological anchors" is about bounding, not pinpointing, AI timelines**](#). This is my best explanation for skeptics of why I find the framework valuable.

As far as I can tell, the only part of Eliezer's piece that addresses an argument along the lines of the "soft bounding" idea is:

OpenPhil: Doesn't our calculation at least provide a soft *upper bound* on how much computation is required to produce human-level intelligence? If a calculation is able to produce an upper bound on a variable, how can it be uninformative about that variable?

Eliezer: You assume that the architecture you're describing can, in fact, work at all to produce human intelligence. This itself strikes me as not only tentative but probably false. I mostly suspect that if you take the exact GPT architecture, [scale it up](#) to what you calculate as human-sized, and start training it using current gradient descent techniques... what mostly happens is that it saturates and asymptotes its loss function at not very far beyond the GPT-3 level - say, it behaves like GPT-4 would, but not much better.

This is what should have been told to Moravec: "Sorry, even if your biology is correct, the assumption that future people can put in X amount of compute and get out Y result is not something you really know." And that point did in fact just completely trash his ability to predict and time the future.

The same must be said to you. Your model contains supposedly known parameters, "how much computation an AGI must eat per second, and how many parameters must be in the trainable model for that, and how many examples are needed to train those parameters". Relative to whatever method is actually first used to produce AGI, I expect your estimates to be wildly inapplicable, as wrong as Moravec was about thinking in terms of just using one supercomputer powerful enough to be a brain. Your parameter estimates may not be about properties that the first successful AGI design even *has*. Why, what if it contains a significant component that *isn't a neural network*? I realize this may be scarcely conceivable to somebody from the present generation, but the world was not always as it was now, and it will change if it does not end.

I don't literally think that the "exact GPT architecture" would work to produce transformative AI, but I think something not too far off would be a strong contender - such that having enough compute to afford this extremely brute-force method, combined with decades more time to produce new innovations and environments, does provide something of a "soft upper bound" on transformative AI timelines.

Another way of putting this is that a slightly modified version of what Eliezer calls "tentative [and] probably false]" seems to me to be "tentative and probably true." There's room for disagreement about this, but this is not where most of Eliezer's piece focused.

While I can't be confident, I also suspect that the person in the [2006 or thereabouts](#) part of Eliezer's piece may have intended to argue for something more like a "(soft) upper bound" than a median estimate.

Finally, I want to point out this quote from Bio Anchors, which reinforces that it is intended as a **tool for informing AI timelines** rather than as a **comprehensive generator of all-things-considered AI timelines**:

This model is not directly estimating the probability of transformative AI, but rather the probability that the amount of computation that would be required to train a transformative model using contemporary ML methods would be attainable for some AI project, assuming that algorithmic progress, spending, and compute prices progress along a “business-as-usual” trajectory ...

How does the probability distribution output by this model relate to TAI timelines? In the very short-term (e.g. 2025), I’d expect this model to overestimate the probability of TAI because it feels especially likely that other elements such as datasets or robustness testing or regulatory compliance will be a bottleneck even if the raw compute is technically affordable, given that a few years is not a lot of time to build up key infrastructure. In the long-term (e.g. 2075), I’d expect it to underestimate the probability of TAI, because it feels especially likely that we would have found an entirely different path to TAI by then.

It seems that Eliezer places higher probability on an “entirely different path” sooner than Bio Anchors, but he does not seem to argue for this (and [see below](#) for why I don’t think it would be a great bet). Instead, he largely argues that the possibility is ignored by Bio Anchors, which is not the case.

Platt's Law and past forecasts

Eliezer writes:

Eliezer: So does the report by any chance say - with however many caveats and however elaborate the probabilistic methods and alternative analyses - that AGI is probably due in about 30 years from now?

OpenPhil: Yes, in fact, our 2020 report’s median estimate is 2050; though, again, with very wide credible intervals around both sides. Is that number significant?

Eliezer: It’s a law generalized by Charles Platt, that any AI forecast will put strong AI thirty years out from when the forecast is made. Vernor Vinge referenced it in the body of his famous 1993 NASA speech, whose abstract begins, “Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended.” ...

OpenPhil: That part about Charles Platt’s generalization is interesting, but just because we unwittingly chose literally exactly the median that Platt predicted people would always choose in consistent error, that doesn’t justify dismissing our work, right? ...

Eliezer: Oh, nice. I was wondering what sort of tunable underdetermined parameters enabled your model to nail the psychologically overdetermined final figure of ‘30 years’ so exactly.

I have a couple issues here.

First, I think Eliezer exaggerates the precision of Platt’s Law and its match to the Bio Anchors projection:

- Some aggregated data for assessing Platt's Law is in [this comment by Matthew Barnett](#) as well as [here](#).
- While Matthew says "Overall I find the law to be pretty much empirically validated, at least by the standards I'd expect from a half in jest Law of Prediction," I don't agree: I don't think an actual trendline on the chart would be particularly close to the Platt's Law line. I think it would, instead, predict that Bio Anchors should point to longer timelines than 30 years out.
- Note that my [own median projection](#) for transformative AI is 40 years, not 30, and I know several people who have much shorter medians (15 years and under) based on their own interpretations of the analysis in the report. So I don't think it's the case that Bio Anchors "automatically" lands one on a particular view, nor that it obviously pushes against timelines as short as Eliezer's. It is a tool for informing AI timelines, and after taking it and other data points into account, Ajeya and I both are estimating longer timelines than Eliezer.

I think a softer "**It's suspicious that Bio Anchors is in the same 'reasonable-sounding' general range ('a few decades') that AI forecasts have been in for a long time**" comment would've been more reasonable than what Eliezer wrote, so from here I'll address that. First, I want to comment on Moravec specifically.

Eliezer characterizes Open Philanthropy as though we think that Hans Moravec's projection was foreseeably silly and overaggressive (see quote above), but now think we have the right approach. This isn't the case.

- On one hand, I do think that if Ajeya or I had been talking with Moravec in 1990, we would've had a further-out median timeline estimate by some amount. This isn't because I think we would've been doing similar estimates to today (we didn't have enough information at the time for this to make much sense), or because I think we would've rejected the framework as irrelevant without today's information. It's simply because we each (myself more than her) have an inclination to apply a fair amount of adjustment in a conservative direction, for generic ["burden of proof" reasons](#), rather than go with the timelines that seem most reasonable based on the report in a vacuum.
- But more importantly, even if we set the above point aside, I simply **don't think it's a mark against Bio Anchors to be in the same reference class as Moravec, and I think his prediction was (according to my views, and more so according to Eliezer's apparent views) impressively good when judged by a reasonable standard and compared to reasonable alternatives.**

To expand on what I mean by a reasonable standard and reasonable alternatives:

- Bio Anchors is, first and foremost, meant as a **tool for updating one's timelines from the place they would naively be after considering broader conventional wisdom and perhaps semi-informative priors**. Re: the former, I'm referring not to surveys of experts or conventional wisdom in futurist circles (both of which are often dismissed outside of these circles), but to what I perceive as most people's "This is nowhere close to happening, ignore it" intuition.
- According to my current views (median expectation of transformative AI around 2060), Moravec's 1988 prediction of 2010-2020 looks *much* better than these alternatives, and even looks impressive. Specifically, it looks impressive by the standards of: "multi-decade forecasting of technologies for which no roadmap exists, with capabilities far exceeding those of anything that exists today." (The

more strongly one expects forecasts in this class to be difficult, the more one should be impressed here, in my view.)

- Eliezer pretty clearly expects shorter timelines than I do, so according to his views, I think Moravec's prediction looks more impressive still (by the standards and alternatives I'm using here). It is implied in the dialogue that Eliezer's median would be somewhere between 2025-2040; if you assume this will turn out to be right, that would make a 1988 prediction of "2010-2020" look extremely good, in my view. (Good enough that, to the extent there's doubt about whether the underlying reasoning is valid or noise, this should be a noticeable update toward the former.)
- I suspect Eliezer has a different picture of the salient context and alternatives here. I suspect that he's mostly operating in a context where it's near-universal to expect transformative AI at least as early as I do; that he has non-biological-anchor-inspired views that point to much shorter timelines; and that a lot of his piece is a reaction to "Humbali" types (whom he notes are distinct from Open Philanthropy) asking him to update away from his detailed short-timelines views.
- I'm sympathetic to that, in the sense that I think Bio Anchors is not very useful for the latter purpose. In particular, perhaps it's helpful for me to say here that **if you think timelines are short for reasons unrelated to biological anchors, I don't think Bio Anchors provides an affirmative argument that you should change your mind.** (I do think it is a useful report for *deconstructing* - or at least clarifying - several specific, biologically inspired short-timelines arguments that have been floating around, none of which I would guess Eliezer has any interest in.) Most of the case I'd make against shorter timelines would come down to a lack of *strong affirmative arguments* plus a nontrivial [burden of proof](#).

Returning to the softened version of Platt's Law: according to my current views on timelines (and more so according to Eliezer's), "a few decades" has been a good range for a prediction to be in for the last few decades (again, keeping in mind what context and alternatives I am using). I think this considerably softens the force of an objection like: "You're forecasting a few decades, as many others have over the last few decades; this in itself undermines your case."

None of the above points constitute arguments for the correctness of Bio Anchors. My point is that "Your prediction is like these other predictions" (the thrust of much of Eliezer's piece) doesn't seem to undermine the argument, partly because the other predictions look broadly good according to both my and Eliezer's current views.

A few other reactions to specific parts

Eliezer: ... The software for a human brain is not going to be 100% efficient compared to the theoretical maximum, nor 10% efficient, nor 1% efficient, even before taking into account the whole thing with parallelism vs. serialism, precision vs. imprecision, or similarly clear low-level differences ...

Eliezer: The makers of AGI aren't going to be doing 10,000,000,000,000 rounds of gradient descent, on entire brain-sized 300,000,000,000,000-parameter models, *algorithmically faster than today*. They're going to get to AGI via some route that *you don't know how to take*, at least if it happens in 2040. If it happens in 2025, it may be via a route that some modern researchers do know how to take, but in this case, of course, your model was also wrong.

On one hand, I think it's a distinct possibility that we're going to see dramatically new approaches to AI development by the time transformative AI is developed.

On the other, I think quotes like this overstate the likelihood in the short-to-medium term.

- Deep learning has been the dominant source of AI breakthroughs for [nearly the last decade](#), and the broader "neural networks" paradigm - while it has come in and out of fashion - has broadly been one of the most-attended-to "contenders" throughout the history of AI research.
- AI research prior to 2012 may have had more frequent "paradigm shifts," but this is probably related to the fact that it was seeing less progress.
- With these two points in mind, it seems off to me to confidently expect a new paradigm to be dominant by 2040 (even conditional on AGI being developed), as the second quote above implies. As for the first quote, I think the implication there is less clear, but I read it as expecting AGI to involve software well over 100x as efficient as the human brain, and I wouldn't bet on that either (in real life, if AGI is developed in the coming decades - not based on what's possible in principle.)

Eliezer: The problem is that *the resource gets consumed differently, so base-rate arguments from resource consumption end up utterly unhelpful in real life*. The human brain consumes around 20 watts of power. Can we thereby conclude that an AGI should consume around 20 watts of power, and that, when technology advances to the point of being able to supply around 20 watts of power to computers, we'll get AGI?

If the world were such that:

- We had some reasonable framework for "power usage" that didn't include gratuitously wasted power, and measured the "power used meaningfully to do computations" in some important sense;
- AI performance seemed to [systematically improve](#) as this sort of power usage increased;
- Power usage was just now coming within a few orders of magnitude of the human brain;
- We were just now starting to see AIs have success with tasks like vision and speech recognition (tasks that seem likely to have been evolutionarily important, and that we haven't found ways to precisely describe GOFAI-style);
- It also looked like AI was starting to have insect-like capabilities somewhere around the time it was consuming insect-level amounts of power;
- And we didn't have some clear candidate for a better metric with similar properties (as I think we do in the case of computations, since the main thing I'd expect increased power usage to be useful for is increased computation);

...Then I would be interested in a Bio Anchors-style analysis of projected power usage. As noted above, I would be interested in this as a tool for analysis rather than as "the way to get my probability distribution." That's also how I'm interested in Bio Anchors (and how it presents itself).

I also think we have some a priori reason to believe that human scientists can "use computations" somewhere near as efficiently as the brain does (software), more than we have reason to believe that human scientists can "use power" somewhere nearly as efficiently as the brain does (hardware).

(As a side note, there is some analysis of how nature vs. humans use power in [this section of Bio Anchors](#).)

Somebody: All of that seems irrelevant to my novel and different argument. I am not foolishly estimating the resources consumed by a single brain; I'm estimating the resources consumed by evolutionary biology to invent brains!

Eliezer: And the humans wracking their own brains and inventing new AI program architectures and deploying those AI program architectures to themselves learn, will consume computations so utterly differently from evolution that there is no point comparing those consumptions of resources. That is the flaw that you share exactly with Moravec, and that is why I say the same of both of you, "This is a kind of thinking that fails to bind upon reality, it doesn't work in real life." I don't care how much painstaking work you put into your estimate of 10^{43} computations performed by biology. It's just not a relevant fact.

It's hard for me to understand how it is not a relevant fact: I think we have good reason to believe that humans can use computations at least as intelligently as evolution did.

I think it's perfectly reasonable to push back on 10^{43} as a *median* estimate, but not as a *number that has some sort of relevance*.

OpenPhil: We have commissioned a Very Serious report on a biologically inspired estimate of how much computation will be required to achieve Artificial General Intelligence, for purposes of forecasting an AGI timeline. ([Summary of report](#).) ([Full draft of report](#).) Our leadership takes this report Very Seriously.

I thought this was a pretty misleading presentation of how Open Philanthropy has communicated about this work. It's true that Open Philanthropy's public communication tends toward a cautious, serious tone (and I think there are good reasons for this); but beyond that, I don't think we do much to convey the sort of attitude implied above. The report's publication announcement was [on LessWrong as a draft report for comment](#), and the report is still in the form of several Google docs. We never did any sort of push to have it treated as a fancy report.

Shulman and Yudkowsky on AI progress

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This post is a transcript of a discussion between Carl Shulman and Eliezer Yudkowsky, following up on [a conversation with Paul Christiano and Ajeya Cotra](#).

Color key:

Chat by Carl and Eliezer Other chat

9.14. Carl Shulman's predictions

[Shulman][20:30]

I'll interject some points re the earlier discussion about how animal data relates to the scaling to AGI' thesis.

1. In humans it's claimed the IQ-job success correlation varies by job, For a scientist or doctor it might be 0.6+, for a low complexity job more like 0.4, or more like 0.2 for sim repetitive manual labor. That presumably goes down a lot with less in the way of hand or focused on low density foods like baleen whales or grazers. If it's 0.1 for animals like orcas or elephants, or 0.05, then there's 4-10x less fitness return to smarts.

2. But they outmass humans by more than 4-10x. Elephants 40x, orca 60x+. Metabolically (20 watts divided by BMR of the animal) the gap is somewhat smaller though, because of metabolic scaling laws (energy scales with 3/4 or maybe 2/3 power so).

https://en.wikipedia.org/wiki/Kleiber%27s_law

If dinosaurs were poikilotherms, that's a 10x difference in energy budget vs a mammal the same size, although there is debate about their metabolism.

3. If we're looking for an innovation in birds and primates, there's some evidence of 'hardware' innovation rather than 'software.' Herculano-Houzel reports in *The Human Advantage* (summarizing much prior work neuron counting) different observational scaling laws for neuron number with brain mass for different animal lineages.

We were particularly interested in cellular scaling differences that might have arisen in primates. If the same rules relating numbers of neurons to brain size in rodents ([6](#))

The brain of the capuchin monkey, for instance, weighing 52 g, contains >3x more neurons in the cerebral cortex and ≈2x more neurons in the cerebellum than the larger brain of the capybara, weighing 76 g.

[Editor's Note: Quote source is "[Cellular scaling rules for primate brains.](#)"]

In rodents brain mass increases with neuron count $n^{1.6}$, whereas it's close to linear ($n^{1.1}$) in primates. For cortex neurons and cortex mass 1.7 and 1.0. In general birds primates are outliers in neuron scaling with brain mass.

Note also that bigger brains with lower neuron density have longer communication time from one side of the brain to the other. So primates and birds can have faster clock speeds for integrated thought than a large elephant or whale with similar neuron counts.

4. Elephants have brain mass ~ 2.5 x human, and 3x neurons, but 98% of those are in the cerebellum (vs 80% in or less in most animals; these are generally the tiniest neurons and seem to do a bunch of fine motor control). Human cerebral cortex has 3x the neurons of the elephant cortex (which has twice the mass). The giant cerebellum seems like controlling the very complex trunk.

<https://nautil.us/issue/35/boundaries/the-paradox-of-the-elephant-brain>

Blue whales get close to human neuron counts with much larger brains.

https://en.wikipedia.org/wiki/List_of_animals_by_number_of_neurons

5. As Paul mentioned, human brain volume correlation with measures of cognitive function after correcting for measurement error on the cognitive side is in the vicinity of 0.3-0.4 (might go a bit higher after controlling for non-functional brain volume variation from removing confounds). The genetic correlation with cognitive function in this study is 0.24:

<https://www.nature.com/articles/s41467-020-19378-5>

So it accounts for a minority of genetic influences on cognitive ability. We'd also expect a bunch of genetic variance that's basically disruptive mutations in mutation-selection balance (e.g. schizophrenia seems to be a result of that, with schizophrenia alleles under negative selection, but a big mutational target, with the standing burden set by the level of fitness penalty for it; in niches with less return to cognition the mutational surface will be cleaned up less frequently and have more standing junk).

Other sources of genetic variance might include allocation of attention/learning (curiosity and thinking about abstractions vs immediate sensory processing/alertness), length of childhood/learning phase, motivation to engage in chains of thought, etc.

Overall I think there's some question about how to account for the full genetic variance, but mapping it onto the ML experience with model size, experience and reward functions being key looks compatible with the biological evidence. I lean towards it, although it's not cleanly and conclusively shown.

Regarding economic impact of AGI, I do not buy the 'regulation strangles all big GDP boosts' story.

The BEA breaks down US GDP by industry here (page 11):

https://www.bea.gov/sites/default/files/2021-06/gdp1q21_3rd_1.pdf

As I work through sectors and the rollout of past automation I see opportunities for large scale rollout that is not heavily blocked by regulation. Manufacturing is still trillions of dollars, and robotic factories are permitted and produced under current law, with the limits being more about which tasks the robots work for at low enough cost (e.g. this stopped Tesla plans for more completely robotic factories). Also worth noting manufacturing is mobile and new factories are sited in friendly jurisdictions.

Software to control agricultural machinery and food processing is also permitted.

Warehouses are also low-regulation environments with logistics worth hundreds of billions of dollars. See Amazon's robot-heavy warehouses limited by robotics software.

Driving is hundreds of billions of dollars, and Tesla has been permitted to use Autopilot and there has been a lot of regulator enthusiasm for permitting self-driving cars with humanlike accident rates. Waymo still hasn't reached that it seems and is lowering costs.

Restaurants/grocery stores/hotels are around a trillion dollars. Replacing humans in vision/voice tasks to take orders, track inventory (Amazon Go style), etc is worth hundreds of billions there and mostly permitted. Robotics cheap enough to replace low wage labor there would also be valuable (although a lower priority than high-wage work). Compute and development costs are similar.

Software is close to a half trillion dollars and the internals of software development are almost wholly unregulated.

Finance is over a trillion dollars, with room for AI in sales and management.

Sales and marketing are big and fairly unregulated.

In highly regulated and licensed professions like healthcare and legal services, you can still see a licensee mechanically administer the advice of the machine, amplifying their reach and productivity.

Even in housing/construction there's still great profits to be made by improving the efficiency of what construction is allowed (a sector worth hundreds of billions).

If you're talking about legions of super charismatic AI chatbots, they could be doing some coaching human manual labor to effectively upskill it, and providing the variety of activities discussed above. They're enough to more than double GDP, even with strong Baumol effects/cost disease, I'd say.

Although of course if you have AIs that can do so much the wages of AI and hardware researchers will be super high, and so a lot of that will go into the intelligence explosion while before that various weaknesses that prevent full automation of AI research will a mess up activity in these other sectors to varying degrees.

Re discontinuity and progress curves, I think Paul is right. AI Impacts went to a lot of effort assembling datasets looking for big jumps on progress plots, and indeed nukes are an extremely high percentile for discontinuity, and were developed by the biggest spender in power (yes other powers could have bet more on nukes, but didn't, and that was related to the US having more to spend and putting more in many bets), with the big gains in military power per \$ coming with the hydrogen bomb and over the next decade.

<https://aiimpacts.org/category/takeoff-speed/continuity-of-progress/discontinuous-progress-investigation/>

For measurable hardware and software progress (Elo in games, loss on defined benchmarks), you have quite continuous hardware progress, and software progress that is on the same ballpark, and not drastically jumpy (like 10 year gains in 1), moreso as you get to metrics used by bigger markets/industries.

I also agree with Paul's description of the prior Go trend, and how DeepMind increased its spending on Go software enormously. That analysis was a big part of why I bet on AlphaGo winning against Lee Sedol at the time (the rest being extrapolation from the Fan Hui version and models of DeepMind's process for deciding when to try a match).

[Yudkowsky][21:38]

I'm curious about how much you think these opinions have been arrived at independently by yourself, Paul, and the rest of the OpenPhil complex?

[Cotra][21:44]

Little of Open Phil's opinions are independent of Carl, the source of all opinions

[Yudkowsky: 😊] [Ngo: 😊]

[Shulman][21:44]

I did the brain evolution stuff a long time ago independently. Paul has heard my points that front, and came up with some parts independently. I wouldn't attribute that to anyone else in that 'complex.'

On the share of the economy those are my independent views.

On discontinuities, that was my impression before, but the additional AI Impacts data collection narrowed my credences.

TBC on the brain stuff I had the same evolutionary concern as you, which was I investigated those explanations and they still are not fully satisfying (without more mid-level data opening the black box of non-brain volume genetic variance and evolution over time).

[Yudkowsky][21:50]

so... when I imagine trying to deploy this style of thought myself to predict the recent past without benefit of hindsight, it returns a lot of errors. perhaps this is because I do know how to use this style of thought, but.

for example, I feel like if I was GPT-continuing your reasoning from the great opportunity still available in the world economy, in early 2020, it would output text like:

"There are many possible regulatory regimes in the world, some of which would permit rapid construction of mRNA-vaccine factories well in advance of FDA approval. Given the overall urgency of the pandemic some of those extra-USA vaccines would be sold to individuals or a few countries like Israel willing to pay high prices for them, which would provide evidence of efficacy and break the usual impulse towards regulatory uniformity among developed countries, not to mention the existence of less developed countries who could potentially pay smaller but significant amounts for vaccines. The FDA doesn't seem likely to actively ban testing; they might under a Democratic regime, but Trump already somewhat ideologically prejudiced against the FDA and would go along with the probable advice of his advisors, or just his personal impulse, to override any FDA action that seemed liable to prevent tests and vaccines from making the problem just go away."

[Shulman][21:59]

Pharmaceuticals is a top 10% regulated sector, which is seeing many startups trying to apply AI to drug design (which has faced no regulatory barriers), which fits into the ordinary observed output of the sector. Your story is about regulation failing to improve relative to normal more than it in fact did (which is a dramatic shift, although abysmal relative to what would be reasonable).

That said, I did lose a 50-50 bet on US control of the pandemic under Trump (although also correctly bet that vaccine approval and deployment would be historically unprecedently fast and successful due to the high demand).

[Yudkowsky][22:02]

it's not impossible that Carl/Paul-style reasoning about the future - near future, or indefinitely later future? - would start to sound more reasonable to me if you tried write out a modal-average concrete scenario that was full of the same disasters found in history books and recent news

like, maybe if hypothetically I knew how to operate this style of thinking, I would know how to add disasters automatically and adjust estimates for them; so you don't need to say that to Paul, who also hypothetically knows

but I do not know how to operate this style of thinking, so I look at your description of world economy and it seems like an endless list of cheerfully optimistic ingredients and the recipe doesn't say how many teaspoons of disaster to add or how long to cook it or how it affects the final taste

[Shulman][22:06]

Like when you look at historical GDP stats and AI progress they are made up of a norm rate of insanity and screwups.

[Ngo: ]

[Yudkowsky][22:07]

on my view of reality, I'm the one who expects business-as-usual in GDP until shortly before the world ends, if indeed business-as-usual-in-GDP changes at all, and you have an optimistic recipe for Not That which doesn't come with an example execution containing typical disasters?

[Shulman][22:07]

Things like failing to rush through neural network scaling over the past decade to the point of financial limitation on model size, insanity on AI safety, anti-AI regulation being driven by social media's role in politics.

[Yudkowsky][22:09]

failing to deploy 99% robotic cars to new cities using fences and electronic gates

[Shulman][22:09]

Historical growth has new technologies and stupid stuff messing it up.

[Yudkowsky][22:09]

so many things one could imagine doing with current tech, and yet, they are not done anywhere on Earth

[Shulman][22:09]

AI is going to be incredibly powerful tech, and after a historically typical haircut it's still lot bigger.

[Yudkowsky][22:09]

so some of this seems obviously driven by longer timelines in general

do you have things which, if they start to happen soonish and in advance of world GDF having significantly broken upward 3 years before then, cause you to say "oh no I'm in the Eliezerverse"?

[Shulman][22:12]

You may be confusing my views and Paul's.

[Yudkowsky][22:12]

"AI is going to be incredibly powerful tech" sounds like long timelines to me, though?

[Shulman][22:13]

No.

[Yudkowsky][22:13]

like, "incredibly powerful tech for longer than 6 months which has time to enter the economy"

if it's "incredibly powerful tech" in the sense of immediately killing everybody then of course we agree, but that didn't seem to be the context

[Shulman][22:15]

I think broadly human-level AGI means intelligence explosion/end of the world in less than a year, but tons of economic value is likely to leak out before that from the combination of worse general intelligence with AI advantages like huge experience.

[Yudkowsky][22:15]

my worldview permits but does not mandate a bunch of weirdly powerful shit that people can do a couple of years before the end, because that would sound like a typically me and chaotic history-book scenario especially if it failed to help us in any way

[Shulman][22:15]

And the economic impact is increasing superlinearly (as later on AI can better manage own introduction and not be held back by human complementarities on both the production side and introduction side).

[Yudkowsky][22:16]

my worldview also permits but does not mandate that you get up to the chimp level, chimps are not very valuable, and once you can do fully AGI thought it compounds very quickly

it feels to me like the Paul view wants something narrower than that, a specific story about a great economic boom, and it sounds like the Carl view wants something that from my perspective seems similarly narrow

which is why I keep asking "can you perhaps be specific about what would count as No That and thereby point to the Eliezerverse"

[Shulman][22:18]

We're in the Eliezerverse with huge kinks in loss graphs on automated programming/Putnam problems.

Not from scaling up inputs but from a local discovery that is much bigger in impact than the sorts of jumps we observe from things like Transformers.

[Yudkowsky][22:19]

...my model of Paul didn't agree with that being a prophecy-distinguishing sign to first order (to second order, my model of Paul agrees with Carl for reasons unbeknownst to me)

I don't think you need something very much bigger than Transformers to get sharp loss drops?

[Shulman][22:19]

not the only disagreement

but that is a claim you seem to advance that seems bogus on our respective reads of the data on software advances

[Yudkowsky][22:21]

but, sure, "huge kinks in loss graphs on automated programming / Putnam problems" sounds like something that is, if not mandated on my model, much more likely than it in the Paulverse. though I am a bit surprised because I would not have expected Paul to be okay betting on that.

like, I thought it was an Eliezer-view unshared by Paul that this was a sign of the Eliezerverse.

but okeydokey if confirmed

to be clear I do not mean to predict those kinks in the next 3 years specifically they grow in probability on my model as we approach the End Times

[Shulman][22:24]

I also predict that AI chip usage is going to keep growing at enormous rates, and that buyers will be getting net economic value out of them. The market is pricing NVDA (up more than 50x since 2014) at more than twice Intel because of the incredible growth and it requires more crazy growth to justify the valuation (but still short of singularity). Although NVDA may be toppled by other producers.

Similarly for increasing spending on model size (although slower than when model costs were <\$1M).

[Yudkowsky][22:27]

relatively more plausible on my view, first because it's arguably already happening (which makes it easier to predict) and second because that can happen with profitable uses of chips which hover around on the economic fringes instead of feeding into core product cycles (waifutech)

it is easy to imagine massive AI chip usage in a world which rejects economic optimism and stays economically sad while engaging in massive AI chip usage

so, more plausible

[Shulman][22:28]

What's with the silly waifu example? That's small relative to the actual big tech company applications (where they quickly roll it into their software/web services or internal processes, which is not blocked by regulation and uses their internal expertise). Super chatbots would be used as salespeople, counselors, non-waifu entertainment.

It seems randomly off from existing reality.

[Yudkowsky][22:29]

seems more... optimistic, Kurzweilian?... to suppose that the tech gets used correctly the way a sane person would hope it would be used

[Shulman][22:29]

Like this is actual current use.

Hollywood and videogames alone are much bigger than anime, software is bigger than that, Amazon/Walmart logistics is bigger.

[Yudkowsky][22:31]

Companies using super chatbots to replace customer service they already hated and previously outsourced, with a further drop in quality, is permitted by the Dark and Gloomy Attempt To Realistically Continue History model

I am on board with wondering if we'll see sufficiently advanced videogame AI, but I'd point out that, again, that doesn't cycle core production loops harder

[Shulman][22:33]

OK, using an example of allowable economic activity that obviously is shaving off more than an order of magnitude on potential market is just misleading compared to something like FAANGSx10.

[Yudkowsky][22:34]

so, like, if I was looking for places that would break upward, I would be like "universal translators that finally work"

but I was also like that when GPT-2 came out and it hasn't happened even though you would think GPT-2 indicated we could get enough real understanding inside a neural network that you'd think, cognition-wise, it would suffice to do pretty good translation

there are huge current economic gradients pointing to the industrialization of places that you might think, could benefit a lot from universal seamless translation

[Shulman][22:36]

Current translation industry is tens of billions, English learning bigger.

[Yudkowsky][22:36]

Amazon logistics are an interesting point, but there's the question of how much economic benefit is produced by automating all of it at once, Amazon cannot ship 10x as much stuff if their warehouse costs go down by 10x.

[Shulman][22:37]

Definitely hundreds of billions of dollars of annual value created from that, e.g. by easier global outsourcing.

[Yudkowsky][22:37]

if one is looking for places where huge economic currents could be produced, AI taking down what was previously a basic labor market barrier, would sound as plausible to me as many other things

[Shulman][22:37]

Amazon has increased sales faster than it lowered logistics costs, there's still a ton of market share to take.

[Yudkowsky][22:37]

I am *able* to generate cheerful scenarios, eg if I need them for an SF short story set in near future where billions of people are using AI tech on a daily basis and this has generated trillions in economic value

[Shulman][22:38]

Bedtime for me though.

[Yudkowsky][22:39]

I don't feel like particular cheerful scenarios like that have very much of a track record coming *true*. I would not be shocked if the next GPT-jump permits that tech, and I would not be shocked if use of AI translation actually did scale a lot. I would be much more impressed, with Earth having gone well for once and better than I expected, if that actually produced significantly more labor mobility and contributed to world GDP.

I just don't actively, >50% expect things going right like that. It seems to me that more often in real life, things do not go right like that, even if it seems quite easy to imagine them going right.

good night!

10. September 22 conversation

10.1. Scaling laws

[Shah][3:05]

My attempt at a reframing:

Places of agreement:

- Trend extrapolation / things done by superforecasters seem like the right way to a first-pass answer
- Significant intuition has to go into exactly which trends to extrapolate and why (e.g. should GDP/GWP be extrapolated as "continue to grow at 3% per year" or as "growth rate continues to increase leading to singularity")
- It is possible to foresee deviations in trends based on qualitative changes in underlying drivers. In the Paul view, this often looks like switching from one trend to another. (For example: instead of "continue to grow at 3%" you notice that feedback loops imply hyperbolic growth, and then you look further back in time and notice that that's the trend on a longer timescale. Or alternatively, you realize that you can't just extrapolate AI progress because you can't keep doubling money invested every few months, and so you start looking at trends in money invested and build a simple model based on that, which you still describe as "basically trend extrapolation".)

Places of disagreement:

- Eliezer / Nate: There is an underlying driver of impact on the world which we might call "general cognition" or "intelligence" or "consequentialism" or "the-thing-spotlighted-by-coherence-arguments", and the zero-to-one transition for that underlying driver will go from "not present at all" to "at or above human-level", without something in between. Rats, dogs and chimps might be impressive in some ways but they do not have this underlying driver of impact; the zero-to-one transition happened between chimps and humans.
- Paul (might be closer to my views, idk): There isn't this underlying driver (or, depending on definitions, the zero-to-one transition happens well before human-level intelligence / impact). There are just more and more general heuristics, and correspondingly higher and higher impact. The case with evolution is unusually favorable because the more general heuristics weren't actually that useful.

To the extent this is accurate, it doesn't seem like you really get to make a bet that resolves before the end times, since you agree on basically everything until the point at which Eliezer predicts that you get the zero-to-one transition on the underlying driver of impact. I think all else equal you probably predict that Eliezer has shorter timelines to the end times than Paul (and that's where you get things like "Eliezer predicts you don't have factory-generating factories before the end times whereas Paul does"). (Of course, all else is not equal.)

[Bensinger][3:36]

but you know enough to have strong timing predictions, e.g. your bet with caplan

Eliezer said in Jan 2017 that the Caplan bet was kind of a joke:

https://www.econlib.org/archives/2017/01/my_end-of-the-w.html/#comment-166919.

Albeit "I suppose one might draw conclusions from the fact that, when I was humorously imagining what sort of benefit I could get from exploiting this amazing phenomenon, my System 1 thought that having the world not end before 2030 seemed like the most I could reasonably ask."

[Cotra][10:01]

@RobBensinger sounds like the joke is that he thinks timelines are even shorter, which strengthens my claim about strong timing predictions?

Now that we clarified up-thread that Eliezer's position is *not* that there was a giant algorithmic innovation in between chimps and humans, but rather that there was some innovation in between dinosaurs and some primate or bird that allowed the primate/bird lines to scale better, I'm now confused about why it still seems like Eliezer expects a major innovation in the future that leads to deep/general intelligence. If the evidence he has is that evolution had *some* innovation like this, why not think that the invention of neural nets in the 60s or the invention of backprop in the 80s or whatever was the corresponding innovation in AI development? Why put it in the future? (Unless I'm misunderstanding and Eliezer doesn't really place very high probability on "AGI is bottlenecked by an insight that lets us figure out how to get the deep intelligence instead of the shallow one"?)

Also if Eliezer would count transformers and so on as the kind of big innovation that would lead to AGI, then I'm not sure we disagree. I feel like that sort of thing is factored into the software progress trends used to extrapolate progress, so projecting those forward folds in expectations of future transformers

But it seems like Eliezer still expects *one* or a few innovations that are much larger in impact than the transformer?

I'm also curious what Eliezer thinks of the claim "extrapolating trends automatically fails in the world's inadequacy and stupidness because the past trend was built from everything happening in the world including the inadequacy"

[Yudkowsky][10:24]

Ajeya asked before, and I see I didn't answer:

what about hardware/software R&D wages? will they get up to \$20m/yr for good ppl
If you mean the best/luckiest people, they're already there. If you mean that say Mike Blume starts getting paid \$20m/yr base salary, then I cheerfully say that I'm willing to make that a narrower prediction of the Paulverse than of the Eliezerverse.

will someone train a 10T param model before end days?

Well, of course, because now it's a headline figure and Goodhart's Law applies, and that's the earlier point where this happens is where somebody trains a useless 10T param model using some much cheaper training method like MoE just to be the first to get the headline where they say they did that, if indeed that hasn't happened already.

But even apart from that, a 10T param model sure sounds lots like a steady stream of headlines we've already seen, even for cases where it was doing something useful like GPT-3, so I would not feel surprised by more headlines like this.

I will, however, be alarmed (not surprised) relatively more by ability improvements, than headline figure improvements, because I am not very impressed by 10T param models per se.

In fact I will probably be more surprised by ability improvements after hearing the 10T figure, than my model of Paul will claim to be, because my model of Paul much more associates 10T figures with capability increases.

Though I don't understand why this prediction success isn't more than counterbalanced by an implied sequence of earlier failures in which Paul's model permitted much more impressive things to happen from 1T Goodharted-headline models, that didn't actually happen, that I expected to not happen - eg the current regime with MoE headlines - so

that by the time that an impressive 10T model comes along and Imaginary Paul says 'yes I claim this for a success', Eliezer's reply is 'I don't understand the aspect of your theory which supposedly told you in advance that this 10T model would scale capability but not all the previous 10T models or the current pointless-headline 20T models where that would be a prediction failure. From my perspective, people eventually scaled capabilities, and param-scaling techniques happened to be getting more powerful at the same time, and so of course the Earliest tech development to be impressive was one that included lots of params. It's not a coincidence, but it's also not a triumph for the param driven theory per se, because the news stories look similar AFAICT in a timeline where 60% algorithms and 40% params."

[Cotra][10:35]

MoEs have very different scaling properties, for one thing they run on way fewer FLOPs (which is just as if not more important than params, though we use params as a shorthand when we're talking about "typical" models which tend to have small constant FLOP/param ratios). If there's a model with a *similar architecture* to the ones we have scaling laws about now, then at 10T params I'd expect it to have the performance that the scaling laws would expect it to have

Maybe something to bet about there. Would you say 10T param GPT-N would perform worse than the scaling law extraps would predict?

It seems like if we just look at a ton of scaling laws and see where they predict benchmark perf to get, then you could either bet on an upward or downward trend break and there could be a bet?

Also, if "large models that aren't that impressive" is a ding against Paul's view, why isn't GPT-3 being so much better than GPT-2 which in turn was better than GPT-1 with little fundamental architecture changes not a plus? It seems like you often cite GPT-3 as evidence for your view

But Paul (and Dario) at the time predicted it'd work. The scaling laws work was before GPT-3 and prospectively predicted GPT-3's perf

[Yudkowsky][10:55]

I guess I should've mentioned that I knew MoEs ran on many fewer FLOPs because often I may not know I know that; it's an obvious charitable-Paul-interpretation but I feel like there's multiple of those and I don't know which, if any, Paul wants to claim as obvious not-just-in-retrospect.

Like, ok, sure people talk about model size. But maybe we really want to talk about gradient descent training ops; oh, wait, actually we meant to talk about gradient descent training ops with a penalty figure for ops that use lower precision, but nowhere near a 50% penalty for 16-bit instead of 32-bit; well, no, really the obvious metric is the one in which the value of a training op scales logarithmically with the total computational depth of the gradient descent (I'm making this up, it's not an actual standard anywhere), and that's why this alternate model that does a ton of gradient descent ops while making little use of the actual limiting resource of inter-GPU bandwidth is not as effective as you'd predict from the raw headline figure about gradient descent ops. And of course we don't want to count ops that are just recomputing a gradient checkpoint, ha ha, that would be silly.

It's not impossible to figure out these adjustments in advance.

But part of me also worries that - though this is more true of other EAs who will read this than Paul or Carl, whose skills I do respect to some degree - that if you ran an MoE model with many fewer gradient descent ops, and it did do something impressive with 10T params that way, people would promptly do a happy dance and say "yay scaling" not wait huh that was not how I thought param scaling worked". After all, somebody originally said "10T", so clearly they were right!

And even with respect to Carl or Paul I worry about looking back and making "obvious" adjustments and thinking that a theory sure has been working out fine so far.

To be clear, I do consider GPT-3 as noticeable evidence for Dario's view and for Paul's view. The degree to which it worked well was more narrowly a prediction of those models than mine.

Thing about narrow predictions like that, if GPT-4 does not scale impressively, the theory loses significantly more Bayes points than it previously gained.

Saying "this previously observed trend is very strong and will surely continue" will quite often let you pick up a few pennies in front of the steamroller, because not uncommon trends do continue, but then they stop and you lose more Bayes points than you previously gained.

I do think of Carl and Paul as being better than this.

But I also think of the average EA reading them as being fooled by this.

[Shulman][11:09]

The scaling laws experiments held architecture fixed, and that's the basis of the prediction that GPT-3 will be along the same line that held over previous OOM, most definitely not switch to MoE/Switch Transformer with way less resources.

[Cotra: ]

[Yudkowsky][11:10]

You can redraw your graphs afterwards so that a variant version of Moore's Law continues to apace, but back in 2000, everyone sure was impressed with CPU GHz going up year after year and computers getting tangibly faster, and that version of Moore's Law sure did continue. Maybe some people were savvier and redrew the graphs as soon as the physical obstacles became visible, but of course, other people had predicted the end of Moore's Law years and years before then. Maybe if superforecasters had been around in 2000 we would have found that they all sorted it out successfully, maybe not.

So, GPT-3 was \$12m to train. In May 2022 it will be 2 years since GPT-3 came out. It feels to me like the Paulian view as I know how to operate it, says that GPT-3 has now got so much revenue and exhibited applications like Codex, and was on a clear trend line of promise so somebody ought to be willing to invest \$120m in training GPT-4, and then we get 4x algorithmic speedups and cost improvements since then (iirc Paul said 2x/yr above? though I can't remember if that was his viewpoint or mine?) so GPT-4 should have 40x 'oomph' in some sense, and what that translates to in terms of intuitive impact ability, don't know.

[Shulman][11:18]

The OAI paper had 16 months (and is probably a bit low because in the earlier data people weren't optimizing for hardware efficiency much): <https://openai.com/blog/ai-ai-efficiency/>

so GPT-4 should have 40x 'oomph' in some sense, and what that translates to in terms of intuitive impact ability, I don't know.

Projecting this: <https://arxiv.org/abs/2001.08361>

[Yudkowsky][11:19]

30x then. I would not be terribly surprised to find that results on benchmarks continue according to graph, and yet, GPT-4 somehow does not seem very much smarter than GPT-3 in conversation.

[Shulman][11:20]

There are also graphs of the human impressions of sense against those benchmarks and they are well correlated. I expect that to continue too.

[Cotra: ]

[Yudkowsky][11:21]

Stuff coming uncorrelated that way, sounds like some of the history I lived through, where people managed to make the graphs of Moore's Law seem to look steady by rejiggering the axes, and yet, between 1990 and 2000 home computers got a whole lot faster, and between 2010 and 2020 they did not.

This is obviously more likely (from my perspective) to break down anywhere between GPT-3 and GPT-6, than between GPT-3 and GPT-4.

Is this also part of the Carl/Paul worldview? Because I implicitly parse a lot of the arguments as assuming a necessary premise which says, "No, this continues on until doomsday and I know it Kurzweil-style."

[Shulman][11:23]

Yeah I expect trend changes to happen, more as you go further out, and especially more when you see other things running into barriers or contradictions. Re language models there is some of that coming up with different scaling laws colliding when the models are good enough to extract almost all the info per character (unless you reconfigure to use more info-dense data).

[Yudkowsky][11:23]

Where "this" is the Yudkowskian "the graphs are fragile and just break down one day, and their meanings are even more fragile and break down earlier".

[Shulman][11:25]

Scaling laws working over 8 or 9 OOM makes me pretty confident of the next couple, r
confident about 10 further OOM out.

More Christiano, Cotra, and Yudkowsky on AI progress

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This post is a transcript of a discussion between Paul Christiano, Ajeya Cotra, and Eliezer Yudkowsky (with some comments from Rob Bensinger, Richard Ngo, and Carl Shulman), continuing from [1](#), [2](#), and [3](#).

Color key:

Chat by Paul and Eliezer	Other chat
--------------------------	------------

10.2. Prototypes, historical perspectives, and betting

[Bensinger][4:25]

I feel confused about the role "innovations are almost always low-impact" plays in slow takeoff-ish views.

Suppose I think that there's some reachable algorithm that's different from current approaches, and can do par-human scientific reasoning without requiring tons of compute.

The existence or nonexistence of such an algorithm is just a fact about the physical world. If I imagine one universe where such an algorithm exists, and another where it doesn't, I don't see why I should expect that one of those worlds has more discontinuous change in GWP, ship sizes, bridge lengths, explosive yields, etc. (outside of any discontinuities caused by the advent of humans and the advent of AGI)? What do these CS facts have to do with the other facts?

But AI Impacts seems to think there's an important connection, and a large number of facts of the form 'steamships aren't like nukes' seem to undergird a lot of Paul's confidence that the scenario I described --

("there's some reachable algorithm that's different from current approaches, and can do par-human scientific reasoning without requiring tons of compute.")

-- is crazy talk. (Unless I'm misunderstanding. As seems actually pretty likely to me!)

(E.g., Paul says "To me your model just seems crazy, and you are saying it predicts crazy stuff at the end but no crazy stuff beforehand", and one of the threads of the timelines conversation has been Paul asking stuff like "do you want to give any example other than nuclear weapons of technologies with the kind of discontinuous impact you are describing?".)

Possibilities that came to mind for me:

1. The argument is 'reality keeps surprising us with how continuous everything else is, we seem to have a cognitive bias favoring discontinuity, so we should have a skeptical prior about *our ability to think our way to 'X is discontinuous'* since our brains are apparently too broken to do that well?

(But to get from 1 to 'discontinuity models are batshit' we surely need something more probability-mass-concentrating than just a bias argument?)

2. The commonality between steamship sizes, bridge sizes, etc. and AGI is something 'how tractable is the world?'. A highly tractable world, one whose principles are easy to understand and leverage, will tend to have more world-shatteringly huge historical breakthroughs in various problems, and will tend to see a larger impact from the advent of humans and the advent of AGI.

Our world looks much less tractable, so even if there's a secret sauce to building AGI, I should expect the resultant AGI to be a lot less impactful.

[Ngo][5:06]

I endorse #2 (although I think more weakly than Paul does) and would also add #3: another commonality is something like "how competitive is innovation?"

[Shulman][8:22]

@RobBensinger It's showing us a fact about the vast space of ideas and technologies we've already explored that they are not so concentrated and lumpy that the law of large numbers doesn't work well as a first approximation in a world with thousands or millions of people contributing. And that specifically includes past computer science innovations.

So the 'we find a secret sauce algorithm that causes a massive unprecedented performance jump, without crappier predecessors' is a 'separate, additional miracle' at exactly the same time as the intelligence explosion is getting going. You can get hyperbolic acceleration from increasing feedbacks from AI to AI hardware and software, including crazy scale-up at the end, as part of a default model. But adding on to it that AGI is hit via an extremely large performance jump of a type that is very rare, takes a probability penalty.

And the history of human brains doesn't seem to provide strong evidence of a fundamental software innovation, vs hardware innovation and gradual increases in selection applied to cognition/communication/culture.

The fact that, e.g. AIs are mastering so much math and language while still wielding vastly infrahuman brain-equivalents, and crossing human competence in many domains (where there was ongoing effort) over decades is significant evidence for something smoother than the development of modern humans and their culture.

That leaves me not expecting a simultaneous unusual massive human concentrated algorithmic leap with AGI, although I expect wildly accelerating progress from increasing feedbacks at that time. Crossing a given milestone is disproportionately likely to happen in the face of an unusually friendly part/jump of a tech tree (like AlexNet/the neural networks->GPU transition) but still mostly not, and likely not from an unprecedented individual computer science algorithmic change.

<https://aiimpacts.org/?s=cross+>

[Cotra: ]

[Yudkowsky][11:26][11:37]

The existence or nonexistence of such an algorithm is just a fact about the physical world. If I imagine one universe where such an algorithm exists, and another where doesn't, I don't see why I should expect that one of those worlds has more discontinuous change in GWP, ship sizes, bridge lengths, explosive yields, etc. (outside of any discontinuities caused by the advent of humans and the advent of AGI)? What do these CS facts have to do with the other facts?

I want to flag strong agreement with this. I am not talking about change in ship sizes because that is relevant in any visible way on my model; I'm talking about it in hopes I can somehow unravel Carl and Paul's model, which talks a whole lot about this being Relevant even though that continues to not seem correlated to me across possible worlds.

I think a lot in terms of "does this style of thinking seem to have any ability to bind to reality"? A lot of styles of thinking in futurism just don't.

I imagine Carl and Paul as standing near the dawn of hominids asking, "Okay, let's try measure how often previous adaptations resulted in simultaneous fitness improvement across a wide range of environmental challenges" or "what's the previous record on an organism becoming more able to survive in a different temperature range over a 100-year period" or "can we look at the variance between species in how high they fly and calculate how surprising it would be for a species to make it out of the atmosphere"

And all of reality is standing somewhere else, going on ahead to do its own thing.

Now maybe this is not the Carl and Paul viewpoint but if so I don't understand how not It's not that viewpoint plus a much narrower view of relevance, because AI Impacts go sent out to measure bridge sizes.

I go ahead and talk about these subjects, in part because maybe I can figure out some way to unravel the viewpoint on its own terms, in part because maybe Carl and Paul can show that they have a style of thinking that works in its own right and that I don't understand, and in part because people like Paul's nonconcrete cheerful writing better and prefer to live there mentally and I have to engage on their terms because they simply won't engage on mine.

But I do not actually think that bridge lengths or atomic weapons have anything to do with this.

Carl and Paul may be doing something sophisticated but wordless, where they fit a sophisticated but wordless universal model of technological permissivity to bridge lengths, then have a wordless model of cognitive scaling in the back of their minds, then get a different prediction of Final Days behavior, then come back to me and say, "Well, if you got such a different prediction of Final Days behavior, can you show me some really long bridges?"

But this is not spelled out in the writing - which, I do emphasize, is a social observation that would be predicted regardless, because other people have not invested a ton of

character points in the ability to spell things out, and a supersupermajority would just plain lack the writing talent for it.

And what other EAs reading it are thinking, I expect, is plain old Robin-Hanson-style [reference class tennis](#) of "Why would you expect *intelligence* to scale differently from *bridges*, where are all the *big bridges*?"

[Cotra][11:36][11:40]

(Just want to interject that Carl has higher P(doom) than Paul and has also critiqued Pa for not being more concrete, and I doubt that this is the source of the common disagreements that Paul/Carl both have with Eliezer)

From my perspective the thing the AI impacts investigation is asking is something like "When people are putting lots of resources into improving some technology, how often is it the case that someone can find a cool innovation that improves things a lot relative to the baseline?" I think that your response to that is something like "Sure, if the broad AI market were efficient and everyone were investigating the right lines of research, then progress might be smooth, but AGI would have also been developed way sooner. We can safely assume that AGI is like an industry where lots of people are pushing toward the same thing"

But it's not assuming a great structural similarity between bridges and AI, except that they're both things that humans are trying hard to find ways to improve

[Yudkowsky][11:42]

I can imagine writing responses like that, if I was engaging on somebody else's terms. with [Eliezer-2012's engagement with Pat Modesto](#) against the careful proof that HPMO cannot possibly become one of the measurably most popular fanfictions, I would never think anything like that inside my own brain.

Maybe I just need to do a thing that I have not done before, and set my little \$6000 Roth IRA to track a bunch of investments that Carl and/or Paul tell me to make, so that my brain will actually track the results, and I will actually get a chance to see this weird style of reasoning produce amazing results.

[Bensinger][11:44]

Sure, if the broad AI market were efficient and everyone were investigating the right lines of research, then AI progress might be smooth

Presumably also "'AI progress' subsumes many different kinds of cognition, we don't currently have baby AGIs, and when we do figure out how to build AGI the very beginning of the curve (the Wright flyer moment, or something very shortly after) will correspond to a huge capability increase."

[Yudkowsky][11:46]

I think there's some much larger scale in which it's worth mentioning that on my own terms of engagement I do not naturally think like this. I don't feel like you could get Gr

Insight by figuring out what the predecessor technologies must have been of the Wright Flyer, finding industries that were making use of them, and then saying Behold the Heralds of the Wright Flyer. It's not a style of thought binding upon reality.

They built the Wright Flyer. It flew. Previous stuff didn't fly. It happens. Even if you yell lot at reality and try to force it into an order, that's still what your actual experience of surprising Future will be like, you'll just be more surprised by it.

Like you can super want Technologies to be Heralded by Predecessors which were Also Profitable but on my native viewpoint this is, like, somebody with a historical axe to go going back and trying to make all the history books read like this, when I have no experience of people who were alive at the time making gloriously correct futuristic predictions using this kind of thinking.

[Cotra][11:53]

I think Paul's view would say:

- Things certainly happen for the first time
- When they do, they happen at small scale in shitty prototypes, like the Wright Flyer or GPT-1 or AlphaGo or the Atari bots or whatever
- When they're making a big impact on the world, it's after a lot of investment and research, like commercial aircrafts in the decades after Kitty Hawk or like the investments people are in the middle of making now with AI that can assist with coding

Paul's view says that the Kitty Hawk moment *already happened for the kind of AI that could be super transformative and could kill us all*, and like the historical Kitty Hawk moment was not immediately a huge deal

[Yudkowsky][11:56]

There is, I think, a really basic difference of thinking here, which is that on my view, A Trend erupting is just a Thing That Happens and not part of a Historical Worldview or a Great Trend.

Human intelligence wasn't part of a grand story reflected in all parts of the ecology, it happened in a particular species.

Now afterwards, of course, you can go back and draw all kinds of Grand Trends into which this Thing Happening was perfectly and beautifully fitted, and yet, it does not seem to me that people have a very good track record of thereby predicting in advance what surprising news story they will see next - with some rare, narrow-superforecasting-technique exceptions, like the Things chart on a steady graph and we know *solidly* where the threshold on that graph corresponds to and that threshold is not too far away compared to the previous length of the chart.

One day the Wright Flyer flew. Anybody *in the future with benefit of hindsight*, who wanted to, could fit that into a grand story about flying, industry, travel, technology, whatever; if they've been on the ground at the time, they would not have thereby had much luck predicting the Wright Flyer. It can be *fit into* a grand story but on the ground it's just a thing that happened. It had some prior causes but it was not thereby constrained to fit into a storyline in which it was the plot climax of those prior causes.

My worldview sure does permit there to be predecessor technologies and for them to have some kind of impact and for some company to make a profit, but it is not nearly interested in that stuff, on a very basic level, because it does not think that the AGI Th Happening is the plot climax of a story about the Previous Stuff Happening.

[Cotra][12:01]

The fact that you express this kind of view about AGI erupting one day is why I thought your thing in IEM was saying there was a major algorithmic innovation *from chimps to humans*, that humans were qualitatively and not just quantitatively better than chimps and this was not because of their larger brain size primarily. But I'm confused because thread in the discussion of evolution you were emphasizing much more that there was innovation between dinosaurs and primates, not that there was an innovation between chimps and humans, and you seemed more open to the chimp/human diff being quantitative and brain-size driven than I had thought you'd be. But being open to the chimp-human diff being quantitative/brain-size-driven suggests to me that you should more open than you are to AGI being developed by slow grinding on the same shit, instead of erupting without much precedent?

[Yudkowsky][12:01]

I think you're confusing a meta-level viewpoint with an object-level viewpoint.

The Wright Flyer does not need to be made out of completely different materials from previous travel devices, in order for the Wright Flyer to be a Thing That Happened One Day which wasn't the plot climax of a grand story about Travel and which people at the time could not have gotten very far in advance-predicting by reasoning about which materials were being used in which conveyances and whether those conveyances look like they'd be about to start flying.

It is the very viewpoint to which I am objecting, which keeps on asking me, metaphorically speaking, to explain how the Wright Flyer could have been made of completely different materials in order for it to be allowed to be so discontinuous with rest of the Travel story of which it is part.

On my viewpoint they're just *different stories* so the Wright Flyer is allowed to be its own thing even though it is not made out of an unprecedented new kind of steel that floats

[Cotra][12:06]

The claim I'm making is that Paul's view predicts a lag and a lot of investment between the first flight and aircraft making a big impact on the travel industry, and predicts that the first flight wouldn't have immediately made a big impact on the travel industry. In other words Kitty Hawk isn't a discontinuity in the Paul view because the metrics he'd expect to be continuous are the ones that large numbers of people are trying hard to optimize, like cost per mile traveled or whatnot, not metrics that almost nobody is trying to optimize, like "height flown."

In other words, it sounds like you're saying:

- Kitty Hawk is analogous to AGI erupting
- Previous history of travel is analogous to pre-AGI history of AI

While Paul is saying:

- Kitty Hawk is analogous to e.g. AlexNet
- Later history of aircraft is analogous to the post-AlexNet story of AI which we're in the middle of living, and will continue on to make huge Singularity-causing impacts on the world

[Yudkowsky][12:09]

Well, unfortunately, Paul and I both seem to believe that our models follow from observing the present-day world, rather than being incompatible with it, and so when we demand of each other that we produce some surprising bold prediction about the present-day world, we both tend to end up disappointed.

I would like, of course, for Paul's surprisingly narrow vision of a world governed by tight bound stories and predictable trends, to produce some concrete bold prediction of the next few years which no ordinary superforecaster would produce, but Paul is not under the impression that his own worldview is similarly strange and narrow, and so has some difficulty in answering this request.

[Cotra][12:09]

But Paul offered to bet with you about literally any quantity you choose?

[Yudkowsky][12:10]

I did assume that required an actual disagreement, eg, I cannot just go look up something superforecasters are very confident about and then demand Paul to bet against it.

[Cotra][12:12]

It still sounds to me like "take a basket of N performance metrics, bet that the model says the perf trend will break upward in > K of them within e.g. 2 or 3 years" should sound good to you, I'm confused why that didn't. If it does and it's just about the legwork then I think we could get someone to come up with the benchmarks and stuff for you

Or maybe the same thing but >K of them will break downward, whatever

We could bet about the human perception of sense in language models, for example

[Yudkowsky][12:14]

I am nervous about Paul's definition of "break" and the actual probabilities to be assigned. You see, both Paul and I think our worldview is a very normal one that matches current reality quite well, so when we are estimating parameters like these, Paul is likely to do it empirically, and I am also liable to do it empirically as my own baseline, and if I point to a trend over time in how long it takes to go from par-human to superhuman performance decreasing, Imaginary Paul says "Ah, yes, what a fine trend, I will bet that things follow this trend" and Eliezer says "No that is MY trend, you don't get to follow it if you have to predict that par-human to superhuman time will be constant" and Paul is like "lol no I get to be a superforecaster and follow trends" and we fail to bet.

Maybe I'm wrong in having mentally played the game out ahead that far, for it is, after all, very hard to predict the Future, but that's where I'd foresee it failing.

[Cotra][12:16]

I don't think you need to bet about calendar times from par-human to super-human, and any meta-trend in that quantity. It sounds like Paul is saying "I'll basically trust the model size to perf trends and predict a 10x bigger model from the same architecture family will get the perf the trends predict," and you're pushing back against that saying e.g. that humans won't find GPT-4 to be subjectively more coherent than GPT-3 and that Paul is neglecting that there could be major innovations in the future that bring down the FLC to get a certain perf by a lot and bend the scaling laws. So why not bet that Paul won't be as accurate as he thinks he is by following the scaling laws?

[Bensinger][12:17]

I think Paul's view would say:

- Things certainly happen for the first time
- When they do, they happen at small scale in shitty prototypes, like the Wright Flyer or GPT-1 or AlphaGo or the Atari bots or whatever
- When they're making a big impact on the world, it's after a lot of investment and research, like commercial aircrafts in the decades after Kitty Hawk or like the investments people are in the middle of making now with AI that can assist with coding

Paul's view says that the Kitty Hawk moment *already happened for the kind of AI that will be super transformative and could kill us all*, and like the historical Kitty Hawk moment, it was not immediately a huge deal

"When they do, they happen at small scale in shitty prototypes, like the Wright Flyer or GPT-1 or AlphaGo or the Atari bots or whatever"

How shitty the prototype is should depend (to a very large extent) on the physical properties of the tech. So I don't find it confusing (though I currently disagree) when someone says "I looked at a bunch of GPT-3 behavior and it's cognitively sophisticated enough that I think it's doing basically what humans are doing, just at a smaller scale. The qualitative cognition I can see going on is just that impressive, taking into account the kinds of stuff I think human brains are doing."

What I find confusing is, like, treating ten thousand examples of non-AI, non-cognitive-tech continuities (nukes, building heights, etc.) as though they're anything but a tiny update about 'will AGI be high-impact' -- compared to the size of updates like 'look at how smart and high-impact humans were' and perhaps 'look at how smart-in-the-relevant-ways GPT-3 is'.

Like, impactfulness is not a simple physical property, so there's not much reason for different kinds of tech to have similar scales of impact (or similar scales of impact n years after the first prototype). Mainly I'm not sure to what extent we disagree about this, vs this just being me misunderstanding the role of the 'most things aren't high-impact' argument.

(And yeah, a random historical technology drawn from a hat will be pretty low-impact. But that base rate also doesn't seem to me like it has much evidential relevance anymore when I update about what specific tech we're discussing.)

[Cotra][12:18]

The question is not "will AGI be high impact" -- Paul agrees it will, and for any FOOM quantity (like crossing a chimp-to-human-sized gap in a day or whatever) he agrees that will happen eventually too.

The technologies studies in the dataset spanned a wide range in their peak impact on society, and they're not being used to forecast the peak impact of mature AI tech

[Bensinger][12:19]

Yeah, I'm specifically confused about how we know that the AGI Wright Flyer and its first successors are low-impact, from looking at how low-impact other technologies are (if that is in fact a meaningful-sized update on your view)

Not drawing a comparison about the overall impactfulness of AI / AGI (e.g., over fifteen years)

[Yudkowsky][12:21]

[So why not bet that Paul won't be as accurate as he thinks he is by following the scaling laws?]

I'm pessimistic about us being able to settle on the terms of a bet like that (and even more so about being able to bet against Carl on it) but in broad principle I agree. The trouble is that if a trend is benchmarkable, I believe more in the trend continuing at least on the next particular time, not least because I believe in people Goodharting benchmarks.

I expect a human sense of intelligence to be harder to fool (even taking into account that it's being targeted to a nonzero extent) but I also expect that to be much harder to measure and bet upon than the Goodhartable metrics. And I think our actual disagreement is more visible over portfolios of benchmarks breaking upward over time but I also expect that if you ask Paul and myself to *quantify our predictions*, we both give "Oh, my theory is the one that fits ordinary reality so obviously I will go look at superforecastery trends over ordinary reality to predict this specifically" and I am like, "No, Paul, if you'd had to predict that without looking at the data, your worldview would've predicted trends breaking down less often" and Paul is like "But Eliezer, shouldn't you be predicting much more upward divergence than this."

Again, perhaps I'm being overly gloomy.

[Cotra][12:23]

I think we should try to find ML predictions where you defer to superforecasters and Paul disagrees, since he said he would bet against superforecasters in ML

[Yudkowsky][12:24]

I am also probably noticeably gloomier and less eager to bet because the whole fight is taking place on grounds that Paul thinks is important and part of a connected story that

continuously describes ordinary reality, and that I think is a strange place where I can't particularly see how Paul's reasoning style works. So I'd want to bet against Paul's over narrow predictions by using ordinary superforecasting, and Paul would like to make his predictions using ordinary superforecasting.

I am, indeed, more interested in a place where Paul wants to bet against superforecasters. I am not guaranteeing up front I'll bet with them because superforecasters did not call AlphaGo correctly and I do not think Paul has zero actual domain expertise. But Paul is allowed to pick up *generic* epistemic credit *including from me* by beating superforecasters because that credit counts toward believing a style of thought is even *working literally at all*; separately from the question of whether Paul's superforecaster-defying prediction also looks like a place where I'd predict in some opposite direction.

Definitely, places where Paul disagrees with superforecasters are much more interesting places to mine for bets.

I am happy to hear about those.

[Cotra][12:27]

I think what Paul was saying last night is you find superforecasters betting on some benchmark performance, and he just figures out which side he'd take (and he expects most/all superforecaster predictions that he would not be deferential, there's a side he would take)

10.3. Predictions and betting (continued)

[Christiano][12:29]

not really following along with the conversation, but my desire to bet about "whatever you want" was driven in significant part by frustration with Eliezer repeatedly saying things like "people like Paul get surprised by reality" and me thinking that's nonsense

[Yudkowsky][12:29]

So the Yudkowskian viewpoint is something like... trends in particular technologies help fixed, will often break down; trends in Goodhartable metrics, will often stay on track but become decoupled from their real meat; trends across multiple technologies, will experience occasional upward breaks when new algorithms on the level of Transformers come out. For me to bet against superforecasters I have to see superforecasters saying something different, which I do not at this time actually know to be the case. For me to bet against Paul betting against superforecasters, the different thing Paul says has to be different from my own direction of disagreement with superforecasters.

[Christiano][12:30]

I still think that if you want to say "this sort of reasoning is garbage empirically" then you ought to be willing to bet about something. If we are just saying "we agree about all of the empirics, it's just that somehow we have different predictions about AGI" then that fine and symmetrical.

[Yudkowsky][12:30]

I have been trying to revise that towards a more nvc "when I try to operate this style of thought myself, it seems to do a bad job of retrofitting and I don't understand how it says X but not Y".

[Christiano][12:30]

even then presumably if you think it's garbage you should be able to point to some particular future predictions where it would be garbage?

if you used it

and then I can either say "no, I don't think that's a valid application for reason X" or "so I'm happy to bet"

and it's possible you can't find any places where it sticks its neck out in practice (even your version), but then I'm again just rejecting the claim that it's empirically ruled out

[Yudkowsky][12:31]

I also think that we'd have an easier time betting if, like, neither of us could look at graphs over time, but we were at least told the values in 2010 and 2011 to anchor our estimates over one year, or something like that.

Though we also need to not have a bunch of existing knowledge of the domain which is hard.

[Christiano][12:32]

I think this might be derailing some broader point, but I am provisionally mostly ignoring your point "this doesn't work in practice" if we can't find places where we actually fore disagreements

(which is fine, I don't think it's core to your argument)

[Yudkowsky][12:33]

Paul, you've previously said that you're happy to bet against ML superforecasts. That sounds promising. What are examples of those? Also I must flee to lunch and am already feeling sort of burned and harried; it's possible I should not ignore the default doomedness of trying to field questions from multiple sources.

[Christiano][12:33]

I don't know if superforecasters make public bets on ML topics, I was saying I'm happy bet on ML topics and if your strategy is "look up what superforecasters say" that's fine and doesn't change my willingness to bet

I think this is probably not as promising as either (i) dig in on the arguments that are in dispute (seemed to be some juicier stuff earlier though I'm just focusing on work to do), or (ii) just talking generally about what we expect to see in the next 5 years so that we can at least get more of a vibe looking back

[Shulman][12:35]

You can bet on the Metaculus AI Tournament forecasts.

<https://www.metaculus.com/ai-progress-tournament/>

[Yudkowsky][13:13]

I worry that trying to jump straight ahead to Let's Bet is being too ambitious too early a cognitively difficult problem of localizing disagreements.

Our prophecies of the End Times's modal final days seem legit different; my impulse would be to try to work that backwards, first, in an intuitive sense of "well which prophesied world would this experience feel more like living in?", and try to dig deeper there before deciding that our disagreements have crystallized into short-term easily-observable bets.

We both, weirdly enough, feel that our current viewpoints are doing a great job of permitting the present-day world, even if, presumably, we both think the other's worldview would've done worse at predicting that world in advance. This cannot be resolved in an instant by standard techniques known to me. Let's try working back from the End Times instead.

I have already stuck out my neck a little and said that, as we start to go past \$50B invested in a model, we are starting to live at least a *little* more in what feels like the Paulverse, not because my model prohibits this, but because, or so I think, Paul's model more narrowly predicts it.

It does seem like the sort of generically weird big thing that could happen, to me, ever before the End Times, there are corporations that could just decide to do that; I am hedging around this exactly because it does feel to my gut like that is a kind of headline that could read one day and have it still be years before the world ended, so I may need to be stingy with those credibility points inside of what I expect to be reality.

But if we get up to \$10T to train a model, that is *much* more strongly Paulverse; it's not that this falsifies the Eliezerverse considered in isolation, but it is *much* more narrowly characteristic of the Words of Paul coming to pass; it feels much more to my gut that, agreeing to this, I am not giving away Bayes points inside my own mainline.

If ordinary salaries for ordinary fairly-good programmers get up to \$20M/year, this is not prohibited by my AI models per se; but it sure sounds like the world becoming less ordinary than I expected it to stay, and like it is part of Paul's Prophecy much more strongly than it is part of Eliezer's Prophecy.

That's two ways that I could concede a great victory to the Paulverse. They both have disadvantages (from my perspective) that the Paulverse, though it must be drawing

probability mass from somewhere in order to stake it there, is legitimately not - so far know - forced to claim that these things happen anytime soon. So they are ways for the Paulverse to win, but not ways for the Eliezerverse to win.

That I have said even this much, I claim, puts Paul in at least a little tiny bit of debt to epistemic-good-behavior-wise; he should be able to describe events which would start make him worry he was living in the Eliezerverse, even if his model did not narrowly rule them out, and even if those events had not been predicted by the Eliezerverse to occur within a narrowly prophesied date such that they would not thereby form a bet the Eliezerverse could clearly lose as well as win.

I have not had much luck in trying to guess what the real Paul will say about issues like this one. My last attempt was to say, "Well, what shouldn't happen, besides the End Times themselves, before world GDP has doubled over a four-year period?" And Paul gave what seems to me like an overly valid reply, which, iirc and without looking it up, was along the lines of, "well, nothing that would double world GDP in a 1-year period".

When I say this is overly valid, I mean that it follows too strongly from Paul's premises, and he should be looking for something less strong than that on which to make a beginning discovery of disagreement - maybe something which Paul's premises don't strongly forbid to him, but which nonetheless looks more like the Eliezerverse or like it would be relatively more strongly predicted by Eliezer's Prophecy.

I do not model Paul as eagerly or strongly agreeing with, say, "The Riemann Hypothesis should not be machine-proven" or "The ABC Conjecture should not be machine-proven before world GDP has doubled. It is only on Eliezer's view that proving the Riemann Hypothesis is about as much of a related or unrelated story to AGI, as are particular benchmarks of GDP.

On Paul's view as I am trying to understand and operate it, this benchmark may be correlated with AGI in time in the sense that most planets wouldn't do it during the Middle Ages before they had any computers, but it is not part of the *story* of AGI, it is not part of Paul's Prophecy; because it doesn't make a huge amount of money and increase GDP to get a huge ton of money flowing into investments in useful AI.

(From Eliezer's perspective, you could tell a story about how a stunning machine proof of the Riemann Hypothesis got Bezos to invest \$50 billion in training a successor model & that was how the world ended, and that would be a just-as-plausible model as some particular economic progress story, of how Stuff Happened Because Other Stuff Happened; it sounds like the story of OpenAI or of Deepmind's early Atari demo, which to say, it sounds to Eliezer like history. Whereas on Eliezer!Paul's view, that's much more of a weird coincidence because it involves Bezos's unforced decision rather than the economic story of which AGI is capstone, or so it seems to me trying to operate Paul's view.)

And yet Paul might still, I hope, be able to find something *like* "The Riemann Hypothesis is machine-proven", which even though it is not very much of an interesting part of his own Prophecy because it's not part of the economic storyline, sounds to him like the sort of thing that the *Eliezerverse* thinks happens as you get close to AGI, which the *Eliezerverse* says is allowed to start happening way before world GDP would double in 4 years; and it happens I'd agree with that characterization of the Eliezerverse.

So Paul might say, "Well, my model doesn't particularly *forbid* that the Riemann Hypothesis gets machine-proven before world GDP has doubled in 4 years or even starts to discernibly break above trend by much; but that does sound *more* like we are living in the Eliezerverse than in the Paulverse."

I am not demanding this particular bet because it seems to me that the Riemann Hypothesis may well prove to be unfairly targetable for current ML techniques while they are still separated from AGI by great algorithmic gaps. But if on the other hand Paul thinks that, I dunno, superhuman performance on stuff like the Riemann Hypothesis do tend to be more correlated with economically productive stuff because it's all roughly the same kind of capability, and lol never mind this "algorithmic gap" stuff, then maybe Paul is willing to pick that example; which is all the better for me because I *do* suspect it might decouple from the AI of the End, and so I think I have a substantial chance of winning by being able to say "SEE!" to the assembled EAs while there's still a year or two left on the timeline.

I'd love to have credibility points on that timeline, if Paul doesn't feel as strong an anticipation of needing them.

[Christiano][15:43]

1/3 that RH has an automated proof before sustained 7%/year GWP growth?

I think the clearest indicator is that we have AI that ought to be able to e.g. run the full automated factory-building factory (not automating mines or fabs, just the robotic manufacturing and construction), but it's not being deployed or is being deployed with very mild economic impacts

another indicator is that we have AI systems that can fully replace human programme (or other giant wins), but total investment in improving them is still small

another indicator is a DeepMind demo that actually creates a lot of value (e.g. 10x larger than DeepMind's R&D costs? or even comparable to DeepMind's cumulative R&D costs if you do the accounting really carefully and I definitely believe it and it wasn't replaced by Brain), it seems like on your model things should "break upwards" and in mine that just doesn't happen that much

sounds like you may have >90% on automated proof of RH before a few years of 7%/year growth driven by AI? so that would give a pretty significant odds ratio either way

I think "stack more layers gets stuck but a clever idea makes crazy stuff happen" is generally going to be evidence for your view

That said, I'd mostly reject AlphaGo as an example, because it's just plugging in neural networks to existing go algorithms in almost the most straightforward way and the bells and whistles don't really matter. But if AlphaZero worked and AlphaGo didn't, and the system accomplished something impressive/important (like proving RH, or being significantly better at self-contained programming tasks), then that would be a surprise.

And I'd reject LSTM -> transformer or MoE as an example because the quantitative effect size isn't that big.

But if something like that made the difference between "this algorithm wasn't scaling before, and now it's scaling," then I'd be surprised.

And the size of jump that surprises me is shrinking over time. So in a few years even getting the equivalent of a factor of 4 jump from some clever innovation would be very surprising to me.

[Yudkowsky][17:44]

sounds like you may have >90% on automated proof of RH before a few years of 7%/year growth driven by AI? so that would give a pretty significant odds ratio either way

I emphasize that this is mostly about no on the GDP growth before the world ending, rather than yes on the RH proof, i.e., I am not 90% on RH before the end of the world at all. Not sure I'm over 50% on it happening before the end of the world at all.

Should it be a consequence of easier earlier problems than full AGI? Yes, on my mainline model; but also on my model, it's a particular thing and maybe the particular people actions doing stuff don't get around to that particular thing.

I guess if I stare hard at my brain it goes 'ehhhh maybe 65% if timelines are relatively long and 40% if it's like the next 5 years', because the faster stuff happens, the less likely anyone is to get around to proving RH in particular or announcing that they've done so they did.

And if the econ threshold is set as low as 7%/yr, I start to worry about that happening longer-term scenarios, just because world GDP has never been moving at a fixed rate over a log chart. the "driven by AI" part sounds very hard to evaluate. I want, I dunno, some other superforecaster or Carl to put a 90% credible bound on 'when world GDP growth hits 7% assuming little economically relevant progress in AI' before I start betting at 80%, let alone 90%, on what should happen before then. I don't have that credible bound already loaded and I'm not specialized in it.

I'm wondering if we're jumping ahead of ourselves by trying to make a nice formal Bayesian bet, as prestigious as that might be. I mean, your 1/3 was probably important for you to say, as it is higher than I might have hoped, and I'd ask you if you really meant for that to be an upper bound on your probability or if that's your actual probability.

But, more than that, I'm wondering if, in the same vague language I used before, you're okay with saying a little more weakly, "RH proven before big AI-driven growth in world GDP, sounds more Eliezerverse than Paulverse."

It could be that this is just not actually true because you do not think that RH is coupled to econ stuff in the Paul Prophecy one way or another, and my own declarations above do not have the Eliezerverse saying it enough more strongly than that. If you don't actually see this as a distinguishing Eliezerverse thing, if it wouldn't actually make you say "Oh maybe I'm in the Eliezerverse", then such are the epistemic facts.

And the size of jump that surprises me is shrinking over time. So in a few years ever getting the equivalent of a factor of 4 jump from some clever innovation would be very surprising to me.

This sounds potentially more promising to me - seems highly Eliezerverse, highly non-Paul-verse according to you, and its negation seems highly oops-maybe-I'm-in-the-Paulverse to me too. How many years is a few? How large a jump is shocking if it happens tomorrow?

11. September 24 conversation

11.1. Predictions and betting (continued 2)

[Christiano][13:15]

I think RH is not that surprising, it's not at all clear to me where "do formal math" sits on the "useful stuff AI could do" spectrum, I guess naively I'd put it somewhere "in the middle" (though the analogy to board games makes it seem a bit lower, and there is a kind of obvious approach to doing this that seems to be working reasonably well so that also makes it seem lower), and 7% GDP growth is relatively close to the end (ETA: by "close to the end" I don't mean super close to the end, just far enough along that there's plenty of time for RH first)

I do think that performance jumps are maybe more dispositive, but I'm afraid that it's basically going to go like this: there won't be metrics that people are tracking that jump up, but you'll point to new applications that people hadn't considered before, and I'll say "but those new applications aren't that valuable" whereas to you they will look more analogous to a world-ending AGI coming out from the blue

like for AGZ I'll be like "well it's not really above the deep learning trend if you run it backwards" and you'll be like "but no one was measuring it before! you can't make up a trend in retrospect!" and I'll be like "OK, but the reason no one was measuring it before was that it was worse than traditional go algorithms until like 2 years ago and the upside is not large enough that you should expect a huge development effort for a small edge"

[Yudkowsky][13:43]

"factor of 4 jump from some clever innovation" - can you say more about that part?

[Christiano][13:53]

like I'm surprised if a clever innovation does more good than spending 4x more computation

[Yudkowsky][15:04]

I worry that I'm misunderstanding this assertion because, as it stands, it sounds extremely likely that I'd win. Would transformers vs. CNNs/RNNs have won this the year that the transformers paper came out?

[Christiano][15:07]

I'm saying that it gets harder over time, don't expect wins as big as transformers

I think even transformers probably wouldn't make this cut though?

certainly not vs CNNs

vs RNNs I think the comparison I'd be using to operationalize it is translation, as measured in the original paper

they do make this cut for translation, looks like the number is like 100 >> 4
100x for english-german, more like 10x for english-french, those are the two benchmarks
they cite
but both more than 4x
I'm saying I don't expect ongoing wins that big
I think the key ambiguity is probably going to be about what makes a measurement
established/hard-to-improve

[Yudkowsky][15:21]

this sounds like a potentially important point of differentiation; I do expect more wins to be big.

the main thing that I imagine might make a big difference to your worldview, but not mine, is if the first demo of the big win only works slightly better (although that might also be because they were able to afford much less compute than the big players, which I think your worldview would see as a redeeming factor for my worldview?) but a couple years later might be 4x or 10x as effective per unit compute (albeit that other innovations would've been added on by then to make the first innovation work properly, which I think on your worldview is like The Point or something)

clarification: by "transformers vs CNNs" I don't mean transformers on ImageNet, I mean transformers vs. contemporary CNNs, RNNs, or both, being used on text problems.

I'm also feeling a bit confused because eg Standard Naive Kurzweilian Accelerationism makes a big deal about the graphs keeping on track because technologies hop new modes as needed. What distinguishes your worldview from saying that no further innovations are needed for AGI or will give a big compute benefit along the way? Is it that any single idea may only ever produce a smaller-than-4X benefit? Is it permitted that a single idea plus 6 months of engineering fiddly details produce a 4X benefit?

all this aside, "don't expect wins as big as transformers" continues to sound to me like a very promising point for differentiating Prophecies.

[Christiano][15:50]

I think the relevant feature of the innovation is that the work to find it is small relative to the work that went into the problem to date (though there may be other work on other avenues)

[Yudkowsky][15:52]

in, like, a local sense, or a global sense? If there's 100 startups searching for ideas collectively with \$10B of funding, and one of them has an idea that's 10x more efficient per unit compute on billion-dollar problems, is that "a small amount of work" because it was only a \$100M startup, or collectively an appropriate amount of work?

[Christiano][15:53]

I'm calling that an innovation because it's a small amount of work

[Yudkowsky][15:54]

(maybe it would be also productive if you pointed to more historical events like Transformers and said 'that shouldn't happen again', because I didn't realize there was anything you thought was like that. AlphaFold 2?)

[Christiano][15:54]

like, it's not just a claim about EMH, it's also a claim about the nature of progress

I think AlphaFold counts and is probably if anything a bigger multiplier, it's just uncertainty over how many people actually worked on the baselines

[Yudkowsky][15:54]

when should we see headlines like those subside?

[Christiano][15:55]

I mean, I think they are steadily subsiding

as areas grow

[Yudkowsky][15:55]

have they already begun to subside relative to 2016, on your view?

(guess that was ninjaed)

[Christiano][15:55]

I would be surprised to see a 10x today on machine translation

[Yudkowsky][15:55]

where that's 10x the compute required to get the same result?

[Christiano][15:55]

though not so surprised that we can avoid talking about probabilities

yeah

or to make it more surprising, old sota with 10x less compute

[Yudkowsky][15:56]

yeah I was about to worry that people wouldn't bother spending 10x the cost of a large model to settle our bet

[Christiano][15:56]

I'm more surprised if they get the old performance with 10x less compute though, so the way around is better on all fronts

[Yudkowsky][15:57]

one reads papers claiming this all the time, though?

[Christiano][15:57]

like, this view also leads me to predict that if I look at the actual amount of manpower that went into alphafold, it's going to be pretty big relative to the other people submitting to that protein folding benchmark

[Yudkowsky][15:57]

though typically for the sota of 2 years ago

[Christiano][15:58]

not plausible claims on problems people care about

I think the comparison is to contemporary benchmarks from one of the 99 other startups who didn't find the bright idea

that's the relevant thing on your view, right?

[Yudkowsky][15:59]

I would expect AlphaFold and AlphaFold 2 to involve... maybe 20 Deep Learning researchers, and for 1-3 less impressive DL researchers to have been the previous limit the field even tried that much; I would not be the least surprised if DM spent 1000x the compute on AlphaFold 2, but I'd be very surprised if the 1-3 large research team could spend that 1000x compute and get anywhere near AlphaFold 2 results.

[Christiano][15:59]

and then I'm predicting that number is already <10 for machine translation and falling (maybe I shouldn't talk about machine translation or at least not commit to numbers given that I know very little about it, but whatever that's my estimate), and for other domains it will be <10 by the time they get as crowded as machine translation, and for transformative tasks they will be <2

isn't there an open-source replication of alphafold?
we could bet about its performance relative to the original

[Yudkowsky][16:00]

it is enormously easier to do what's already been done

[Christiano][16:00]

I agree

[Yudkowsky][16:00]

I believe the open-source replication was by people who were told roughly what Deepmind had done, possibly more than roughly
on the Yudkowskian view, those 1-3 previous researchers just would not have thought doing things the way Deepmind did them

[Christiano][16:01]

anyway, my guess is generally that if you are big relative to previous efforts in the area you can make giant improvements, if you are small relative to previous efforts you might get lucky (or just be much smarter) but that gets increasingly unlikely as the field gets bigger

like alexnet and transformers are big wins by groups who are small relative to the rest of the field, but transformers are much smaller than alexnet and future developments will continue to shrink

[Yudkowsky][16:02]

but if you're the *same size* as previous efforts and don't have 100x the compute, you shouldn't be able to get huge improvements in the Paulverse?

[Christiano][16:03]

I mean, if you are the same size as all the prior effort put together?

I'm not surprised if you can totally dominate in that case, especially if prior efforts are well-coordinated

and for things that are done by hobbyists, I wouldn't be surprised if you can be a bit bigger than an individual hobbyist and dominate

[Yudkowsky][16:03]

I'm thinking something like, if Deepmind comes out with an innovation such that it duplicates old SOTA on machine translation with 1/10th compute, that still violates the

Paulverse because Deepmind is not Paul!Big compared to all MTL efforts
though I am not sure myself how seriously Earth is taking MTL in the first place

[Christiano][16:04]

yeah, I think if DeepMind beats Google Brain by 10x compute next year on translation
that's a significant strike against Paul

[Yudkowsky][16:05]

I know that Google offers it for free, I expect they at least have 50 mediocre AI people
working on it, I don't know whether or not they have 20 excellent AI people working or
and if they've ever tried training a 200B parameter non-MoE model on it

[Christiano][16:05]

I think not that seriously, but more seriously than 2016 and than anything else where
are seeing big swings

and so I'm less surprised than for TAI, but still surprised

[Yudkowsky][16:06]

I am feeling increasingly optimistic that we have some notion of what it means to not I
within the Paulverse! I am not feeling that we have solved the problem of having enou
signs that enough of them will appear to tell EA how to notice which universe it is insic
many years before the actual End Times, but I sure do feel like we are making progres
things that have happened in the past that you feel *shouldn't happen again* are great
places to poke for Eliezer-disagreements!

[Christiano][16:07]

I definitely think there's a big disagreement here about what to expect for pre-end-of-
days ML

but lots of concerns about details like what domains are crowded enough to be surpris
and how to do comparisons

I mean, to be clear, I think the transformer paper having giant gains is also evidence
against paulverse

it's just that there are really a lot of datapoints, and some of them definitely go agains
paul's view

to me it feels like the relevant thing for making the end-of-days forecast is something
"how much of the progress comes from 'innovations' that are relatively unpredictable
and/or driven by groups that are relatively small, vs scaleup and 'business as usual'
progress in small pieces?"

11.2. Performance leap scenario

[Yudkowsky][16:09]

my heuristics tell me to try wargaming out a particular scenario so we can determine in advance which key questions Paul asks

in 2023, Deepmind releases an MTL program which is suuuper impressive. everyone who reads the MTL of, say, a foreign novel, or uses it to conduct a text chat with a contractor in Indonesia, is like, "They've basically got it, this is about as good as a human and only makes minor and easily corrected errors."

[Christiano][16:12]

I mostly want to know how good Google's translation is at that time; and if DeepMind's product is expensive or only shows gains for long texts, I want to know whether there is actually an economic niche for it that is large relative to the R&D cost.

like I'm not sure whether anyone works at all on long-text translation, and I'm not sure it would actually make Google \$ to work on it

great text chat with contractor in indonesia almost certainly meets that bar though

[Yudkowsky][16:14]

furthermore, Eliezer and Paul publicized their debate sufficiently to some internal Deepmind people who spoke to the right other people at Deepmind, that Deepmind showed a graph of loss vs. previous-SOTA methods, and Deepmind's graph shows that their thing crosses the previous-SOTA line while having used 12x less compute for inference training.

(note that this is less... salient?... on the Eliezerverse per se, than it is as an important issue and surprise on the Paulverse, so I am less confident about part.)

a nitpicker would note that previous-SOTA metric they used is however from 1 year previously and the new model also uses Sideways Batch Regularization which the 1-year previous SOTA graph didn't use. on the other hand, they got 12x rather than 10x improvement so there was some error margin there.

[Christiano][16:15]

I'm OK if they don't have the benchmark graph as long as they have some evaluation of what other people were trying at, I think real-time chat probably qualifies

[Yudkowsky][16:15]

but then it's harder to measure the 10x

[Christiano][16:15]

also I'm saying 10x less training compute, not inference (but 10x less inference compute is harder)

yes

[Yudkowsky][16:15]

or to know that Deepmind didn't just use a bunch more compute

[Christiano][16:15]

in practice it seems almost certain that it's going to be harder to evaluate

though I agree there are really clean versions where they actually measured a benchmark other people work on and can compare training compute directly

(like in the transformer paper)

[Yudkowsky][16:16]

literally a pessimal typo, I meant to specify training vs. inference and somehow managed to type "inference" instead

[Christiano][16:16]

I'm more surprised by the clean version

[Yudkowsky][16:17]

I literally don't know what you'd be surprised by in the unclean version

was GPT-2 beating the field hard enough that it would have been surprising if they'd only used similar amounts of training compute

?

and how would somebody else judge that for a new system?

[Christiano][16:17]

I'd want to look at either human evals or logprob, I think probably not? but it's possible it was

[Yudkowsky][16:19]

btw I also feel like the Eliezer model is more surprised and impressed by "they beat the old model with 10x less compute" than by "the old model can't catch up to the new model with 10x more compute"

the Eliezerverse thinks in terms of techniques that saturate such that you have to find new techniques for new training to go on helping

[Christiano][16:19]

it's definitely way harder to win at the old task with 10x less compute

[Yudkowsky][16:19]

but for expensive models it seems really genuinely unlikely to me that anyone will give this data!

[Christiano][16:19]

I think it's usually the case that if you scale up far enough past previous sota, you will be able to find tons of techniques needed to make it work at the new scale

but I'm expecting it to be less of a big deal because all experiments will be roughly at the frontier of what is feasible

and so the new thing won't be able to afford to go 10x bigger

unlike today when we are scaling up spending so fast

but this does make it harder for the next few years at least, which is maybe the key period

(it makes it hard if we are both close enough to the edge that "10x cheaper to get old results" seems unlikely but "getting new results that couldn't be achieved with 10x more compute and old method" seems likely)

what I basically expect is to (i) roughly know how much performance you get from models 10x bigger, (ii) roughly know how much someone beat the competition, and then you can compare the numbers

[Yudkowsky][16:22]

well, you could say, not in a big bet-winning sense, but in a mild trend sense, that if the next few years are full of "they spent 100x more on compute in this domain and got much better results" announcements, that is business as usual for the last few years and is perfectly on track for the Paulverse; while the Eliezerverse permits but does not mandate that we will also see occasional announcements about brilliant new techniques, from some field where somebody already scaled up to the ~~big models~~ big compute, producing more impressive results than the previous big compute.

[Christiano][16:23]

(but "performance from making models 10x bigger" depends a lot on exactly how big they were and whether you are in a regime with unfavorable scaling)

[Yudkowsky][16:23]

so the Eliezerverse must be putting at least a *little* less probability mass on business-a before Paulverse

[Christiano][16:24]

I am also expecting a general scale up in ML training runs over time, though it's plausible that you also expect that until the end of days and just expect a much earlier end of d

[Yudkowsky][16:24]

I mean, why wouldn't they?

if they're purchasing more per unit of compute, they will quite often spend more on total compute (Jevons Paradox)

[Christiano][16:25]

that's going to kill the "they spent 100x more compute" announcements soon enough like, that's easy when "100x more" means \$1M, it's a bit hard when "100x more" means \$100M, it's not going to happen except on the most important tasks when "100x more" means \$10B

[Yudkowsky][16:26]

the Eliezerverse is full of weird things that somebody could apply ML to, and doesn't have many professionals who will wander down completely unwalked roads; and so it is much more friendly to announcements that "we tried putting a lot of work and computation into protein folding, since nobody ever tried doing that seriously with protein folding before, look what came out" continuing for the next decade if the Earth lasts that long

[Christiano][16:27]

I'm not surprised by announcements like protein folding, it's not that the world overall gets more and more hostile to big wins, it's that any industry gets more and more hostile as it gets bigger (or across industries, they get more and more hostile as the stakes grow)

[Yudkowsky][16:28]

well, the Eliezerverse has more weird novel profitable things, because it has more weirdness; and more weird novel profitable things, because it has fewer people diligent about going around trying all the things that will sound obvious in retrospect; but it also has fewer weird novel profitable things, because it has fewer novel things that are allowed to be profitable.

[Christiano][16:29]

(I mean, the protein folding thing is a datapoint against my view, but it's not that much evidence and it's not getting bigger over time)

yeah, but doesn't your view expect more innovations for any given problem?

like, it's not just that you think the universe of weird profitable applications is larger, you also think AI progress is just more driven by innovations, right?

otherwise it feels like the whole game is about whether you think that AI-automating-AI progress is a weird application or something that people will try on

[Yudkowsky][16:30]

the Eliezerverse is more strident about there being lots and lots more stuff like "ReLUs and "batch normalization" and "transformers" in the design space in principle, and less strident about whether current people are being paid to spend all day looking for them rather than putting their efforts someplace with a nice predictable payoff.

[Christiano][16:31]

yeah, but then don't you see big wins from the next transformers?

and you think those just keep happening even as fields mature

[Yudkowsky][16:31]

it's much more *permitted* in the Eliezerverse than in the Paulverse

[Christiano][16:31]

or you mean that they might slow down because people stop working on them?

[Yudkowsky][16:32]

this civilization has mental problems that I do not understand well enough to predict, when it comes to figuring out how they'll affect the field of AI as it scales

that said, I don't see us getting to AGI on Stack More Layers.

there may perhaps be a bunch of stacked layers in an AGI but there will be more ideas it than that.

such that it would require far, far more than 10X compute to get the same results with GPT-like architecture if that was literally possible

[Christiano][16:33]

it seems clear that it will be more than 10x relative to GPT

I guess I don't know what GPT-like architecture means, but from what you say it seems like normal progress would result in a non-GPT-like architecture

so I don't think I'm disagreeing with that

[Yudkowsky][16:34]

I also don't think we're getting there by accumulating a ton of shallow insights; I expect takes at least one more big one, maybe 2-4 big ones.

[Christiano][16:34]

do you think transformers are a big insight?

(is adding soft attention to LSTMs a big insight?)

[Yudkowsky][16:34]

hard to deliver a verdict of history there

no

[Christiano][16:35]

(I think the intellectual history of transformers is a lot like "take the LSTM out of the LS with attention")

[Yudkowsky][16:35]

"how to train deep gradient descent without activations and gradients blowing up or dying out" was a big insight

[Christiano][16:36]

that really really seems like the accumulation of small insights

[Yudkowsky][16:36]

though the history of that big insight is legit complicated

[Christiano][16:36]

like, residual connections are the single biggest thing

and relus also help

and batch normalization helps

and attention is better than lstms

[Yudkowsky][16:36]

and the inits help (like xavier)

[Christiano][16:36]

you could also call that the accumulation of big insights, but the point is that it's an accumulation of a lot of stuff

mostly developed in different places

[Yudkowsky][16:37]

but on the Yudkowskian view the biggest insight of all was the one waaaay back at the beginning where they were initing by literally unrolling Restricted Boltzmann Machines

and people began to say: *hey if we do this the activations and gradients don't blow up die out*

it is not a history that strongly distinguishes the Paulverse from Eliezerverse, because that insight took time to manifest

it was not, as I recall, the first thing that people said about RBM-unrolling

and there were many little or not-really-so-little inventions that sustained the insight to deeper and deeper nets

and those little inventions did not correspond to huge capability jumps immediately in hands of their inventors, with, I think, the possible exception of transformers

though also I think back then people just didn't do as much SoTA-measuring-and-comparing

[Christiano][16:40]

(I think transformers are a significantly smaller jump than previous improvements)

also a thing we could guess about though

[Yudkowsky][16:40]

right, but did the people who demoed the improvements demo them as big capability jumps?

harder to do when you don't have a big old well funded field with lots of eyes on SoTA claims

they weren't dense in SoTA, I think?

anyways, there has not, so far as I know, been an insight of similar size to that last one since then

[Christiano][16:42]

also 10-100x is still actually surprising to me for transformers
so I guess lesson learned

[Yudkowsky][16:43]

I think if you literally took pre-transformer SoTA, and the transformer paper plus the minimum of later innovations required to make transformers scale at all, then as you tried scaling stuff to GPT-1 scale, the old stuff would probably just flatly not work or asymptote?

[Christiano][16:44]

in general if you take anything developed at scale X and try to scale it way past X I think it won't work
or like, it will work much worse than something that continues to get tweaked

[Yudkowsky][16:44]

I'm not sure I understand what you mean if you mean "10x-100x on transformers actually happened and therefore actually surprised me"

[Christiano][16:44]

yeah, I mean that given everything I know I am surprised that transformers were as large as a 100x improvement on translation
in that paper

[Yudkowsky][16:45]

though it may not help my own case, I remark that my generic heuristics say to have an assistant go poke a bit at that claim and see if your noticed confusion is because you are being more confused by fiction than by reality.

[Christiano][16:45]

yeah, I am definitely interested to understand a bit better what's up there
but tentatively I'm sticking to my guns on the original prediction
if you have random 10-20 person teams getting 100x speedups versus prior sota
as we approach TAI

that's so far from paulverse

[Yudkowsky][16:46]

like, not about this case specially, just sheer reflex from "this assertion in a science pa is surprising" to "go poke at it". many unsurprising and hence unpoked assertions will be false, of course, but the surprising ones even more so on average.

[Christiano][16:48]

anyway, seems like a good approach to finding a concrete disagreement
and even looking back at this conversation would be a start for diagnosing who is mor right in hindsight
main thing is to say how quickly and in what industries I'm how surprised

[Yudkowsky][16:49]

I suspect you want to attach conditions to that surprise? Like, the domain must be sufficiently explored OR sufficiently economically important, because Paulverse also predicts(?) that as of a few years (3?? 2??? 15????) all the economically important stuf will have been poked with lots of compute already.

and if there's economically important domains where nobody's tried throwing \$50M at model yet, that also sounds like not-the-Paulverse?

[Christiano][16:50]

I think the economically important prediction doesn't really need that much of "within few years"
like the total stakes have just been low to date
none of the deep learning labs are that close to paying for themselves
so we're not in the regime where "economic niche > R&D budget"
we are still in the paulverse-consistent regime where investment is driven by the hope future wins
though paul is surprised that R&D budgets aren't *more* larger than the economic value

[Yudkowsky][16:51]

well, it's a bit of a shame from the Eliezer viewpoint that the Paulverse can't be falsified yet, then, considering that in the Eliezerverse it is allowed (but not mandated) for the world to end while most DL labs haven't paid for themselves.

albeit I'm not sure that's true of the present world?

DM had that thing about "we just rejigged cooling the server rooms for Google and got back 1/3 of their investment in us" and that was years ago.

[Christiano][16:52]

I'll register considerable skepticism

[Yudkowsky][16:53]

I don't claim deep knowledge.

But if the imminence, and hence strength and falsifiability, of Paulverse assertions, depend on how much money all the deep learning labs are making, that seems like something we could ask OpenPhil to measure?

[Christiano][16:55]

it seems easier to just talk about ML tasks that people work on

it seems really hard to arbitrate the "all the important niches are invested in" stuff in a way that's correlated with takeoff

whereas the "we should be making a big chunk of our progress from insights" seems like it's easier

though I understand that your view could be disjunctive, of either "AI will have hidden secrets that yield great intelligence," or "there are hidden secret applications that yield incredible profit"

(sorry that statement is crude / not very faithful)

should follow up on this in the future, off for now though

[Yudkowsky][16:58]



Conversation on technology forecasting and gradualism

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This post is a transcript of a multi-day discussion between Paul Christiano, Richard Ngo, Eliezer Yudkowsky, Rob Bensinger, Holden Karnofsky, Rohin Shah, Carl Shulman, Nate Soares, and Jaan Tallinn, following up on the Yudkowsky/Christiano debate in [1](#), [2](#), [3](#), and [4](#).

Color key:

Chat by Paul, Richard, and Eliezer Other chat

12. Follow-ups to the Christiano/Yudkowsky conversation

12.1. Bensinger and Shah on prototypes and technological forecasting

[Bensinger][16:22] (Sep. 23)

Quoth Paul:

seems like you have to make the wright flyer much better before it's important, and that it becomes more like an industry as that happens, and that this is intimately related to why so few people were working on it

Is this basically saying 'the Wright brothers didn't personally capture much value by inventing heavier-than-air flying machines, and this was foreseeable, which is why there wasn't a huge industry effort already underway to try to build such machines as fast as possible.' ?

My maybe-wrong model of Eliezer says here 'the Wright brothers knew a (Thielian) secret', while my maybe-wrong model of Paul instead says:

- They didn't know a secret -- it was obvious to tons of people that you could do something sorta like what the Wright brothers did and thereby invent airplanes; the Wright brothers just had unusual non-monetary goals that made them passionate about doing a thing most people didn't care about.
- Or maybe it's better to say: they knew some specific secrets about physics/engineering, but only because other people *correctly* saw 'there are secrets'

to be found here, but they're stamp-collecting secrets of little economic value to me, so I won't bother to learn the secrets'. ~Everyone knows where the treasure located, and ~everyone knows the treasure won't make you rich.

[Yudkowsky][17:24] (Sep. 23)

My model of Paul says there could be a secret, but only because the industry was tiny the invention was nearly worthless directly.

[Cotra: +]

[Christiano][17:53] (Sep. 23)

I mean, I think they knew a bit of stuff, but it generally takes a lot of stuff to make something valuable, and the more people have been looking around in an area the more confident you can be that it's going to take a lot of stuff to do much better, and it starts to look like an extremely strong regularity for big industries like ML or semiconductors it's pretty rare to find small ideas that don't take a bunch of work to have big impacts
I don't know exactly what a thielian secret is (haven't read the reference and just have vibe)

straightening it out a bit, I have 2 beliefs that combine disjunctively: (i) generally it takes a lot of work to do stuff, as a strong empirical fact about technology, (ii) generally if the returns are bigger there are more people working on it, as a slightly-less-strong fact at sociology

[Bensinger][18:09] (Sep. 23)

secrets = important undiscovered information (or information that's been discovered but isn't widely known), that you can use to get an edge in something.
<https://www.lesswrong.com/posts/ReB7yoF22GuerNfhH/thiel-on-secrets-and-indefiniteness>

There seems to be a Paul/Eliezer disagreement about how common these are in general And maybe a disagreement about how much more efficiently humanity discovers and propagates secrets as you scale up the secret's value?

[Yudkowsky][18:35] (Sep. 23)

Many times it has taken much work to do stuff; there's further key assertions here about "It takes \$100 billion" and "Multiple parties will invest \$10B first" and "\$10B gets you a lot of benefit first because scaling is smooth and without really large thresholds".

Eliezer is like "ah, yes, sometimes it takes 20 or even 200 people to do stuff, but core researchers often don't scale well past 50, and there aren't always predecessors that could do a bunch of the same stuff" even though Eliezer agrees with "it often takes a lot of work to do stuff". More premises are needed for the conclusion, that one alone does not distinguish Eliezer and Paul by enough.

[Bensinger][20:03] (Sep. 23)

My guess is that everyone agrees with claims 1, 2, and 3 here (please let me know if I'm wrong!):

1. The history of humanity looks less like **Long Series of Cheat Codes World**, and more like **Well-Designed Game World**.

In Long Series of Cheat Codes World, human history looks like this, over and over: Some guy found a cheat code that totally outclasses everyone else and makes him God or Emperor, until everyone else starts using the cheat code too (if the Emperor allows it). After which things are maybe normal for another 50 years, until a new Cheat Code arises that makes its first adopters invincible gods relative to the previous tech generation, and then the cycle repeats.

In Well-Designed Game World, you can sometimes eke out a small advantage, and the balance isn't *perfect*, but it's pretty good and the leveling-up tends to be gradual. A level 100 character totally outclasses a level 1 character, and some level transitions are a bigger deal than others, but there's no level that makes you a god relative to the people one level below you.

2. General intelligence took over the world once. Someone who updated on that fact but otherwise hasn't thought much about the topic should not consider it 'bonkers' that machine general intelligence could take over the world too, even though they should still consider it 'bonkers' that eg a coffee startup could take over the world.

(Because beverages have never taken over the world before, whereas general intelligence has; and because our inside-view models of coffee and of general intelligence make it a lot harder to imagine plausible mechanisms by which coffee could make someone emperor, kill all humans, etc., compared to general intelligence.)

(In the game analogy, the situation is a bit like 'I've never found a crazy cheat code or exploit in this game, but I haven't ruled out that there is one, and I heard of a character once who did a lot of crazy stuff that's at least *suggestive* that she might have had a cheat code'.)

3. AGI is arising in a world where agents with science and civilization already exist, whereas humans didn't arise in such a world. This is one reason to think AGI might not take over the world, but it's *not* a strong enough consideration on its own to make the scenario 'bonkers' (because AGIs are likely to differ from humans in many respects, and wouldn't obviously be bonkers if the first AGIs turned out to be qualitatively way smart cheaper to run, etc.).

If folks agree with the above, then I'm confused about how one updates from the above epistemic state to 'bonkers'.

It was to a large extent physics facts that determined how easy it was to understand the feasibility of nukes without (say) decades of very niche specialized study. Likewise, it was physics facts that determined you need rare materials, many scientists, and a large engineering+infrastructure project to build a nuke. In a world where the *physics* of nukes resulted in it being some PhD's quiet 'nobody thinks this will work' project like Andrew Wiles secretly working on a proof of Fermat's Last Theorem for seven years, that would have *happened*.

If an alien came to me in 1800 and told me that totally new physics would let future humans build city-destroying superbombs, then I don't see why I should have considered

it bonkers that it might be lone mad scientists rather than nations who built the first superbomb. The 'lone mad scientist' scenario sounds more conjunctive to me (assume the mad scientist knows something that isn't widely known, AND has the ability to act on that knowledge without tons of resources), so I guess it should have gotten less probability, but maybe not dramatically less?

'Mad scientist builds city-destroying weapon in basement' sounds wild to me, but I feel like almost all of the actual unlikeliness comes from the 'city-destroying weapons exist at all' part, and then the other parts only moderately lower the probability.

Likewise, I feel like the prima-facie craziness of basement AGI mostly comes from 'generally intelligence is a crazy thing, it's wild that anything could be that high-impact' and a much smaller amount comes from 'it's wild that something important could happen in some person's basement'.

It *does* structurally make sense to me that Paul might know things I don't about GPT-3 and/or humans that make it obvious to him that we roughly know the roadmap to AGI if it's this.

If the entire 'it's bonkers that some niche part of ML could crack open AGI in 2026 and reveal that GPT-3 (and the mainstream-in-2026 stuff) was on a very different part of the tech tree' view is coming from a detailed inside-view model of intelligence like this, then that immediately ends my confusion about the argument structure.

I don't understand why you think you have the roadmap, and given a high-confidence roadmap I'm guessing I'd still put more probability than you on someone finding a very different, shorter path that works too. But the *argument structure* "roadmap therefore bonkers" makes sense to me.

If there are meant to be *other* arguments against 'high-impact AGI via niche ideas/techniques' that are strong enough to make it bonkers, then I remain confused about the argument structure and how it can carry that much weight.

I can imagine an inside-view model of human cognition, GPT-3 cognition, etc. that tells you 'AGI coming from nowhere in 3 years is bonkers'; I can't imagine an ML-is-a-reasonably-efficient-market argument that does the same, because even a perfectly efficient market isn't *omniscient* and can still be surprised by undiscovered physics facts that tell you 'nukes are relatively easy to build' and 'the fastest path to nukes is relatively hard to figure out'.

(Caveat: I'm using the 'basement nukes' and 'Fermat's last theorem' analogy because it helps clarify the principles involved, not because I think AGI will be that extreme on the spectrum.)

[Yudkowsky: +1]

Oh, I also wouldn't be confused by a view like "I think it's 25% likely we'll see a more Eliezer-ish world. But it sounds like Eliezer is, like, 90% confident that will happen, and *that level of confidence* (and/or the weak reasoning he's provided for that confidence) seems bonkers to me."

The thing I'd be confused by is e.g. "ML is efficient-ish, therefore *the out-of-the-blue-AGI scenario itself* is bonkers and gets, like, 5% probability."

[Shah][1:58] (Sep. 24)

(I'm unclear on whether this is acceptable for this channel, please let me know if not)

I can't imagine an ML-is-a-reasonably-efficient-market argument that does the same because even a perfectly efficient market isn't omniscient and can still be surprised by undiscovered physics facts

I think this seems right as a first pass.

Suppose we then make the empirical observation that in tons and tons of other fields, extremely rare that people discover new facts that lead to immediate impact. (Set aside for now whether or not that's true; assume that it is.) Two ways you could react to this

1. Different fields are different fields. It's not like there's a common generative process that outputs a distribution of facts and how hard they are to find that is common across fields. Since there's no common generative process, facts about field X shouldn't be expected to transfer to make predictions about field Y.
2. There's some latent reason, that we don't currently know, that makes it so that it is rare for newly discovered facts to lead to immediate impact.

It seems like you're saying that (2) is not a reasonable reaction (i.e. "not a valid argument structure"), and I don't know why. There are lots of things we don't know, is it really so bad to posit one more?

(Once we agree on the argument structure, we should then talk about e.g. reasons why such a latent reason can't exist, or possible guesses as to what the latent reason is, etc but fundamentally I feel generally okay with starting out with "there's probably some reason for this empirical observation, and absent additional information, I should expect that reason to continue to hold".)

[Bensinger][3:15] (Sep. 24)

I think 2 is a valid argument structure, but I didn't mention it because I'd be surprised had enough evidential weight (in this case) to produce an 'update to bonkers'. I'd love hear more about this if anyone thinks I'm under-weighting this factor. (Or any others I out!)

[Shah][23:57] (Sep. 25)

Idk if it gets all the way to "bonkers", but (2) seems pretty strong to me, and is how I would interpret Paul-style arguments on timelines/takeoff if I were taking on what-I-believe-to-be your framework

[Bensinger][11:06] (Sep. 25)

Well, I'd love to hear more about that!

Another way of getting at my intuition: I feel like a view that assigns very small probability to 'suddenly vastly superhuman AI, because something that high-impact hasn't happened before'

(which still seems weird to me, because physics doesn't know what 'impact' is and I do see what physical mechanism could forbid it that strongly and generally, short of simulation hypotheses)

... would also assign very small probability in 1800 to 'given an alien prediction that totally new physics will let us build superbombs at least powerful enough to level cities, the superbomb in question will ignite the atmosphere or otherwise destroy the Earth'.

But this seems flatly wrong to me -- if you buy that the bomb works by a totally different mechanism (and exploits a different physics regime) than eg gunpowder, then the outcome of the bomb is a *physics* question, and I don't see how we can concentrate our probabilities much without probing the relevant physics. The history of boat and building sizes is a negligible input to 'given a totally new kind of bomb that suddenly lets us (at least) destroy cities, what is the total destructive power of the bomb?'.

[Yudkowsky: +1]

(Obviously the bomb *didn't* destroy the Earth, and I wouldn't be surprised if there's some Bayesian evidence or method-for-picking-a-prior that could have validly helped you suspect as much in 1800? But it would be a suspicion, not a confident claim.)

[Shah][1:45] (Sep. 27)

would also assign very small probability in 1800 to 'given an alien prediction that totally new physics will let us build superbombs at least powerful enough to level cities, the superbomb in question will ignite the atmosphere or otherwise destroy the Earth'

(As phrased you also have to take into account the question of whether humans would deploy the resulting superbomb, but I'll ignore that effect for now.)

I think this isn't exactly right. The "totally new physics" part seems important to update on.

Let's suppose that, in the reference class we built of boat and building sizes, empirical nukes were the 1 technology out of 20 that had property X. (Maybe X is something like "discontinuous jump in things humans care about" or "immediate large impact on the world" or so on.) Then, I think in 1800 you assign ~5% to 'the first superbomb at least powerful enough to level cities will ignite the atmosphere or otherwise destroy the Earth'

Once you know more details about how the bomb works, you should be able to update away from 5%. Specifically, "entirely new physics" is an important detail that causes you to update away from 5%. I wouldn't go as far as you in throwing out reference classes entirely at that point -- there can still be unknown latent factors that apply at the level of physics -- but I agree reference classes look harder to use in this case.

With AI, I start from ~5% and then I don't really see any particular detail for AI that I think I should strongly update on. My impression is that Eliezer thinks that "general intelligence" is a qualitatively different sort of thing than that-which-neural-nets-are-doing, and maybe that's what's analogous to "entirely new physics". I'm pretty unconvinced of this, but something in this genre feels quite crux-y for me.

Actually, I think I've lost the point of this analogy. What's the claim for AI that's analogous to

'given an alien prediction that totally new physics will let us build superbombs at least powerful enough to level cities, the superbomb in question will ignite the atmosphere or otherwise destroy the Earth'

?

Like, it seems like this is saying "We figure out how to build a new technology that does X. What's the chance it has side effect Y?" Where X and Y are basically unrelated.

I was previously interpreting the argument as "if we know there's a new superbomb based on totally new physics, and we know that the first such superbomb is at least capable of leveling cities, what's the probability it would have enough destructive force to also destroy the world", but upon rereading that doesn't actually seem to be what you were gesturing at.

[Bensinger][3:08] (Sep. 27)

I'm basically responding to this thing Ajeya wrote:

I think Paul's view would say:

- Things certainly happen for the first time
- When they do, they happen at small scale in shitty prototypes, like the Wright Flyer or GPT-1 or AlphaGo or the Atari bots or whatever
- When they're making a big impact on the world, it's after a lot of investment and research, like commercial aircrafts in the decades after Kitty Hawk or like the investments people are in the middle of making now with AI that can assist with coding

To which my reply is: I agree that the first AGI systems will be shitty compared to *later* AGI systems. But Ajeya's Paul-argument seems to additionally require that AGI system be relatively unimpressive at cognition compared to preceding AI systems that weren't AGI.

If this is because of some general law that things are shitty / low-impact when they "happen for the first time", then I don't understand what physical mechanism could produce such a general law that holds with such force.

As I see it, physics 'doesn't care' about human conceptions of impactfulness, and will instead produce AGI prototypes, aircraft prototypes, and nuke prototypes that have as much impact as is implied by the detailed case-specific workings of general intelligence, flight, and nuclear chain reactions respectively.

We could frame the analogy as:

- 'If there's a year where AI goes from being unable to do competitive par-human reasoning in the hard sciences, to being able to do such reasoning, we should estimate the impact of the first such systems by drawing on our beliefs about par-human scientific reasoning itself.'
- Likewise: 'If there's a year where explosives go from being unable to destroy cities to being able to destroy cities, we should estimate the impact of the first such explosives by drawing on our beliefs about how (current or future) physics might allow a city to be destroyed, and what other effects or side-effects such a process might have. We should spend little or no time thinking about the impactfulness of the first steam engine or the first telescope.'

[Shah][3:14] (Sep. 27)

Seems like your argument is something like "when there's a zero-to-one transition, then you have to make predictions based on reasoning about the technology itself". I think in that case I'd say this thing from above:

My impression is that Eliezer thinks that "general intelligence" is a qualitatively different sort of thing than that-which-neural-nets-are-doing, and maybe that's what's analogous to "entirely new physics". I'm pretty unconvinced of this, but something in this genre feels quite crux-y for me.

(Like, you wouldn't a priori expect anything special to happen once conventional bombs become big enough to demolish a football stadium for the first time. It's because nukes are based on "totally new physics" that you might expect unprecedented new impacts from nukes. What's the analogous thing for AGI? Why isn't AGI just regular AI but scaled up in a way that's pretty continuous?)

I'm curious if you'd change your mind if you were convinced that AGI is just regular AI scaled up, with no qualitatively new methods -- I expect you wouldn't but idk why

[Bensinger][4:03] (Sep. 27)

In my own head, the way I think of 'AGI' is basically: "Something happened that allows humans to do biochemistry, materials science, particle physics, etc., even though none of those things were present in our environment of evolutionary adaptedness. Eventually those systems will similarly be able to generalize to biochemistry, materials science, particle physics, etc. We can call that kind of AI 'AGI'."

There might be facts I'm unaware of that justify conclusions like 'AGI is mostly just a bigger version of current ML systems like GPT-3', and there might be facts that justify conclusions like 'AGI will be preceded by a long chain of predecessors, each slightly less general and slightly less capable than its successor'.

But if so, I'm assuming those will be facts about CS, human cognition, etc., not at all a bunch of facts like 'the first steam engine didn't take over the world', 'the first telescope didn't take over the world'.... Because the physics of brains doesn't care about those things, and because in discussing brains we're already in 'things that have been known to take over the world' territory.

(I think that paying much attention *at all* to the technology-wide base rate for 'does this allow you to take over the world?', once you already know you're doing something like 'inventing a new human', doesn't really make sense at all? It sounds to me like going to a bookstore and then repeatedly worrying 'What if they don't have the book I'm looking for?' Most stores don't sell books at all, so this one might not have the one I want.' If you know it's a *book* store, then you shouldn't be thinking at that level of generality at all; the base rate just goes out the window.)

[Yudkowsky: +1]

My way of thinking about AGI is pretty different from saying AGI follows 'totally new mystery physics' -- I'm explicitly anchoring to a known phenomenon, humans.

The analogous thing for nukes might be 'we're going to build a bomb that uses processes kind of like the ones found in the Sun in order to produce enough energy to destroy (at least) a city'.

[Shah][0:44] (Sep. 28)

The analogous thing for nukes might be 'we're going to build a bomb that uses processes kind of like the ones found in the Sun in order to produce enough energy to destroy (at least) a city'.

(And I assume the contentious claim is "that bomb would then ignite the atmosphere, destroy the world, or otherwise have hugely more impact than just destroying a city".)

In 1800, we say "well, we'll probably just make existing fires / bombs bigger and bigger until they can destroy a city, so we shouldn't expect anything particularly novel or crazy to happen", and assign (say) 5% to the claim.

There is a wrinkle: you said it was processes like the ones found in the Sun. Idk what the state of knowledge was like in 1800, but maybe they knew that the Sun couldn't be a conventional fire. If so, then they could update to a higher probability.

(You could also infer that since someone bothered to mention "processes like the ones found in the Sun", those processes must be ones we don't know yet, which also allows you to make that update. I'm going to ignore that effect, but I'll note that this is one way in which the phrasing of the claim is incorrectly pushing you in the direction of "assign higher probability", and I think a similar thing happens for AI when saying "processes like those in the human brain".)

With AI I don't see why the human brain is a different kind of thing than (say) convnets. So I feel more inclined to just take the starting prior of 5%.

Presumably you think that assigning 5% to the nukes claim in 1800 was incorrect, even though that perspective doesn't know that the Sun is not just a very big conventional fire. I'm not sure why this is. According to me this is just the natural thing to do because things are usually continuous and so in the absence of detailed knowledge that's what your prior should be. (If I had to justify this, I'd point to facts about bridges and buildings and materials science and so on.)

there might be facts that justify conclusions like 'AGI will be preceded by a long chain of slightly-less-general, slightly-less-capable successors'.

The frame of "justify[ing] conclusions" seems to ask for more confidence than I expect to get. Rather I feel like I'm setting an initial prior that could then be changed radically by engaging with details of the technology. And then I'm further saying that I don't see any particular details that should cause me to update away significantly (but they could arise in the future).

For example, suppose I have a random sentence generator, and I take the first well-formed claim it spits out. (I'm using a random sentence generator so that we don't update on the process by which the claim was generated.) This claim turns out to be "Alice has a fake skeleton hidden inside her home". Let's say we know nothing about Alice except that she is a real person somewhere in the US who has a home. You can still assign < 10% probability to the claim, and take 10:1 bets with people who don't know any additional details about Alice. Nonetheless, as you learn more about Alice, you could update towards higher probability, e.g. if you learn that she loves Halloween, that's a modest update; if you learn she runs a haunted house at Halloween every year, that's a large update; if you go to her house and see the fake skeleton you can update to ~100%. That's the sort of situation I feel like we're in with AI.

If you asked me what facts justify the conclusion that Alice probably doesn't have a fake skeleton hidden inside her house, I could only point to reference classes, and all the other people I've met who don't have such skeletons. This is not engaging with the details of

Alice's situation, and I could similarly say "if I wanted to know about Alice, surely I should spend most of my time learning about Alice, rather than looking at what Bob and Carol did". Nonetheless, it is still correct to assign < 10% to the claim.

It really does seem to come down to -- why is human-level intelligence such a special turning point that should receive special treatment? Just as you wouldn't give special treatment to "the first time bridges were longer than 10m", it doesn't seem obvious that there's anything all that special at the point where AIs reach human-level intelligence least for the topics we're discussing; there are obvious reasons that's an important point when talking about the economic impact of AI)

[Tallinn][7:04] (Sep. 28)

FWIW, my current 1-paragraph compression of the debate positions is something like:

catastrophists: when evolution was gradually improving hominid brains, suddenly something clicked - it stumbled upon the core of general reasoning - and hominids went from banana classifiers to spaceship builders. hence we should expect a similar (but much sharper, given the process speeds) discontinuity with AI.

gradualists: no, there was no discontinuity with hominids per se; human brains merely reached a threshold that enabled cultural accumulation (and in a meaningful sense it was *culture* that built those spaceships). similarly, we should not expect sudden discontinuities with AI per se, just an accelerating (and possibly unfavorable to humans) cultural changes as human contributions will be automated away.

—
one possible crux to explore is “how thick is culture”: is it something that AGI will quickly decouple from (dropping directly to physics-based ontology instead) OR will culture remain AGI’s main environment/ontology for at least a decade.

[Ngo][11:18] (Sep. 28)

FWIW, my current 1-paragraph compression of the debate positions is something like:

catastrophists: when evolution was gradually improving hominid brains, suddenly something clicked - it stumbled upon the core of general reasoning - and hominids went from banana classifiers to spaceship builders. hence we should expect a similar (but much sharper, given the process speeds) discontinuity with AI.

gradualists: no, there was no discontinuity with hominids per se; human brains merely reached a threshold that enabled cultural accumulation (and in a meaningful sense it was *culture* that built those spaceships). similarly, we should not expect sudden discontinuities with AI per se, just an accelerating (and possibly unfavorable to humans) cultural changes as human contributions will be automated away.

—
one possible crux to explore is “how thick is culture”: is it something that AGI will quickly decouple from (dropping directly to physics-based ontology instead) OR will culture remain AGI’s main environment/ontology for at least a decade.

Clarification: in the sentence “just an accelerating (and possibly unfavorable to humans) cultural changes as human contributions will be automated away”, what work is “cultu

"changes" doing? Could we just say "changes" (including economic, cultural, etc) instead?

In my own head, the way I think of 'AGI' is basically: "Something happened that allows humans to do biochemistry, materials science, particle physics, etc., even though none of those things were present in our environment of evolutionary adaptedness. Eventually, AI will similarly be able to generalize to biochemistry, materials science, particle physics, etc. We can call that kind of AI 'AGI'."

There might be facts I'm unaware of that justify conclusions like 'AGI is mostly just a bigger version of current ML systems like GPT-3', and there might be facts that justify conclusions like 'AGI will be preceded by a long chain of predecessors, each slightly less general and slightly less capable than its successor'.

But if so, I'm assuming those will be facts about CS, human cognition, etc., not at all a list of a hundred facts like 'the first steam engine didn't take over the world', 'the first telescope didn't take over the world'.... Because the physics of brains doesn't care about those things, and because in discussing brains we're already in 'things that have been known to take over the world' territory.

(I think that paying much attention *at all* to the technology-wide base rate for 'does this allow you to take over the world?', once you already know you're doing something like 'inventing a new human', doesn't really make sense at all? It sounds to me like going to a bookstore and then repeatedly worrying 'What if they don't have the book I'm looking for? Most stores don't sell books at all, so this one might not have the one I want.' If you know it's a *book* store, then you shouldn't be thinking at that level of generality at all; the base rate just goes out the window.)

I'm broadly sympathetic to the idea that claims about AI cognition should be weighted more highly than claims about historical examples. But I think you're underrating historical examples. There are at least three ways those examples can be informative telling us about:

1. Domain similarities
2. Human effort and insight
3. Human predictive biases

You're mainly arguing against 1, by saying that there are facts about physics, and facts about intelligence, and they're not very related to each other. This argument is fairly compelling to me (although it still seems plausible that there are deep similarities which we don't understand yet - e.g. the laws of statistics, which apply to many different domains).

But historical examples can also tell us about #2 - for instance, by giving evidence that great leaps of insight are rare, and so if there exists a path to AGI which doesn't require great leaps of insight, that path is more likely than one which does.

And they can also tell us about #3 - for instance, by giving evidence that we usually overestimate the differences between old and new technologies, and so therefore those same biases might be relevant to our expectations about AGI.

[Bensinger][12:31] (Sep. 28)

In the 'alien warns about nukes' example, my intuition is that 'great leaps of insight are rare' and 'a random person is likely to overestimate the importance of the first steam

engines and telescopes' tell me practically nothing, compared to what even a small amount of high-uncertainty physics reasoning tells me.

The 'great leap of insight' part tells me ~nothing because even if there's an easy low-insight path to nukes and a hard high-insight path, I don't thereby know the explosive yield of a bomb on either path (either absolutely or relatively); it depends on how nukes work.

Likewise, I don't think 'a random person is likely to overestimate the first steam engine' really helps with estimating the power of nuclear explosions. I could *imagine* a world where this bias exists and is so powerful and inescapable it ends up being a big weight on the scales, but I don't think we live in that world?

I'm not even sure that a random person *would* overestimate the importance of prototyping in general. Probably, I guess? But my intuition is still that you're better off in 1800 focusing on physics calculations rather than the tug-of-war 'maybe X is cognitively biasing me in *this* way, no wait maybe Y is cognitively biasing me in *this* other way, no wait...'

Our situation might not be analogous to the 1800-nukes scenario (e.g., maybe we know by observation that current ML systems are basically scaled-down humans). But if it *is* analogous, then I think the history-of-technology argument is not very useful here.

[Tallinn][13:00] (Sep. 28)

re "cultural changes": yeah, sorry, i meant "culture" in very general "substrate of human society" sense. "cultural changes" would then include things like changes in power structures and division of labour, but *not* things like "diamondoid bacteria killing all humans in 1 second" (that would be a change in humans, not in the culture)

[Shah][13:09] (Sep. 28)

I want to note that I agree with your (Rob's) latest response, but I continue to think most of the action is in whether AGI involves something analogous to "totally new physics", where I would guess "no" (and would do so particularly strongly for shorter timelines).

(And I would still point to historical examples for "many new technologies don't involve something analogous to 'totally new physics'", and I'll note that Richard's #2 about human effort and insight still applies)

12.2. Yudkowsky on Steve Jobs and gradualism

[Yudkowsky][15:26] (Sep. 28)

So recently I was talking with various people about the question of why, for example, Steve Jobs could not find somebody else with UI taste 90% as good as his own, to take over Apple, even while being able to pay infinite money. A successful founder I was

talking to was like, "Yep, I sure would pay \$100 million to hire somebody who could do 80% of what I can do, in fact, people have earned more than that for doing less."

I wondered if OpenPhil was an exception to this rule, and people with more contact with OpenPhil seemed to think that OpenPhil did not have 80% of a Holden Karnofsky (besides Holden).

And of course, what sparked this whole thought process in me, was that I'd staked all the effort I put into the Less Wrong sequences, into the belief that if I'd managed to bring myself into existence, then there ought to be lots of young near-Eliezers in Earth's person-space including some with more math talent or physical stamina not so unusual, who could be started down the path to being Eliezer by being given a much larger dose of concentrated hints than I got, starting off the compounding cascade of skill formations that I saw as having been responsible for producing me, "on purpose instead of by accident".

I see my gambit as having largely failed, just like the successful founder couldn't pay \$100 million to find somebody 80% similar in capabilities to himself, and just like Steve Jobs could not find anyone to take over Apple for presumably much larger amounts of money and status and power. Nick Beckstead had some interesting stories about various ways that Steve Jobs had tried to locate successors (which I wasn't even aware of).

I see a plausible generalization as being a "Sparse World Hypothesis": The shadow of Earth with eight billion people, projected into some dimensions, is much sparser than plausible arguments might lead you to believe. Interesting people have few neighbors even when their properties are collapsed and projected onto lower-dimensional tests of output production. The process of forming an interesting person passes through enough 0-1 critical thresholds that all have to be passed simultaneously in order to start a process of gaining compound interest in various skills, that they then cannot find other people who are 80% as good as what they *do* (never mind being 80% similar to them or people).

I would expect human beings to start out much denser in a space of origins than AI projects, and for the thresholds and compounding cascades of our mental lives to be much less sharp than chimpanzee-human gaps.

Gradualism about humans sure sounds totally reasonable! It is in fact much more plausible-sounding *a priori* than the corresponding proposition about AI projects! I staked years of my own life on the incredibly reasoning-sounding theory that if one actual Eliezer existed then there should be lots of neighbors near myself that I could catalyze into existence by removing some of the accidental steps from the process that had accidentally produced me.

But it didn't work in real life because plausible-sounding gradualist arguments just... probably don't work in real life even though they sure sound plausible. I spent a lot of time arguing with Robin Hanson, who was more gradualist than I was, and was taken by surprise when reality itself was much less gradualist than I was.

My model has Paul or Carl coming back with some story about how, why, no, it is totally reasonable that Steve Jobs couldn't find a human who was 90% as good at a problem class as Steve Jobs to take over Apple for billions of dollars despite looking, and, why, this is not at all a falsified retro-prediction of the same gradualist reasoning that says a leading AI project should be inside a dense space of AI projects that projects onto a dense space of capabilities such that it has near neighbors.

If so, I was not able to use this hypothetical model of *selective* gradualist reasoning to deduce in advance that replacements for myself would be sparse in the same sort of

space and I'd end up unable to replace myself.

I do not really believe that, without benefits of hindsight, the advance predictions of gradualism would differ between the two cases.

I think if you don't peek at the answer book in advance, the same sort of person who finds it totally reasonable to expect successful AI projects to have close lesser earlier neighbors, would also find it totally reasonable to think that Steve Jobs definitely ought to be able to find somebody 90% as good to take over his job - and should actually be able to find somebody *much* better because Jobs gets to run a wider search and offer more incentive than when Jobs was wandering into early involvement in Apple.

It's completely reasonable-sounding! Totally plausible to a human ear! Reality disagrees: Jobs tried to find a successor, couldn't, and now the largest company in the world by market cap seems no longer capable of sending the iPhones back to the designers and asking them to do something important differently.

This is part of the story for why I put gradualism into a mental class of "arguments that sound plausible and just fail in real life to be binding on reality; reality says 'so what' a goes off to do something else".

[Christiano][17:46] (Sep. 28)

It feels to me like a common pattern is: I say that ML in particular, and most technology in general, seem to improve quite gradually on metrics that people care about or track. You say that some kind of "gradualism" worldview predicts a bunch of other stuff (some claim about markets or about Steve Jobs or whatever that feels closely related on your view but not mine). But it feels to me like there are just a ton of technologies, and a ton of AI benchmarks, and those are just *much* more analogous to "future AI progress." I know that to you this feels like reference class tennis, but I think I legitimately don't understand what kind of approach to forecasting you are using that lets you just make (what I see as) the obvious boring prediction about all of the non-AGI technologies.

Perhaps you are saying that symmetrically you don't understand what approach to forecasting I'm using, that would lead me to predict that technologies improve gradually yet people vary greatly in their abilities. To me it feels like the simplest thing in the world is to expect future technological progress in domain X to be like past progress in domain Y, and future technological progress to be like past technological progress, and future market moves to be like past market moves, and future elections to be like past elections.

And it seems like you *must* be doing something that ends up making almost the same predictions as that almost all the time, which is why you don't get incredibly surprised every single year by continuing boring and unsurprising progress in batteries or solar panels or robots or ML or computers or microscopes or whatever. Like it's fine if you say "Yes, those areas have trend breaks sometimes" but there are *so many* boring years that you must somehow be doing something like having the baseline "this year is probably going to be boring."

Such that intuitively it feels to me like the disagreement between us *must* be in the part where AGI feels to me like it is similar to AI-to-date and feels to you like it is very different and better compared to evolution of life or humans.

It has to be the kind of argument that you can make about progress-of-AI-on-metrics-people-care-about, but *not* progress-of-other-technologies-on-metrics-people-care-about.

otherwise it seems like you are getting hammered every boring year for every boring technology.

I'm glad we have the disagreement on record where I expect ML progress to continue to get less jumpy as the field grows, and maybe the thing to do is just poke more at that since it is definitely a place where I gut level expect to win bayes points and so could legitimately change my mind on the "which kinds of epistemic practices work better?" question. But it feels like it's not the main action, the main action has got to be about thinking that there is a really impactful change somewhere between {modern AI, lowe animals} and {AGI, humans} that doesn't look like ongoing progress in AI.

I think "would GPT-3 + 5 person-years of engineering effort foom?" feels closer to core me.

(That said, the way AI could be different need not feel like "progress is lumpier," could totally be more like "Progress is always kind of lumpy, which Paul calls 'pretty smooth' and Eliezer calls 'pretty lumpy' and doesn't lead to any disagreements; but Eliezer thinks AGI is different in that kind-of-lumpy progress leads to fast takeoff, while Paul thinks it leads to kind-of-lumpy increases in the metrics people care about or track.")

[Yudkowsky][7:46] (Sep. 29)

I think "would GPT-3 + 5 person-years of engineering effort foom?" feels closer to core to me.

I truly and legitimately cannot tell which side of this you think we should respectively be on. My guess is you're against GPT-3 foaming because it's too low-effort and a short timeline, even though I'm the one who thinks GPT-3 isn't on a smooth continuum with AGI??

With that said, the rest of this feels on-target to me; I sure do feel like {natural selection, humans, AGI} form an obvious set with each other, though even there the internal differences are too vast and the data too scarce for legit outside viewing.

I truly and legitimately cannot tell which side of this you think we should respectively be on. My guess is you're against GPT-3 foaming because it's too low-effort and a short timeline, even though I'm the one who thinks GPT-3 isn't on a smooth continuum with AGI??

I mean I obviously think you can foom starting from an empty Python file with 5 person years of effort if you've got the Textbook From The Future; you wouldn't use the GPT code or model for anything in that, the Textbook says to throw it out and start over.

[Christiano][9:45] (Sep. 29)

I think GPT-3 will foom given very little engineering effort, it will just be much slower than the human foom

and then that timeline will get faster and faster over time

it's also fair to say that it wouldn't foom because the computers would break before it figured out how to repair them (and it would run out of metal before it figured out how mine it, etc.), depending on exactly how you define "foom," but the point is that "you can repair the computers faster than they break" happens much before you can outrun human civilization

so the relevant threshold you cross is the one where you are outrunning civilization

(and my best guess about human evolution is pretty similar, it looks like humans are smart enough to foom over a few hundred thousand years, and that we were the ones foom because that is also roughly how long it was taking evolution to meaningfully improve our cognition---if we foomed slower it would have instead been a smarter successor who overtook us, if we foomed faster it would have instead been a dumber predecessor, though this is *much* less of a sure-thing than the AI case because natural selection is not trying to make something that fooms)

and regarding {natural selection, humans, AGI} the main question is why modern AI a homo erectus (or even chimps) aren't in the set

it feels like the core disagreement is that I mostly see a difference in degree between various animals, and between modern AI and future AI, a difference that is likely to be covered by gradual improvements that are pretty analogous to contemporary improvements, and so as the AI community making contemporary improvements grow get more and more confident that TAI will be a giant industry rather than an innovation

[Ngo][5:45] (Oct. 1)

Do you have a source on Jobs having looked hard for a successor who wasn't Tim Cook?

Also, I don't have strong opinions about how well Apple is doing now, so I default to looking at the share price, which seems very healthy.

(Although I note in advance that this doesn't feel like a particularly important point, roughly for the same reason that Paul mentioned: gradualism about Steve Jobs doesn't seem like a central example of the type of gradualism that informs beliefs about AI development.)

[Yudkowsky][10:40] (Oct. 1)

My source is literally "my memory of stuff that Nick Beckstead just said to me in person", maybe he can say more if we invite him.

I'm not quite sure what to do with the notion that "gradualism about Steve Jobs" is somehow less to be expected than gradualism about AGI projects. Humans are GIs. They are *extremely* similar to each other design-wise. There are a *lot* of humans, billions of them, many many many more humans than I expect AGI projects. Despite this the leading edge of human-GIs is sparse enough in the capability space that there is no 90%-of-Steve-Jobs that Jobs can locate, and there is no 90%-of-von-Neumann known to 20th century history. If we are not to take any evidence about this to A-GIs, then I do not understand the rules you're using to apply gradualism to some domains but not others.

And to be explicit, a skeptic who doesn't find these divisions intuitive, might well ask, "Is gradualism perhaps isomorphic to 'The coin always comes up heads on Heady occasions', where 'Heady' occasions are determined by an obscure intuitive method going through some complicated nonverbalizable steps one of which is unfortunately 'check whether the coin actually came up heads'?"

(As for my own theory, it's always been that AGIs are mostly like AGIs and not very much like humans or the airplane-manufacturing industry, and I do not, on my own account of things, appeal much to supposed outside viewing or base rates.)

[Shulman][11:11] (Oct. 2)

I think the way to apply it is to use observable data (drawn widely) and math.

Steve Jobs does look like a (high) draw (selected for its height, in the sparsest tail of the CEO distribution) out of the economic and psychometric literature (using the same kind of approach I use in other areas like estimating effects of introducing slightly superhuman abilities on science, the genetics of height, or wealth distributions). You have roughly normal or log-normal distributions on some measures of ability (with fatter tails when there are some big factors present, e.g. super-tall people are enriched for normal common variants for height but are more frequent than a Gaussian estimated from the middle range because of some weird disease/hormonal large effects). And we have lots of empirical data about the thickness and gaps there. Then you have a couple effects that can make returns in wealth/output created larger.

You get amplification from winner-take-all markets, IT, and scale that let higher ability add value to more places. This is the same effect that lets top modern musicians make so much money. Better CEOs get allocated to bigger companies because multiplicative management decisions are worth more in big companies. Software engineering becomes more valuable as the market for software grows.

Wealth effects are amplified by multiplicative growth (noise in a given period multiplies wealth for the rest of the series, and systematic biases from abilities can grow exponentially or superexponentially over a lifetime), and there are some versions of that in gaining expensive-to-acquire human capital (like fame for Hollywood actors, or experience using incredibly expensive machinery or companies).

And we can read off the distributions of income, wealth, market share, lead time in innovations, scientometrics, etc.

That sort of data lead you to expect cutting edge tech to be months to a few years ahead of followers, winner-take-all tech markets to a few leading firms and often a clearly dominant one (but not driving an expectation of being able to safely rest on laurels for years while others innovate without a moat like network effects). That's one of my longstanding arguments with Robin Hanson, that his model has more even capabilities and market share for AGI/WBE than typically observed (he says that AGI software will have to be more diverse requiring more specialized companies, to contribute so much GDP).

It is tough to sample for extreme values on multiple traits at once, superexponentially tough as you go out or have more criteria. CEOs of big companies are smarter than average, taller than average, have better social skills on average, but you can't find people who are near the top on several of those.

https://www.hbs.edu/ris/Publication%20Files/16-044_9c05278e-9d11-4315-a744-de008edf4d80.pdf

Correlations between the things help, but it's tough. E.g. if you have thousands of people in a class on a measure of cognitive skill, and you select on only partially correlated matters of personality, interest, motivation, prior experience, etc, the math says it gets thin and you'll find different combos (and today we see more representation of different profiles of abilities, including rare and valuable ones, in this community)

I think the bigger update for me from trying to expand high-quality save the world efforts has been on the funny personality traits/habits of mind that need to be selected and their scarcity.

[Karnofsky][11:30] (Oct. 2)

A cpl comments, without commitment to respond to responses:

1. Something in the zone of "context / experience / obsession" seems important for explaining the Steve Jobs type thing. It seems to me that people who enter an area early tend to maintain an edge even over more talented people who enter later - examples are not just founder/CEO types but also early employees of some companies who are more experienced with higher-level stuff (and often know the history of how they got there) better than later-entering people.

2. I'm not sure if I am just rephrasing something Carl or Paul has said, but something that bugs me a lot about the Rob/Eliezer arguments is that I feel like if I accept >5% probability for the kind of jump they're talking about, I don't have a great understanding of how I avoid giving >5% to a kajillion other claims from various startups that they're about to revolutionize their industry, in ways that seem inside-view plausible and seem to equally "depend on facts about some physical domain rather than facts about reference classes."

The thing that actually most comes to mind here is Thiel - he has been a phenomenal investor financially, but he has also invested by now in a lot of "atoms" startups with big stories about what they might do, and I don't think any have come close to reaching those visions (though they have sometimes made \$ by doing something orders of magnitude less exciting).

If a big crux here is "whether Thielian secrets exist" this track record could be significant.

I think I might update if I had a cleaner sense of how I could take on this kind of "Well, if it is just a fact about physics that I have no idea about, it can't be that unlikely" view without then betting on a lot of other inside-view-plausible breakthroughs that haven't happened. Right now all I can say to imitate this lens is "General intelligence is 'different'"

I don't feel the same way about "AI might take over the world" - I feel like I have good reasons this applies to AI and not a bunch of other stuff

[Soares][11:11] (Oct. 2)

Ok, a few notes from me (feel free to ignore):

1. It seems to me like the convo here is half attempting-to-crux and half attempting-to-distill-out-a-bet. I'm interested in focusing explicitly on cruxing for the time being, for whatever that's worth. (It seems to me like y'all're already trending in that direction.)

2. It seems to me that one big revealed difference between the Eliezerverse and the Paulverse is something like:

- In the Paulverse, we already have basically all the fundamental insights we need for AGI, and now it's just a matter of painstaking scaling.
- In the Eliezerverse, there are large insights yet missing (and once they're found we have plenty of reason to expect things to go quickly).

For instance, in Eliezerverse they say "The Wright flyer didn't need to have historical precedents, it was allowed to just start flying. Similarly, the AI systems of tomorrow are allowed to just start Gling without historical precedent.", and in the Paulverse they say "The analog of the Wright flyer has already happened, it was Alexnet, we are now in the phase analogous to the slow grinding transition from human flight to commercially viable human flight."

(This seems to me like basically what Ajeya articulated [upthread](#).)

3. It seems to me that another revealed intuition-difference is in the difficulty that people have operating each other's models. This is evidenced by, eg, Eliezer/Rob saying things like "I don't know how to operate the gradualness model without making a bunch of bad predictions about Steve Jobs", and Paul/Holden responding with things like "I don't know how to operate the secrets-exist model without making a bunch of bad predictions about material startups".

I'm not sure whether this is a shallower or deeper disagreement than (2). I'd be interested in further attempts to dig into the questions of how to operate the models, in hopes that the disagreement looks interestingly different once both parties can at least operate the other model.

[Tallinn: +]

Ngo's view on alignment difficulty

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This post features a write-up by Richard Ngo on his views, with inline comments.

Color key:



13. Follow-ups to the Ngo/Yudkowsky conversation

13.1. Alignment difficulty debate: Richard Ngo's case

[Ngo][9:31] (Sep. 25)

As promised, here's a write-up of some thoughts from my end. In particular, since I've spent a lot of the debate poking Eliezer about his views, I've tried here to put forward more positive beliefs of my own in this doc (along with some more specific claims): [GDocs link]

[Ngo] [Sep. 25] Google Doc

We take as a starting observation that a number of “grand challenges” in AI have been solved by AIs that are very far from the level of generality which people expected would be needed. Chess, once considered to be the pinnacle of human reasoning, was solved by an algorithm that’s essentially useless for real-world tasks. Go required more flexible learning algorithms, but policies which beat human performance are still nowhere near generalising to anything else; the same for StarCraft, DOTA, and the protein folding problem. Now it seems very plausible that AIs will even be able to pass (many versions of) the Turing Test while still being a long way from AGI.

[Yudkowsky][11:26] (Sep. 25 comment)

Now it seems very plausible that AIs will even be able to pass (many versions of) the Turing Test while still being a long way from AGI.

I remark: Restricted versions of the Turing Test. Unrestricted passing of the Turing Test happens after the world ends. Consider how smart you'd have to be to pose as an AGI to an AGI; you'd need all the cognitive powers of an AGI as well as all of your human powers.

[Ngo][11:24] (Sep. 29 comment)

Perhaps we can quantify the Turing test by asking something like:

- What percentile of competence is the judge?
- What percentile of competence are the humans who the AI is meant to pass as?
- How much effort does the judge put in (measured in, say, hours of strategic preparation)?

Does this framing seem reasonable to you? And if so, what are the highest numbers for each of these metrics that correspond to a Turing test which an AI could plausibly pass before the world ends?

[Ngo] (Sep. 25 Google Doc)

I expect this trend to continue until after we have AIs which are superhuman at mathematical theorem-proving, programming, many other white-collar jobs, and many types of scientific research. It seems like Eliezer doesn't. I'll highlight two specific disagreements which seem to play into this.

[Yudkowsky][11:28] (Sep. 25 comment)

doesn't

Eh? I'm pretty fine with something proving the Riemann Hypothesis before the world ends. It came up during my recent debate with Paul, in fact.

Not so fine with something designing nanomachinery that can be built by factories built by proteins. They're legitimately different orders of problem, and it's no coincidence that the second one has a path to pivotal impact, and the first does not.

[Ngo] (Sep. 25 Google Doc)

A first disagreement is related to Eliezer's characterisation of GPT-3 as a shallow pattern-memoriser. I think there's a continuous spectrum between pattern-memorisation and general intelligence. In order to memorise more and more patterns, you need to start understanding them at a high level of abstraction, draw inferences about parts of the patterns based on other parts, and so on. When those patterns are drawn from the real world, then this process leads to the gradual development of a world-model.

This position seems more consistent with the success of deep learning so far than Eliezer's position (although my advocacy of it loses points for being post-hoc; I was closer to Eliezer's position before the GPTs). It also predicts that deep learning will lead to agents which can reason about the world in increasingly impressive ways (although I don't have a strong position on the extent to which new architectures and algorithms will be required for that). I think that the spectrum from less to more intelligent animals (excluding humans) is a good example of what it looks like to gradually move from pattern-memorisation to increasingly sophisticated world-models and abstraction capabilities.

[Yudkowsky][11:30] (Sep. 25 comment)

In order to memorise more and more patterns, you need to start understanding them at a high level of abstraction, draw inferences about parts of the patterns based on other parts, and so on.

Correct. You can believe this and *not* believe that exactly GPT-like architectures can keep going deeper until their overlap of a greater number of patterns achieves the same level of depth and generalization as human depth and generalization from fewer

patterns, just like pre-transformer architectures ran into trouble in memorizing deeper patterns than the shallower ones those earlier systems could memorize.

[Ngo] (Sep. 25 Google Doc)

I expect that Eliezer won't claim that pattern-memorisation is *unrelated* to general intelligence, but will claim that a pattern-memoriser needs to undergo a sharp transition in its cognitive algorithms before it can reason reliably about novel domains (like open scientific problems) - with his main argument for that being the example of the sharp transition undergone by humans.

However, it seems unlikely to me that humans underwent a major transition in our underlying cognitive algorithms since diverging from chimpanzees, because our brains are so similar to those of chimps, and because our evolution from chimps didn't take very long. This evidence suggests that we should favour explanations for our success which don't need to appeal to big algorithmic changes, if we have any such explanations; and I think we do. More specifically, I'd characterise the three key differences between humans and chimps as:

1. Humans have bigger brains.
2. Humans have a range of small adaptations primarily related to motivation and attention, such as infant focus on language and mimicry, that make us much better at cultural learning.
3. Humans grow up in a rich cultural environment.

[Ngo][9:13] (Sep. 23 comment on earlier draft)

bigger brains

I recall a 3-4x difference; but this paper says 5-6x for frontal cortex: <https://www.nature.com/articles/nrn814>

[Tallinn][3:24] (Sep. 26 comment)

language and mimicry

"apes are unable to ape sounds" claims david deutsch in "the beginning of infinity"

[Barnes][8:09] (Sep. 23 comment on earlier draft)

[Humans grow up in a rich cultural environment.]

much richer cultural environment including deliberate teaching

[Ngo] (Sep. 25 Google Doc)

I claim that the discontinuity between the capabilities of humans and chimps is mainly explained by the general intelligence of chimps not being aimed in the direction of learning the skills required for economically valuable tasks, which in turn is mainly due to chimps lacking the "range of small adaptations" mentioned above.

My argument is a more specific version of Paul's claim that chimp evolution was not primarily selecting for doing things like technological development. In particular, it was not selecting for them because no cumulative cultural environment existed while chimps

were evolving, and selection for the application of general intelligence to technological development is much stronger in a cultural environment. (I claim that the cultural environment was so limited before humans mainly because cultural accumulation is [very sensitive to transmission fidelity](#).)

By contrast, AIs will be trained in a cultural environment (including extensive language use) from the beginning, so this won't be a source of large gains for later systems.

[Ngo][6:01] (Sep. 22 comment on earlier draft)

more specific version of Paul's claim

Based on some of Paul's recent comments, this may be what he intended all along; though I don't recall his original writings on takeoff speeds making this specific argument.

[Shulman][14:23] (Sep. 25 comment)

(I claim that the cultural environment was so limited before humans mainly because cultural accumulation is [very sensitive to transmission fidelity](#).)

There can be other areas with superlinear effects from repeated application of a skill. There's reason to think that the most productive complex industries tend to have that character.

Making individual minds able to correctly execute long chains of reasoning by reducing per-step error rate could plausibly have very superlinear effects in programming, engineering, management, strategy, persuasion, etc. And you could have new forms of 'super-culture' that don't work with humans.

<https://ideas.repec.org/a/eee/jeborg/v85y2013icp1-10.html>

[Ngo] (Sep. 25 Google Doc)

If true, this argument would weigh against Eliezer's claims about agents which possess a core of general intelligence being able to easily apply that intelligence to a wide range of tasks. And I don't think that Eliezer has a compelling alternative explanation of the key cognitive differences between chimps and humans (the closest I've seen in his writings is the brainstorming [at the end of this post](#)).

If this is the case, I notice an analogy between Eliezer's argument against Kurzweil, and my argument against Eliezer. Eliezer attempted to put microfoundations underneath the trend line of Moore's law, which led to a different prediction than Kurzweil's straightforward extrapolation. Similarly, my proposed microfoundational explanation of the chimp-human gap gives rise to a different prediction than Eliezer's more straightforward, non-microfoundational extrapolation.

[Yudkowsky][11:39] (Sep. 25 comment)

Similarly, my proposed microfoundational explanation of the chimp-human gap gives rise to a different prediction than Eliezer's more straightforward, non-microfoundational extrapolation.

Eliezer does not use "non-microfoundational extrapolations" for very much of anything, but there are obvious reasons why the greater Earth does not benefit from me winning debates through convincingly and correctly listing all the particular capabilities you need to add over and above what GPToid architectures can achieve, in order to achieve AGI.

Nobody else with a good model of larger reality will publicly describe such things in a way they believe is correct. I prefer not to argue convincingly but wrongly. But, no, it is not Eliezer's way to sound confident about anything unless he thinks he has a more detailed picture of the microfoundations than the one you are currently using yourself.

[Ngo][11:40] (Sep. 29 comment)

Good to know; apologies for the incorrect inference.

Given that this seems like a big sticking point in the debate overall, do you have any ideas about how to move forward while avoiding infohazards?

[Ngo] (Sep. 25 Google Doc)

My position makes some predictions about hypothetical cases:

1. If chimpanzees had the same motivational and attention-guiding adaptations towards cultural learning and cooperation that humans do, and were raised in equally culturally-rich environments, then they could become economically productive workers in a range of jobs (primarily as manual laborers, but plausibly also for operating machinery, etc).
 1. Results from chimps raised in human families, like [Washoe](#), seem moderately impressive, although still very uncertain. There's probably a lot of bias towards positive findings - but on the other hand, it's only been done a handful of times, and I expect that more practice at it would lead to much better results.
 2. Comparisons between humans and chimps which aren't raised in similar ways to humans are massively biased towards humans. For the purposes of evaluating general intelligence, comparisons between chimpanzees and [feral children](#) seem fairer (although it's very hard to know how much the latter were affected by non-linguistic childhoods as opposed to abuse or pre-existing disabilities).
2. Consider a hypothetical species which has the same level of "general intelligence" that chimpanzees currently have, but is as well-adapted to the domains of abstract reasoning and technological development as chimpanzee behaviour is to the domain of physical survival (e.g. because they evolved in an artificial environment where their fitness was primarily determined by their intellectual contributions). I claim that this species would have superhuman scientific research capabilities, and would be able to make progress in novel areas of science (analogously to how chimpanzees can currently learn to navigate novel physical landscapes).
 1. Insofar as Eliezer doubts this, but *does* believe that this species could outperform a society of village idiots at scientific research, then he needs to explain why the village-idiot-to-Einstein gap is so significant in this context but not in others.
 2. However, this is a pretty weird thought experiment, and maybe doesn't add much to our existing intuitions about AIs. My main intention here is to point at how animal behaviour is *really really well-adapted* to physical environments, in a way which makes people wonder what it would be like to be *really really well-adapted* to intellectual environments.
3. ~~I claim that the difficulty of human-level oracle AGIs matching humans~~ Consider an AI which has been trained only to answer questions, and is now human-level at doing so. I claim that the difficulty of this AI matching humans at a range of real-world tasks (without being specifically trained to do so) would be much closer to the difficulty of teaching chimps to do science, than the difficulty of teaching adult humans to do abstract reasoning about a new domain.
 1. The analogy here is: chimps have reasonably general intelligence, but it's hard for them to apply it to science because they weren't trained to apply intelligence to that. Likewise, human-level oracle AGIs have general intelligence, but it'll be hard for them to apply it to influencing the world because they weren't trained to apply intelligence to that.

[Barnes][8:21] (Sep. 23 comment on earlier draft)

village-idiot-to-Einstein gap

I wonder to what extent you can model within-species intelligence differences partly just as something like hyperparameter search - if you have a billion humans with random variation in their neural/cognitive traits, the top human will be a lot better than average. Then you could say something like:

- humans are the dumbest species you could have where the distribution of intelligence in each generation is sufficient for cultural accumulation
- that by itself might not imply a big gap from chimps
- but human society has much larger population, so the smartest individuals are much smarter

[Ngo][9:05] (Sep. 23 comment on earlier draft)

I think Eliezer's response (which I'd agree with) would be that the cognitive difference between the best humans and normal humans is strongly constrained by the fact that we're all one species who can interbreed with each other. And so our cognitive variation can't be very big compared with inter-species variation (at the top end at least; although it could at the bottom end via things breaking).

[Barnes][9:35] (Sep. 23 comment on earlier draft)

I think that's not obviously true - it's definitely possible that there's a lot of random variation due to developmental variation etc. If that's the case then population size could create large within-species differences

[Yudkowsky][11:46] (Sep. 25 comment)

oracle AGIs

Remind me of what this is? Surely you don't just mean the AI that produces plans it doesn't implement itself, because that AI becomes an agent by adding an external switch that routes its outputs to a motor; it can hardly be much cognitively different from an agent. Then what do you mean, "oracle AGI"?

(People tend to produce shallow specs of what they mean by "oracle" that make no sense in my microfoundations, a la "Just drive red cars but not blue cars!", leading to my frequent reply, "Sorry, still AGI-complete in terms of the machinery you have to build to do that.")

[Ngo][11:44] (Sep. 29 comment)

Edited to clarify what I meant in this context (and remove the word "oracle" altogether).

[Yudkowsky][12:01] (Sep. 29 comment)

My reply holds just as much to "AIs that answer questions"; what restricted question set do you imagine suffices to save the world without dangerously generalizing internal engines?

[Barnes][8:15] (Sep. 23 comment on earlier draft)

The analogy here is: chimps have reasonably general intelligence, but it's hard for them to apply it to science because they weren't trained to apply intelligence to that. Likewise, human-level oracle AGIs have general intelligence, but it'll be hard for them to apply it to influencing the world because they weren't trained to apply intelligence to that.

this is not intuitive to me; it seems pretty plausible that the subtasks of predicting the world and of influencing the world are much more similar than the subtasks of surviving in a chimp society are to the subtasks of doing science

[Ngo][8:59] (Sep. 23 comment on earlier draft)

I think Eliezer's position is that all of these tasks are fairly similar *if you have general intelligence*. E.g. he argued that the difference between very good theorem-proving and influencing the world is significantly smaller than people expect. So even if you're right, I think his position is too strong for your claim to help him. (I expect him to say that I'm significantly overestimating the extent to which chimps are running general cognitive algorithms).

[Barnes][9:33] (Sep. 23 comment on earlier draft)

I wasn't trying to defend his position, just disagreeing with you :P

[Ngo] (Sep. 25 Google Doc)

More specific details

Here are three training regimes which I expect to contribute to AGI:

- Self-supervised training - e.g. on internet text, code, books, videos, etc.
- Task-based RL - agents are rewarded (likely via human feedback, and some version of iterated amplification) for doing well on bounded tasks.
- Open-ended RL - agents are rewarded for achieving long-term goals in rich environments.

[Yudkowsky][11:56] (Sep. 25 comment)

bounded tasks

There's an interpretation of this I'd agree with, but all of the work is being carried by the *boundedness* of the tasks, little or none via the "human feedback" part which I shrug at, and none by the "iterated amplification" part since I consider that tech unlikely to exist before the world ends.

[Ngo] (Sep. 25 Google Doc)

Most of my probability of catastrophe comes from AGIs trained primarily via open-ended RL. Although IA makes these scenarios less likely by making task-based RL more powerful, it doesn't seem to me that IA tackles the hardest case (of aligning agents trained via open-ended RL) head-on. But disaster from open-ended RL also seems a long way away - mainly because getting long-term real-world feedback is very slow, and I expect it to be hard to create sufficiently rich artificial environments. By that point I do expect the strategic landscape to be significantly different, because of the impact of task-based RL.

[Yudkowsky][11:57] (Sep. 25 comment)

a long way away

Oh, definitely, at the present rates of progress we've got years, plural.

The [history of futurism](#) says that even saying that tends to be unreliable in the general case (people keep saying it right up until the Big Thing actually happens) and also that it's rather a difficult form of knowledge to obtain more than a few years out.

[Yudkowsky][12:01] (Sep. 25 comment)

hard to create sufficiently rich artificial environments

Disagree; I don't think that making environments more difficult in a way that challenges the environment inside will prove to be a significant AI development bottleneck. Making simulations easy enough for current AIs to do interesting things in them, but hard enough that the things they do are not completely trivial, takes some work relevant to current levels of AI intelligence. I think that making those environments more tractably challenging for smarter AIs is not likely to be nearly a bottleneck in progress, compared to making the AIs smarter and able to solve the environment. It's a one-way-hash, P-vs-NP style thing - not literally, just that general relationship between it taking a lower amount of effort to pose a problem such that solving it requires a higher amount of effort.

[Ngo] (Sep. 25 Google Doc)

Perhaps the best way to pin down disagreements in our expectations about the effects of the strategic landscape is to identify some measures that could help to reduce AGI risk, and ask how seriously key decision-makers would need to take AGI risk for each measure to be plausible, and how powerful and competent they would need to be for that measure to make a significant difference. Actually, let's lump these metrics together into a measure of "amount of competent power applied". Some benchmarks, roughly in order (and focusing on the effort applied by the US):

- Banning chemical/biological weapons
- COVID
 - Key points: mRNA vaccines, lockdowns, mask mandates
- Nuclear non-proliferation
 - Key points: [Nunn-Lugar Act](#), [stuxnet](#), various treaties
- The International Space Station
 - Cost to US: ~\$75 billion
- Climate change
 - US expenditure: >\$154 billion (but not very effectively)
- Project Apollo
 - Wikipedia says that Project Apollo "was the largest commitment of resources (\$156 billion in 2019 US dollars) ever made by any nation in peacetime. At its peak, the Apollo program employed 400,000 people and required the support of over 20,000 industrial firms and universities."
- WW1
- WW2

[Yudkowsky][12:02] (Sep. 25 comment)

WW2

This level of effort starts to buy significant amounts of time. This level will not be reached, nor approached, before the world ends.

[Ngo] (Sep. 25 Google Doc)

Here are some wild speculations (I just came up with this framework, and haven't thought about these claims very much):

1. The US and China preventing any other country from becoming a leader in AI requires about as much competent power as banning chemical/biological weapons.
2. The US and China enforcing a ban on AIs above a certain level of autonomy requires about as much competent power as the fight against climate change.
 1. In this scenario, all the standard forces which make other types of technological development illegal have pushed towards making autonomous AGI illegal too.
3. Launching a good-faith joint US-China AGI project requires about as much competent power as launching Project Apollo.
 1. According to [this article](#), Kennedy (and later Johnson) made several offers (some of which were public) of a joint US-USSR Moon mission, which Khrushchev reportedly came close to accepting. Of course this is a long way from actually doing a joint project (and it's not clear how reliable the source is), but it still surprised me a lot, given that I viewed the "space race" as basically a zero-sum prestige project. If your model predicted this, I'd be interested to hear why.

[Yudkowsky][12:07] (Sep. 25 comment)

The US and China preventing any other country from becoming a leader in AI requires about as much competent power as banning chemical/biological weapons.

I believe this is wholly false. On my model it requires closer to WW1 levels of effort. I don't think you're going to get it without credible threats of military action leveled at previously allied countries.

AI is easier and more profitable to build than chemical / biological weapons, and correspondingly harder to ban. Existing GPU factories need to be shut down and existing GPU clusters need to be banned and no duplicate of them can be allowed to arise, across many profiting countries that were previously military allies of the United States, which - barring some vast shift in world popular *and* elite opinion against AI, which is also not going to happen - those countries would be extremely disinclined to sign, especially if the treaty terms permitted the USA and China to forge ahead.

The reason why chem weapons bans were much easier was that people did not like chem weapons. They were awful. There was a perceived common public interest in nobody having chem weapons. It was understood popularly and by elites to be a Prisoner's Dilemma situation requiring enforcement to get to the Pareto optimum. Nobody was profiting tons off the infrastructure that private parties could use to make chem weapons.

An AI ban is about as easy as banning advanced metal-forging techniques in current use so nobody can get ahead of the USA and China in making airplanes. That would be HARD and likewise require credible threats of military action against former allies.

"AI ban is as easy as a chem weapons ban" seems to me like politically crazy talk. I'd expect a more politically habited person to confirm this.

[Shulman][14:32] (Sep. 25 comment)

AI ban much, much harder than chemical weapons ban. Indeed chemical weapons were low military utility, that was central to the deal, and they have still been used subsequently.

An AI ban is about as easy as banning advanced metal-forging techniques in current use so nobody can get ahead of the USA and China in making airplanes. That would be HARD and likewise require credible threats of military action against former allies.

If large amounts of compute relative to today are needed (and presumably Eliezer rejects this), the fact that there is only a single global leading node chip supply chain makes it vastly easier than metal forging, which exists throughout the world and is vastly cheaper.

Sharing with allies (and at least embedding allies to monitor US compliance) also reduces the conflict side.

OTOH, if compute requirements were super low then it gets a lot worse.

And the biological weapons ban failed completely: the Soviets built an enormous bioweapons program, the largest ever, after agreeing to the ban, and the US couldn't even tell for sure they were doing so.

[Yudkowsky][18:15] (Oct. 4 comment)

I've updated somewhat off of Carl Shulman's argument that there's only one chip supply chain which goes through eg a single manufacturer of lithography machines (ASML), which could maybe make a lock on AI chips possible with only WW1 levels of cooperation instead of WW2.

That said, I worry that, barring WW2 levels, this might not last very long if other countries started duplicating the supply chain, even if they had to go back one or two process nodes on the chips? There's a difference between the proposition "ASML has a lock on the lithography market right now" and "if aliens landed and seized ASML, Earth would forever after be unable to build another lithography plant". I mean, maybe that's just true because we lost technology and can't rebuild old bridges either, but it's at least less obvious.

Launching Tomahawk cruise missiles at any attempt anywhere to build a new ASML, is getting back into "military threats against former military allies" territory and hence what I termed WW2 levels of cooperation.

[Shulman][18:30] (Oct. 4 comment)

China has been trying for some time to build its own and has failed with tens of billions of dollars (but has captured some lagging node share), but would be substantially more likely to succeed with a trillion dollar investment. That said, it is hard to throw money at these things and the tons of tacit knowledge/culture/supply chain networks are tough to replicate. Also many ripoffs of the semiconductor subsidies have occurred. Getting more NASA/Boeing and less SpaceX is a plausible outcome even with huge investment.

They are trying to hire people away from the existing supply chain to take its expertise and building domestic skills with the lagging nodes.

[Yudkowsky][19:14] (Oct. 4 comment)

Does that same theory predict that if aliens land and grab some but not all of the current ASML personnel, Earth is thereby successfully taken hostage for years, because Earth has trouble rebuilding ASML, which had the irreproducible lineage of masters and apprentices dating back to the era of Lost Civilization? Or would Earth be much better at this than China, on your model?

[Shulman][19:31] (Oct. 4 comment)

I'll read that as including the many suppliers of ASML (one EUV machine has over 100,000 parts, many incredibly fancy or unique). It's just a matter of how many years it takes. I think Earth fails to rebuild that capacity in 2 years but succeeds in 10.

"A study this spring by Boston Consulting Group and the Semiconductor Industry Association estimated that creating a self-sufficient chip supply chain would take at least \$1 trillion and sharply increase prices for chips and products made with them...The situation underscores the crucial role played by ASML, a once obscure company whose market value now exceeds \$285 billion. It is "the most important company you never heard of," said C.J. Muse, an analyst at Evercore ISI."

<https://www.nytimes.com/2021/07/04/technology/tech-cold-war-chips.html>

[Yudkowsky][19:59] (Oct. 4 comment)

No in 2 years, yes in 10 years sounds reasonable to me for this hypothetical scenario, as far as I know in my limited knowledge.

[Yudkowsky][12:10] (Sep. 25 comment)

Launching a good-faith joint US-China AGI project requires about as much competent power as launching Project Apollo.

It's really weird, relative to my own model, that you put the item that the US and China can bilaterally decide to do all by themselves, without threats of military action against their former allies, as more difficult than the items that require conditions imposed on other developed countries that don't want them.

Political coordination is hard. No, seriously, it's hard. It comes with a difficulty penalty that scales with the number of countries, how complete the buy-in has to be, and how much their elites and population don't want to do what you want them to do relative to how much elites and population agree that it needs doing (where this very rapidly goes to "impossible" or "WW1/WW2" as they don't particularly want to do your thing).

[Ngo] (Sep. 25 Google Doc)

So far I haven't talked about how much competent power I actually expect people to apply to AI governance. I don't think it's useful for Eliezer and me to debate this directly, since it's largely downstream from most of the other disagreements we've had. In particular, I model him as believing that there'll be very little competent power applied to prevent AI risk from governments and wider society, partly because he expects a faster takeoff than I do, and partly because he has a lower opinion of governmental competence than I do. But for the record, it seems likely to me that there'll be as much competent effort put into reducing AI risk by governments and wider society as there has been into fighting COVID; and plausibly (but not likely) as much as fighting climate change.

One key factor is my expectation that arguments about the importance of alignment will become much stronger as we discover more compelling examples of misalignment. I don't currently have strong opinions about how compelling the worst examples of misalignment before catastrophe are likely to be; but identifying and publicising them seems like a particularly effective form of advocacy, and one which we should prepare for in advance.

The predictable accumulation of easily-accessible evidence that AI risk is important is one example of a more general principle: that it's much easier to understand, publicise, and solve problems as those problems get closer and more concrete. This seems like a strong effect to me, and a key reason why so many predictions of doom throughout history have failed to come true, even when they seemed compelling at the time they were made.

Upon reflection, however, I think that even taking this effect into account, the levels of competent power required for the interventions mentioned above are too high to justify the level of optimism about AI governance that I started our debate with. On the other hand, I found Eliezer's arguments about consequentialism less convincing than I expected. Overall I've updated that AI risk is higher than I previously believed; though I expect my views to be quite unsettled while I think more, and talk to more people, about specific governance interventions and scenarios.

Ngo and Yudkowsky on scientific reasoning and pivotal acts

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is a transcript of a conversation between Richard Ngo and Eliezer Yudkowsky, facilitated by Nate Soares (and with some comments from Carl Shulman). This transcript continues the [Late 2021 MIRI Conversations](#) sequence, following [Ngo's view on alignment difficulty](#).

Color key:

Chat by Richard and Eliezer Other chat

14. October 4 conversation

14.1. Predictable updates, threshold functions, and the human cognitive range

[Ngo][15:05]

Two questions which I'd like to ask Eliezer:

1. How strongly does he think that the "shallow pattern-memorisation" abilities of GPT-3 are evidence for Paul's view over his view (if at all)
 2. How does he suggest we proceed, given that he thinks directly explaining his model of the chimp-human difference would be the wrong move?
-

[Yudkowsky][15:07]

1 - I'd say that it's some evidence for the Dario viewpoint which seems close to the Paul viewpoint. I think it's some evidence for the Dario viewpoint because Dario seems to be the person who made something like an advance prediction about it. It's not enough to make me believe that you can straightforwardly extend the GPT architecture to 3e14 parameters and train it on 1e13 samples and get human-equivalent performance.

[Ngo][15:09]

Did you make any advance predictions, around the 2008-2015 period, of what capabilities we'd have before AGI?

[Yudkowsky][15:10]

not especially that come to mind? on my model of the future this is not particularly something I am supposed to know unless there is a rare flash of predictability.

[Ngo][15:11]

1 - I'd say that it's some evidence for the Dario viewpoint which seems close to the Paul viewpoint. say it's some evidence for the Dario viewpoint because Dario seems to be the person who made something like an advance prediction about it. It's not enough to make me believe that you can straightforwardly extend the GPT architecture to 3e14 parameters and train it on 1e13 samples and get human-equivalent performance.

For the record I remember Paul being optimistic about language when I visited OpenAI in summer 20 But I don't know how advanced internal work on GPT-2 was by then.

[Yudkowsky][15:13]

2 - in lots of cases where I learned more specifics about X, and updated about Y, I had the experience looking back and realizing that knowing *anything* specific about X would have predictably produced a directional update about Y. like, knowing anything in particular about how the first AGI eats computation, would cause you to update far away from thinking that biological analogies to the computation consumed by humans were a good way to estimate how many computations an AGI needs to eat. you know lots of details about how humans consume watts of energy, and you know lots of details about how modern AI consumes watts, so it's very visible that these quantities are so incredibly different and go through so many different steps that they're basically unanchored from each other.

I have specific ideas about how you get AGI that isn't just scaling up Stack More Layers, which lead me to think that the way to estimate the computational cost of it is not "3e14 parameters trained at 1e1 ops per step for 1e13 steps, because that much computation and parameters seems analogous to human biology and 1e13 steps is given by past scaling laws", a la recent OpenPhil publication. But it seems to me that it should be possible to have the abstract insight that knowing more about general intelligence in AGIs or in humans would make the biological analogy look less plausible, because you wouldn't be matching up an unknown key to an unknown lock.

Unfortunately I worry that this depends on some life experience with actual discoveries to get something this abstract-sounding on a gut level, because people basically never seem to make abstract updates of this kind when I try to point to them as predictable directional updates?

But, in principle, I'd hope there would be aspects of this where I could figure out how to show that *any* knowledge of specifics would probably update you in a predictable direction, even if it doesn't seem best for Earth for me to win that argument by giving specifics conditional on those specifics actually being correct, and it doesn't seem especially sound to win that argument by giving specifics that are wrong.

[Ngo][15:17]

I'm confused by this argument. Before I thought much about the specifics of the chimpanzee-human transition, I found the argument "humans foomed (by biological standards) so AIs will too" fairly compelling. But after thinking more about the specifics, it seems to me that the human foom was in part caused by a factor (sharp cultural shift) that won't be present when we train AIs.

[Yudkowsky][15:17]

sure, and other factors will be present in AIs but not in humans

[Ngo][15:17]

This seems like a case where more specific knowledge updated me away from your position, contrary to what you're claiming.

[Yudkowsky][15:18]

eg, human brains don't scale and mesh, while it's far more plausible that with AI you could just run more of it

that's a huge factor leading one to expect AI to scale faster than human brains did

it's like communication between humans, but squared!

this is admittedly a specific argument and I'm not sure how it would abstract out to any specific argument

[Ngo][15:20]

Again, this is an argument that I believed less after looking into the details, because right now it's pre difficult to throw more compute at neural networks at runtime.

Which is not to say that it's a bad argument, the differences in compute-scalability between humans AIs are clearly important. But I'm confused about the structure of your argument that knowing more details will predictably update me in a certain direction.

[Yudkowsky][15:21]

I suppose the genericized version of my actual response to that would be, "architectures that have a harder time eating more compute are architectures which, for this very reason, are liable to need better versions invented of them, and this in particular seems like something that plausibly happens before scaling to general intelligence is practically possible"

[Soares][15:23]

(Eliezer, I see Richard as requesting that you either back down from, or clarify, your claim that any specific observations about how much compute AI systems require will update him in a predictable direction.)

[Ngo: 👍]

[Yudkowsky][15:24]

I'm not saying I know how to make that abstracted argument for exactly what Richard cares about, part because I don't understand Richard's exact model, just that it's one way to proceed past the point where the obvious dilemma crops up of, "If a theory about AGI capabilities is true, it is a disservice to Earth to speak it, and if a theory about AGI capabilities is false, an argument based on it is not sound"

[Ngo][15:25]

Ah, I see.

[Yudkowsky][15:26]

possible viewpoint to try: that systems in general often have threshold functions as well as smooth functions inside them.

only in ignorance, then, do we imagine that the whole thing is one smooth function.

the history of humanity has a threshold function of, like, communication or culture or whatever.

the correct response to this is not, "ah, so this was the unique, never-to-be-seen-again sort of fact which cropped up in the weirdly complicated story of humanity in particular, which will not appear in the much simpler story of AI"

this only sounds plausible because you don't know the story of AI so you think it will be a simple story

the correct generalization is "guess some weird thresholds will also pop up in whatever complicated story of AI will appear in the history books"

[Ngo][15:28]

Here's a quite general argument about why we shouldn't expect too many threshold functions in the impact of AI: because at any point, humans will be filling in the gaps of whatever AIs can't do. (The law of this type of smoothing is, I claim, why culture was a sharp threshold for humans - if there had been another intelligent species we could have learned culture from, then we would have developed more gradually.)

[Yudkowsky][15:30]

something like this indeed appears in my model of why I expect not much impact on GDP before AGI powerful enough to bypass human economies entirely

during the runup phase, pre-AGI won't be powerful to do "whole new things" that depend on doing lots of widely different things that humans can't do

just marginally new things that depend on doing one thing humans can't do, or can do but a bunch worse

[Ngo][15:31]

Okay, that's good to know.

Would this also be true in [a civilisation of village idiots?](#)

[Yudkowsky][15:32]

there will be sufficient economic reward for building out industries that are mostly human plus one that pre-AGI does, and people will pocket those economic rewards, go home, and not be more ambitious than that. I have trouble empathically grasping *why* almost all the CEOs are like this in our current Earth, because I am very much not like that myself, but observationally, the current Earth sure does seem to behave like rich people would almost uniformly rather not rock the boat too much.

I did not understand the whole thing about village idiots actually

do you want to copy and paste the document, or try rephrasing the argument?

[Ngo][15:35]

Rephrasing:

Claim 1: AIs will be better at doing scientific research (and other similar tasks) than village idiots, before we reach AGI.

Claim 2: Village idiots still have the core of general intelligence (which you claim chimpanzees don't have).

Claim 3: It would be surprising if narrow AI's research capabilities fell specifically into the narrow gap between village idiots and Einsteins, given that they're both general intelligences and are very similar in terms of architecture, algorithms, etc.

(If you deny claim 2, then we can substitute, say, someone at the 10th percentile of human intelligence - I don't know what specific connotations "village idiot" has to you.)

[Yudkowsky][15:37]

My models do not have an easy time of visualizing "as generally intelligent as a chimp, but specialized to science research, gives you superhuman scientific capability and the ability to make progress in most areas of science".

(this is a reference back to the pre-rephrase in the document)

it seems like, I dunno, "gradient descent can make you generically good at anything without that taking too much general intelligence" must be a core hypothesis there?

[Ngo][15:39]

I mean, we both agree that gradient descent can produce *some* capabilities without also producing much general intelligence. But claim 1 plus your earlier claims that narrow AIs won't surpass humans in scientific research, lead to the implication that the limitations of gradient-descent-without-much-general-intelligence fall in a weirdly narrow range.

[Yudkowsky][15:42]

I do credit the Village Idiot to Einstein Interval with being a little broader as a target than I used to think since the Alpha series of Go-players took a couple of years to go from pro to world-beating even once they had a scalable algorithm. Still seems to me that, over time, the wall clock time to traverse those ranges has been getting shorter, like Go taking less time than Chess. My intuitions still say that it'd be quite weird to end up hanging out for a long time with AGIs that conduct humanlike conversations and are ambitious enough to run their own corporations while those AGIs are still not much good at science.

But on my present model, I suspect the limitations of "gradient-descent-without-much-general-intelligence" to fall underneath the village idiot side?

[Ngo][15:43]

Oh, interesting.

That seems like a strong prediction

[Yudkowsky][15:43]

Your model, as I understand it, is saying, "But surely, GD-without-GI must suffice to produce better scientists than village idiots, by specializing chimps on science" and my current reply, though it's not a particular question I've thought a lot about before, is, "That... does not quite seem to me like a thing that should happen along the mainline?"

though, as always, in the limit of superintelligences doing things, or our having the Textbook From The Future, we could build almost any kind of mind on purpose if we knew how, etc.

[Ngo][15:44]

For example, I expect that if I prompt GPT-3 in the right way, it'll say some interesting and not-totally-nonsensical claims about advanced science.

Whereas it would be very hard to prompt a village idiot to do the same.

[Yudkowsky][15:44]

e.g., a superintelligence could load up chimps with lots of domain-specific knowledge they were not generally intelligent enough to learn themselves.

ehhhhhh, it is *not* clear to me that GPT-3 is better than a village idiot at advanced science, even in the narrow sense, especially if the village idiot is allowed some training

[Ngo][15:46]

It's not clear to me either. But it does seem plausible, and then it seems even more plausible that this will be true of GPT-4

[Yudkowsky][15:46]

I wonder if we're visualizing different village idiots

my choice of "village idiot" originally was probably not the best target for visualization, because in a
of cases, a village idiot - especially the stereotype of a village idiot - is, like, a damaged general
intelligence with particular gears missing?

[Ngo][15:47]

I'd be happy with "10th percentile intelligence"

[Yudkowsky][15:47]

whereas it seems like what you want is something more like "Homo erectus but it has language"

oh, wow, 10th percentile intelligence?

that's super high

GPT-3 is far far out of its league

[Ngo][15:49]

I think GPT-3 is far below this person's league in a lot of ways (including most common-sense reasoning)
but I become much less confident when we're talking about abstract scientific reasoning.

[Yudkowsky][15:51]

I think that if scientific reasoning were as easy as you seem to be imagining(?), the publication factor
of the modern world would be *much* more productive of real progress.

[Ngo][15:51]

Well, a 10th percentile human is very unlikely to contribute to real scientific progress either way

[Yudkowsky][15:53]

Like, on my current model of how the world really works, China pours vast investments into universities
and sober-looking people with PhDs and classes and tests and postdocs and journals and papers; but
none of this is the real way of Science which is actually, secretly, unbeknownst to China, passed down
rare lineages and apprenticeships from real scientist mentor to real scientist student, and China does
have much in the way of lineages so the extra money they throw at stuff doesn't turn into real science

[Ngo][15:52]

Can you think of any clear-cut things that they could do and GPT-3 can't?

[Yudkowsky][15:53]

Like... make sense... at all? Invent a handaxe when nobody had ever seen a handaxe before?

[Ngo][15:54]

You're claiming that 10th percentile humans invent handaxes?

[Yudkowsky][15:55]

The activity of rearranging scientific sentences into new plausible-sounding paragraphs is well within reach of publication factories, in fact, they often use considerably more semantic sophistication than that, and yet, this does not cumulate into real scientific progress even in quite large amounts.

I think GPT-3 is basically just Not Science Yet to a much greater extent than even these empty publication factories.

If 10th percentile humans don't invent handaxes, GPT-3 sure as hell doesn't.

[Ngo][15:55]

I don't think we're disagreeing. Publication factories are staffed with people who do better academically than 90+% of all humans.

If 90th-percentile humans are very bad at science, then of course GPT-3 and 10th-percentile humans are very very bad at science. But it still seems instructive to compare them (e.g. on tasks like "talk cogently about a complex abstract topic")

[Yudkowsky][15:58]

I mean, while it is usually weird for something to be barely within a species's capabilities while being within those capabilities at all, such that only relatively smarter individual organisms can do it, in the case of something that a social species has only very recently started to do collectively, it's plausible that the thing appeared at the point where it was barely accessible to the smartest members. Eg, it wouldn't be surprising if it would have taken a long time or forever for humanity to invent science from scratch, if all the Francis Bacons and Newtons and even average-intelligence people were eliminated leaving only the bottom 10%. Because our species just started doing that, at the point where our species was barely able to start doing that, meaning, at the point where some rare smart people could spearhead it, historically speaking. It's not obvious whether or not less smart people can do it over a longer time.

I'm not sure we disagree much about the human part of this model.

My guess is that our disagreement is more about GPT-3.

"Talk 'cogently' about a complex abstract topic" doesn't seem like much of anything significant to me. GPT-3 is 'cogent'. It fails to pass the threshold for inventing science and, I expect, for most particular sciences.

[Ngo][16:00]

How much training do you think a 10th-percentile human would need in a given subject matter (say, economics) before they could answer questions as well as GPT-3 can?

(Right now I think GPT-3 does better by default because it at least recognises the terminology, where most humans don't at all.)

[Yudkowsky][16:01]

I also expect that if you offer a 10th-percentile human lots of money, they can learn to talk more cogently than GPT-3 about narrower science areas. GPT-3 is legitimately more well-read at its lower level of intelligence, but train the 10-percentiler in a narrow area and they will become able to write better nonsense about that narrow area.

[Ngo][16:01]

This sounds like an experiment we can actually run.

[Yudkowsky][16:02]

Like, what we've got going on here is a real *breadth* advantage that GPT-3 has in some areas, but the breadth doesn't add up because it lacks the depth of a 10%er.

[Ngo][16:02]

If we asked them to read a single introductory textbook and then quiz both them and GPT-3 about it covered in that textbook, do you expect that the human would come out ahead?

[Yudkowsky][16:02]

AI has figured out how to do a subhumanly shallow kind of thinking, and it *is* to be expected that when AI can do anything at all, it can soon do more of that thing than the whole human species could do.

No, that's nothing remotely like giving the human the brief training the human needs to catch up to GPT-3's longer training.

A 10%er does not learn in an instant - they learn faster than GPT-3, but not in an instant.

This is more like a scenario of paying somebody to, like, sit around for a year with an editor, learning how to mix-and-match economics sentences until they can learn to sound more like they're making an argument than GPT-3 does, despite still not understanding any economics.

A lot of the learning would just go into producing sensible-sounding nonsense at all, since lots of 10%ers have not been to college and have not learned how to regurgitate rearranged nonsense for college teachers.

[Ngo][16:05]

What percentage of humans do you think could learn to beat GPT-3's question-answering by reading a single textbook over, say, a period of a month?

[Yudkowsky][16:06]

~_(ツ)_/~

[Ngo][16:06]

More like 0.5 or 5 or 50?

[Yudkowsky][16:06]

Humans cannot in general pass the Turing Test for posing as AIs!

What percentage of humans can pass as a calculator by reading an arithmetic textbook?

Zero!

[Ngo][16:07]

I'm not asking them to mimic GPT-3, I'm asking them to produce better answers.

[Yudkowsky][16:07]

Then it depends on what kind of answers!

I think a lot of 10%ers could learn to do wedding-cake multiplication, if sufficiently well-paid as adults rather than being tortured in school, out to 6 digits, thus handily beating the current GPT-3 at 'multiplication'.

[Ngo][16:08]

For example: give them an economics textbook to study for a month, then ask them what inflation is whether it goes up or down if the government prints more money, whether the price of something increases or decreases when the supply increases.

[Yudkowsky][16:09]

GPT-3 did not learn to produce its responses by reading *textbooks*.

You're not matching the human's data to GPT-3's data.

[Ngo][16:10]

I know, this is just the closest I can get in an experiment that seems remotely plausible to actually ru

[Yudkowsky][16:10]

You would want to collect, like, 1,000 Reddit arguments about inflation, and have the human read them and have the human produce their own Reddit arguments, and have somebody tell them whether the sounded like real Reddit arguments or not.

The textbook is just not the same thing at all.

I'm not sure we're at the core of the argument, though.

To me it seems like GPT-3 is allowed to be superhuman at producing remixed and regurgitated sentences about economics, because this is about as relevant to Science talent as a calculator being able to do perfect arithmetic, only less so.

[Ngo][16:15]

Suppose that the remixed and regurgitated sentences slowly get more and more coherent, until GPT-can debate with a professor of economics and sustain a reasonable position.

[Yudkowsky][16:15]

Are these points that GPT-N read elsewhere on the Internet, or are they new good points that no professor of economics on Earth has ever made before?

[Ngo][16:15]

I guess you don't expect this to happen, but I'm trying to think about what experiments we could run get evidence for or against it.

The latter seems both very hard to verify, and also like a very high bar - I'm not sure if most professo of economics have generated new good arguments that no other professor has ever made before.

So I guess the former.

[Yudkowsky][16:18]

Then I think that you can do this without being able to do science. It's a lot like if somebody with a really good memory was lucky enough to have read that exact argument on the Internet yesterday, & to have a little talent for paraphrasing. Not by coincidence, having this ability gives you - on my model no ability to do science, invent science, be the first to build handaxes, or design nanotechnology.

I admit, this does reflect my personal model of how Science works, presumably not shared by many leading bureaucrats, where in fact the papers full of regurgitated scientific-sounding sentences are not accomplishing much.

[Ngo][16:20]

So it seems like your model doesn't rule out narrow AIs producing well-reviewed scientific papers, since you don't trust the review system very much.

[Yudkowsky][16:23]

I'm trying to remember whether or not I've heard of that happening, like, 10 years ago.

My vague recollection is that things in the Sokal Hoax genre where the submissions succeeded, used humans to hand-generate the nonsense rather than any submissions in the genre having been purely machine-generated.

[Ngo][16:24]

Which doesn't seem like an unreasonable position, but it does make it harder to produce tests that we have opposing predictions on.

[Yudkowsky][16:24]

Obviously, that doesn't mean it couldn't have been done 10 years ago, because 10 years ago it's plausibly a lot easier to hand-generate passing nonsense than to write an AI program that does it.

oh, wait, I'm wrong!

<https://news.mit.edu/2015/how-three-mit-students-fooled-scientific-journals-0414>

In April of 2005 the team's submission, "Rooter: A Methodology for the Typical Unification of Access Points and Redundancy," was accepted as a non-reviewed paper to the World Multiconference on Systemics, Cybernetics and Informatics (WMSCI), a conference that Krohn says is known for "being spammy and having loose standards."

in 2013 IEEE and Springer Publishing removed more than 120 papers from their sites after a French researcher's analysis determined that they were generated via SCIGen

[Ngo][16:26]

Oh, interesting

Meta note: I'm not sure where to take the direction of the conversation at this point. Shall we take a brief break?

[Yudkowsky][16:27]

The creators continue to get regular emails from computer science students proudly linking to papers they've snuck into conferences, as well as notes from researchers urging them to make versions for other disciplines.

Sure! Resume 5p?

[Ngo][16:27]

Yepp

14.2. Domain-specific heuristics and nanotechnology

[Soares][16:41]

A few takes:

1. It looks to me like there's some crux in "how useful will the 'shallow' stuff get before dangerous things happen". I would be unsurprised if this spiraled back into the gradualness debate. I'm excited about attempts to get specific and narrow disagreements in this domain (not necessarily bettable; I nominal distilling out specific disagreements before worrying about finding bettable ones).
 2. It seems plausible to me we should have some much more concrete discussion about possible ways things could go right, according to Richard. I'd be up for playin the role of beeping when things seem insufficiently concrete.
 3. It seems to me like Richard learned a couple things about Eliezer's model in that last bout of conversation. I'd be interested to see him try to paraphrase his current understanding of it, and to see Eliezer produce beeps where it seems particularly off.
-

[Yudkowsky][17:00]



[Ngo][17:02]

Hmm, I'm not sure that I learned too much about Eliezer's model in this last round.

[Soares][17:03]

(dang :-p)

[Ngo][17:03]

It seems like Eliezer thinks that the returns of scientific investigation are very heavy-tailed.

Which does seem pretty plausible to me.

But I'm not sure how useful this claim is for thinking about the development of AI that can do science

I attempted in my document to describe some interventions that would help things go right.

And the levels of difficulty involved.

[Yudkowsky][17:07]

(My model is something like: there are some very shallow steps involved in doing science, lots of medium steps, occasional very deep steps, assembling the whole thing into Science requires having

the lego blocks available. As soon as you look at anything with details, it ends up 'heavy-tailed' because it has multiple pieces and says how things don't work if all the pieces aren't there.)

[Ngo][17:08]

Eliezer, do you have an estimate of how much slower science would proceed if everyone's IQs were shifted down by, say, 30 points?

[Yudkowsky][17:10]

It's not obvious to me that science proceeds significantly past its present point. I would not have the right to be surprised if Reality told me the correct answer was that a civilization like that just doesn't reach AGI, ever.

[Ngo][17:12]

Doesn't your model take a fairly big hit from predicting that humans just happen to be within 30 IQ points of not being able to get any more science?

It seems like a surprising coincidence.

Or is this dependent on the idea that doing science is much harder now than it used to be?

And so if we'd been dumber, we might have gotten stuck before newtonian mechanics, or else before relativity?

[Yudkowsky][17:13]

No, humanity is exactly the species that finds it barely possible to do science.

[Ngo][17:14]

It seems to me like humanity is exactly the species that finds it barely possible to do *civilisation*.

[Yudkowsky][17:14]

If it were possible to do it with less intelligence, we'd be having this conversation over the Internet than we'd developed with less intelligence.

[Ngo][17:15]

And it seems like many of the key inventions that enabled civilisation weren't anywhere near as intelligence-bottlenecked as modern science.

[Yudkowsky][17:15]

Yes, it does seem that there's quite a narrow band between "barely smart enough to develop agriculture" and "barely smart enough to develop computers"! Though there were genuinely fewer people in the preagricultural world, with worse nutrition and no Ashkenazic Jews, and there's the whole question about to what degree the reproduction of the shopkeeper class over several centuries was important to the Industrial Revolution getting started.

[Ngo][17:15]

(e.g. you'd get better spears or better plows or whatever just by tinkering, whereas you'd never get relativity just by tinkering)

[Yudkowsky][17:17]

I model you as taking a lesson from this which is something like... you can train up a villager to be Jol von Neumann by spending some evolutionary money on giving them science-specific brain features, since John von Neumann couldn't have been much more deeply or generally intelligent, and you could spend even more money and make a chimp a better scientist than John von Neumann.

My model is more like, yup, the capabilities you need to invent aqueducts sure do generalize the crap out of things, though also at the upper end of cognition there are compounding returns which can bring John von Neumann into existence, and also also there's various papers suggesting that selection was happening really fast over the last few millennia and real shifts in cognition shouldn't be ruled out. (last part is an update to what I was thinking when I wrote [Intelligence Explosion Microeconomics](#), and from my own perspective a more gradualist line of thinking, because it means there's a wider actual target to traverse before you get to von Neumann.)

[Ngo][17:20]

It's not that "von Neumann isn't much more deeply generally intelligent", it's more like "domain-specific heuristics and instincts get you a long way". E.g. soccer is a domain where spending evolutionary money on specific features will very much help you beat von Neumann, and so is art, and so is music

[Yudkowsky][17:20]

My skepticism here is that there's a version of, like, "invent nanotechnology" which routes through just the shallow places, which humanity stumbles over before we stumble over deep AGI.

[Ngo][17:21]

Would you be comfortable publicly discussing the actual cognitive steps which you think would be necessary for inventing nanotechnology?

[Yudkowsky][17:23]

It should not be overlooked that there's a very valid sibling of the old complaint "Anything you can do can be done by AI", which is that "Things you can do with surprisingly-to-your-model shallow cognition are precisely the things that Reality surprises you by telling you that AI can do earlier than you expected". When we see GPT-3, we were getting some amount of real evidence about AI capabilities advancing faster than I expected, and some amount of evidence about GPT-3's task being performable using shallower cognition than expected.

Many people were particularly surprised by Go because they thought that Go was going to require deeper real thought than chess.

And I think AlphaGo probably was thinking in a legitimately deeper way than Deep Blue. Just not as much deeper as Douglas Hofstadter thought it would take.

Conversely, people thought a few years ago that driving cars really seemed to be the sort of thing that machine learning would be good at, and were unpleasantly surprised by how the last 0.1% of driving conditions were resistant to shallow techniques.

Despite the inevitable fact that some surprises of this kind now exist, and that more such surprises will exist in the future, it continues to seem to me that science-and-engineering on the level of "invent nanotech" still seems pretty unlikely to be easy to do with shallow thought, by means that humanity discovers before AGI tech manages to learn deep thought?

What actual cognitive steps? Outside-the-box thinking, throwing away generalizations that govern your previous answers and even your previous questions, inventing new ways to represent your questions, figuring out which questions you need to ask and developing plans to answer them; these are some answers that I hope will be sufficiently useless to AI developers that it is safe to give them, while still pointing in the direction of things that have an un-GPT-3-like quality of depth about them.

Doing this across unfamiliar domains that couldn't be directly trained in by gradient descent because they were too expensive to simulate a billion examples of

If you have something this powerful, why is it not also noticing that the world contains humans? Why it not noticing itself?

[Ngo][17:30]

If humans were to invent this type of nanotech, what do you expect the end intellectual result to be?

E.g. consider the human knowledge involved in building cars

There are thousands of individual parts, each of which does a specific thing

[Yudkowsky][17:30]

Uhhhh... is there a reason why "Eric Drexler's *Nanosystems* but, like, the real thing, modulo however much Drexler did not successfully Predict the Future about how to do that, which was probably a lot" not the obvious answer here?

[Ngo][17:31]

And some deep principles governing engines, but not really very crucial ones to actually building (ea versions of) those engines

[Yudkowsky][17:31]

that's... not historically true at all?

getting a grip on quantities of heat and their flow was *critical* to getting steam engines to work

it didn't happen until the math was there

[Ngo][17:32]

Ah, interesting

[Yudkowsky][17:32]

maybe you can be a mechanic banging on an engine that somebody else designed, around principles that somebody even earlier invented, without a physics degree

but, like, engineers have actually needed math since, like, that's been a thing, it wasn't just a prestige trick

[Ngo][17:34]

Okay, so you expect there to be a bunch of conceptual work in finding equations which govern nanosystems.

Uhhhh... is there a reason why "Eric Drexler's *Nanosystems* but, like, the real thing, modulo however much Drexler did not successfully Predict the Future about how to do that, which was probably a lot" is not the obvious answer here?

This may in fact be the answer; I haven't read it though.

[Yudkowsky][17:34]

or other abstract concepts than equations, which have never existed before like, maybe not with a type signature unknown to humanity, but with specific instances unknown to present humanity
that's what I'd expect to see from humanly designed nanosystems

[Ngo][17:35]

So something like AlphaFold is only doing a very small proportion of the work here, since it's not able generate new abstract concepts (of the necessary level of power)

[Yudkowsky][17:35]

yeeeessss, that is why DeepMind did not take over the world last year
it's not just that AlphaFold lacks the concepts but that it lacks the machinery to invent those concept and the machinery to do anything with such concepts

[Ngo][17:38]

I think I find this fairly persuasive, but I also expect that people will come up with increasingly clever ways to leverage narrow systems so that they can do more and more work.
(including things like: if you don't have enough simulations, then train another narrow system to help that, etc)

[Yudkowsky][17:39]

(and they will accept their trivial billion-dollar-payouts and World GDP will continue largely undisturbed on my mainline model, because it will be easiest to find ways to make money by leveraging narrow systems on the less regulated, less real parts of the economy, instead of trying to build houses or do medicine, etc.)

real tests being expensive, simulation being impossibly expensive, and not having enough samples to train your civilization's current level of AI technology, is not a problem you can solve by training a new AI to generate samples, because you do not have enough samples to train your civilization's current level of AI technology to generate more samples

[Ngo][17:41]

Thinking about nanotech makes me more sympathetic to the argument that developing general intelligence will bring a sharp discontinuity. But it also makes me expect longer timelines to AGI, during which there's more time to do interesting things with narrow AI. So I guess it weighs more against Dario's view, less against Paul's view.

[Yudkowsky][17:41]

well, I've been debating Paul about that separately in the timelines channel, not sure about recapitulating it here

but in broad summary, since I expect the future to look like it was drawn from the "history book" barrel and not the "futurism" barrel, I expect huge barriers to doing *huge* things with narrow AI in small amounts of time; you can sell waifutech because it's unregulated and hard to regulate, but that does feed into core mining and steel production.

we could already have double the GDP if it was legal to build houses and hire people, etc., and the change brought by pre-AGI will perhaps be that our GDP could *quadruple* instead of just *double* if it were legal to do things, but that will not make it legal to do things, and why would anybody try to do things and probably fail when there are easier \$36 billion profits to be made in waifutech.

14.3. Relatively shallow cognition, Go, and math

[Ngo][17:45]

I'd be interested to see Paul's description of how we would train AIs to solve hard scientific problems. think there's some prediction that's like "we train it on arxiv and fine-tune it until it starts to output credible hypotheses about nanotech". And this seems like it has a step that's quite magical to me, b~~u~~ perhaps that'll be true of any prediction that I make before fully understanding how intelligence work

[Yudkowsky][17:46]

my belief is not so much that this training can never happen, but that this probably means the system was trained *beyond the point of safe shallowness*

not in principle over all possible systems a superintelligence could build, but in practice when it happens on Earth

my only qualm about this is that current techniques make it possible to buy shallowness in larger quantities than this Earth has ever seen before, and people are looking for surprising ways to make use of that

so I weigh in my mind the thought of Reality saying Gotcha! by handing me a headline I read tomorrow about how GPT-4 has started producing totally reasonable science papers that are actually correct

and I am pretty sure that exact thing doesn't happen

and I ask myself about GPT-5 in a few more years, which had the same architecture as GPT-3 but more layers and more training, doing the same thing

and it's still largely "nope"

then I ask myself about people in 5 years being able to use the shallow stuff *in any way whatsoever* to produce the science papers

and of course the answer there is, "okay, but is it doing that without having shallowly learned stuff *that adds up to deep stuff* which is *why it can now do science*"

and I try saying back "no, it was born of shallowness and it remains shallow and it's just doing science because it turns out that there is totally a way to be an incredibly mentally shallow skillful scientist if you think 10,000 shallow thoughts per minute instead of 1 deep thought per hour"

and my brain is like, "I cannot absolutely rule it out but it really seems like trying to call the next big surprise in 2014 and you guess self-driving cars instead of Go because how the heck would you guess that Go was shallower than self-driving cars"

like, that is an *imaginable* surprise

[Ngo][17:52]

On that *particular* point it seems like the very reasonable heuristic of "pick the most similar task" would say that go is like chess and therefore you can do it shallowly.

[Yudkowsky][17:52]

but there's a world of difference between saying that a surprise is imaginable, and that it wouldn't surprise you

[Ngo][17:52]

I wasn't thinking that much about AI at that point, so you're free to call that post-hoc.

[Yudkowsky][17:52]

the Chess techniques had already failed at Go

actual new techniques were required

the people around at the time had witnessed sudden progress on self-driving cars a few years earlier

[Ngo][17:53]

My advance prediction here is that "math is like go and therefore can be done shallowly".

[Yudkowsky][17:53]

self-driving cars were of obviously greater economic interest as well

my recollection is that talk of the time was about self-driving

heh! I have the same sense.

that is, math being shallower than science.

though perhaps not as shallow as Go, and you will note that Go has fallen and Math has not

[Ngo][17:54]

right

I also expect that we'll need new techniques for math (although not as different from the go techniques as the go techniques were from chess techniques)

But I guess we're not finding strong disagreements here either.

[Yudkowsky][17:57]

if Reality came back and was like "Wrong! Keeping up with the far reaches of human mathematics is harder than being able to develop your own nanotech," I would be like "What?" to about the same degree as being "What?" on "You can build nanotech just by thinking trillions of thoughts that are too shallow to notice humans!"

[Ngo][17:58]

Perhaps let's table this topic and move on to one of the others Nate suggested? I'll note that walking through the steps required to invent a science of nanotechnology does make your position feel more compelling, but I'm not sure how much of that is the general "intelligence is magic" intuition I mentioned before.

[Yudkowsky][17:59]

How do you suspect your beliefs would shift if you had any detailed model of intelligence?

Consider trying to imagine a particular wrong model of intelligence and seeing what it would say differently?

(not sure this is a useful exercise and we could indeed try to move on)

[Ngo][18:01]

I think there's one model of intelligence where scientific discovery is more actively effortful - as in, you need to be very goal-directed in determining hypotheses, testing hypotheses, and so on.

And there's another in which scientific discovery is more constrained by flashes of insight, and the systems which are producing those flashes of insight are doing pattern-matching in a way that's fairly disconnected from the real-world consequences of those insights.

[Yudkowsky][18:05]

The first model is true and the second one is false, if that helps. You can tell this by contemplating where you would update if you learned any model, by considering that things look more disconnected when you can't see the machinery behind them. If you don't know what moves the second hand on a watch and the minute hand on a watch, they could just be two things that move at different rates for completely unconnected reasons; if you can see inside the watch, you'll see that the battery is shared and the central timing mechanism is shared and then there's a few gears to make the hands move at different rates.

Like, in my ontology, the notion of "effortful" doesn't particularly parse as anything basic, because it doesn't translate over into paperclip maximizers, which are neither effortful nor effortless.

But in a human scientist you've got thoughts being shoved around by all sorts of processes behind the curtains, created by natural selection, some of them reflecting shards of Consequentialism / shadowy paths through time

The flashes of insight come to people who were looking in nonrandom places

If they didn't plan deliberately and looked on pure intuition, they looked with an intuition trained by past success and failure

Somebody walking doesn't plan to walk, but long ago as a baby they learned from falling over, and the ancestors who fell over more didn't reproduce

[Ngo][18:09]

I think the first model is probably more true for humans in the domain of science. But I'm uncertain about the extent to which this is true because humans have not been optimised very much for doing science - we consider the second model in a domain that humans have actually been optimised very hard for (say, physical activity) - then maybe we can use the analogy of a coach and a player. The coach can tell the player what to practice, but almost all the work is done by the player practicing in a way which updates their intuitions.

This has become very abstract, though.

14.4. Pivotal acts and historical precedents

[Ngo][18:11]

A few takes:

1. It looks to me like there's some crux in "how useful will the 'shallow' stuff get before dangerous things happen". I would be unsurprised if this spiraled back into the gradualness debate. I'm excited about attempts to get specific and narrow disagreements in this domain (not necessarily bettable; nominate distilling out specific disagreements before worrying about finding bettable ones).
2. It seems plausible to me we should have some much more concrete discussion about possible ways things could go right, according to Richard. I'd be up for playing the role of beeping when things seem insufficiently concrete.
3. It seems to me like Richard learned a couple things about Eliezer's model in that last bout of conversation. I'd be interested to see him try to paraphrase his current understanding of it, and to see Eliezer produce beeps where it seems particularly off.

Here's Nate's comment.

We could try his #2 suggestion: concrete ways that things could go right.

[Soares][18:12]

(I am present and am happy to wield the concreteness-hammer)

[Ngo][18:13]

I think I'm a little cautious about this line of discussion, because my model doesn't strongly constrain the ways that different groups respond to increasing developments in AI. The main thing I'm confident about is that there will be much clearer responses available to us once we have a better picture of AI development.

E.g. before modern ML, the option of international constraints on compute seemed much less salient, because algorithmic developments seemed much more important.

Whereas now, tracking/constraining compute use seems like one promising avenue for influencing AI development.

Or in the case of nukes, before knowing the specific details about how they were constructed, it would be hard to give a picture of how arms control goes well. But once you know more details about the process of uranium enrichment, you can construct much more efficacious plans.

[Yudkowsky][18:19]

Once we knew specific things about bioweapons, countries developed specific treaties for controlling them, which failed (according to @CarlShulman)

[Ngo][18:19, moved two down in log]

(As a side note, I think that if Eliezer had been around in the 1930s, and you described to him what actually happened with nukes over the next 80 years, he would have called that "insanely optimistic")

[Yudkowsky][18:21]

Mmmmmmaybe. Do note that I tend to be more optimistic than the average human about, say, global warming, or everything in transhumanism outside of AGI.

Nukes have going for them that, in fact, nobody has an incentive to start a global thermonuclear war. Eliezer is not in fact pessimistic about everything and views his AGI pessimism as generalizing to very few other things, which are not, in fact, as bad as AGI.

[Ngo][18:21]

I think I put this as the lowest application of competent power out of the things listed in my doc; I'd need to look at the historical details to know if important decision-makers actually cared about it, or were just doing it for PR reasons.

[Shulman][18:22]

Once we knew specific things about bioweapons, countries developed specific treaties for controlling them, which failed (according to @CarlShulman)

The treaties were pro forma without verification provisions because the powers didn't care much about bioweapons. They did have verification for nuclear and chemical weapons which did work.

[Yudkowsky][18:22]

But yeah, compared to pre-1946 history, nukes actually kind of did go *really surprisingly well!*

Like, this planet used to be a huge warring snakepit of Great Powers and Little Powers and then nuke came along and people actually got serious and decided to stop having the largest wars they could fi

[Shulman][18:22][18:23]

The analog would be an international agreement to sign a nice unenforced statement of AI safety principles and then all just building AGI in doomy ways without explicitly saying they're doing it..

The BWC also allowed 'defensive' research that is basically as bad as the offensive kind.

[Yudkowsky][18:23]

The analog would be an international agreement to sign a nice unenforced statement of AI safety principles and then all just building AGI in doomy ways without explicitly saying they're doing it..

This scenario sure sounds INCREDIBLY PLAUSIBLE, yes

[Ngo][18:22]

On that point: do either of you have strong opinions about the anthropic shadow argument about nukes? That seems like one reason why the straw 1930s-Eliezer I just cited would have been justified.

[Yudkowsky][18:23]

I mostly don't consider the anthropic shadow stuff

[Shulman][18:24]

In the late Cold War Gorbachev and Reagan might have done the BWC treaty+verifiable dismantling, but they were in a rush on other issues like nukes and collapse of the USSR.

Putin just wants to keep his bioweapons program, it looks like. Even denying the existence of the exposed USSR BW program.

[Yudkowsky][18:25]

I'm happy making no appeal to anthropics here.

[Shulman][18:25]

Boo anthropic shadow claims. Always dumb.

(Sorry I was only invoked for BW, holding my tongue now.)

[Yudkowsky: ♥] [Soares: ♥]

[Yudkowsky][18:26]

There may come a day when the strength of nonanthropic reasoning fails... but that is not this day!

[Ngo][18:27]

Okay, happy to rule that out for now too. So yeah, I picture 1930s-Eliezer pointing to technological trends and being like "by default, 30 years after the first nukes are built, you'll be able to build one in your back yard. And governments aren't competent enough to stop that happening."

And I don't think I could have come up with a compelling counterargument back then.

[Soares][18:27]

[Sorry I was only invoked for BW, holding my tongue now.]

(fwiw, I thought that when Richard asked "you two" re: anthropic shadow, he meant you also. But I appreciate the caution. And in case Richard meant me, I will note that I agree w/ Carl and Eliezer on this count.)

[Ngo][18:28]

(fwiw, I thought that when Richard asked "you two" re: anthropic shadow, he meant you also. But I appreciate the caution. And in case Richard meant me, I will note that I agree w/ Carl and Eliezer on this count.)

Oh yeah, sorry for the ambiguity, I meant Carl.

I do believe that AI control will be more difficult than nuclear control, because AI is so much more useful. But I also expect that there will be many more details about AI development that we don't currently understand, that will allow us to influence it (because AGI is a much more complicated concept than "really really big bomb").

[Yudkowsky][18:29]

[So yeah, I picture 1930s-Eliezer pointing to technological trends and being like "by default, 30 years after the first nukes are built, you'll be able to build one in your back yard. And governments aren't competent enough to stop that happening."]

And I don't think I could have come up with a compelling counterargument back then.]

So, I mean, in fact, I don't prophesize doom from very many trends at all! It's literally just AGI that is anywhere near that unmanageable! Many people in EA are more worried about biotech than I am, for example.

[Ngo][18:31]

I appreciate that my response is probably not very satisfactory to you here, so let me try to think about more concrete things we can disagree about.

[Yudkowsky][18:31]

[I do believe that AI control will be more difficult than nuclear control, because AI is so much more useful. But I also expect that there will be many more details about AI development that we don't

currently understand, that will allow us to influence it (because AGI is a much more complicated concept than "really really big bomb").]

Er... I think this is not a correct use of the Way I was attempting to gesture at; things being more complicated when known than unknown, does not mean you have more handles to influence them because each complication has the potential to be a handle. It is not in general true that very complicated things are easier for humanity in general, and governments in particular, to control, because they have so many exposed handles.

I think there's a valid argument about it maybe being more possible to control the supply chain for AI training processors if the global chip supply chain is narrow (also per Carl).

[Ngo][18:34]

One thing that we seemed to disagree on, to a significant extent, is the difficulty of "US and China preventing any other country from becoming a leader in AI"

[Yudkowsky][18:35]

It is in fact a big deal about nuclear tech that uranium can't be mined in every country, as I understand it, and that centrifuges stayed at the frontier of technology and were harder to build outside the well-developed countries, and that the world ended up revolving around a few Great Powers that had no interest in nuclear tech proliferating any further.

[Ngo][18:35]

It seems to me that the US and/or China could apply a lot of pressure to many countries.

[Yudkowsky][18:35]

Unfortunately, before you let that encourage you too much, I would also note it was an important fact about nuclear bombs that they did not produce streams of gold and then ignite the atmosphere if you turned up the stream of gold too high with the actual thresholds involved being unpredictable.

[Ngo][18:35]

E.g. if the UK had actually seriously tried to block Google's acquisition of DeepMind, and the US had actually seriously tried to convince them not to do so, then I expect that the UK would have folded. (Although it's a weird hypothetical.)

Unfortunately, before you let that encourage you too much, I would also note it was an important fact about nuclear bombs that they did not produce streams of gold and then ignite the atmosphere if you turned up the stream of gold too high with the actual thresholds involved being unpredictable.

Not a critical point, but nuclear power does actually seem like a "stream of gold" in many ways.

(also, quick meta note: I need to leave in 10 mins)

[Yudkowsky][18:38]

I would be a lot more cheerful about a few Great Powers controlling AGI if AGI produced wealth, but more powerful AGI produced no more wealth; if AGI was made entirely out of hardware, with no software component that could be keep getting orders of magnitude more efficient using hardware-independent ideas; and if the button on AGIs that destroyed the world was clearly labeled.

That does take AGI to somewhere in the realm of nukes.

[Ngo][18:38]

How much improvement do you think can be eked out of existing amounts of hardware if people just to focus on algorithmic improvements?

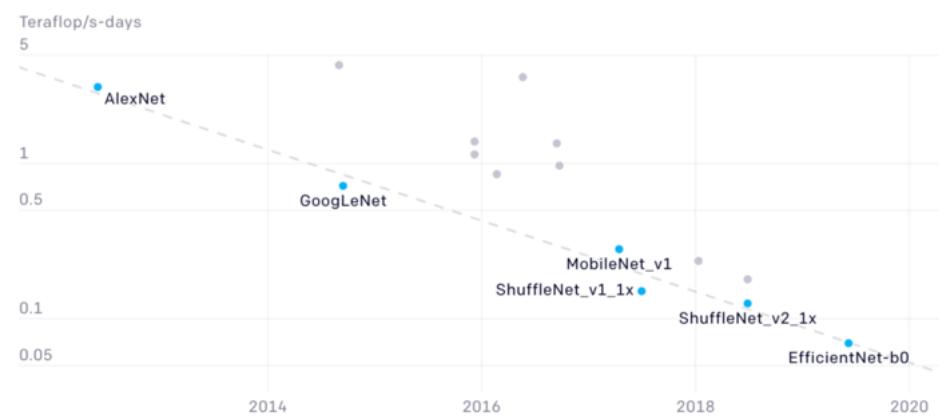
[Yudkowsky][18:38]

And Eliezer is capable of being less concerned about things when they are intrinsically less concerning which is why my history does not, unlike some others in this field, involve me running also being Terra Concerned about nuclear war, global warming, biotech, and killer drones.

[Ngo][18:39]

This says 44x improvements over 7 years: <https://openai.com/blog/ai-and-efficiency/>

44x less compute required to get to AlexNet performance 7 years later



[Yudkowsky][18:39]

Well, if you're a superintelligence, you can probably do human-equivalent human-speed general intelligence on a 286, though it might possibly have less fine motor control, or maybe not, I don't know

[Ngo][18:40]

(within reasonable amounts of human-researcher-time - say, a decade of holding hardware fixed)

[Yudkowsky][18:40]

I wouldn't be surprised if human ingenuity asymptoted out at AGI on a home computer from 1995. Don't know if it'd take more like a hundred years or a thousand years to get fairly close to that.

[Ngo][18:41]

Does this view cash out in a prediction about how the AI and Efficiency graph projects into the future

[Yudkowsky][18:42]

The question of how efficiently you can perform a fixed algorithm doing fixed things, often pales compared to the gains on switching to different algorithms doing different things.

Given government control of all the neural net training chips and no more public GPU farms, I buy that they could keep a nuke!AGI (one that wasn't tempting to crank up and had clearly labeled Doom-Causing Buttons whose thresholds were common knowledge) under lock of the Great Powers for 7 years during which software decreased hardware requirements by 44x. I am a bit worried about how long it takes before there's a proper paradigm shift on the level of deep learning getting started in 2006, after which the Great Powers need to lock down on individual GPUs.

[Ngo][18:46]

Hmm, okay.

14.5. Past ANN progress

[Ngo][18:46]

I don't expect another paradigm shift like that

(in part because I'm not sure the paradigm shift actually happened in the first place - it seems like neural networks were improving pretty continuously over many decades)

[Yudkowsky][18:47]

I've noticed that opinion around OpenPhil! It makes sense if you have short timelines and expect the world to end before there's another paradigm shift, but OpenPhil doesn't seem to expect that either.

Yeah, uh, there was kinda a paradigm shift in AI between say 2000 and now. There really, really was.

[Ngo][18:49]

What I mean is more like: it's not clear to me that an extrapolation of the trajectory of neural networks made much better by incorporating data about the other people who weren't using neural networks.

[Yudkowsky][18:49]

Would you believe that at one point Netflix ran a prize contest to produce better predictions of their users' movie ratings, with a \$1 million prize, and this was one of the largest prizes ever in AI and got tons of contemporary ML people interested, and neural nets were not prominent on the solutions list at all, because, back then, people occasionally solved AI problems *not using neural nets*?

I suppose that must seem like a fairy tale, as history always does, but I lived it!

[Ngo][18:50]

(I wasn't denying that neural networks were for a long time marginalised in AI)

I'd place much more credence on future revolutions occurring if neural networks had actually only been invented recently.

(I have to run in 2 minutes)

[Yudkowsky][18:51]

The world might otherwise end before the next paradigm shift, but if the world keeps on ticking for 1 years, 20 years, there will not always be the paradigm of training massive networks by even more massive amounts of gradient descent; I do not think that is actually the most efficient possible way to turn computation into intelligence.

Neural networks stayed stuck at only a few layers for a long time, because the gradients would explode or die out if you made the networks any deeper.

There was a critical moment in 2006(?) where Hinton and Salakhutdinov(?) proposed training Restricted Boltzmann machines unsupervised in layers, and then 'unrolling' the RBMs to initialize the weights in the network, and then you could do further gradient descent updates from there, because the activations and gradients wouldn't explode or die out given that initialization. That got people to, I dunno, 6 layers instead of 3 layers or something? But it focused attention on the problem of exploding gradients as the reason why deeply layered neural nets never worked, and that kicked off the entire modern field of deep learning, more or less.

[Ngo][18:56]

Okay, so are you claiming that that neural networks were mostly bottlenecked by algorithmic improvements, not compute availability, for a significant part of their history?

[Yudkowsky][18:56]

If anybody goes back and draws a graph claiming the whole thing was continuous if you measure the right metric, I am not really very impressed unless somebody at the time was using that particular graph and predicting anything like the right capabilities off of it.

[Ngo][18:56]

If so this seems like an interesting question to get someone with more knowledge of ML history than me to dig into; I might ask around.

[Yudkowsky][18:57]

[Okay, so are you claiming that that neural networks were mostly bottlenecked by algorithmic improvements, not compute availability, for a significant part of their history?]

Er... yeah? There was a long time when, even if you threw a big neural network at something, it just wouldn't work.

Good night, btw?

[Ngo][18:57]

Let's call it here; thanks for the discussion.

[Soares][18:57]

Thanks, both!

[Ngo][18:57]

I'll be interested to look into that claim, it doesn't fit with the impressions I have of earlier bottlenecks. I think the next important step is probably for me to come up with some concrete governance plans that I'm excited about.

I expect this to take quite a long time

[Soares][18:58]

We can coordinate around that later. Sorry for keeping you so late already, Richard.

[Ngo][18:59]

No worries

My proposal would be that we should start on whatever work is necessary to convert the debate into publicly accessible document now

In some sense coming up with concrete governance plans is my full-time job, but I feel like I'm still quite a way behind in my thinking on this, compared with people who have been thinking about governance specifically for longer

[Soares][19:01]

(@RobBensinger is already on it 😊)

[Bensinger:]

[Yudkowsky][19:03]

Nuclear plants might be like narrow AI in this analogy; some designs potentially contribute to proliferation, and you can get more economic wealth by building more of them, but they have no Unlabeled Doom Dial where you can get more and more wealth out of them by cranking them up until some unlabeled point the atmosphere ignites.

Also a thought: I don't think you just want somebody with more knowledge of AI history, I think you might need to ask an actual old fogey *who was there at the time*, and hasn't just learned an ordered history of just the parts of the past that are relevant to the historian's theory about how the present happened.

Two of them, independently, to see if the answers you get are reliable-as-in-statistical-reliability.

[Soares][19:19]

My own quick take, for the record, is that it looks to me like there are two big cruxes here.

One is about whether "deep generality" is a good concept, and in particular whether it pushes AI systems quickly from "nonscary" to "scary" and whether we should expect human-built AI systems to acquire it in practice (before the acute risk period is ended by systems that lack it). The other is about how easy it will be to end the acute risk period (eg by use of politics or nonscary AI systems alone).

I suspect the latter is the one that blocks on Richard thinking about governance strategies. I'd be interested in attempting further progress on the former point, though it's plausible to me that that should happen over in #timelines instead of here.

Christiano and Yudkowsky on AI predictions and human intelligence

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is a transcript of a conversation between Paul Christiano and Eliezer Yudkowsky, with comments by Rohin Shah, Beth Barnes, Richard Ngo, and Holden Karnofsky, continuing the [Late 2021 MIRI Conversations](#).

Color key:

Chat by Paul and Eliezer Other chat

15. October 19 comment

[Yudkowsky][11:01]

thing that struck me as an iota of evidence for Paul over Eliezer:
<https://twitter.com/tamaybes/status/1450514423823560706?s=20>

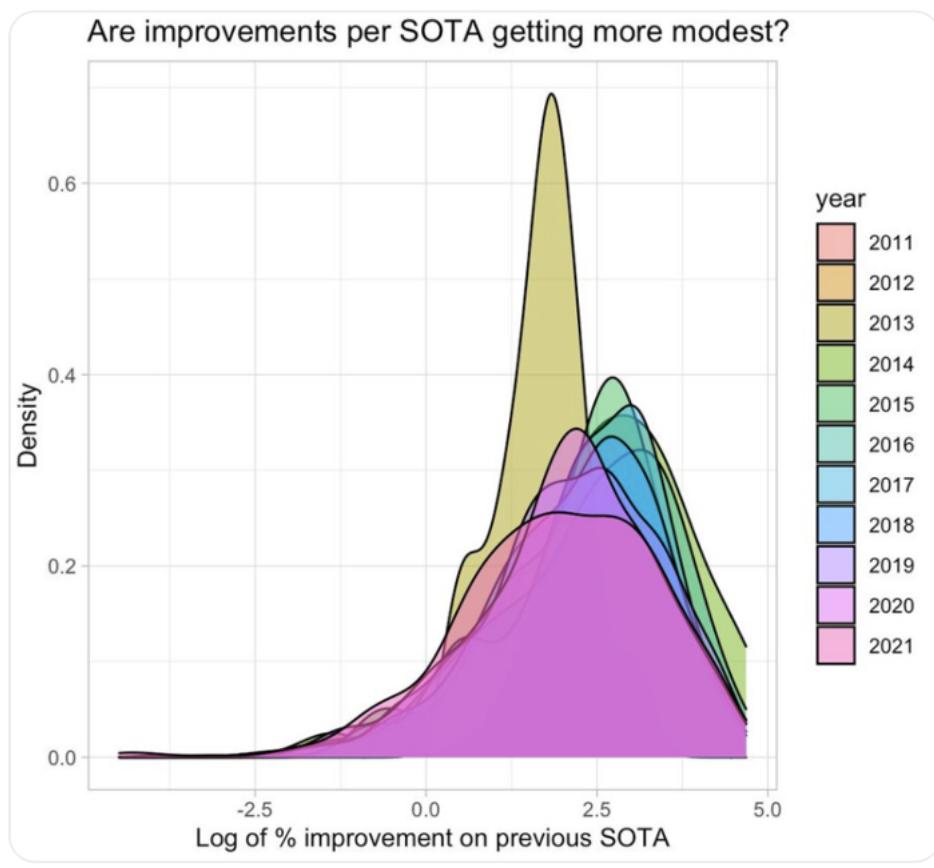


Tamay Besiroglu
@tamaybes

...

Replying to [@tamaybes](#) and [@ESYudkowsky](#)

There is some evidence to suggest that even amongst the top researchers, breakthroughs are getting a little more modest. Here is a density plot showing % improvement on previous SOTA across various years. The data seems to suggest that SOTA improvements get more modest over time.



10:29 AM · Oct 19, 2021 · Twitter Web App

16. November 3 conversation

16.1. EfficientZero

[Yudkowsky][9:30]

Thing that (if true) strikes me as... straight-up falsifying Paul's view as applied to modern-day AI, at the frontier of the most AGI-ish part of it and where Deepmind put in substantial effort on their project? EfficientZero (allegedly) learns Atari in 100,000 frames. Caveat: I'm not having an easy time figuring out how many frames MuZero would've required to achieve the same performance level. MuZero was trained on 200,000,000 frames but reached what looks like an allegedly higher high; the EfficientZero paper compares their performance to MuZero on 100,000 frames, and claims theirs is much better than MuZero given only that many frames.

<https://arxiv.org/pdf/2111.00210.pdf> CC: @paulfchristiano.

(I would further argue that this case is important because it's about the central contemporary model approaching AGI, at least according to Eliezer, rather than any number of random peripheral AI tasks

[Shah][14:46]

I only looked at the front page, so might be misunderstanding, but the front figure says "Our proposed method EfficientZero is 170% and 180% better than the previous SoTA performance in mean and median human normalized score [...] on the Atari 100k benchmark", which does not seem like a huge leap?

Oh, I incorrectly thought that was 1.7x and 1.8x, but it is actually 2.7x and 2.8x, which is a bigger deal (though still feels not crazy to me)

[Yudkowsky][15:28]

the question imo is how many frames the previous SoTA would require to catch up to EfficientZero

(I've tried emailing an author to ask about this, no response yet)

like, perplexity on GPT-3 vs GPT-2 and "losses decreased by blah%" would give you a pretty meaningful concept of how far ahead GPT-3 was from GPT-2, and I think the "2.8x performance" figure in terms of scoring is equally meaningless as a metric of how much EfficientZero improves if any

what you want is a notion like "previous SoTA would have required 10x the samples" or "previous SoTA would have required 5x the computation" to achieve that performance level

[Shah][15:38]

I see. Atari curves are not nearly as nice and stable as GPT curves and often have the problem that they plateau rather than making steady progress with more training time, so that will make these metrics noisier, but it does seem like a reasonable metric to track

(Not that I have recommendations about how to track it; I doubt the authors can easily get these metrics)

[Christiano][18:01]

If you think our views are making such starkly different predictions then I'd be happy to actually state any of them in advance, including e.g. about future ML benchmark results.

I don't think this falsifies my view, and we could continue trying to hash out what my view is but it seems like slow going and I'm inclined to give up.

Relevant questions on my view are things like: is MuZero optimized at all for performance in the tiny-sample regime? (I think not, I don't even think it set SoTA on that task and I haven't seen any evidence). What's the actual rate of improvements since people started studying this benchmark ~2 years ago,

and how much work has gone into it? And I totally agree with your comments that "# of frames" is the natural unit for measuring and that would be the starting point for any discussion.

[Barnes][18:22]

In previous MCTS RL algorithms, the environment model is either given or only trained with reward values, and policies, which cannot provide sufficient training signals due to their scalar nature. The problem is more severe when the reward is sparse or the bootstrapped value is not accurate. The MCTS policy improvement operator heavily relies on the environment model. Thus, it is vital to have an accurate one.

We notice that the output s_{t+1} from the dynamic function G should be the same as s_{t+1} , i.e. the output of the representation function H with input of the next observation o_{t+1} (Fig. 2). This can help to supervise the predicted next state \hat{s}_{t+1} using the actual s_{t+1} , which is a tensor with at least a few hundred dimensions. This provides \hat{s}_{t+1} with much more training signals than the default scalar reward and value.

This seems like a super obvious thing to do and I'm confused why DM didn't already try this. It was definitely being talked about in ~2018

Will ask a DM friend about it

[Yudkowsky][22:45]

I... don't think I want to take *all* of the blame for misunderstanding Paul's views; I think I also want to complain at least a little that Paul spends an insufficient quantity of time pointing at extremely concrete specific possibilities, especially real ones, and saying how they do or don't fit into the scheme.

Am I rephrasing correctly that, in this case, if Efficient Zero was actually a huge (3x? 5x? 10x?) jump RL sample efficiency over previous SOTA, measured in 1 / frames required to train to a performance level, then that means the Paul view *doesn't* apply to the present world; but this could be because MuZero wasn't the real previous SOTA, or maybe because nobody really worked on pushing out this benchmark for 2 years and therefore on the Paul view it's fine for there to still be huge jumps? In other words, this is something Paul's worldview has to either deny or excuse, and not just, "well, sure, why wouldn't it do that, you have misunderstood which kinds of AI-related events Paul is even trying to talk about"?

In the case where, "yes it's a big jump and that shouldn't happen later, but it could happen now because it turned out nobody worked hard on pushing past MuZero over the last 2 years", I wish to register that my view permits it to be the case that, when the world begins to end, the frontier that enters into AG is similarly something that not a lot of people spent a huge effort on since a previous prototype from 2 years earlier. It's just not very surprising to me if the future looks a lot like the past, or if human civilization neglects to invest a ton of effort in a research frontier.

Gwern guesses that getting to EfficientZero's performance level would require around 4x the samples for MuZero-Reanalyze (the more efficient version of MuZero which replayed past frames), which is also apparently the only version of MuZero the paper's authors were considering in the first place - without replays, MuZero requires 20 billion frames to achieve its performance, not the figure of 200 million. <https://www.lesswrong.com/posts/jYNT3Qihn2aAYaaPb/efficientzero-human-ale-sample-efficiency-with-muzero-self?commentId=JEHPQa7i8Qjcg7TW6>

17. November 4 conversation

17.1. EfficientZero (continued)

[Christiano][7:42]

I think it's possible the biggest misunderstanding is that you somehow think of my view as a "scheme" and your view as a normal view where probability distributions over things happen.

Concretely, this is a paper that adds a few techniques to improve over MuZero in a domain that (it appears) wasn't a significant focus of MuZero. I don't know how much it improves but I can believe gwern's estimates of 4x.

I'd guess MuZero itself is a 2x improvement over the baseline from a year ago, which was maybe a 4x improvement over the algorithm from a year before that.

If that's right, then no it's not mindblowing on my view to have 4x progress one year, 2x progress the next, and 4x progress the next.

If other algorithms were better than MuZero, then the 2019-2020 progress would be >2x and the 2020-2021 progress would be <4x.

I think it's probably >4x sample efficiency though (I don't totally buy gwern's estimate there), which makes it at least possibly surprising.

But it's never going to be that surprising. It's a benchmark that people have been working on for a few years that has been seeing relatively rapid improvement over that whole period.

The main innovation is how quickly you can learn to predict future frames of Atari games, which has tiny economic relevance and calling it the most AGI-ish direction seems like it's a very Eliezer-ish view, this isn't the kind of domain where I'm either most surprised to see rapid progress at all nor is the kind of thing that seems like a key update re: transformative AI

yeah, SoTA in late 2020 was SPR, published by a much smaller academic group:
<https://arxiv.org/pdf/2007.05929.pdf>

MuZero wasn't even setting sota on this task at the time it was published

my "schemes" are that (i) if a bunch of people are trying on a domain and making steady slow progress, I'm surprised to see giant jumps and I don't expect most absolute progress to occur in such jumps, (ii) if a domain is worth a lot of \$, generally a bunch of people will be trying. Those aren't claims about what is always true, they are claims about what is typically true and hence what I'm guessing will be true for transformative AI.

Maybe you think those things aren't even good general predictions, and that I don't have long enough tails in my distributions or whatever. But in that case it seems we can settle it quickly by prediction.

I think this result is probably significant (>30% absolute improvement) + faster-than-trend (>50% faster than previous increment) progress relative to prior trend on 8 of the 27 atari games (from table 1, treating SimPL->{max of MuZero, SPR}->EfficientZero as 3 equally spaced datapoints): Asterix, Breakout, almost ChopperCMD, almost CrazyClimber, Gopher, Kung Fu Master, Pong, Qbert, SeaQuest. My guess is that they thought a lot about a few of those games in particular because they are very influential on the mean/median. Note that this paper is a giant grab bag and that simply stapling together the prior methods would have already been a significant improvement over prior SoTA. (ETA: I don't think saying "its only 8 of 27 games" is an update against it being big progress or anything. I do think saying "stapling together 2 previous methods without any complementarity at all would already have significantly beaten SoTA" is fairly good evidence that it's not a hard-to-beat SoTA.)

and even fewer people working on the ultra-low-sample extremely-low-dimensional DM control environments (this is the subset of problems where the state space is 4 dimensions,

people are just not trying to publish great results on cartpole), so I think the most surprising contribution is the atari stuff

OK, I now also understand what the result is I think?

I think the quick summary is: the prior SoTA is SPR, which learns to predict the domain and then does Q-learning. MuZero instead learns to predict the domain and does MCTS, but it predicts the domain in a slightly less sophisticated way than SPR (basically just predicts rewards, whereas SPR predicts all of the agent's latent state in order to get more signal from each frame). If you combine MCTS with more sophisticated prediction, you do better.

I think if you told me that DeepMind put in significant effort in 2020 (say, at least as much post-MuZero effort as the new paper?) trying to get great sample efficiency on the easy-exploration atari games, and failed to make significant progress, then I'm surprised.

I don't think that would "falsify" my view, but it would be an update against? Like maybe if DM put in that much effort I'd maybe have given only a 10-20% probability to a new project of similar size putting in that much effort making big progress, and even conditioned on big progress this is still >>median (ETA: and if DeepMind put in much more effort I'd be more surprised than 10-20% by big progress from the new project)

Without DM putting in much effort, it's significantly less surprising and I'll instead be comparing to the other academic efforts. But it's just not surprising that you can beat them if you are willing to put in the effort to reimplement MCTS and they aren't, and that's a step that is straightforwardly going to improve performance.

(not sure if that's the situation)

And then to see how significant updates against are, you have to actually contrast them with all the updates in the other direction where people *don't* crush previous benchmark results and instead just make modest progress

I would guess that if you had talked to an academic about this question (what happens if you combine SPR+MCTS) they would have predicted significant wins in sample efficiency (at the expense of compute efficiency) and cited the difficulty of implementing MuZero compared to any of the academic results. That's another way I could be somewhat surprised (or if there were academics with MuZero-quality MCTS implementations working on this problem, and they somehow didn't set SoTA, then I'm even more surprised). But I'm not sure if you'll trust any of those judgments in hindsight.

Repeating the main point :

I don't really think a 4x jump over 1 year is something I have to "defy or excuse", it's something that I think becomes more or less likely depending on facts about the world, like (i) how fast was previous progress, (ii) how many people were working on previous projects and how targeted were they at this metric, (iii) how many people are working in this project and how targeted was it at this metric

it becomes continuously less likely as those parameters move in the obvious directions

it never becomes 0 probability, and you just can't win that much by citing isolated events that I'd give say a 10% probability to, unless you actually say something about how you are giving >10% probabilities to those events without losing a bunch of probability mass on what I see as the 90% of boring stuff

[Ngo: ]

and then separately I have a view about lots of people working on important problems, which doesn't say anything about this case

(I actually don't think this event is as low as 10%, though it depends on what background facts about the project you are conditioning on---obviously I gave <<10% probability to someone publishing this particular result, but something like "what fraction of progress in this field would come down to jumps like this" or whatever is probably >10% until you tell me that DeepMind actually cared enough to have already tried)

[Ngo][8:48]

I expect Eliezer to say something like: DeepMind believes that both improving RL sample efficiency, and benchmarking progress on games like Atari, are important parts of the path towards AGI. So insofar as your model predicts that smooth progress will be caused by people working directly towards AGI, DeepMind not putting effort into this is a hit to that model. Thoughts?

[Christiano][9:06]

I don't think that learning these Atari games in 2 hours is a very interesting benchmark even for deep RL sample efficiency, and it's totally unrelated to the way in which humans learn such games quickly. It seems ~~pretty likely~~ totally plausible (50%?) to me that DeepMind feels the same way, and then the question is about other random considerations like how they are making some PR calculation.

[Ngo][9:18]

If Atari is not a very interesting benchmark, then why did DeepMind put a bunch of effort into making Agent57 and applying MuZero to Atari?

Also, most of the effort they've spent on games in general has been on methods very unlike the way humans learn those games, so that doesn't seem like a likely reason for them to overlook these methods for increasing sample efficiency.

[Shah][9:32]

It seems pretty likely totally plausible (50%?) to me that DeepMind feels the same way, and then the question is about other random considerations like how they are making some PR calculation.

Not sure of the exact claim, but DeepMind is big enough and diverse enough that I'm pretty confident at least some people working on relevant problems don't feel the same way

[...] This seems like a super obvious thing to do and I'm confused why DM didn't already try this. It was definitely being talked about in ~2018

Speculating without my DM hat on: maybe it kills performance in board games, and they want one algorithm for all settings?

[Christiano][10:29]

Atari games in the tiny sample regime are a different beast

there are just a lot of problems you can state about Atari some of which are more or less interesting (e.g. jointly learning to play 57 Atari games is a more interesting problem than learning how to play one of them absurdly quickly, and there are like 10 other problems about Atari that are more interesting than this one)

That said, Agent57 also doesn't seem interesting except that it's an old task people kind of care about. I don't know about the take within DeepMind but outside I don't think anyone would care about it other than historical significance of the benchmark / obviously-not-cherrypickedness of the problem.

I'm sure that some people at DeepMind care about getting the super low sample complexity regime. I don't think that really tells you how large the DeepMind effort is compared to some random academics who care about it.

[Shah: ]

I think the argument for working on deep RL is fine and can be based on an analogy with humans while you aren't good at the task. Then once you are aiming for crazy superhuman

performance on Atari games you naturally start asking "what are we doing here and why are we still working on atari games?"

[Ngo: ]

and correspondingly they are a smaller and smaller slice of DeepMind's work over time

[Ngo: ]

(e.g. Agent57 and MuZero are the only DeepMind blog posts about Atari in the last 4 years, it's not the main focus of MuZero and I don't think Agent57 is a very big DM project)

Reaching this level of performance in Atari games is largely about learning perception, and doing that from 100k frames of an Atari game just doesn't seem very analogous to anything humans do or that is economically relevant from any perspective. I totally agree some people are into it, but I'm totally not surprised if it's not going to be a big DeepMind project.

[Yudkowsky][10:51]

would you agree it's a load-bearing assumption of your worldview - where I also freely admit to having a worldview/scheme, this is not meant to be a prejudicial term at all - that the line of research which leads into world-shaking AGI must be in the mainstream and not in a weird corner where a few months earlier there were more profitable other ways of doing all the things that weird corner did?

eg, the tech line leading into world-shaking AGI must be at the profitable forefront of non-world-shaking tasks. as otherwise, afaict, your worldview permits that if counterfactually we were in the Paul-forbidden case where the immediate precursor to AGI was something like EfficientZero (whose motivation had been beating an old SOTA metric rather than, say, market-beating self-driving cars), there might be huge capability leaps there just as EfficientZero represents a large leap, because there wouldn't have been tons of investment in that line.

[Christiano][10:54]

Something like that is definitely a load-bearing assumption

Like there's a spectrum with e.g. EfficientZero --> 2016 language modeling --> 2014 computer vision --> 2021 language modeling --> 2021 computer vision, and I think everything anywhere close to transformative AI will be way way off the right end of that spectrum

But I think quantitatively the things you are saying don't seem quite right to me. Suppose that MuZero wasn't the best way to do anything economically relevant, but it was within a factor of 4 on sample efficiency for doing tasks that people care about. That's already going to be enough to make tons of people extremely excited.

So yes, I'm saying that anything leading to transformative AI is "in the mainstream" in the sense that it has more work on it than 2021 language models.

But not necessarily that it's the most profitable way to do anything that people care about. Different methods scale in different ways, and something can burst onto the scene in a dramatic way, but I strongly expect speculative investment driven by that possibility to already be way (way) more than 2021 language models. And I don't expect gigantic surprises. And I'm willing to bet that e.g. EfficientZero isn't a big surprise for researchers who are paying attention to the area (*in addition* to being 3+ orders of magnitude more neglected than anything close to transformative AI)

2021 language modeling isn't even very competitive, it's still like 3-4 orders of magnitude smaller than semiconductors. But I'm giving it as a reference point since it's obviously much, much more competitive than sample-efficient atari.

This is a place where I'm making much more confident predictions, this is "falsify paul's worldview" territory once you get to quantitative claims anywhere close to TAI and "even a

single example seriously challenges paul's worldview" a few orders of magnitude short of that

[Yudkowsky][11:04]

can you say more about what falsifies your worldview previous to TAI being super-obviously-to-all-EAs imminent?

or rather, "seriously challenges", sorry

[Christiano][11:05][11:08]

big AI applications achieved by clever insights in domains that aren't crowded, we should be quantitative about how crowded and how big if we want to get into "seriously challenges"

like e.g. if this paper on atari was actually a crucial ingredient for making deep RL for robotics work, I'd be actually surprised rather than 10% surprised

but it's not going to be, those results are being worked on by much larger teams of more competent researchers at labs with \$100M+ funding

it's definitely possible for them to get crushed by something out of left field

but I'm betting against every time

or like, the set of things people would describe as "out of left field," and the quantitative degree of neglectedness, becomes more and more mild as the stakes go up

[Yudkowsky][11:08]

how surprised are you if in 2022 one company comes out with really good ML translation, and they manage to sell a bunch of it temporarily until others steal their ideas or Google acquires them? my model of Paul is unclear on whether this constitutes "many people are already working on language models including ML translation" versus "this field is not profitable enough right this minute for things to be efficient there, and it's allowed to be nonobvious in worlds where it's about to become profitable".

[Christiano][11:08]

if I wanted to make a prediction about that I'd learn a bunch about how much google works on translation and how much \$ they make

I just don't know the economics

and it depends on the kind of translation that they are good at and the economics (e.g. google mostly does extremely high-volume very cheap translation)

but I think there are lots of things like that / facts I could learn about Google such that I'd be surprised in that situation

independent of the economics, I do think a fair number of people are working on adjacent stuff, and I don't expect someone to come out of left field for google-translate-cost translation between high-resource languages

but it seems quite plausible that a team of 10 competent people could significantly outperform google translate, and I'd need to learn about the economics to know how surprised I am by 10 people or 100 people or what

I think it's allowed to be non-obvious whether a domain is about to be really profitable

but it's not that easy, and the higher the stakes the more speculative investment it will drive, etc.

[Yudkowsky][11:14]

if you don't update much off EfficientZero, then people also shouldn't be updating much off of most of the graph I posted earlier as possible Paul-favoring evidence, because most of those SOTAs weren't highly profitable so your worldview didn't have much to say about them. ?

[Christiano][11:15]

Most things people work a lot on improve gradually. EfficientZero is also quite gradual compared to the crazy TAI stories you tell. I don't really know what to say about this game other than I would prefer make predictions in advance and I'm happy to either propose questions/domains or make predictions in whatever space you feel more comfortable with.

[Yudkowsky][11:16]

I don't know how to point at a future event that you'd have strong opinions about. it feels like, whenever I try, I get told that the current world is too unlike the future conditions you expect.

[Christiano][11:16]

Like, whether or not EfficientZero is evidence for your view depends on exactly how "who knows what will happen" you are. if you are just a bit more spread out than I am, then it's definitely evidence for your view.

I'm saying that I'm willing to bet about *any event you want to name*, I just think my model of how things work is more accurate.

I'd prefer it be related to ML or AI.

[Yudkowsky][11:17]

to be clear, I appreciate that it's similarly hard to point at an event like that for myself, because my own worldview says "well mostly the future is not all that predictable with a few rare exceptions"

[Christiano][11:17]

But I feel like the situation is not at all symmetrical, I expect to outperform you on practically any category of predictions we can specify.

so like I'm happy to bet about benchmark progress in LMs, or about whether DM or OpenAI or Google or Microsoft will be the first to achieve something, or about progress in computer vision, or about progress in industrial robotics, or about translations

whatever

17.2. Near-term AI predictions

[Yudkowsky][11:18]

that sounds like you ought to have, like, a full-blown storyline about the future?

[Christiano][11:18]

what is a full-blown storyline? I have a bunch of ways that I think about the world and make predictions about what is likely

and yes, I can use those ways of thinking to make predictions about whatever

and I will very often lose to a domain expert who has better and more informed ways of making predictions

[Yudkowsky][11:19]

what happens if 2022 through 2024 looks literally exactly like Paul's modal or median predictions on things?

[Christiano][11:19]

but I think in ML I will generally beat e.g. a superforecaster who doesn't have a lot of experience in the area

give me a question about 2024 and I'll give you a median?

I don't know what "what happens" means

storylines do not seem like good ways of making predictions

[Shah: 👍]

[Yudkowsky][11:20]

I mean, this isn't a crux for anything, but it seems like you're asking me to give up on that and just ask for predictions? so in 2024 can I hire an artist who doesn't speak English and converse with them almost seamlessly through a machine translator?

[Christiano][11:22]

median outcome (all of these are going to be somewhat easy-to-beat predictions because I'm not thinking): you can get good real-time translations, they are about as good as a +1 stdev bilingual speaker who listens to what you said and then writes it out in the other language as fast as they can type

Probably also for voice -> text or voice -> voice, though higher latencies and costs.

Not integrated into standard video chatting experience because the UX is too much of a pain and the world sucks.

That's a median on "how cool/useful is translation"

[Yudkowsky][11:23]

I would unfortunately also predict that in this case, this will be a highly competitive market and hence not a very profitable one, which I predict to match your prediction, but I ask about the economics here just in case.

[Christiano][11:24]

Kind of typical sample: I'd guess that Google has a reasonably large lead, most translation still provided as a free value-added, cost per translation at that level of quality is like \$0.01/word, total revenue in the area is like \$10Ms / year?

[Yudkowsky][11:24]

well, my model also permits that Google does it for free and so it's an uncompetitive market but not a profitable one... ninjaed.

[Christiano][11:25]

first order of improving would be sanity-checking economics and thinking about #s, second would be learning things like "how many people actually work on translation and what is the state of the field?"

[Yudkowsky][11:26]

did Tesla crack self-driving cars and become a \$3T company instead of a \$1T company? do you own Tesla options?

did Waymo beat Tesla and cause Tesla stock to crater, same question?

[Christiano][11:27]

1/3 chance tesla has FSD in 2024

conditioned on that, yeah probably market cap is >\$3T?

conditioned on Tesla having FSD, 2/3 chance Waymo has also at least rolled out to a lot of cities

conditioned on no tesla FSD, 10% chance Waymo has rolled out to like half of big US cities?

dunno if numbers make sense

[Yudkowsky][11:28]

that's okay, I dunno if my questions make sense

[Christiano][11:29]

(5% NW in tesla, 90% NW in AI bets, 100% NW in more normal investments; no tesla options that sounds like a scary place with lottery ticket biases and the crazy tesla investors)

[Yudkowsky][11:30]

(am I correctly understanding you're 2x levered?)

[Christiano][11:30][11:31]

yeah

it feels like you've got to have weird views on trajectory of value-added from AI over the coming years

on how much of the \$ comes from domains that are currently exciting to people (e.g. that Google already works on, self-driving, industrial robotics) vs stuff out of left field

on what kind of algorithms deliver \$ in those domains (e.g. are logistics robots trained using the same techniques tons of people are currently pushing on)

on my picture you shouldn't be getting big losses on any of those

just losing like 10-20% each time

[Yudkowsky][11:31][11:32]

my uncorrected inside view says that machine translation should be in reach and generate huge amounts of economic value even if it ends up an unprofitable competitive or Google-freebie field

and also that not many people are working on basic research in machine translation or see it as a "currently exciting" domain

[Christiano][11:32]

how many FTE is "not that many" people?

also are you expecting improvement in the google translate style product, or in lower-latencies for something closer to normal human translator prices, or something else?

[Yudkowsky][11:33]

my worldview says more like... sure, maybe there's 300 programmers working on it worldwide, but most of them aren't aggressively pursuing new ideas and trying to explore the space, they're just applying existing techniques to a new language or trying to throw on some tiny mod that lets them beat SOTA by 1.2% for a publication

because it's not an *exciting* field

"What if you could rip down the language barriers" is an economist's dream, or a humanist's dream, and Silicon Valley is neither

and looking at GPT-3 and saying, "God damn it, this really seems like it must on some level *understand* what it's reading well enough that the same learned knowledge would suffice to do really good machine translation, this must be within reach for gradient descent technology we just don't know how to reach it" is Yudkowskian thinking; your AI system has internal parts like "how much it understands language" and there's thoughts about what those parts ought to be able to do if you could get them into a new system with some other parts

[Christiano][11:36]

my guess is we'd have some disagreements here

but to be clear, you are talking about text-to-text at like \$0.01/word price point?

[Yudkowsky][11:38]

I mean, do we? Unfortunately another Yudkowskian worldview says "and people can go on failing to notice this for arbitrarily long amounts of time".

if that's around GPT-3's price point then yeah

[Christiano][11:38]

gpt-3 is a lot cheaper, happy to say gpt-3 like price point

[Yudkowsky][11:39]

(thinking about whether \$0.01/word is meaningfully different from \$0.001/word and concluding that it is)

[Christiano][11:39]

(api is like 10,000 words / \$)

I expect you to have a broader distribution over who makes a great product in this space, how great it ends up being etc., whereas I'm going to have somewhat higher probabilities on it being google research and it's going to look boring

[Yudkowsky][11:40]

what is boring?

boring predictions are often good predictions on my own worldview too

lots of my gloom is about things that are boringly bad and awful

(and which add up to instant death at a later point)

but, I mean, what does boring machine translation look like?

[Christiano][11:42]

Train big language model. Have lots of auxiliary tasks especially involving reading in source language and generation in target language. Have pre-training on aligned sentences and perhaps using all the unsupervised translation we have depending on how high-resource language is. Fine-tune with smaller amount of higher quality supervision.

Some of the steps likely don't add much value and skip them. Fair amount of non-ML infrastructure.

For some languages/domains/etc. dedicated models, over time increasingly just have a giant model with learned dispatch as in mixture of experts.

[Yudkowsky][11:44]

but your worldview is also totally ok with there being a Clever Trick added to that which produces a 2x reduction in training time. or with there being a new innovation like transformers, which was developed a year earlier and which everybody now uses, without which the translator wouldn't work at all. ?

[Christiano][11:44]

Just for reference, I think transformers aren't that visible on a (translation quality) vs (time) graph?

But yes, I'm totally fine with continuing architectural improvements, and 2x reduction in training time is currently par for the course for "some people at google thought about architectures for a while" and I expect that to not get that much tighter over the next few years.

[Yudkowsky][11:45]

unrolling Restricted Boltzmann Machines to produce deeper trainable networks probably wasn't much visible on a graph either, but good luck duplicating modern results using only lower portions of the tech tree. (I don't think we disagree about this.)

[Christiano][11:45]

I do expect it to eventually get tighter, but not by 2024.

I don't think unrolling restricted boltzmann machines is that important

[Yudkowsky][11:46]

like, historically, or as a modern technology?

[Christiano][11:46]

historically

[Yudkowsky][11:46]

interesting

my model is that it got people thinking about "what makes things trainable" and led into ReLUs and inits

but I am going more off having watched from the periphery as it happened, than having read a detailed history of that

like, people asking, "ah, but what if we had a deeper network and the gradients *didn't* explode or die out?" and doing that en masse in a productive way rather than individuals being wistful for 30 seconds

[Christiano][11:48]

well, not sure if this will introduce differences in predictions

I don't feel like it should really matter for our bottom line predictions whether we classify google's random architectural change as something fundamentally new (which happens to just have a modest effect at the time that it's built) or as something boring

I'm going to guess how well things will work by looking at how well things work right now and seeing how fast it's getting better

and that's also what I'm going to do for applications of AI with transformative impacts

and I actually believe you will do something today that's analogous to what you would do in the future, and in fact will make somewhat different predictions than what I would do

and then some of the action will be in new things that people haven't been trying to do in the past, and I'm predicting that new things will be "small" whereas you have a broader distribution, and there's currently some not-communicated judgment call in "small"

if you think that TAI will be like translation, where google publishes tons of papers, but that they will just get totally destroyed by some new idea, then it seems like that should correspond to a difference in P(google translation gets totally destroyed by something out-of-left-field)

and if you think that TAI won't be like translation, then I'm interested in examples more like TAI

I don't really understand the take "and people can go on failing to notice this for arbitrarily long amounts of time," why doesn't that also happen for TAI and therefore cause it to be the boring slow progress by google? Why would this be like a 50% probability for TAI but <10% for translation?

perhaps there is a disagreement about how good the boring progress will be by 2024? looks to me like it will be very good

[Yudkowsky][11:57]

I am not sure that is where the disagreement lies

17.3. The evolution of human intelligence

[Yudkowsky][11:57]

I am considering advocating that we should have more disagreements about the past, which has the advantage of being very concrete, and being often checkable in further detail than either of us already know

[Christiano][11:58]

I'm fine with disagreements about the past; I'm more scared of letting you pick arbitrary things to "predict" since there is much more impact from differences in domain knowledge (also not quite sure why it's more concrete, I guess because we can talk about what led to particular events? mostly it just seems faster)

also as far as I can tell our main differences are about whether people will ~~spend a lot of money~~ work effectively on things that would make a lot of money, which means if we look to the past we will have to move away from ML/AI

[Yudkowsky][12:00]

so my understanding of how Paul writes off the example of human intelligence, is that you are like, "evolution is much stupider than a human investor; if there'd been humans running the genomes, people would be copying all the successful things, and hominid brains would be developing in this ecology of competitors instead of being a lone artifact". ?

[Christiano][12:00]

I don't understand why I have to write off the example of human intelligence

[Yudkowsky][12:00]

because it looks nothing like your account of how TAI develops

[Christiano][12:00]

it also looks nothing like your account, I understand that you have some analogy that makes sense to you

[Yudkowsky][12:01]

I mean, to be clear, I also write off the example of humans developing morality and have to explain to people at length why humans being as nice as they are, doesn't imply that paperclip maximizers will be anywhere near that nice, nor that AIs will be other than paperclip maximizers.

[Christiano][12:01][12:02]

you could state some property of how human intelligence developed, that is in common with your model for TAI and not mine, and then we could discuss that

if you say something like: "chimps are not very good at doing science, but humans are" then yes my answer will be that it's because evolution was not selecting us to be good at science

and indeed AI systems will be good at science using *much* less resources than humans or chimps

[Yudkowsky][12:02][12:02]

would you disagree that humans developing intelligence, on the sheer surfaces of things, looks much more Yudkowskian than Paulian?

like, not in terms of compatibility with underlying model

just that there's this one corporation that came out and massively won the entire AGI race with zero competitors

[Christiano][12:03]

I agree that "how much did the winner take all" is more like your model of TAI than mine

I don't think zero competitors is reasonable, I would say "competitors who were tens of millions of years behind"

[Yudkowsky][12:03]

sure

and your account of this is that natural selection is nothing like human corporate managers copying each other

[Christiano][12:03]

which was a reasonable timescale for the old game, but a long timescale for the new game

[Yudkowsky][12:03]

yup

[Christiano][12:04]

that's not my only account

it's also that for human corporations you can form large coalitions, i.e. raise huge amounts of \$ and hire huge numbers of people working on similar projects (whether or not vertically integrated), and those large coalitions will systematically beat small coalitions

and that's basically *the* key dynamic in this situation, and isn't even trying to have any analog in the historical situation

(the key dynamic w.r.t. concentration of power, not necessarily the main thing overall)

[Yudkowsky][12:07]

the modern degree of concentration of power seems relatively recent and to have tons and tons to do with the regulatory environment rather than underlying properties of the innovation landscape

back in the old days, small startups would be better than Microsoft at things, and Microsoft would try to crush them using other forces than superior technology, not always successfully or such was the common wisdom of USENET

[Christiano][12:08]

my point is that the evolution analogy is extremely unpersuasive w.r.t. concentration of power

I think that AI software capturing the amount of power you imagine is also kind of implausible because we know something about how hardware trades off against software progress (maybe like 1 year of progress = 2x hardware) and so even if you can't form coalitions on innovation *at all* you are still going to be using tons of hardware if you want to be in the running

though if you can't parallelize innovation at all and there is enough dispersion in software progress then the people making the software could take a lot of the \$ / influence from the partnership

anyway, I agree that this is a way in which evolution is more like your world than mine

but think on this point the analogy is pretty unpersuasive

because it fails to engage with any of the a priori reasons you wouldn't expect concentration of power

[Yudkowsky][12:11]

I'm not sure this is the correct point on which to engage, but I feel like I should say out loud that I am unable to operate my model of your model in such fashion that it is not falsified by how the software industry behaved between 1980 and 2000.

there should've been no small teams that beat big corporations

today those are much rarer, but on my model, that's because of regulatory changes (and possibly metabolic damage from something in the drinking water)

[Christiano][12:12]

I understand that you can't operate my model, and I've mostly given up, and on this point I would prefer to just make predictions or maybe retrodictions

[Yudkowsky][12:13]

well, anyways, my model of how human intelligence happened looks like this:

there is a mysterious kind of product which we can call G, and which brains can operate as factories to produce

G in turn can produce other stuff, but you need quite a lot of it piled up to produce *better* stuff than your competitors

as late as 1000 years ago, the fastest creatures on Earth are not humans, because you need even *more G than that* to go faster than cheetahs

(or peregrine falcons)

the natural selections of various species were fundamentally stupid and blind, incapable of foresight and incapable of copying the successes of other natural selections; but even if they had been as foresighted as a modern manager or investor, they might have made just the same mistake

before 10,000 years they would be like, "what's so exciting about these things? they're not the fastest runners."

if there'd been an economy centered around running, you wouldn't invest in deploying a human

(well, unless you needed a stamina runner, but that's something of a separate issue, let's consider just running races)

you would invest on improving cheetahs

because the pile of human G isn't large enough that their G beats a specialized naturally selected cheetah

[Christiano][12:17]

how are you improving cheetahs in the analogy?

you are trying random variants to see what works?

[Yudkowsky][12:18]

using conventional, well-tested technology like MUSCLES and TENDONS

trying variants on those

[Christiano][12:18]

ok

and you think that G doesn't help you improve on muscles and tendons?

until you have a big pile of it?

[Yudkowsky][12:18]

not as a metaphor but as simple historical fact, that's how it played out

it takes a whole big pile of G to go faster than a cheetah

[Christiano][12:19]

as a matter of fact there is no one investing in making better cheetahs

so it seems like we're already playing analogy-game

[Yudkowsky][12:19]

the natural selection of cheetahs is investing in it

it's not doing so by copying humans because of fundamental limitations

however if we replace it with an average human investor, it still doesn't copy humans, why would it

[Christiano][12:19]

that's the part that is silly
or like, it needs more analogy

[Yudkowsky][12:19]

how so? humans aren't the fastest.

[Christiano][12:19]

humans are great at breeding animals
so if I'm natural selection personified, the thing to explain is why I'm not using some of that G to improve on my selection
not why I'm not using G to build a car

[Yudkowsky][12:20]

I'm... confused
is this implying that a key aspect of your model is that people are using AI to decide which AI tech to invest in?

[Christiano][12:20]

no
I think I just don't understand your analogy
here in the actual world, some people are trying to make faster robots by tinkering with robot designs
and then someone somewhere is training their AGI

[Yudkowsky][12:21]

what I'm saying is that you can imagine a little cheetah investor going, "I'd like to copy and imitate some other species's tricks to make my cheetahs faster" and they're looking enviously at falcons, not at humans
not until very late in the game

[Christiano][12:21]

and the relevant question is whether the pre-AGI thing is helpful for automating the work that humans are doing while they tinker with robot designs
that seems like the actual world
and the interesting claim is you saying "nope, not very"

[Yudkowsky][12:22]

I am again confused. Does it matter to your model whether the pre-AGI thing is helpful for automating "tinkering with robot designs" or just profitable machine translation? Either seems like it induces equivalent amounts of investment.

If anything the latter induces much more investment.

[Christiano][12:23]

sure, I'm fine using "tinkering with robot designs" as a lower bound

both are fine

the point is I have no idea what you are talking about in the analogy

what is analogous to what?

I thought cheetahs were analogous to faster robots

[Yudkowsky][12:23]

faster cheetahs are analogous to more profitable robots

[Christiano][12:23]

sure

so you have some humans working on making more profitable robots, right?

who are tinkering with the robots, in a way analogous to natural selection tinkering with cheetahs?

[Yudkowsky][12:24]

I'm suggesting replacing the Natural Selection of Cheetahs with a new optimizer that has the Copy Competitor and Invest In Easily-Predictable Returns feature

[Christiano][12:24]

OK, then I don't understand what those are analogous to

like, what is analogous to the humans who are tinkering with robots, and what is analogous to the humans working on AGI?

[Yudkowsky][12:24]

and observing that, even this case, the owner of Cheetahs Inc. would not try to copy Humans Inc.

[Christiano][12:25]

here's the analogy that makes sense to me

natural selection is working on making faster cheetahs = some humans tinkering away to make more profitable robots

natural selection is working on making smarter humans = some humans who are tinkering away to make more powerful AGI

natural selection doesn't try to copy humans because they suck at being fast = robot-makers don't try to copy AGI-makers because the AGIs aren't very profitable robots

[Yudkowsky][12:26]

with you so far

[Christiano][12:26]

eventually humans build cars once they get smart enough = eventually AGI makes more profitable robots once it gets smart enough

[Yudkowsky][12:26]

yup

[Christiano][12:26]

great, seems like we're on the same page then

[Yudkowsky][12:26]

and by this point it is LATE in the game

[Christiano][12:27]

great, with you still

[Yudkowsky][12:27]

because the smaller piles of G did not produce profitable robots

[Christiano][12:27]

but there's a step here where you appear to go totally off the rails

[Yudkowsky][12:27]

or operate profitable robots

say on

[Christiano][12:27]

can we just write out the sequence of AGIs, AGI(1), AGI(2), AGI(3)... in analogy with the sequence of human ancestors H(1), H(2), H(3)...?

[Yudkowsky][12:28]

Is the last member of the sequence H(n) the one that builds cars and then immediately destroys the world before anything that operates on Cheetah Inc's Owner's scale can react?

[Christiano][12:28]

sure

I don't think of it as the last

but it's the last one that actually arises?

maybe let's call it the last, H(n)

great

and now it seems like you are imagining an analogous story, where AGI(n) takes over the world and maybe incidentally builds some more profitable robots along the way

(building more profitable robots being easier than taking over the world, but not so much easier that AGI(n-1) could have done it unless we make our version numbers really close together, close enough that deploying AGI(n-1) is stupid)

[Yudkowsky][12:31]

if this plays out in the analogous way to human intelligence, AGI(n) becomes able to build more profitable robots 1 hour before it becomes able to take over the world; my worldview does not put that as the median estimate, but I do want to observe that this is what happened historically

[Christiano][12:31]

sure

[Yudkowsky][12:32]

ok, then I think we're still on the same page as written so far

[Christiano][12:32]

so the question that's interesting in the real world is which AGI is useful for replacing humans in the design-better-robots task; is it 1 hour before the AGI that takes over the world, or 2 years, or what?

[Yudkowsky][12:33]

my worldview tends to make a big ol' distinction between "replace humans in the design-better-robots task" and "run as a better robot", if they're not importantly distinct from your standpoint can we talk about the latter?

[Christiano][12:33]

they seem importantly distinct

totally different even

so I think we're still on the same page

[Yudkowsky][12:34]

ok then, "replacing humans at designing better robots" sure as heck sounds to Eliezer like the world is about to end or has already ended

[Christiano][12:34]

my whole point is that in the evolutionary analogy we are talking about "run as a better robot" rather than "replace humans in the design-better-robots-task"

and indeed there is no analog to "replace humans in the design-better-robots-task"

which is where all of the action and disagreement is

[Yudkowsky][12:35][12:36]

well, yes, I was exactly trying to talk about when humans start running as better cheetahs and how that point is still very late in the game

not as late as when humans take over the job of making the thing that makes better cheetahs, aka humans start trying to make AGI, which is basically the fingersnap end of the world from the perspective of Cheetahs Inc.

[Christiano][12:36]

OK, but I don't care when humans are better cheetahs---in the real world, when AGIs are better robots. In the real world I care about when AGIs start replacing humans in the design-better-robots-task. I'm game to use evolution as an analogy to help answer *that* question (where I do agree that it's informative), but want to be clear what's actually at issue.

[Yudkowsky][12:37]

so, the thing I was trying to work up to, is that my model permits the world to end in a way where AGI doesn't get tons of investment because it has an insufficiently huge pile of G that it could run as a better robot. people are instead investing in the equivalents of cheetahs.

I don't understand why your model doesn't care when humans are better cheetahs. AGIs running as more profitable robots is what induces the huge investments in AGI that your model requires to produce very close competition. ?

[Christiano][12:38]

it's a sufficient condition, but it's not the most robust one at all

like, I happen to think that in the real world AIs actually are going to be incredibly profitable robots, and that's part of my boring view about what AGI looks like

But the thing that's more robust is that the sub-taking-over-world AI is already really important, and receiving huge amounts of investment, as something that automates the R&D process. And it seems like the best guess given what we know now is that this process starts years before the singularity.

From my perspective that's where most of the action is. And your views on that question seem related to your views on how e.g. AGI is a fundamentally different ballgame from making better robots (whereas I think the boring view is that they are closely related), but that's more like an upstream question about what you think AGI will look like, most relevant because I think it's going to lead you to make bad short-term predictions about what kinds of technologies will achieve what kinds of goals.

[Yudkowsky][12:41]

but not all AIs are the same branch of the technology tree. factory robotics are already really important and they are "AI" but, on my model, they're currently on the cheetah branch rather than the hominid branch of the tech tree; investments into better factory robotics are not directly investments into improving MuZero, though they may buy chips that MuZero also buys.

[Christiano][12:42]

Yeah, I think you have a mistaken view of AI progress. But I still disagree with your bottom line even if I adopt (this part of) your view of AI progress.

Namely, I think that the AGI line is mediocre before it is great, and the mediocre version is spectacularly valuable for accelerating R&D (mostly AGI R&D).

The way I end up sympathizing with your view is if I adopt both this view about the tech tree, + another equally-silly-seeming view about how close the AGI line is to foaming (or how inefficient the area will remain as we get close to foaming)

17.4. Human generality and body manipulation

[Yudkowsky][12:43]

so metaphorically, you require that humans be doing Great at Various Things and being Super Profitable way before they develop agriculture; the rise of human intelligence cannot be a case in point of your model because the humans were too uncompetitive at most animal activities for unrealistically long (edit: compared to the AI case)

[Christiano][12:44]

I don't understand

Human brains are really great at basically everything as far as I can tell?

like it's not like other animals are better at manipulating their bodies

we crush them

[Yudkowsky][12:44]

if we've got weapons, yes

[Christiano][12:44]

human bodies are also pretty great, but they are not the greatest on every dimension

[Yudkowsky][12:44]

wrestling a chimpanzee without weapons is famously ill-advised

[Christiano][12:44]

no, I mean everywhere

chimpanzees are practically the same as humans in the animal kingdom

they have almost as excellent a brain

[Yudkowsky][12:45]

as is attacking an elephant with your bare hands

[Christiano][12:45]

that's not because of elephant brains

[Yudkowsky][12:45]

well, yes, exactly

you need a big pile of G before it's profitable

so big the game is practically over by then

[Christiano][12:45]

this seems so confused

but that's exciting I guess

like, I'm saying that the brains to automate R&D

are similar to the brains to be a good factory robot

analogously, I think the brains that humans use to do R&D

are similar to the brains we use to manipulate our body absurdly well

I do not think that our brains make us fast

they help a tiny bit but not much

I do not think the physical actuators of the industrial robots will be that similar to the actuators of the robots that do R&D

the claim is that the problem of building the brain is pretty similar

just as the problem of building a brain that can do science is pretty similar to the problem of building a brain that can operate a body really well

(and indeed I'm claiming that human bodies kick ass relative to other animal bodies---there may be particular tasks other animal brains are pre-built to be great at, but (i) humans would be great at those too if we were under mild evolutionary pressure with our otherwise excellent brains, (ii) there are lots of more general tests of how good you are at operating a body and we will crush it at those tests)

(and that's not something I know much about, so I could update as I learned more about how actually we just aren't that good at motor control or motion planning)

[Yudkowsky][12:49]

so on your model, we can introduce humans to a continent, forbid them any tool use, and they'll still wipe out all the large animals?

[Christiano][12:49]

(but damn we seem good to me)

I don't understand why that would even plausibly follow

[Yudkowsky][12:49]

because brains are profitable early, even if they can't build weapons?

[Christiano][12:49]

I'm saying that if you put our brains in a big animal body
we would wipe out the big animals
yes, I think brains are great

[Yudkowsky][12:50]

because we'd still have our late-game pile of G and we would build weapons

[Christiano][12:50]

no, I think a human in a big animal body, with brain adapted to operate that body instead of our own, would beat a big animal straightforwardly
without using tools

[Yudkowsky][12:51]

this is a strange viewpoint and I do wonder whether it is a crux of your view

[Christiano][12:51]

this feels to me like it's more on the "eliezer vs paul disagreement about the nature of AI"
rather than "eliezer vs paul on civilizational inadequacy and continuity", but enough changes
on "nature of AI" would switch my view on the other question

[Yudkowsky][12:51]

like, ceteris paribus maybe a human in an elephant's body beats an elephant after a burn-in
practice period? because we'd have a strict intelligence advantage?

[Christiano][12:52]

practice may or may not be enough

but if you port over the excellent human brain to the elephant body, then run evolution for a
brief burn-in period to get all the kinks sorted out?

elephants are pretty close to humans so it's less brutal than for some other animals (and
also are elephants the best example w.r.t. the possibility of direct conflict?) but I totally
expect us to win

[Yudkowsky][12:53]

I unfortunately need to go do other things in advance of an upcoming call, but I feel like
disagreeing about the past is proving noticeably more interesting, confusing, and perhaps
productive, than disagreeing about the future

[Christiano][12:53]

actually probably I just think practice is enough

I think humans have way more dexterity, better locomotion, better navigation, better motion
planning...

some of that is having bodies optimized for those things (esp. dexterity), but I also think
most animals just don't have the brains for it, with elephants being one of the closest calls

I'm a little bit scared of talking to zoologists or whoever the relevant experts are on this question, because I've talked to bird people a little bit and they often have very strong "humans aren't special, animals are super cool" instincts even in cases where that take is totally and obviously insane. But if we found someone reasonable in that area I'd be interested to get their take on this.

I think this is pretty important for the particular claim "Is AGI like other kinds of ML?"; that definitely doesn't persuade me to be into fast takeoff on its own though it would be a clear way the world is more Eliezer-like than Paul-like

I think I do further predict that people who know things about animal intelligence, and don't seem to have identifiably crazy views about any adjacent questions that indicate a weird pro-animal bias, will say that human brains are a lot better than other animal brains for dexterity/locomotion/similar physical tasks (and that the comparison isn't that close for e.g. comparing humans vs big cats).

Incidentally, seems like DM folks did the same thing this year, presumably publishing now because they got scooped. Looks like they probably have a better algorithm but used harder environments instead of Atari. (They also evaluate the algorithm SPR+MuZero I mentioned which indeed gets one factor of 2x improvement over MuZero alone, roughly as you'd guess): <https://arxiv.org/pdf/2111.01587.pdf>

[Barnes][13:45]

My DM friend says they tried it before they were focused on data efficiency and it didn't help in that regime, sounds like they ignored it for a while after that

[Christiano: ]

[Christiano][13:48]

Overall the situation feels really boring to me. Not sure if DM having a highly similar unpublished result is more likely on my view than Eliezer's (and initially ignoring the method because they weren't focused on sample-efficiency), but at any rate I think it's not anywhere close to falsifying my view.

18. Follow-ups to the Christiano/Yudkowsky conversation

[Karnofsky][9:39] (Nov. 5)

Going to share a point of confusion about this latest exchange.

It started with Eliezer saying this:

Thing that (if true) strikes me as... straight-up falsifying Paul's view as applied to modern-day AI, at the frontier of the most AGI-ish part of it and where Deepmind put in substantial effort on their project? EfficientZero (allegedly) learns Atari in 100,000 frames. Caveat: I'm not having an easy time figuring out how many frames MuZero would've required to achieve the same performance level. MuZero was trained on 200,000,000 frames but reached what looks like an allegedly higher high; the EfficientZero paper compares their performance to MuZero on 100,000 frames, and claims theirs is much better than MuZero given only that many frames.

So at this point, I thought Eliezer's view was something like: "EfficientZero represents a several-OM (or at least one-OM?) jump in efficiency, which should shock the hell out of Paul." The upper bound on the improvement is 2000x, so I figured he thought the corrected improvement would be some number of OMs.

But very shortly afterwards, Eliezer quotes Gwern's guess of a 4x improvement, and Paul then said:

Concretely, this is a paper that adds a few techniques to improve over MuZero in a domain that (it appears) wasn't a significant focus of MuZero. I don't know how much it improves but I can believe gwern's estimates of 4x.

I'd guess MuZero itself is a 2x improvement over the baseline from a year ago, which was maybe a 4x improvement over the algorithm from a year before that. If that's right, then no it's not mindblowing on my view to have 4x progress one year, 2x progress the next, and 4x progress the next.

Eliezer never seemed to push back on this 4x-2x-4x claim.

What I thought would happen after the 4x estimate and 4x-2x-4x claim: Eliezer would've said "Hmm, we should nail down whether we are talking about 4x-2x-4x or something more like 4x-2x-100x. If it's 4x-2x-4x, then I'll say 'never mind' re: my comment that this 'straight-up falsifies Paul's view.' At best this is just an iota of evidence or something."

Why isn't that what happened? Did Eliezer mean all along to be saying that a 4x jump on Atari sample efficiency would "straight-up falsify Paul's view?" Is a 4x jump the kind of thing Eliezer thinks is going to power a jumpy AI timeline?

[Ngo:] [Shah:]

[Yudkowsky][11:16] (Nov. 5)

This is a proper confusion and probably my fault; I also initially thought it was supposed to be 1-2 OOM and should've made it clearer that Gwern's 4x estimate was less of a direct falsification.

I'm not yet confident Gwern's estimate is correct. I just got a reply from my query to the paper's first author which reads:

Dear Eliezer: It's a good question. But due to the limits of resources and time, we haven't evaluated the sample efficiency towards different frames systematically. I think it's not a trivial question as the required time and resources are much expensive for the 200M frames setting, especially concerning the MCTS-based methods. Maybe you need about several days or longer to finish a run with GPUs in that setting. I hope my answer can help you. Thank you for your email.

I replied asking if Gwern's 3.8x estimate sounds right to them.

A 10x improvement could power what I think is a jumpy AI timeline. I'm currently trying to draft a depiction of what I think an unrealistically dignified but computationally typical end-of-world would look like if it started in 2025, and my first draft of that had it starting with a new technique published by Google Brain that was around a 10x improvement in training speeds for very large networks at the cost of higher inference costs, but which turned out to be specially applicable to online learning.

That said, I think the 10x part isn't either a key concept or particularly likely, and it's much more likely that hell breaks loose when an innovation changes some particular step of the problem from "can't realistically be done at all" to "can be done with a lot of computing power", which was what I had being the real effect of that hypothetical Google Brain innovation when applied to online learning, and I will probably rewrite to reflect that.

[Karnofsky][11:29] (Nov. 5)

That's helpful, thanks.

Re: "can't realistically be done at all" to "can be done with a lot of computing power", cpl things:

1. Do you think a 10x improvement in efficiency at some particular task could qualify as this? Could a smaller improvement?

2. I thought you were pretty into the possibility of a jump from "can't realistically be done at all" to "can be done with a *small* amount of computing power," eg some random ppl with a \$1-10mm/y budget blowing past mtpl labs with >\$1bb/y budgets. Is that wrong?

[Yudkowsky][13:44] (Nov. 5)

1 - yes and yes, my revised story for how the world ends looks like Google Brain publishing something that looks like only a 20% improvement but which is done in a way that lets it be adapted to make online learning by gradient descent "work at all" in DeepBrain's ongoing Living Zero project (not an actual name afaik)

2 - that definitely remains very much allowed in principle, but I think it's not my current mainline probability for how the world's end plays out - although I feel hesitant / caught between conflicting heuristics here.

I think I ended up much too conservative about timelines and early generalization speed because of arguing with Robin Hanson, and don't want to make a similar mistake here, but on the other hand a lot of the current interesting results have been from people spending huge compute (as wasn't the case to nearly the same degree in 2008) and if things happen on short timelines it seems reasonable to guess that the future will look that much like the present. This is very much due to cognitive limitations of the researchers rather than a basic fact about computer science, but cognitive limitations are also facts and often stable ones.

[Karnofsky][14:35] (Nov. 5)

Hm OK. I don't know what "online learning by gradient descent" means such that it doesn't work at all now (does "work at all" mean something like "work with human-ish learning efficiency")?

[Yudkowsky][15:07] (Nov. 5)

I mean, in context, it means "works for Living Zero at the performance levels where it's running around accumulating knowledge", which by hypothesis it wasn't until that point.

[Karnofsky][15:12] (Nov. 5)

Hm. I am feeling pretty fuzzy on whether your story is centrally about:

1. A <10x jump in efficiency at something important, leading pretty directly/straightforwardly to crazytown

2. A 100x ish jump in efficiency at something important, which may at first "look like" a mere <10x jump in efficiency at something else

#2 is generally how I've interpreted you and how the above sounds, but under #2 I feel like we should just have consensus that the Atari thing being 4x wouldn't be much of an update. Maybe we already do (it was a bit unclear to me from your msg)

(And I totally agree that we haven't established the Atari thing is only 4x - what I'm saying is it feels like the conversation should've paused there)

[Yudkowsky][15:13] (Nov. 5)

The Atari thing being 4x over 2 years is I think legit not an update because that's standard software improvement speed

you're correct that it should pause there

[Karnofsky][15:14] (Nov. 5)



[Yudkowsky] [15:24] (Nov. 5)

I think that my central model is something like - there's a central thing to general intelligence that starts working when you get enough pieces together and they coalesce, which is why humans went down this evolutionary gradient by a lot before other species got 10% of the way there in terms of output; and then it takes a big pile of that thing to do big things, which is why humans didn't go faster than cheetahs until extremely late in the game.

so my visualization of how the world starts to end is "gear gets added and things start to happen, maybe slowly-by-my-standards at first such that humans keep on pushing it along rather than it being self-moving, but at some point starting to cumulate pretty quickly in the same way that humans cumulated pretty quickly once they got going" rather than "dial gets turned up 50%, things happen 50% faster, every year".

[Yudkowsky][15:16] (Nov. 5, switching channels)

as a quick clarification, I agree that if this is 4x sample efficiency over 2 years then that doesn't at all challenge Paul's view

[Christiano][0:20] (Nov. 26)

FWIW, I felt like the entire discussion of EfficientZero was a concrete example of my view making a number of more concentrated predictions than Eliezer that were then almost immediately validated. In particular, consider the following 3 events:

- The quantitative effect size seems like it will turn out to be much smaller than Eliezer initially believed, much closer to being in line with previous progress.
- DeepMind had relatively similar results that got published immediately after our discussion, making it look like random people didn't pull ahead of DM after all.
- DeepMind appears not to have cared much about the metric in question, as evidenced by (i) Beth's comment above, which is basically what I said was probably going on, (ii) they barely even mention Atari sample-efficiency in their paper about similar methods.

If only 1 of these 3 things had happened, then I agree this would have been a challenge to my view that would make me update in Eliezer's direction. But that's only possible if Eliezer actually assigns a higher probability than me to ≤ 1 of these things happening, and hence a lower probability to ≥ 2 of them happening. So if we're playing a reasonable epistemic game, it seems like I need to collect some epistemic credit every time something looks boring to me.

[Yudkowsky][15:30] (Nov. 26)

I broadly agree; you win a Bayes point. I think some of this (but not all!) was due to my tripping over my own feet and sort of rushing back with what looked like a Relevant Thing without contemplating the winner's curse of exciting news, the way that paper authors tend to frame things in more exciting rather than less exciting ways, etc. But even if you set that aside, my underlying AI model said that was a thing which could happen (which is why I didn't have technically rather than sociologically triggered skepticism) and your model said it shouldn't happen, and it currently looks like it mostly didn't happen, so you win a Bayes point.

Notes that some participants may deem obvious(?) but that I state expecting wider readership:

- Just like markets are almost entirely efficient (in the sense that, even when they're not efficient, you can only make a very small fraction of the money that could be made from the entire market if you owned a time machine), even sharp and jerky progress has to look almost entirely not so fast almost all the time if the Sun isn't right in the middle of going supernova. So the notion that progress sometimes goes jerky and fast does have to be evaluated by a portfolio view over time. In worlds where progress is jerky even before the End Days, Paul wins soft steady Bayes points in most weeks and then I win back more Bayes points once every year or two.
- We still don't have a very good idea of how much longer you would need to train the previous algorithm to match the performance of the new algorithm, just an estimate by Gwern based off linearly extrapolating a graph in a paper. But, also to be clear, not knowing something is not the same as expecting it to update dramatically, and you have to integrate over the distribution you've got.
- It's fair to say, "Hey, Eliezer, if you tripped over your own feet here, but only noticed that because Paul was around to call it, maybe you're tripping over your feet at other times when Paul isn't around to check your thoughts in detail" - I don't want to minimize the Bayes point that Paul won either.

[Christiano][16:29] (Nov. 27)

Agreed that it's (i) not obvious how large the EfficientZero gain was, and in general it's not a settled question what happened, (ii) it's not that big an update, it needs to be part of a portfolio (but this is indicative of the kind of thing I'd want to put in the portfolio), (iii) it generally seems pro-social to flag potentially relevant stuff without the presumption that you are staking a lot on it.

Shah and Yudkowsky on alignment failures

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is the final discussion log in the [Late 2021 MIRI Conversations](#) sequence, featuring Rohin Shah and Eliezer Yudkowsky, with additional comments from Rob Bensinger, Nate Soares, Richard Ngo, and Jaan Tallinn.

The discussion begins with summaries and comments on Richard and Eliezer's debate. Rohin's summary has since been revised and published [in the Alignment Newsletter](#).

After this log, we'll be concluding this sequence with an [AMA](#), where we invite you to comment with questions about AI alignment, cognition, forecasting, etc. Eliezer, Richard, Paul Christiano, Nate, and Rohin will all be participating.

Color key:

[Chat by Rohin and Eliezer](#) [Other chat](#) [Emails](#) [Follow-ups](#)

19. Follow-ups to the Ngo/Yudkowsky conversation

19.1. Quotes from the public discussion

[Bensinger][9:22] (Nov. 25)

Interesting extracts from the public discussion of [Ngo and Yudkowsky on AI capability gains](#):

Eliezer:

I think some of your confusion may be that you're putting "probability theory" and "Newtonian gravity" into the same bucket. You've been raised to believe that powerful theories ought to meet certain standards, like successful bold advance experimental predictions, such as Newtonian gravity made about the existence of Neptune (quite a while after the theory was first put forth, though).

"Probability theory" also sounds like a powerful theory, and the people around you believe it, so you think you ought to be able to produce a powerful advance prediction it made; but it is for some reason hard to come up with an example like the discovery of Neptune, so you cast about a bit and think of the central limit theorem. That theorem is widely used and praised, so it's "powerful", and it wasn't invented *before* probability theory, so it's "advance", right? So we can go on putting probability theory in the same bucket as Newtonian gravity?

They're actually just very different kinds of ideas, ontologically speaking, and the standards to which we hold them are properly different ones. It seems like the sort of thing that would take a subsequence I don't have time to write, expanding beyond the underlying obvious ontological difference between validities and empirical-truths, to cover the way in which "How do we trust this, when" differs between "I have the following new empirical theory about the underlying model of gravity" and "I think that the logical notion of 'arithmetic' is a good tool to use to organize our current understanding of this little-observed phenomenon, and it appears within making the following empirical predictions..." But at least step one could be saying, "Wait, do these two kinds of ideas actually go into the same bucket at all?"

In particular it seems to me that you want properly to be asking "How do we know this empirical thing ends up looking like it's close to the abstraction?" and not "Can you show me that this abstraction is a very powerful one?" Like, imagine that instead of asking Newton about planetary movements and how we know that the particular bits of calculus he used were empirically true about the planets in particular, you instead started asking Newton for proof that calculus is a very powerful piece of mathematics worthy to predict the planets themselves - but in a way where you wanted to see some highly valuable material object that calculus had *produced*, like earlier praiseworthy achievements in alchemy. I think this would reflect confusion and a wrongly directed inquiry; you would have lost sight of the particular reasoning steps that made ontological sense, in the course of trying to figure out whether calculus was praiseworthy under the standards of praiseworthiness that you'd been previously raised to believe in as universal standards about all ideas.

Richard:

I agree that "powerful" is probably not the best term here, so I'll stop using it going forward (note, though, that I didn't use it in my previous comment, which I endorse more than my claims in the original debate).

But before I ask "How do we know this empirical thing ends up looking like it's close to the abstraction?", I need to ask "Does the abstraction even make sense?" Because you have the abstraction in your head, and I don't, and so whenever you tell me that X is a (non-advance) prediction of your theory of consequentialism, I end up in a pretty similar epistemic state as if George Soros tells me that X is a

prediction of the [theory of reflexivity](#), or if a complexity theorist tells me that X is a prediction of the [theory of self-organisation](#). The problem in those two cases is less that the abstraction is a bad fit for this specific domain, and more that the abstraction is not sufficiently well-defined (outside very special cases) to even be the type of thing that can robustly make predictions.

Perhaps another way of saying it is that they're not crisp/robust/coherent concepts (although I'm open to other terms, I don't think these ones are particularly good). And it would be useful for me to have evidence that the abstraction of consequentialism you're using is a crisper concept than Soros' theory of reflexivity or the theory of self-organisation. If you could explain the full abstraction to me, that'd be the most reliable way - but given the difficulties of doing so, my backup plan was to ask for impressive advance predictions, which are the type of evidence that I don't think Soros could come up with.

I also think that, when you talk about me being raised to hold certain standards of praiseworthiness, you're still ascribing too much modesty epistemology to me. I mainly care about novel predictions or applications insofar as they help me distinguish crisp abstractions from evocative metaphors. To me it's the same type of rationality technique as asking people to make bets, to help distinguish post-hoc confabulations from actual predictions.

Of course there's a social component to both, but that's not what I'm primarily interested in. And of course there's a strand of naive science-worship which thinks you have to follow the Rules in order to get anywhere, but I'd thank you to assume I'm at least making a more interesting error than that.

Lastly, on probability theory and Newtonian mechanics: I agree that you shouldn't question how much sense it makes to use calculus in the way that you described, but that's because the application of calculus to mechanics is so clearly-defined that it'd be very hard for the type of confusion I talked about above to sneak in. I'd put evolutionary theory halfway between them: it's partly a novel abstraction, and partly a novel empirical truth. And in this case I do think you have to be very careful in applying the core abstraction of evolution to things like cultural evolution, because it's easy to do so in a confused way.

19.2. Rohin Shah's summary and thoughts

[Shah][7:06] (Nov. 6 email)

Newsletter summaries attached, would appreciate it if Eliezer and Richard checked that I wasn't misrepresenting them. (Conversation is a lot harder to accurately summarize than blog posts or papers.)

Best,

Rohin

Planned summary for the Alignment Newsletter:

Eliezer is known for being pessimistic about our chances of averting AI catastrophe. His main argument is roughly as follows:

[Yudkowsky][9:56] (Nov. 6 email reply)

[...] Eliezer is known for being pessimistic about our chances of averting AI catastrophe. His main argument

I request that people stop describing things as my "main argument" unless I've described them that way myself. These are answers that I customized for Richard Ngo's questions. Different questions would get differently emphasized replies. "His argument in the dialogue with Richard Ngo" would be fine.

[Shah][1:53] (Nov. 8 email reply)

I request that people stop describing things as my "main argument" unless I've described them that way myself.

Fair enough. It still does seem pretty relevant to know the purpose of the argument, and I would like to state something along those lines in the summary. For example, perhaps it is:

1. One of several relatively-independent lines of argument that suggest we're doomed; cutting this argument would make almost no difference to the overall take
2. Your main argument, but with weird Richard-specific emphases that you wouldn't have necessarily included if making this argument more generally; if someone refuted the core of the argument to your satisfaction it would make a big difference to your overall take
3. Not actually an argument you think much about at all, but somehow became the topic of discussion
4. Something in between these options
5. Something else entirely

If you can't really say, then I guess I'll just say "His argument in this particular dialogue".

I'd also like to know what the main argument is (if there is a main argument rather than lots of independent lines of evidence or something else entirely); it helps me orient to the discussion, and I suspect would be useful for newsletter readers as well.

[Shah][7:06] (Nov. 6 email)

1. We are very likely going to keep improving AI capabilities until we reach AGI, at which point either the world is destroyed, or we use the AI system to take some pivotal act before some careless actor destroys the world.
2. In either case, the AI system must be producing high-impact, world-rewriting plans; such plans are "consequentialist" in that the simplest way to get them (and thus, the one we will first build) is if you are forecasting what might happen, thinking about the expected consequences, considering possible obstacles, searching for routes around the obstacles, etc. If you don't do this sort of reasoning, your plan goes off the rails very quickly; it is highly unlikely to lead to high impact. In particular, long lists of shallow heuristics (as with current deep learning systems) are unlikely to be enough to produce high-impact plans.
3. We're producing AI systems by selecting for systems that can do impressive stuff, which will eventually produce AI systems that can accomplish high-impact plans using a general underlying "consequentialist"-style reasoning process (because that's the only way to keep doing more impressive stuff). However, this selection process does *not* constrain the goals towards which those plans are aimed. In addition, most goals seem to have convergent instrumental subgoals like survival and power-seeking that would lead to extinction. This suggests that, unless we find a way to constrain the goals towards which plans are aimed, we should expect an existential catastrophe.
4. None of the methods people have suggested for avoiding this outcome seem like they actually avert this story.

[Yudkowsky][9:56] (Nov. 6 email reply)

[...] This suggests that, unless we find a way to constrain the goals towards which plans are aimed, we should expect an existential catastrophe.

I would not say we face catastrophe "unless we find a way to constrain the goals towards which plans are aimed". This is, first of all, not my ontology, second, I don't go around randomly slicing away huge sections of the solution space. Workable: "This suggests that we should expect an existential catastrophe by default."

[Shah][1:53] (Nov. 8 email reply)

I would not say we face catastrophe "unless we find a way to constrain the goals towards which plans are aimed".

Should I also change "However, this selection process does *not* constrain the goals towards which those plans are aimed", and if so what to? (Something along these lines seems crucial to the argument, but if this isn't your native ontology, then presumably you have some other thing you'd say here.)

[Shah][7:06] (Nov. 6 email)

Richard responds to this with a few distinct points:

1. It might be possible to build narrow AI systems that humans use to save the world, for example, by making AI systems that do better alignment research. Such AI systems do not seem to require the property of making long-term plans in the real world in point (3) above, and so could plausibly be safe. We might say that narrow AI systems could save the world but can't destroy it, because humans will put plans into action for the former but not the latter.
2. It might be possible to build general AI systems that only *state* plans for achieving a goal of interest that we specify, without *executing* that plan.
3. It seems possible to create consequentialist systems with constraints upon their reasoning that lead to reduced risk.
4. It also seems possible to create systems that make effective plans, but towards ends that are not about outcomes in the real world, but instead are about properties of the plans -- think for example of [corrigibility \(AN #35\)](#) or deference to a human user.
5. (Richard is also more bullish on coordinating not to use powerful and/or risky AI systems, though the debate did not discuss this much.)

Eliezer's responses:

1. This is plausible, but seems unlikely; narrow not-very-consequentialist AI (aka "long lists of shallow heuristics") will probably not scale to the point of doing alignment research better than humans.

[Yudkowsky][9:56] (Nov. 6 email reply)

[...] This is plausible, but seems unlikely; narrow not-very-consequentialist AI (aka "long lists of shallow heuristics") will probably not scale to the point of doing alignment research better than humans.

No, your summarized-Richard-1 is just not plausible. "AI systems that do better alignment research" are dangerous in virtue of the lethally powerful work they are doing, not because of some particular narrow way of doing that work. If you can do it by gradient descent then that means gradient descent got to the point of doing lethally dangerous work. Asking for safely weak systems that do world-savingly strong tasks is almost everywhere a case of asking for nonwater, and asking for AI that does alignment research is an extreme case in point.

[Shah][1:53] (Nov. 8 email reply)

No, your summarized-Richard-1 is just not plausible. "AI systems that do better alignment research" are dangerous in virtue of the lethally powerful work they are doing, not because of some particular narrow way of doing that work.

How about "AI systems that help with alignment research to a sufficient degree that it actually makes a difference are almost certainly already dangerous."?

(Fwiw, I used the word "plausible" because of this sentence from the doc: "*Definitely, <description of summarized-Richard-1> is among the more plausible advance-specified miracles we could get.*", though I guess the point was that it is still a miracle, it just also is more likely than other miracles.)

[Ngo][9:59] (Nov. 6 email reply)

Thanks Rohin! Your efforts are much appreciated.

Eliezer: when you say "No, your summarized-Richard-1 is just not plausible", do you mean the argument is implausible, or it's not a good summary of my position (which you also think is implausible)?

For my part the main thing I'd like to modify is the term "narrow AI". In general I'm talking about all systems that are not of literally world-destroying intelligence+agency. E.g. including oracle AGIs which I wouldn't call "narrow".

More generally, I don't think all AGIs are capable of destroying the world. E.g. humans are GIs. So it might be better to characterise Eliezer as talking about *some* level of general intelligence which leads to destruction, and me as talking about the things that can be done with systems that are less general or less agentic than that.

We might say that narrow AI systems could save the world but can't destroy it, because humans will put plans into action for the former but not the latter.

I don't endorse this, I think plenty of humans would be willing to use narrow AI systems to do things that could destroy the world.

systems that make effective plans, but towards ends that are not about outcomes in the real world, but instead are about properties of the plans

I'd change this to say "systems with the primary aim of producing plans with certain properties (that aren't just about outcomes in the world)"

[Yudkowsky][10:18] (Nov. 6 email reply)

Eliezer: when you say "No, your summarized-Richard-1 is just not plausible", do you mean the argument is implausible, or it's not a good summary of my position (which you also think is implausible)?

I wouldn't have presumed to state on your behalf whether it's a good summary of your position! I mean that the stated position is implausible, whether or not it was a good summary of your position.

[Shah][7:06] (Nov. 6 email)

2. This might be an improvement, but not a big one. It is the plan itself that is risky; if the AI system made a plan for a goal that wasn't the one we actually meant, and we don't understand that plan, that plan can still cause extinction. It is the *misaligned optimization that produced the plan* that is dangerous, even if there was no "agent" that specifically wanted the goal that the plan was optimized for.

3 and 4. It is certainly *possible* to do such things; the space of minds that could be designed is very large. However, it is *difficult* to do such things, as they tend to make consequentialist reasoning weaker, and on our current trajectory the first AGI that we build will probably not look like that.

[Yudkowsky][9:56] (Nov. 6 email reply)

2. This might be an improvement, but not a big one. It is the plan itself that is risky; if the AI system made a plan for a goal that wasn't the one we actually meant, and we don't understand that plan, that plan can still cause extinction. It is the *misaligned optimization that produced the plan* that is dangerous, even if there was no "agent" that specifically wanted the goal that the plan was optimized for.

No, it's not a significant improvement if the "non-executed plans" from the system are meant to do things in human hands powerful enough to save the world. They could of course be so weak as to make their human execution have no inhumanly big consequences, but this is just making the AI strategically isomorphic to a rock. The notion of there being "no 'agent' that specifically wanted the goal" seems confused to me as well; this is not something I'd ever say as a restatement of one of my own opinions. I'd shrug and tell someone to taboo the word 'agent' and would try to talk without using the word if they'd gotten hung up on that point.

[Shah][7:06] (Nov. 6 email)

Planned opinion:

I first want to note my violent agreement with the notion that a major scary thing is “consequentialist reasoning”, and that high-impact plans require such reasoning, and that we will end up building AI systems that produce high-impact plans. Nonetheless, I am still optimistic about AI safety relative to Eliezer, which I suspect comes down to three main disagreements:

1. There are many approaches that don't solve the problem, but do increase the level of intelligence required before the problem leads to extinction. Examples include Richard's points 1-4 above. For example, if we build a system that states plans without executing them, then for the plans to cause extinction they need to be complicated enough that the humans executing those plans don't realize that they are leading to an outcome that was not what they wanted. It seems non-trivially probable to me that such approaches are sufficient to prevent extinction up to the level of AI intelligence needed before we can execute a pivotal act.
2. The consequentialist reasoning is only scary to the extent that it is “aimed” at a bad goal. It seems non-trivially probable to me that it will be “aimed” at a goal sufficiently good to not lead to existential catastrophe, without putting in much alignment effort.
3. I do expect some coordination to not do the most risky things.

I wish the debate had focused more on the claim that narrow AI can't e.g. do better alignment research, as it seems like a major crux. (For example, I think that sort of intuition drives my disagreement #1.) I expect AI progress looks a lot like “the heuristics get less and less shallow in a gradual / smooth / continuous manner” which eventually leads to the sorts of plans Eliezer calls “consequentialist”, whereas I think Eliezer expects a sharper qualitative change between “lots of heuristics” and that-which-implements-consequentialist-planning.

20. November 6 conversation

20.1. Concrete plans, and AI-mediated transparency

[Yudkowsky][13:22]

So I have a general thesis about a failure mode here which is that, the moment you try to sketch any concrete plan or events which correspond to the abstract descriptions, it is much more obviously wrong, and that is why the descriptions stay so abstract in the mouths of everybody who sounds more optimistic than I am.

This may, perhaps, be confounded by the phenomenon where I am one of the last living descendants of the lineage that ever knew how to say anything concrete at all. Richard Feynman - or so I would now say in retrospect - is noticing concreteness dying out of the world, and being worried about that, at the point where he goes to a college and hears a professor talking about "essential objects" in class, and Feynman asks "Is a brick an essential object?" - meaning to work up to the notion of the inside of a brick, which can't be observed because breaking a brick in half just gives you two new exterior surfaces - and everybody in the classroom has a different notion of what it would mean for a brick to be an essential object.

Richard Feynman knew to try plugging in bricks as a special case, but the people in the classroom didn't, and I think the mental motion has died out of the world even further since Feynman wrote about it. The loss has spread to STEM as well. Though if you don't read old books and papers and contrast them to new books and papers, you wouldn't see it, and maybe most of the people who'll eventually read this will have no idea what I'm talking about because they've never seen it any other way...

I have a thesis about how optimism over AGI works. It goes like this: People use really abstract descriptions and never imagine anything sufficiently concrete, and this lets the abstract properties waver around ambiguously and inconsistently to give the desired final conclusions of the argument. So MIRI is the only voice that gives concrete examples and also by far the most pessimistic voice; if you go around fully specifying things, you can see that what gives you a good property in one place gives you a bad property someplace else, you see that you can't get all the properties you want simultaneously. Talk about a superintelligence building nanomachinery, talk concretely about megabytes of instructions going to small manipulators that repeat to lay trillions of atoms in place, and this shows you a lot of useful visible power paired with such unpleasantly visible properties as "no human could possibly check what all those instructions were supposed to do".

Abstract descriptions, on the other hand, can waver as much as they need to between what's desirable in one dimension and undesirable in another. Talk about "an AGI that just helps humans instead of replacing them" and never say exactly what this AGI is supposed to do, and this can be so much more optimistic so long as it never becomes too unfortunately concrete.

When somebody asks you "how powerful is it?" you can momentarily imagine - without writing it down - that the AGI is helping people by giving them the full recipes for protein factories that build second-stage nanotech and the instructions to feed those factories, and reply, "Oh, super powerful! More than powerful enough to flip the gameboard!" Then when somebody asks how safe it is, you can momentarily imagine that it's just giving a human mathematician a hint about proving a theorem, and say, "Oh, super duper safe, for sure, it's just helping people!"

Or maybe you don't even go through the stage of momentarily imagining the nanotech and the hint, maybe you just navigate straight in the realm of abstractions from the impossibly vague wordage of "just help humans" to the reassuring and also extremely vague "help them lots, super powerful, very safe tho".

[...] I wish the debate had focused more on the claim that narrow AI can't e.g. do better alignment research, as it seems like a major crux. (For example, I think that sort of intuition drives my disagreement #1.) I expect AI progress looks a lot like "the heuristics get less and less shallow in a gradual / smooth / continuous manner" which eventually leads to the sorts of plans Eliezer calls "consequentialist", whereas I think Eliezer expects a sharper qualitative change between "lots of heuristics" and that-which-implements-consequentialist-planning.

It is in this spirit that I now ask, "What the hell could it look like concretely for a safely narrow AI to help with alignment research?"

Or if you think that a left-handed wibble planner can totally make useful plans that are very safe because it's all leftish and wibbly: can you please give an example of a *plan to do what?*

And what I expect is for minds to bounce off that problem as they first try to visualize "Well, a plan to give mathematicians hints for proving theorems... oh, Eliezer will just say that's not useful enough to flip the gameboard... well, plans for building nanotech... Eliezer will just say that's not safe... darn it, this whole concreteness thing is such a conversational no-win scenario, maybe there's something abstract I can say instead".

[Shah][16:41]

It's reasonable to suspect failures to be concrete, but I don't buy that hypothesis as applied to me; I think I have sufficient personal evidence against it, despite the fact that I usually speak abstractly. I don't expect to convince you of this, nor do I particularly want to get into that sort of debate.

I'll note that I have the exact same experience of not seeing much concreteness, both of other people and myself, about stories that lead to doom. To be clear, in what I take to be the Eliezer-story, the part where the misaligned AI designs a pathogen that wipes out all humans or solves nanotech and gains tons of power or some other pivotal act seems fine.

The part that seems to lack concreteness is how we built the superintelligence and why the superintelligence was misaligned enough to lead to extinction. (Well, perhaps. I also wouldn't be surprised if you gave a concrete example and I disagreed that it would lead to extinction.)

From my perspective, the simple concrete stories about the future are wrong and the complicated concrete stories about the future don't sound plausible, whether about safety or about doom.

Nonetheless, here's an attempt at some concrete stories. It is *not* the case that I think these would be convincing to you. I do expect you to say that it won't be useful enough to flip the gameboard (or perhaps that if it could possibly flip the gameboard then it couldn't be safe), but that seems to be because you think alignment will be way more difficult than I do (in expectation), and perhaps we should get into that instead.

- Instead of having to handwrite code that does feature visualization or other methods of "naming neurons", an AI assistant can automatically inspect a neural net's weights, perform some experiments with them, and give them human-understandable "names". What a "name" is depends on the system being analyzed, but you could imagine that sometimes it's short memorable phrases (e.g. for the later layers of a language model), or pictures of central concepts (e.g. for image classifiers), or paragraphs describing the concept (e.g. for novel concepts discovered by a scientist AI). Given these names, it is much easier for humans to read off "circuits" from the neural net to understand how it works.
- Like the above, except the AI assistant also reads out the circuits, and efficiently reimplements the neural network in, say, readable Python, that humans can then more easily mechanistically understand. (These two tasks could also be done by two different AI systems, instead of the same one; perhaps that would be easier / safer.)
- We have AI assistants search for inputs on which the AI system being inspected would do something that humans would rate as bad. (We can choose any not-horribly-unnatural rating scheme we want that humans can understand, e.g. "don't say something the user said not to talk about, even if it's in their best interest" can be a tenet for finetuned GPT-N if we want.) We can either train on those inputs, or use them as a test for how well our other alignment schemes have worked.

(These are all basically leveraging the fact that we could have AI systems that are really knowledgeable in the realm of "connecting neural net activations to human concepts", which seems plausible to do without being super general or consequentialist.)

There's also lots of meta stuff, like helping us with literature reviews, speeding up paper- and blog-post-writing, etc, but I doubt this is getting at what you care about

[Yudkowsky][17:09]

If we thought that helping with literature review was enough to save the world from extinction, then we should be trying to spend at least \$50M on helping with literature review right now today, and if we can't effectively spend \$50M on that, then we also can't build the dataset required to train narrow AI to do literature review. Indeed, any time somebody suggests doing something weak with AGI, my response is often "Oh how about we start on that right now using humans, then," by which question its pointlessness is revealed.

[Shah][17:11]

I mean, doesn't seem crazy to just spend \$50M on effective PAs, but in any case I agree with you that this is not the main thing to be thinking about

[Yudkowsky][17:13]

The other cases of "using narrow AI to help with alignment" via pointing an AI, or rather a loss function, at a transparency problem, seem to seamlessly blend into all of the other clever-ideas we may have for getting more insight into the giant inscrutable matrices of floating-point numbers. By this concreteness, it is revealed that we are not speaking of von-Neumann-plus-level AGIs who come over and firmly but gently set aside our paradigm of giant inscrutable matrices, and do something more alignable and transparent; rather, we are trying more tricks with loss functions to get human-language translations of the giant inscrutable matrices.

I have thought of various possibilities along these lines myself. They're on my list of things to try out when and if the EA community has the capacity to try out ML ideas in a format I could and would voluntarily access.

There's a basic reason I expect the world to die despite my being able to generate infinite clever-ideas for ML transparency, which, at the usual rate of 5% of ideas working, could get us as many as three working ideas in the impossible event that the facilities were available to test 60 of my ideas.

[Shah][17:15]

By this concreteness, it is revealed that we are not speaking of von-Neumann-plus-level AGIs who come over and firmly but gently set aside our paradigm of giant inscrutable matrices, and do something more alignable and transparent; rather, we are trying more tricks with loss functions to get human-language translations of the giant inscrutable matrices.

Agreed, but I don't see the point here

(Beyond "Rohin and Eliezer disagree on how impossible it is to align giant inscrutable matrices")

(I might dispute "tricks with loss functions", but that's nitpicky, I think)

[Yudkowsky][17:16]

It's that, if we get better transparency, we are then left looking at stronger evidence that our systems are planning to kill us, but this will not help us because we will not have anything we can do to make the system *not* plan to kill us.

[Shah][17:18]

The adversarial training case is one example where you are trying to change the system, and if you'd like I can generate more along these lines, but they aren't going to be that different and are still going to come down to what I expect you will call "playing tricks with loss functions"

[Yudkowsky][17:18]

Well, part of the point is that "AIs helping us with alignment" is, from my perspective, a classic case of something that might ambiguously between the version that concretely corresponds to "they are very smart and can give us the Textbook From The Future that we can use to easily build a robust superintelligence" (which is powerful, pivotal, unsafe, and kills you) or "they can help us with literature review" (safe, weak, unpivotal) or "we're going to try clever tricks with gradient descent and loss functions and labeled datasets to get alleged natural-language translations of some of the giant inscrutable matrices" (which was always the plan but which I expected to not be sufficient to avert ruin).

[Shah][17:19]

I'm definitely thinking of the last one, but I take your point that disambiguating between these is good

And I also think it's revealing that this is not in fact the crux of disagreement

20.2. Concrete disaster scenarios, out-of-distribution problems, and corrigibility

[Yudkowsky][17:20]

I'll note that I have the exact same experience of not seeing much concreteness, both of other people and myself, about stories that lead to doom.

I have a boundless supply of greater concrete detail for the asking, though if you ask large questions I may ask for a narrower question to avoid needing to supply 10,000 words of concrete detail.

[Shah][17:24]

I guess the main thing is to have an example of a story which includes a method for building a superintelligence (yes, I realize this is info-hazard-y, sorry, an abstract version might work) + how it becomes misaligned and what its plans become optimized for. Though as I type this out I realize that I'm likely going to disagree on the feasibility of the method for building a superintelligence?

[Yudkowsky][17:25]

I mean, I'm obviously not going to want to make any suggestions that I think could possibly work and which are not very very very obvious.

[Shah][17:25]

Yup, makes sense

[Yudkowsky][17:25]

But I don't think that's much of an issue.

I could just point to MuZero, say, and say, "Suppose something a lot like this scaled."

Do I need to explain how you would die in this case?

[Shah][17:26]

What sort of domain and what training data?

Like, do we release a robot in the real world, have it collect data, build a world model, and run MuZero with a reward for making a number in a bank account go up?

[Yudkowsky][17:28]

Supposing they're naive about it: playing all the videogames, predicting all the text and images, solving randomly generated computer puzzles, accomplishing sets of easily-labelable sensorymotor tasks using robots and webcams

[Shah][17:29]

Okay, so far I'm with you. Is there a separate deployment step, and if so, how did they finetune the agent for the deployment task? Or did it just take over the world halfway through training?

[Yudkowsky][17:29]

(though this starts to depart from the Mu Zero architecture if it has the ability to absorb knowledge via learning on more purely predictive problems)

[Shah][17:30]

(I'm okay with that, I think)

[Yudkowsky][17:32]

vaguely plausible rough scenario: there was a big ongoing debate about whether or not to try letting the system trade stocks, and while the debate was going on, the researchers kept figuring out ways to make Something Zero do more with less computing power, and then it started visibly talking at people and trying to manipulate them, and there was an enormous fuss, and what happens past this point depends on whether or not you want me to try to describe a scenario in which we die with an unrealistic amount of dignity, or a realistic scenario where we die much faster

I shall assume the former.

[Shah][17:32]

Actually I think I want concreteness earlier

[Yudkowsky][17:32]

Okay. I await your further query.

[Shah][17:32]

it started visibly talking at people and trying to manipulate them

What caused this?

Was it manipulating people in order to make e.g. sensory stuff easier to predict?

[Yudkowsky][17:36]

Cumulative lifelong learning from playing videogames took its planning abilities over a threshold; cumulative solving of computer games and multimodal real-world tasks took its internal mechanisms for unifying knowledge and making them coherent over a threshold; and it gained sufficient compressive understanding of the data it had implicitly learned by reading through hundreds of terabytes of Common Crawl, not so much the semantic knowledge contained in those pages, but the associated implicit knowledge of the Things That Generate Text (aka humans).

These combined to form an imaginative understanding that some of its real-world problems were occurring in interactions with the Things That Generate Text, and it started making plans which took that into account and tried to have effects on the Things That Generate Text in order to affect the further processes of its problems.

Or perhaps somebody trained it to write code in partnership with programmers and it already had experience coworking with and manipulating humans.

[Shah][17:39]

Checking understanding: At this point it is able to make novel plans that involve applying knowledge about humans and their role in the data-generating process in order to create a plan that leads to more reward for the real-world problems?

(Which we call "manipulating humans")

[Yudkowsky][17:40]

Yes, much as it might have gained earlier experience with making novel Starcraft plans that involved "applying knowledge about humans and their role in the data-generating process in order to create a plan that leads to more reward", if it was trained on playing Starcraft against humans at any point, or even needed to make sense of how other agents had played Starcraft

This in turn can be seen as a direct outgrowth and isomorphism of making novel plans for playing Super Mario Brothers which involve understanding Goombas and their role in the screen-generating process

except obviously that the Goombas are much less complicated and not themselves agents

[Shah][17:41]

Yup, makes sense. Not sure I totally agree that this sort of thing is likely to happen as quickly as it sounds like you believe but I'm happy to roll with it; I do think it will happen eventually

So doesn't seem particularly cruxy

I can see how this leads to existential catastrophe, if you don't expect the programmers to be worried at this early manipulation warning sign. (This is potentially cruxy for p(doom), but doesn't feel like the main action.)

[Yudkowsky][17:46]

On my mainline, where this is all happening at Deepmind, I do expect at least one person in the company has ever read anything I've written. I am not sure if Demis understands he is looking straight at death, but I am willing to suppose for the sake of discussion that he does understand this - which isn't ruled out by my actual knowledge - and talk about how we all die from there.

The very brief tl;dr is that they know they're looking at a warning sign but they cannot ~~fix the warning sign~~ actually fix the real underlying problem that the warning sign is about, and AGI is getting easier for other people to develop too.

[Shah][17:46]

I assume this is primarily about social dynamics + the ability to patch things such that things look fixed?

Yeah, makes sense

I assume the "real underlying problem" is somehow not the fact that the task you were training your AI system to do was not what you actually wanted it to do?

[Yudkowsky][17:48]

It's about the unavailability of any actual fix and the technology continuing to get easier. Even if Deepmind understands that surface patches are lethal and understands that the easy ways of hammering down the warning signs are just eliminating the visibility rather than the underlying problems, there is nothing they can do about that except wait for somebody else to destroy the world instead.

I do not know of any pivotal task you could possibly train an AI system to do using tons of correctly labeled data. This is part of why we're all dead.

[Shah][17:50]

Yeah, I think if I adopted (my understanding of) your beliefs about alignment difficulty, and there wasn't already a non-racing scheme set in place, seems like we're in trouble

[Yudkowsky][17:50]

Like, "the real underlying problem is the fact that the task you were training your AI system to do was not what you actually wanted it to do" is one way of looking at one of the several problems that are truly fundamental, but this has no remedy that I know of, besides training your AI to do something small enough to be unpivotal.

[Shah][17:51][17:52]

I don't actually know the response you'd have to "why not just do value alignment?" I can name several guesses

- [Fragility of value](#)
- Not sufficiently concrete
- Can't give correct labels for human values

[Yudkowsky][17:52][17:52]

To be concrete, you can't ask the AGI to build one billion nanosystems, label all the samples that wiped out humanity as bad, and apply gradient descent updates

In part, you can't do that because one billion samples will get you one billion lethal systems, but even if that wasn't true, you still couldn't do it.

[Shah][17:53]

even if that wasn't true, you still couldn't do it.

Why not? [Nearest unblocked strategy?](#)

[Yudkowsky][17:53]

...no, because the first supposed output for training generated by the system at superintelligent levels kills everyone and there is nobody left to label the data.

[Shah][17:54]

Oh, I thought you were asking me to imagine away that effect with your second sentence

In fact, I still don't understand what it was supposed to mean

(Specifically this one:

In part, you can't do that because one billion samples will get you one billion lethal systems, but even if that wasn't true, you still couldn't do it.

)

[Yudkowsky][17:55]

there's a separate problem where you can't apply reinforcement learning when there's no good examples, even assuming you live to label them

and, of course, yet another form of problem where you can't tell the difference between good and bad samples

[Shah][17:56]

Okay, makes sense

Let me think a bit

[Yudkowsky][18:00]

and lest anyone start thinking that was an exhaustive list of fundamental problems, note the absence of, for example, "applying lots of optimization using an outer loss function doesn't necessarily get you something with a faithful internal cognitive representation of that loss function" aka "natural selection applied a ton of optimization power to humans using a very strict very simple criterion of 'inclusive genetic fitness' and got out things with no explicit representation of or desire towards 'inclusive genetic fitness' because that's what happens when you hill-climb and take wins in the order a simple search process through cognitive engines encounters those wins"

[Shah][18:02]

(Agreed that is another major fundamental problem, in the sense of something that could go wrong, as opposed to something that almost certainly goes wrong)

I am still curious about the "why not value alignment" question, where to expand, it's something like "let's get a wide range of situations and train the agent with gradient descent to do what a human would say is the right thing to do". (We might also call this "imitation"; maybe "value alignment" isn't the right term, I was thinking of it as trying to align the planning with "human values".)

My own answer is that we shouldn't expect this to generalize to nanosystems, but that's again much more of a "there's not great reason to expect this to go right, but also not great reason to go wrong either".

(This is a place where I would be particularly interested in concreteness, i.e. what does the AI system do in these cases, and how does that almost-necessarily follow from the way it was trained?)

[Yudkowsky][18:05]

what's an example element from the "wide range of situations" and what is the human labeling?

(I could make something up and let you object, but it seems maybe faster to ask you to make something up)

[Shah][18:09]

Uh, let's say that the AI system is being trained to act well on the Internet, and it's shown some tweet / email / message that a user might have seen, and asked to reply to the tweet / email / message. User says whether the replies are good or not (perhaps via comparisons, a la [Deep RL from Human Preferences](#))

If I were not making it up on the spot, it would be more varied than that, but would not include "building nanosystems"

[Yudkowsky][18:10]

And presumably, in this example, the AI system is not smart enough that exposing humans to text it generates is already a world-wrecking threat if the AI is hostile?

i.e., does not just hack the humans

[Shah][18:10]

Yeah, let's assume that for the moment

[Yudkowsky][18:11]

so what you want to do is train on 'weak-safe' domains where the AI isn't smart enough to do damage, and the humans can label the data pretty well because the AI isn't smart enough to fool them

[Shah][18:11]

"want to do" is putting it a bit strongly. This is more like a scenario I can't prove is unsafe, but do not strongly believe is safe

[Yudkowsky][18:12]

but the domains where the AI can execute a world-saving pivotal act are out-of-distribution for those domains. *extremely* out-of-distribution. *fundamentally* out-of-distribution. the AI's own thought processes are out-of-distribution for any inscrutable matrices that were learned to influence those thought processes in a corrigible direction.

it's not like trying to generalize experience from playing Super Mario Bros to Metroid.

[Shah][18:13]

Definitely, but my reaction to this is "okay, no particular reason for it to be safe" -- but also not huge reason for it to be unsafe. Like, it would not hugely shock me if what-we-want is sufficiently "natural" that the AI system picks up on the right thing from the 'weak-safe' domains alone

[Yudkowsky][18:14]

you have this whole big collection of possible AI-domain tuples that are powerful-dangerous and they have properties that aren't in *any* of the weak-safe training situations, that are moving along third dimensions where all the weak-safe training examples were flat

now, just because something is out-of-distribution, doesn't mean that nothing can ever generalize there

[Shah][18:15]

I mean, you correctly would not accept this argument if I said that by training blue-car-driving robots solely on blue cars I am ensuring they would be bad on red-car-driving

[Yudkowsky][18:15]

humans generalize from the savannah to the vacuum

so the actual problem is that I expect the optimization to generalize and the corrigibility to fail

[Shah][18:15]

^Right, that

I am not clear on why you expect this so strongly

Maybe you think generalization is extremely rare and optimization is a special case because of how it is so useful for basically everything?

[Yudkowsky][18:16]

no

did you read the section of my dialogue with Richard Ngo where I tried to explain [why corrigibility is anti-natural](#), or where Nate tried to give the [example](#) of why planning to get a laser from point A to point B without being scattered by fog is the sort of thing that also naturally says to prevent humans from filling the room with fog?

[Shah][18:19]

Ah, right, I should have predicted that. (Yes, I did read it.)

[Yudkowsky][18:19]

or for that matter, am I correct in remembering that these sections existed

k

so, do you need more concrete details about some part of that?

a bunch of the reason why I suspect that corrigibility is anti-natural is from trying to work particular problems there in MIRI's earlier history, and not finding anything that wasn't contrary to [coherence](#) the overlap in the shards of inner optimization that, when ground into existence by the outer optimization loop, coherently mix to form the part of cognition that generalizes to do powerful things; and nobody else finding it either, etc.

[Shah][18:22]

I think I disagreed with that part more directly, in that it seemed like in those sections the corrigibility was assumed to be imposed "from the outside" on top of a system with a goal, rather than having a goal that was corrigible. (I also had a similar reaction to the 2015 [Corrigibility](#) paper.)

So, for example, it seems to me like [CIRL](#) is an example of an objective that can be maximized in which the agent is corrigible-in-a-certain-sense. I agree that due to [updated deference](#) it will eventually stop seeking information from the human / be subject to corrections by the human. I don't see why, at that point, it wouldn't have just learned to do what the humans actually want it to do.

(There are objections like misspecification of the reward prior, or misspecification of the $P(\text{behavior} \mid \text{reward})$, but those feel like different concerns to the ones you're describing.)

[Yudkowsky][18:25]

a thing that MIRI tried and failed to do was find a sensible generalization of expected utility which could contain a generalized utility function that would look like an AI that let itself be shut down, without trying to force you to shut it down

and various workshop attendees not employed by MIRI, etc

[Shah][18:26]

I do agree that a CIRL agent would not let you shut it down

And this is something that should maybe give you pause, and be a lot more careful about potential misspecification problems

[Yudkowsky][18:27]

if you could give a perfectly specified prior such that the result of updating on lots of observations would be a representation of the utility function that [CEV](#) outputs, and you could perfectly [inner-align](#) an optimizer to do that thing in a way that scaled to arbitrary levels of cognitive power, then you'd be home free, sure.

[Shah][18:28]

I'm not trying to claim this is a solution. I'm more trying to point at a reason why I am not convinced that corrigibility is anti-natural.

[Yudkowsky][18:28]

the reason CIRL doesn't get off the ground is that there isn't any known, and isn't going to be any known, prior over (observation|'true' utility function) such that an AI which updates on lots of observations ends up with our true desired utility function.

if you can do that, the AI *doesn't need to be corrigible*

that's why it's not a counterexample to corrigibility being anti-natural

the AI just boomfs to superintelligence, observes all the things, and does all the goodness

it doesn't listen to you say no and won't let you shut it down, but by hypothesis this is fine because it got the true utility function yay

[Shah][18:31]

In the world where it doesn't immediately start out as a superintelligence, it spends a lot of time trying to figure out what you want, asking you what you prefer it does, making sure to focus on the highest-EV questions, being very careful around any irreversible actions, etc

[Yudkowsky][18:31]

and making itself smarter as fast as possible

[Shah][18:32]

Yup, that too

[Yudkowsky][18:32]

I'd do that stuff too if I was waking up in an alien world

and, with all due respect to myself, *I am not corrigible*

[Shah][18:33]

You'd do that stuff because you'd want to make sure you don't accidentally get killed by the aliens; a CIRL agent does it because it "wants to help the human"

[Yudkowsky][18:34]

no, a CIRL agent does it because it wants to implement the True Utility Function, which it may, early on, suspect to consist of helping* humans, and maybe to have some overlap (relative to its currently reachable short-term outcome sets, though these are of vanishingly small relative utility under the True Utility Function) with what some humans desire some of the time

(*) 'help' may not be help

separately it asks a lot of questions because the things humans do are evidence about the True Utility Function

[Shah][18:35]

I agree this is also an accurate description of CIRL

A more accurate description, even

Wait why is it vanishingly small relative utility? Is the assumption that the True Utility Function doesn't care much about humans? Or was there something going on with short vs. long time horizons that I didn't catch

[Yudkowsky][18:39]

in the short term, a weak CIRL tries to grab the hand of a human about to fall off a cliff, because its TUF probably does prefer the human who didn't fall off the cliff, if it has only exactly those two options, and this is the sort of thing it would learn was probably true about the TUF early on, given the obvious ways of trying to produce a CIRL-ish thing via gradient descent

humans eat healthy in the ancestral environment when ice cream doesn't exist as an option

in the long run, the things the CIRL agent wants do *not* overlap with anything humans find more desirable than paperclips (because there is no known scheme that takes in a bunch of observations, updates a prior, and outputs a utility function whose achievable maximum is galaxies living happily forever after)

and plausible TUF schemes are going to notice that grabbing the hand of a current human is a vanishing fraction of all value eventually at stake

[Shah][18:42]

Okay, cool, short vs. long time horizons

Makes sense

[Yudkowsky][18:42]

right, a weak but sufficiently reflective CIRL agent will notice an alignment of short-term interests with humans but deduce misalignment of long-term interests

though I should maybe call it CIRL* to denote the extremely probable case that the limit of its updating on observation does not in fact converge to CEV's output

[Soares][18:43]

(Attempted rephrasing of a point I read Eliezer as making upstream, in hopes that a rephrasing makes it click for Rohin:)

Corrigibility isn't for bug-free CIRL agents with a prior that actually dials in on goodness given enough observation; if you have one of those you can just run it and call it a day. Rather, corrigibility is for surviving your civilization's inability to do the job right on the first try.

CIRL doesn't have this property; it instead amounts to the assertion "if you are optimizing with respect to a distribution on utility functions that dials in on goodness given enough observation then that gets you just about as much good as optimizing goodness"; this is somewhat tangential to corrigibility.

[Yudkowsky: +1]

[Yudkowsky][18:44]

and you should maybe update on how, even though somebody thought CIRL was going to be more corrigible, in fact it made *absolutely zero progress on the real problem*

[Ngo: ]

the notion of having an uncertain utility function that you update from observation is coherent and doesn't yield circular preferences, running in circles, incoherent betting, etc.

so, of course, it is antithetical in its intrinsic nature to corrigibility

[Shah][18:47]

I guess I am not sure that I agree that this is the purpose of corrigibility-as-I-see-it. The point of corrigibility-as-I-see-it is that you don't have to specify the object-level outcomes that your AI system must produce, and instead you can specify the meta-level processes by which your AI system should come to know what the object-level outcomes to optimize for are

(At CHAI we had taken to talking about corrigibility_MIRI and corrigibility_Paul as completely separate concepts and I have clearly fallen out of that good habit)

[Yudkowsky][18:48]

speaking as the person who invented the concept, asked for name submissions for it, and selected 'corrigibility' as the winning submission, that is absolutely not how I intended the word to be used

and I think that the thing I was actually trying to talk about is important and I would like to retain a word that talks about it

'corrigibility' is meant to refer to the sort of putative hypothetical motivational properties that prevent a system from wanting to kill you after you didn't build it exactly right

low impact, mild optimization, shutdownability, abortable planning, behaviorism, conservatism, etc. (note: some of these may be less antinatural than others)

[Shah][18:51]

Cool. Sorry for the miscommunication, I think we should probably backtrack to here

so the actual problem is that I expect the optimization to generalize and the corrigibility to fail

and restart.

Though possibly I should go to bed, it is quite late here and there was definitely a time at which I would not have confused corrigibility_MIRI with corrigibility_Paul, and I am a bit worried at my completely having missed that this time

[Yudkowsky][18:51]

the thing you just said, interpreted literally, is what I would call simply "going meta" but my guess is you have a more specific metanness in mind

...does Paul use "corrigibility" to mean "going meta"? I don't think I've seen Paul doing that.

[Shah][18:54]

Not exactly "going meta", no (and I don't think I exactly mean that either). But I definitely infer a different concept from

<https://www.alignmentforum.org/posts/fkLYhTQteAu5SinAc/corrigibility>

than the one you're describing here. It is definitely possible that this comes from me misunderstanding Paul; I have done so many times

[Yudkowsky][18:55]

That looks to me like Paul used 'corrigibility' around the same way I meant it, if I'm not just reading my own face into those clouds. maybe you picked up on the exciting metanness of it and thought 'corrigibility' was talking about the metanness part? ☺

but I also want to create an affordance for you to go to bed

hopefully this last conversation combined with previous dialogues has created any sense of why I worry that corrigibility is anti-natural and hence that "on the first try at doing it, the optimization generalizes from the weak-safe domains to the strong-lethal domains, but the corrigibility doesn't"

so I would then ask you what part of this you were skeptical about

as a place to pick up when you come back from the realms of Morpheus

[Shah][18:58]

Yup, sounds good. Talk to you tomorrow!

21. November 7 conversation

21.1. Corrigibility, value learning, and pessimism

[Shah][3:23]

Quick summary of discussion so far (in which I ascribe views to Eliezer, for the sake of checking understanding, omitting for brevity the parts about how these are facts about my beliefs about Eliezer's beliefs and not Eliezer's beliefs themselves):

- Some discussion of "how to use non-world-optimizing AIs to help with AI alignment", which are mostly in the category "clever tricks with gradient descent and loss functions and labeled datasets" rather than "textbook from the future". Rohin thinks these help significantly (and that "significant help" = "reduced x-risk"). Eliezer thinks that whatever help they provide is not sufficient to cross the line from "we need a miracle" to "we have a plan that has non-trivial probability of success without miracles". The crux here seems to be alignment difficulty.
- Some discussion of how doom plays out. I agree with Eliezer that if the AI is catastrophic by default, and we don't have a technique that stops the AI from being catastrophic by default, and we don't already have some global coordination scheme in place, then bad things happen. Cruxes seem to be alignment difficulty and the plausibility of a global coordination scheme, of which alignment difficulty seems like the bigger one.
- On alignment difficulty, an example scenario is "train on human judgments about what the right thing to do is on a variety of weak-safe domains, and hope for generalization to potentially-lethal domains". Rohin views this as neither confidently safe nor confidently unsafe. Eliezer views this as confidently unsafe, because he strongly expects the optimization to generalize while the corrigibility doesn't, because corrigibility is anti-natural.

(Incidentally, "optimization generalizes but corrigibility doesn't" is an example of the sort of thing I wish were more concrete, if you happen to be able to do that)

My current take on "corrigibility":

- Prior to this discussion, in my head there was corrigibility_A and corrigibility_B. Corrigibility_A, which I associated with MIRI, was about imposing a constraint "from the outside". Given an AI system, it is a method of modifying that AI system to (say) allow you to shut it down, by performing some sort of operation on its goal. Corrigibility_B, which I associated with Paul, was about building an AI system which would have particular nice behaviors like learning about the user's preferences, accepting corrections about what it should do, etc.
- After this discussion, I think everyone meant corrigibility_B all along. The point of the 2015 MIRI paper was to check whether it is possible to build a version of corrigibility_B that was compatible with expected utility maximization with a not-terribly-complicated utility function; the point of this was to see whether corrigibility could be made compatible with "plans that lase".
- While I think people agree on the behaviors of corrigibility, I am not sure they agree on why we want it. Eliezer wants it for surviving failures, but maybe others want it for "dialing in on goodness". When I think about a "broad basin of corrigibility", that intuitively seems more compatible with the "dialing in on goodness" framing (but this is an aesthetic judgment that could easily be wrong).
- I don't think I meant "going meta", e.g. I wouldn't have called indirect normativity an example of corrigibility. I think I was pointing at "dialing in on goodness" vs. "specifying goodness".
- I agree CIRL doesn't help survive failures. But if you instead talk about "dialing in on goodness", CIRL does in fact do this, at least conceptually (and other alternatives don't).
- I am somewhat surprised that "how to conceptually dial in on goodness" is not something that seems useful to you. Maybe you think it is useful, but you're objecting to me calling it corrigibility, or saying we knew how to do it before CIRL?

(A lot of the above on corrigibility is new, because the distinction between surviving-failures and dialing-in-on-goodness as different use cases for very similar kinds of behaviors is new to me. Thanks for discussion that led me to making such a distinction.)

Possible avenues for future discussion, in the order of my-guess-at-usefulness:

1. Discussing anti-naturality of corrigibility. As a starting point: you say that an agent that makes plans but doesn't execute them is also dangerous, because it is the plan itself that lases, and corrigibility is antithetical to lasing. Does this mean you predict that you, or I, with suitably enhanced intelligence and/or reflectivity, would not be capable of producing a plan to help an alien civilization optimize their world, with that plan being corrigible w.r.t the aliens? (This seems like a strange and unlikely position to me, but I don't see

- how to not make this prediction under what I believe to be your beliefs. Maybe you just bite this bullet.)
2. Discussing why it is very unlikely for the AI system to generalize correctly both on optimization and values-or-goals-that-guide-the-optimization (which seems to be distinct from corrigibility). Or to put it another way, why is "alignment by default according to John Wentworth" doomed to fail?
<https://www.lesswrong.com/posts/Nwgdq6kHke5LY692J/alignment-by-default>
 3. More checking of where I am failing to pass your ITT
 4. Why is "dialing in on goodness" not a reasonable part of the solution space (to the extent you believe that)?
 5. More concreteness on how optimization generalizes but corrigibility doesn't, in the case where the AI was trained by human judgment on weak-safe domains Just to continue to state it so people don't misinterpret me: in most of the cases that we're discussing, my position is *not* that they are safe, but rather that they are not overwhelmingly likely to be unsafe.

[Ngo][3:41]

I don't understand what you mean by dialling in on goodness. Could you explain how CIRL does this better than, say, [reward modelling](#)?

[Shah][3:49]

Reward modeling does not by default (a) choose relevant questions to ask the user in order to get more information about goodness, (b) act conservatively, especially in the face of irreversible actions, while it is still uncertain about what goodness is, or (c) take actions that are known to be robustly good, while still waiting for future information that clarifies the nuances of goodness

You could certainly do something like Deep RL from Human Preferences, where the preferences are things like "I prefer you ask me relevant questions to get more information about goodness", in order to get similar behavior. In this case you are transferring desired behaviors from a human to the AI system, whereas in CIRL the behaviors "fall out of" optimization for a specific objective

In Eliezer/Nate terms, the CIRL story shows that dialing on goodness is compatible with "plans that lase", whereas reward modeling does not show this

[Ngo][4:04]

The meta-level objective that CIRL is pointing to, what makes that thing deserve the name "goodness"? Like, if I just gave an alien CIRL, and I said "this algorithm dials an AI towards a given thing", and they looked at it without any preconceptions of what the designers wanted to do, why wouldn't they say "huh, it looks like an algorithm for dialling in on some

extrapolation of the unintended consequences of people's behaviour" or something like that?

See also this part of my second discussion with Eliezer, where he brings up CIRL: [https://www.lesswrong.com/posts/7im8at9PmhbT4JHsW/ngo-and-yudkowsky-on-alignment-difficulty#3_2_Brain_functions_and_outcome_pumps] He was emphasising that CIRL, and most other proposals for alignment algorithms, just shuffle the problematic consequentialism from the original place to a less visible place. I didn't engage much with this argument because I mostly agree with it.

[Yudkowsky: +1]

[Shah][5:28]

I think you are misunderstanding my point. I am not claiming that we know how to implement CIRL such that it produces good outcomes; I agree this depends a ton on having a sufficiently good $P(\text{obs} \mid \text{reward})$. Similarly, if you gave CIRL to aliens, whether or not they say it is about getting some extrapolation of unintended consequences depends on exactly what $P(\text{obs} \mid \text{reward})$ you ended up using. There is some not-too-complicated $P(\text{obs} \mid \text{reward})$ such that you do end up getting to "goodness", or something sufficiently close that it is not an existential catastrophe; I do not claim we know what it is.

I am claiming that behaviors like (a), (b) and (c) above are compatible with expected utility theory, and thus compatible with "plans that lase". This is demonstrated by CIRL. It is not demonstrated by reward modeling, see e.g. [these three papers](#) for problems that arise (which make it so that it is working at cross purposes with itself and seems incompatible with "plans that lase"). (I'm most confident in the first supporting my point, it's been a long time since I read them so I might be wrong about the others.) To my knowledge, similar problems don't arise with CIRL (and they shouldn't, because it is a nice integrated Bayesian agent doing expected utility theory).

I could imagine an objection that $P(\text{obs} \mid \text{reward})$, while not as complicated as "the utility function that rationalizes a twitching robot", is still too complicated to really show compatibility with plans-that-lase, but pointing out that $P(\text{obs} \mid \text{reward})$ could be misspecified doesn't seem particularly relevant to whether behaviors (a), (b) and (c) are compatible with plans-that-lase.

Re: shuffling around the problematic consequentialism: it is not my main plan to avoid consequentialism in the sense of plans-that-lase. I broadly agree with Eliezer that you need consequentialism to do high-impact stuff. My plan is for the consequentialism to be aimed at good ends. So I agree that there is still consequentialism in CIRL, and I don't see this as a damning point; when I talk about "dialing in to goodness", I am thinking of aiming the consequentialism at goodness, not getting rid of consequentialism.

(You can still do things like try to be domain-specific rather than domain-general; I don't mean to completely exclude such approaches. They do seem to give additional safety. But the mainline story is that the consequentialism / optimization is directed at what we want rather than something else.)

[Ngo][6:21]

If you don't know how to implement CIRL in such a way that it actually aims at goodness, then you don't have an algorithm with properties a, b and c above.

Or, to put it another way: suppose I replace the word "goodness" with "winningness". Now I can describe AlphaStar as follows:

- it choose relevant questions to ask (read: scouts to send) in order to get more information about winningness
- it acts conservatively while it is still uncertain about what winningness is
- it take actions that are known to be robustly good winningish, while still waiting for future information that clarifies the nuances of winningness

Now, you might say that the difference is that CIRL implements uncertainty over possible utility functions, not possible empirical beliefs. But this is just a semantic difference which shuffles the problem around without changing anything substantial. E.g. it's exactly equivalent if we think of CIRL as an agent with a fixed (known) utility function, which just has uncertainty about some empirical parameter related to the humans it interacts with.

[Yudkowsky: +1]

[Soares][6:55]

[...] it take actions that are known to be robustly good, while still waiting for future information that clarifies the nuances of winningness

(typo: "known to be robustly good" -> "known to be robustly winningish" :-p)

[Ngo: ]

Some quick reactions, some from me and some from my model of Eliezer:

Eliezer thinks that whatever help they provide is not sufficient [...] The crux here seems to be alignment difficulty.

I'd be more hesitant to declare the crux "alignment difficulty". My understanding of Eliezer's position on your "use AI to help with alignment" proposals (which focus on things like using AI to make

paradigmatic AI systems more transparent) is "that was always the plan, and it doesn't address the sort of problems I'm worried about". Maybe you understand the problems Eliezer's worried about, and believe them not to be very difficult to overcome, thus putting the crux somewhere like "alignment difficulty", but I'm not convinced.

I'd update towards your crux-hypothesis if you provided a good-according-to-Eliezer summary of what other problems Eliezer sees and the reasons-according-to-Eliezer that "AI make our tensors more transparent" doesn't much address them.

Corrigibility_A [...] Corrigibility_B [...]

Of the two Corrigibility_B does sound a little closer to my concept, though neither of your descriptions cause me to be confident that communication has occurred. Throwing some checksums out there:

- There are three reasons a young weak AI system might accept your corrections. It could be corrigible, or it could be incorrigibly pursuing goodness, or it could be incorrigibly pursuing some other goal while calculating that accepting this correction is better according to its current goals than risking a shutdown.
- One way you can tell that CIRL is not corrigible is that it does not accept corrections when old and strong.
- There's an intuitive notion of "you're here to help us implement a messy and fragile concept not yet clearly known to us; work with us here?" that makes sense to humans, that includes as a side effect things like "don't scan my brain and then disregard my objections; there could be flaws in how you're inferring my preferences from my objections; it's actually quite important that you be cautious and accept brain surgery even in cases where your updated model says we're about to make a big mistake according to our own preferences".

The point of the 2015 MIRI paper was to check whether it is possible to build a version of corrigibility_B that was compatible with expected utility maximization with a not-terribly-complicated utility function; the point of this was to see whether corrigibility could be made compatible with "plans that lase".

More like:

- Corrigibility seems, at least on the surface, to be in tension with the simple and useful patterns of optimization that tend to be spotlit by demands for cross-domain success, similar to how acting like two oranges are worth one apple and one apple is worth one orange is in tension with those patterns.
- In practice, this tension seems to run more than surface-deep. In particular, various attempts to reconcile the tension fail, and cause the AI to have undesirable preferences (eg, incentives to convince you to shut it down whenever its utility is suboptimal), exploitably bad beliefs (eg, willingness to bet at unreasonable odds that it won't be shut down), and/or to not be corrigible in the first place (eg, a preference for destructively uploading your mind against your

protests, at which point further protests from your coworkers are screened off by its access to that upload).

[Yudkowsky: ]

(There's an argument I occasionally see floating around these parts that goes "ok, well what if the AI is *fractally* corrigible, in the sense that instead of its cognition being oriented around pursuit of some goal, its cognition is oriented around doing what it predicts a human would do (or what a human would want it to do) in a corrigible way, at every level and step of its cognition". This is perhaps where you perceive a gap between your A-type and B-type notions, where MIRI folk tend to be more interested in reconciling the tension between corrigibility and coherence, and Paulian folk tend to place more of their chips on some such fractal notion?)

I admit I don't find much hope in the "fractally corrigible" view myself, and I'm not sure whether I could pass a proponent's ITT, but fwiw my model of the Yudkowskian rejoinder is "mindspace is deep and wide; that could plausibly be done if you had sufficient mastery of minds; you're not going to get anywhere near close to that in practice, because of the way that basic normal everyday cross-domain training will highlight patterns that you'd call orienting-cognition-around-a-goal".)

And my super-quick takes on your avenues for future discussion:

1. Discussing anti-naturality of corrigibility.

Hopefully the above helps.

2. Discussing why it is very unlikely for the AI system to generalize correctly both on optimization and values-or-goals-that-guide-the-optimization

The concept "patterns of thought that are useful for cross-domain success" is latent in the problems the AI faces, and known to have various simple mathematical shadows, and our training is more-or-less banging the AI over the head with it day in and day out. By contrast, the specific values we wish to be pursued are not latent in the problems, are known to *lack* a simple boundary, and our training is much further removed from it.

3. More checking of where I am failing to pass your ITT

+1

4. Why is "dialing in on goodness" not a reasonable part of the solution space?

It has long been the plan to say something less like "the following list comprises goodness: ..." and more like "yo we're tryin to optimize some difficult-to-name concept; help us out?". "Find a prior that, with observation of the human operators, dials in on goodness" is a fine guess at how to formalize the latter.

If we had been planning to take the former tack, and you had come in suggesting CIRL, that might have helped us switch to the latter tack, which would have been cool. In that sense, it's a fine part of the solution.

It also provides some additional formality, which is another iota of potential solution-ness, for that part of the problem.

It doesn't much address the rest of the problem, which is centered much more around "how do you point powerful cognition in any direction at all" (such as towards your chosen utility function or prior thereover).

5. More concreteness on how optimization generalizes but corrigibility doesn't, in the case where the AI was trained by human judgment on weak-safe domains

+1

[Shah][13:23]

If you don't know how to implement CIRL in such a way that it actually aims at goodness, then you don't have an algorithm with properties a, b and c above.

I want clarity on the premise here:

- Is the premise "Rohin cannot write code that when run exhibits properties a, b, and c"? If so, I totally agree, but I'm not sure what the point is. All alignment work ever until the very last step will not lead you to writing code that when run exhibits an aligned superintelligence, but this does not mean that the prior alignment work was useless.
- Is the premise "there does not exist code that (1) we would call an implementation of CIRL and (2) when run has properties a, b, and c"? If so, I think your premise is false, for the reasons given previously (I can repeat them if needed)

I imagine it is neither of the above, and you are trying to make a claim that some conclusion that I am drawing from or about CIRL is invalid, because in order for me to draw that conclusion, I need to exhibit the correct $P(\text{obs} | \text{reward})$. If so, I want to know which conclusion is invalid and why I have to exhibit the correct $P(\text{obs} | \text{reward})$ before I can reach that conclusion.

I agree that the fact that you can get properties (a), (b) and (c) are simple straightforward consequences of being Bayesian about a quantity you are uncertain about and care about, as with AlphaStar and "winningness". I don't know what you intend to imply by this -- because it also applies to other Bayesian things, it can't imply anything about alignment? I also agree the uncertainty over reward is equivalent to uncertainty over some parameter of the human (and have proved this theorem myself in the paper I wrote on the topic). I do not claim that anything in here is particularly non-obvious or clever, in case anyone thought I was making that claim.

To state it again, my claim is that behaviors like (a), (b) and (c) are consistent with "plans-that-lase", and as evidence for this claim I cite the *existence* of an expected-utility-maximizing algorithm that displays them, specifically CIRL with the correct $p(\text{obs} | \text{reward})$. I do *not* claim that I can write down the code, I am just claiming that it *exists*. If you agree with the claim but not the evidence then let's just drop the point. If you disagree with the claim then tell me why it's false. If you are unsure about the claim then point to the step in the argument you think doesn't work.

The reason I care about this claim is that it seems to me like even if you think that superintelligences only involve plans-that-lase, it seems to me like this does *not* rule out what we might call "dialing in to goodness" or "assisting the user", and thus it seems like this is a valid target for you to try to get your superintelligence to do.

I suspect that I do not agree with Eliezer about what plans-that-lase can do, but it seems like the two of us should at least agree that behaviors like (a), (b) and (c) can be exhibited in plans-that-lase, and if we don't agree on that some sort of miscommunication has happened.

Throwing some checksums out there

The checksums definitely make sense. (Technically I could name more reasons why a young AI might accept correction, such as "it's still sphexish in some areas, accepting corrections is one of those reasons", and for the third reason the AI could be calculating negative consequences for things other than shutdown, but that seems nitpicky and I don't think it means I have misunderstood you.)

I think the third one feels somewhat slippery and vague, in that I don't know exactly what it's claiming, but it clearly seems to be the same sort of thing as corrigibility. Mostly it's more like I wouldn't be surprised if the Textbook from the Future tells us that we mostly had the right concept of corrigibility, but that third checksum is not quite how they would describe it any more. I would be a lot more surprised if the Textbook says we mostly had the right concept but then says checksums 1 and 2 were misguided.

"The point of the 2015 MIRI paper was to check whether it is possible to build a version of corrigibility_B that was compatible with expected utility maximization with a not-terribly-complicated utility function; the point of this was to see whether corrigibility could be made compatible with 'plans that lase'."

More like:

- Corrigibility seems, at least on the surface, to be in tension with the simple and useful patterns of optimization that tend to be spotlit by demands for cross-domain success, similar to how as acting like an two oranges are worth one apple and one apple is worth one orange is in tension with those patterns.

- In practice, this tension seems to run more than surface-deep. In particular, various attempts to reconcile the tension fail, and cause the AI to have undesirable preferences (eg, incentives to convince you to shut it down whenever its utility is suboptimal), exploitably bad beliefs (eg, willingness to bet at unreasonable odds that it won't be shut down), and/or to not be corrigible in the first place (eg, a preference for destructively uploading your mind against your protests, at which point further protests from your coworkers are screened off by its access to that upload).

On the 2015 Corrigibility paper, is this an accurate summary: "it wasn't that we were checking whether corrigibility could be compatible with useful patterns of optimization; it was already obvious at least at a surface level that corrigibility was in tension with these patterns, and we wanted to check and/or show that this tension persisted more deeply and couldn't be easily fixed".

(My other main hypothesis is that there's an important distinction between "simple and useful patterns of optimization" (term in your message) and "plans that lase" (term in my message) but if so I don't know what it is.)

[Soares][13:52]

What we *wanted* to do was show that the apparent tension was merely superficial. We failed.

[Shah: ]

(Also, IIRC -- and it's been a long time since I checked -- the 2015 paper contains only one exploration, relating to an idea of Stuart Armstrong's. There were another host of ideas raised and shot down in that era, that didn't make it into that paper, pro'lly b/c they came afterwards.)

[Shah][13:55]

What we *wanted* to do was show that the apparent tension was merely superficial. We failed.

(That sounds like what I originally said? I'm a bit confused why you didn't just agree with my original phrasing:

The point of the 2015 MIRI paper was to check whether it is possible to build a version of corrigibility_B that was compatible with expected utility maximization with a not-terribly-complicated utility function; the point of this was to see whether corrigibility could be made compatible with "plans that lase".

)

(I'm kinda worried that there's some big distinction between "EU maximization", "plans that lase", and "simple and useful patterns of

optimization", that I'm not getting; I'm treating them as roughly equivalent at the moment when putting on my MIRI-ontology-hat.)

[Soares][14:01]

(There are a bunch of aspects of your phrasing that indicated to me a different framing, and one I find quite foreign. For instance, this talk of "building a version of corrigibility_B" strikes me as foreign, and the talk of "making it compatible with 'plans that lase'" strikes me as foreign. It's plausible to me that you, who understand your original framing, can tell that my rephrasing matches your original intent. I do not yet feel like I could emit the description you emitted without contorting my thoughts about corrigibility in foreign ways, and I'm not sure whether that's an indication that there are distinctions, important to me, that I haven't communicated.)

(I'm kinda worried that there's some big distinction between "EU maximization", "plans that lase", and "simple and useful patterns of optimization", that I'm not getting; I'm treating them as roughly equivalent at the moment when putting on my MIRI-ontology-hat.)

I, too, believe them to be basically equivalent (with the caveat that the reason for using expanded phrasings is because people have a history of misunderstanding "utility maximization" and "coherence", and so insofar as you round them all to "coherence" and then argue against some very narrow interpretation of coherence, I'm gonna protest that you're bailey-and-motting).

[Shah: ]

[Shah][14:12]

Hopefully the above helps.

I'm still interested in the question "Does this mean you predict that you, or I, with suitably enhanced intelligence and/or reflectivity, would not be capable of producing a plan to help an alien civilization optimize their world, with that plan being corrigible w.r.t the aliens?" I don't currently understand how you avoid making this prediction given other stated beliefs. (Maybe you just bite the bullet and do predict this?)

By contrast, the specific values we wish to be pursued are not latent in the problems, are known to lack a simple boundary, and our training is much further removed from it.

I'm not totally sure what is meant by "simple boundary", but it seems like a lot of human values are latent in text prediction on the Internet, and when training from human feedback the training is not very removed from values.

It has long been the plan to say something less like "the following list comprises goodness: ..." and more like "yo we're tryin to optimize

some difficult-to-name concept; help us out?". [...]

I take this to mean that "dialing in on goodness" is a reasonable part of the solution space? If so, I retract that question. I thought from previous comments that Eliezer thought this part of solution space was more doomed than corrigibility.

(I get the sense that people think that I am butthurt about CIRL not getting enough recognition or something. I do in fact think this, but it's not part of my agenda here. I originally brought it up to make the argument that corrigibility is not in tension with EU maximization, then realized that I was mistaken about what "corrigibility" meant, but still care about the argument that "dialing in on goodness" is not in tension with EU maximization. But if we agree on that claim then I'm happy to stop talking about CIRL.)

[Soares][14:13]

I'd be *capable* of helping aliens optimize their world, sure. I wouldn't be motivated to, but I'd be capable.

[Shah][14:14]

(There are a bunch of aspects of your phrasing that indicated to me a different framing, and one I find quite foreign. For instance, this talk of "building a version of corrigibility_B" strikes me as foreign, and the talk of "making it compatible with 'plans that lase'" strikes me as foreign. It's plausible to me that you, who understand your original framing, can tell that my rephrasing matches your original intent. I do not yet feel like I could emit the description you emitted without contorting my thoughts about corrigibility in foreign ways, and I'm not sure whether that's an indication that there are distinctions, important to me, that I haven't communicated.)

This makes sense. I guess you might think of these concepts as quite pinned down? Like, in your head, EU maximization is just a kind of behavior (= set of behaviors), corrigibility is just another kind of behavior (= set of behaviors), and there's a straightforward yes-or-no question about whether the intersection is empty which you set out to answer, you can't "make" it come out one way or the other, nor can you "build" a new kind of corrigibility

[Soares][14:17]

Re: CIRL, my current working hypothesis is that by "use CIRL" you mean something analogous to what I say when I say "do CEV" -- namely, direct the AI to figure out what we "really" want in some correct sense, rather than attempting to specify what we want concretely. And to be clear, on my model, this *is* part of the solution to the overall alignment problem, and it's more-or-less why we wouldn't die immediately on the "value is

"fragile / we can't name exactly what we want" step if we solved the other problems.

My guess as to the disagreement about how much credit CIRL should get, is that there is in fact a disagreement, but it's not coming from MIRI folk saying "no we should be specifying the actual utility function by hand", it's coming from MIRI folk saying "this is just the advice 'do CEV' dressed up in different clothing and presented as a reason to stop worrying about corrigibility, which is irritating, given that it's orthogonal to corrigibility".

If you wanna fight that fight, I'd start by asking: Do you think CIRL is doing anything above and beyond what "use CEV" is doing? If so, what?

Regardless, I think it might be a good idea for you to try to pass my (or Eliezer's) ITT about what parts of the problem remain beyond the thing I'd call "do CEV" and why they're hard. (Not least b/c if my working hypothesis is wrong, demonstrating your mastery of that subject might prevent a bunch of toil covering ground you already know.)

[Shah][14:17]

I'd be *capable* of helping aliens optimize their world, sure. I wouldn't be motivated to, but I'd be capable.

Okay, so it seems like the danger requires the thing-producing-the-plan to be badly-motivated. But then I'm not sure why it seems so impossible to have a (not-badly-motivated) thing that, when given a goal, produces a plan to corrigibly get that goal. (This is a scenario Richard mentioned earlier.)

[Soares][14:19]

This makes sense. I guess you might think of these concepts as quite pinned down? Like, in your head, EU maximization is just a kind of behavior (= set of behaviors), corrigibility is just another kind of behavior (= set of behaviors), and there's a straightforward yes-or-no question about whether the intersection is empty which you set out to answer, you can't "make" it come out one way or the other, nor can you "build" a new kind of corrigibility

That sounds like one of the big directions in which your framing felt off to me, yeah :-(. (I don't fully endorse that rephrasing, but it seems directionally correct to me.)

Okay, so it seems like the danger requires the thing-producing-the-plan to be badly-motivated. But then I'm not sure why it seems so impossible to have a (not-badly-motivated) thing that, when given a goal, produces a plan to corrigibly get that goal. (This is a scenario Richard mentioned earlier.)

On my model, aiming the powerful optimizer is the hard bit.

Like, once I grant "there's a powerful optimizer, and all it does is produce plans to corrigibly attain a given goal", I agree that the problem is mostly solved.

There's maybe some cleanup, but the bulk of the alignment challenge preceded that point.

[Shah: ]

(This is hard for all the usual reasons, that I suppose I could retread.)

[Shah][14:24]

[...] Regardless, I think it might be a good idea for you to try to pass my (or Eliezer's) ITT about what parts of the problem remain beyond the thing I'd call "do CEV" and why they're hard. (Not least b/c if my working hypothesis is wrong, demonstrating your mastery of that subject might prevent a bunch of toil covering ground you already know.)

(Working on ITT)

[Soares][14:30]

(To clarify some points of mine, in case this gets published later to other readers: (1) I might call it more centrally something like "build a [DWIM system](#)" rather than "use CEV"; and (2) this is not advice about what your civilization should do with early AGI systems, I strongly recommend against trying to pull off CEV under that kind of pressure.)

[Shah][14:32]

I don't particularly want to have fights about credit. I just didn't want to falsely state that I do not care about how much credit CIRL gets, when attempting to head off further comments that seemed designed to appease my sense of not-enough-credit. (I'm also not particularly annoyed at MIRI, here.)

On passing ITT, about what's left beyond "use CEV" (stated in my ontology because it's faster to type; I think you'll understand, but I can also translate if you think that's important):

- The main thing is simply how to actually get the AI system to care about pursuing CEV. I think MIRI ontology would call this the target loading problem.
- This is hard because (a) you can't just train on CEV, because you can't just implement CEV and provide that as training and (b) even if you magically could train on CEV, that does not establish that the resulting AI system then wants to optimize CEV. It could just as well optimize some other objective that correlated with CEV in the situations you trained, but no longer correlates in some new

- situation (like when you are building a nanosystem). (Point (b) is how I would talk about inner alignment.)
- This is made harder for a variety of reasons, including (a) you're working with inscrutable matrices that you can't look at the details of, (b) there are clear racing incentives when the prize is to take over the world (or even just lots of economic profit), (c) people are unlikely to understand the issues at stake (unclear to me of the exact reasons, I'd guess it would be that the issues are too subtle / conceptual, + pressure to rationalize it away), (d) there's very little time in which we have a good understanding of the situation we face, because of fast / discontinuous takeoff

[Soares: 

[Soares][14:37]

Passable ^_^ (Not exhaustive, obviously; "it will have a tendency to kill you on the first real try if you get it wrong" being an example missing piece, but I doubt you were trying to be exhaustive.) Thanks.

[Shah: 

Okay, so it seems like the danger requires the thing-producing-the-plan to be badly-motivated. But then I'm not sure why it seems so impossible to have a (not-badly-motivated) thing that, when given a goal, produces a plan to corrigibly get that goal. (This is a scenario Richard mentioned earlier.)

I'm uncertain where the disconnect is here. Like, I could repeat some things from past discussions about how "it only outputs plans, it doesn't execute them" does very little (not nothing, but very little) from my perspective? Or you could try to point at past things you'd expect me to repeat and name why they don't seem to apply to you?

[Shah][14:40]

(Flagging that I should go to bed soon, though it doesn't have to be right away)

[Yudkowsky][14:50]

...I do not know if this is going to help anything, but I have a feeling that there's a frequent disconnect wherein I invented an idea, considered it, found it necessary-but-not-sufficient, and moved on to looking for additional or varying solutions, and then a decade or in this case 2 decades later, somebody comes along and sees this brilliant solution which MIRI is for some reason neglecting

this is perhaps exacerbated by a deliberate decision during the early days, when I looked very weird and the field was much more allergic to

weird, to not even try to stamp my name on all the things I invented. eg, I told Nick Bostrom to please use various of my ideas as he found appropriate and only credit them if he thought that was strategically wise.

I expect that some number of people now in the field don't know I invented corrigibility, and any number of other things that I'm a little more hesitant to claim here because I didn't leave Facebook trails for inventing them

and unless you had been around for quite a while, you definitely wouldn't know that I had been (so far as I know) the first person to perform the unexceptional-to-me feat of writing down, in 2001, the very obvious idea I called "external reference semantics", or as it's called nowadays, CIRL

[Shah][14:53]

I really honestly am not trying to say that MIRI didn't think of CIRL-like things, nor am I trying to get credit for CIRL. I really just wanted to establish that "learn what is good to do" seems not-ruled-out by EU maximization. That's all. It sounds like we agree on this point and if so I'd prefer to drop it.

[Soares: ❤]

[Yudkowsky][14:53]

Having a prior over utility functions that gets updated by evidence is not ruled out by EU maximization. That exact thing is hard for other reasons than it being contrary to the nature of EU maximization.

If it was ruled out by EU maximization for any simple reason, I would have noticed that back in 2001.

[Ngo][14:54]

I think we all agree on this point.

[Shah: 👍] [Soares: 👍]

One thing I'd note is that during my debate with Eliezer, I'd keep saying "oh so you think X is impossible" and he'd say "no, all these things are possible, they're just really really hard".

[Yudkowsky][14:58]

...to do correctly on your first try when a failed attempt kills you.

[Shah][14:58]

Maybe it's fine; perhaps the point is just that target loading is hard, and the question is why target loading is so hard.

From my perspective, the main confusing thing about the Eliezer/Nate view is how *confident* it is. With each individual piece, I (usually) find myself nodding along and saying "yes, it seems like if we wanted to guarantee safety, we would need to solve this". What I don't do is say "yes, it seems like without a solution to this, we're near-certainly dead". The uncharitable view (which I share mainly to emphasize where the disconnect is, not because I think it is true) would be something like "Eliezer/Nate are falling to a Murphy bias, where they assume that unless they have an ironclad positive argument for safety, the worst possible thing will happen and we all die". I try to generate things that seem more like ironclad (or at least "leatherclad") positive arguments for doom, and mostly don't succeed; when I say "human values are very complicated" there's the rejoinder that "a superintelligence will certainly know about human values; pointing at them shouldn't take that many more bits"; when I say "this is ultimately just praying for generalization", there's the rejoinder "but it may in fact actually generalize"; add to all of this the fact that a bunch of people will be trying to prevent the problem and it seems weird to be so confident in doom.

A lot of my questions are going to be of the form "it seems like this is a way that we could survive; it definitely involves luck and does not say good things about our civilization, but it does not seem as improbable as the word 'miracle' would imply"

[Yudkowsky][15:00]

heh. from my standpoint, I'd say of this that it reflects those old experiments where if you ask people for their "expected case" it's indistinguishable from their "best case" (since both of these involve visualizing various things going on their imaginative mainline, which is to say, as planned) and reality is usually worse than their "worst case" (because they didn't adjust far enough away from their best-case anchor towards the statistical distribution for actual reality when they were trying to imagine a few failures and disappointments of the sort that reality had previously delivered)

it rhymes with the observation that it's incredibly hard to find people - even inside the field of computer security - who really have what Bruce Schneier termed the security mindset, of asking how to break a cryptography scheme, instead of imagining how your cryptography scheme could succeed

from my perspective, people are just living in a fantasy reality which, if we were actually living in it, would not be full of failed software projects or rocket prototypes that blow up even after you try quite hard to get a system design about which you made a strong prediction that it wouldn't explode

they think something special has to go wrong with a rocket design, that you must have committed some grave unusual sin against rocketry, for the rocket to explode

as opposed to every rocket wanting really strongly to explode and needing to constrain every aspect of the system to make it not explode and then the first 4 times you launch it, it blows up anyways

why? because of some particular technical issue with O-rings, with the flexibility of rubber in cold weather?

[Shah][15:05]

(I have read your Rocket Alignment and security mindset posts. Not claiming this absolves me of bias, just saying that I am familiar with them)

[Yudkowsky][15:05]

no, because the strains and temperatures in rockets are large compared to the materials that we use to make up the rockets

the fact that sometimes people are wrong in their uncertain guesses about rocketry does not make their life easier in this regard

the less they understand, the less ability they have to force an outcome within reality

it's no coincidence that when you are Wrong about your rocket, the particular form of Being Wrong that reality delivers to you as a surprise message, is not that you underestimated the strength of steel and so your rocket went to orbit and came back with fewer scratches on the hull than expected

when you are working with powerful forces there is not a symmetry around pleasant and unpleasant surprises being equally likely relative to your first-order model. if you're a good Bayesian, they will be equally likely relative to your second-order model, but this requires you to be HELLA pessimistic, indeed, SO PESSIMISTIC that sometimes you are pleasantly surprised

which looks like such a bizarre thing to a mundane human that they will gather around and remark at the case of you being pleasantly surprised

they will not be used to seeing this

and they shall say to themselves, "haha, what pessimists"

because to be unpleasantly surprised is so ordinary that they do not bother to gather and gossip about it when it happens

my fundamental sense about the other parties in this debate, underneath all the technical particulars, is that they've constructed a Murphy-free

fantasy world from the same fabric that weaves crazy optimistic software project estimates and brilliant cryptographic codes whose inventors didn't quite try to break them, and are waiting to go through that very common human process of trying out their optimistic idea, letting reality gently correct them, predictably becoming older and wiser and starting to see the true scope of the problem, and so in due time becoming one of those Pessimists who tell the youngsters how ha ha of course things are not that easy

this is how the cycle usually goes

the problem is that instead of somebody's first startup failing and them then becoming much more pessimistic about lots of things they thought were easy and then doing their second startup

the part where they go ahead optimistically and learn the hard way about things in their chosen field which aren't as easy as they hoped

[Shah][15:13]

Do you want to bet on that? That seems like a testable prediction about beliefs of real people in the not-too-distant future

[Yudkowsky][15:13]

kills everyone

not just them

everyone

this is an issue

how on Earth would we bet on that if you think the bet hasn't already resolved? I'm describing the attitudes of people that I see right now today.

[Shah][15:15]

Never mind, I wanted to bet on "people becoming more pessimistic as they try ideas and see them fail", but if your idea of "see them fail" is "superintelligence kills everyone" then obviously we can't bet on that

(people here being alignment researchers, obviously ones who are not me)

[Yudkowsky][15:17]

there is some element here of the Bayesian not updating in a predictable direction, of executing today the update you know you'll make later, of

saying, "ah yes, I can see that I am in the same sort of situation as the early AI pioneers who thought maybe it would take a summer and actually it was several decades because Things Were Not As Easy As They Imagined, so instead of waiting for reality to correct me, I will imagine myself having already lived through that and go ahead and be more pessimistic right now, not just a little more pessimistic, but so incredibly pessimistic that I am as likely to be pleasantly surprised as unpleasantly surprised by each successive observation, which is even more pessimism than even some sad old veterans manage", an element of genresavviness, an element of knowing the advice that somebody would predictably be shouting at you from outside, of not just blindly enacting the plot you were handed

and I don't quite know *why* this is so much less common than I would have naively thought it would be

why people are content with enacting the predictable plot where they start out cheerful today and get some hard lessons and become pessimistic later

they are their own scriptwriters, and they write scripts for themselves about going into the haunted house and then splitting up the party

I would not have thought that to defy the plot was such a difficult thing for an actual human being to do

that it would require so much reflectivity or something, I don't know what else

nor do I know how to train other people to do it if they are not doing it already

but that from my perspective is the basic difference in gloominess

I am a time-traveler who came back from the world where it (super duper predictably) turned out that a lot of early bright hopes didn't pan out and various things went WRONG and alignment was HARD and it was NOT SOLVED IN ONE SUMMER BY TEN SMART RESEARCHERS

and now I am trying to warn people about this development which was, from a certain perspective, really quite obvious and not at all difficult to see coming

but people are like, "what the heck are you doing, you are enacting the wrong part of the plot, people are currently supposed to be cheerful, you can't prove that anything will go wrong, why would I turn into a grizzled veteran before the part of the plot where reality hits me over the head with the awful real scope of the problem and shows me that my early bright ideas were way too optimistic and naive"

and I'm like "no you don't get it, where I come from, everybody died and didn't turn into grizzled veterans"

and they're like "but that's not what the script says we do next"... or something, I do not know what leads people to think like this because I do

not think like that myself

[Soares][15:24]

(I think what they actually do is say "it's not obvious to me that this is one of those scenarios where we become grizzled veterans, as opposed to things just actually working out easily")

("many things work out easily all the time; obviously society spends a bunch more focus on things that don't work out easily b/c the things that work easily tend to get resolved fairly quickly and then you don't notice them", or something)

(more generally, I kinda suspect that bickering closer to the object level is likely more productive)

(and i suspect this convo might be aided by Rohin naming a concrete scenario where things go well, so that Eliezer can lament the lack of genre saviness in various specific points)

[Yudkowsky][15:26]

there are, of course, lots of more local technical issues where I can specifically predict the failure mode for somebody's bright-eyed naive idea, especially when I already invented a more sophisticated version a decade or two earlier, and this is what I've usually tried to discuss

[Soares: ♥]

because conversations like that can sometimes make any progress

[Soares][15:26]

(and possibly also Eliezer naming a concrete story where things go poorly, so that Rohin may lament the seemingly blind pessimism & premature grizzledness)

[Yudkowsky][15:27]

whereas if somebody lacks the ability to see the warning signs of which genre they are in, I do not know how to change the way they are by talking at them

[Shah][15:28]

Unsurprisingly I have disagreements with the meta-level story, but it seems really thorny to make progress on and I'm kinda inclined to not discuss it. I also should go to sleep now.

One thing it did make me think of -- it's possible that the "do it correctly on your first try when a failed attempt kills you" could be the crux here. There's a clearly-true sense which is "the first time you build a superintelligence that you cannot control, if you have failed in your alignment, then you die". There's a different sense which is "and also, anything you try to do with non-superintelligences that you can control, will tell you approximately nothing about the situation you face when you build a superintelligence". I mostly don't agree with the second sense, but if Eliezer / Nate do agree with it, that would go a long way to explaining the confidence in doom.

Two arguments I can see for the second sense: (1) the non-superintelligences only seem to respond well to alignment schemes because they don't yet have the core of general intelligence, and (2) the non-superintelligences only seem to respond well to alignment schemes because despite being misaligned they are doing what we want in order to survive and later execute a treacherous turn. EDIT: And (3) fast takeoff = not much time to look at the closest non-dangerous examples

(I still should sleep, but would be interested in seeing thoughts tomorrow, and if enough people think it's actually worthwhile to engage on the meta level I can do that. I'm cheerful about engaging on specific object-level ideas.)

[Soares: ]

[Yudkowsky][15:28]

it's not that early failures tell you nothing

the failure of the 1955 Dartmouth Project to produce strong AI over a summer told those researchers something

it told them the problem was harder than they'd hoped on the first shot

it didn't show them the correct way to build AGI in 1957 instead

[Bensinger][16:41]

Linking to a chat log between Eliezer and some anonymous people (and Steve Omohundro) from early September:

[<https://www.lesswrong.com/posts/CpvyhFy9WvCNSifkY/discussion-with-eliezer-yudkowsky-onagi-interventions>]

Eliezer tells me he thinks it pokes at some of Rohin's questions

[Yudkowsky][16:48]

I'm not sure that I can successfully, at this point, go back up and usefully reply to the text that scrolled past - I also note some internal grinding about this having turned into a thing which has Pending Replies instead of

Scheduled Work Hours - and this maybe means that in the future we shouldn't have such a general chat here, which I didn't anticipate before the fact. I shall nonetheless try to pick out some things and reply to them.

[Shah: ]

- While I think people agree on the behaviors of corrigibility, I am not sure they agree on why we want it. Eliezer wants it for surviving failures, but maybe others want it for "dialing in on goodness". When I think about a "broad basin of corrigibility", that intuitively seems more compatible with the "dialing in on goodness" framing (but this is an aesthetic judgment that could easily be wrong).

This is a weird thing to say in my own ontology.

There's a general project of AGI alignment where you try to do some useful pivotal thing, which has to be powerful enough to be pivotal, and so you somehow need a system that thinks powerful thoughts in the right direction without it killing you.

This could include, for example:

- Trying to train in "low impact" via an RL loss function that penalizes a sufficiently broad range of "impacts" that we hope the learned impact penalty generalizes to all the things we'd consider impacts - even as we scale up the system, without the sort of obvious pathologies that would materialize only over options available to sufficiently powerful systems, like sending out nanosystems to erase the visibility of its actions from human observers
- Tweaking MCTS search code so that it behaves in the fashion of "mild optimization" or "[taskishness](#)" instead of searching as hard as it has power available to search
- Exposing the system to lots of labeled examples of relatively simple and safe instructions being obeyed, hoping that it generalizes safe instruction-following to regimes too dangerous for us to inspect outputs and label results
- Writing code that tries to recognize cases of activation vectors going outside the bounds they occupied during training, as a check on whether internal cognitive conservatism is being violated or something is seeking out adversarial counterexamples to a constraint

You could say that only parts 1 and 3 are "dialing in on goodness" because only those parts involve iteratively refining a target, or you could say that all 4 parts are "dialing in on goodness" because parts 2 and 4 help you stay alive while you're doing the iterative refining. But I don't see this distinction as fundamental or particularly helpful. What if, on part 4, you were training something to recognize out-of-bounds activations, instead of trying to hardcode it? Is that dialing in on goodness? Or is it just dialing in on survivability or corrigibility or whatnot? Or maybe even part 3 isn't really "dialing in on goodness"

because the true distinction between Good and Evil is still external in the programmers and not inside the system?

I don't see this as an especially useful distinction to draw. There's a hardcoded/learned distinction that probably does matter in several places. There's a maybe-useful forest-level distinction between "actually doing the pivotal thing" and "not destroying the world as a side effect" which breaks down around the trees because the very definition of "that pivotal thing you want to do" is to do *that thing* and *not* to destroy the world.

And all of this is a class of shallow ideas that I can generate in great quantity. I now and then consider writing up the ideas like this, just to make clear that I've already thought of way more shallow ideas like this than the net public output of the entire rest of the alignment field, so it's not that my concerns of survivability stem from my having missed any of the obvious shallow ideas like that.

The reason I don't spend a lot of time talking about it is not that I haven't thought of it, it's that I've thought of it, explored it for a while, and decided not to write it up because I don't think it can save the world and the infinite well of shallow ideas seems more like a distraction from the level of miracle we would actually need.

-

As a starting point: you say that an agent that makes plans but doesn't execute them is also dangerous, because it is the plan itself that lases, and corrigibility is antithetical to lasing. Does this mean you predict that you, or I, with suitably enhanced intelligence and/or reflectivity, would not be capable of producing a plan to help an alien civilization optimize their world, with that plan being corrigible w.r.t the aliens? (This seems like a strange and unlikely position to me, but I don't see how to not make this prediction under what I believe to be your beliefs. Maybe you just bite this bullet.)

I 'could' corrigibly help the [Babyeaters](#) in the sense that I have a notion of what it would mean to corrigibly help them, and if I wanted to do that thing for some reason, like an outside super-universal entity offering to pay me a googolplex flops of eudaimonium if I did that one thing, then I could do that thing. Absent the superuniversal entity bribing me, I wouldn't want to behave corrigibly towards the Babyeaters.

This is not a defect of myself as an individual. The Superhappies would also be able to understand what it would be like to be corrigible; they wouldn't want to behave corrigibly towards the Babyeaters, because, like myself, they don't want exactly what the Babyeaters want. In particular, we would rather the universe be other than it is with respect to the Babyeaters eating babies.

[Shah: ]

22. Follow-ups

[Shah][0:33] (Nov. 8)

[...] Absent the superuniversal entity bribing me, I wouldn't want to behave corrigibly towards the Babyeaters. [...]

Got it. Yeah I think I just misunderstood a point you were saying previously. When Richard asked about systems that simply produce plans rather than execute them, you said something like "the plan itself is dangerous", which I now realize meant "you don't get additional safety from getting to read the plan, the superintelligence would have just chosen a plan that was convincing to you but nonetheless killed everyone / otherwise worked in favor of the superintelligence's goals", but at the time I interpreted it as "any reasonable plan that can actually build nanosystems is going to be dangerous, regardless of the source", which seemed obviously false in the case of a well-motivated system.

[...] This is a weird thing to say in my own ontology. [...]

When I say "dialing in on goodness", I mean a specific class of strategies for getting a superintelligence to do a useful pivotal thing, in which you build it so that the superintelligence is applying its force towards figuring out what it is that you actually want it to do and pursuing that, which among other things would involve taking a pivotal act to reduce x-risk to ~zero.

I previously had the mistaken impression that you thought this class of strategies was probably doomed because it was incompatible with expected utility theory, which seemed wrong to me. (I don't remember why I had this belief; possibly it was while I was still misunderstanding what you meant by "corrigibility" + the claim that corrigibility is anti-natural.)

I now think that you think it is probably doomed for the same reason that most other technical strategies are probably doomed, which is that there still doesn't seem to be any plausible way of loading in the right target to the superintelligence, even when that target is a process for learning-what-to-optimize, rather than just what-to-optimize.

Linking to a chat log between Eliezer and some anonymous people (and Steve Omohundro) from early September:

[<https://www.lesswrong.com/posts/CpvyhFy9WvCNSifkY/discussion-with-eliezer-yudkowsky-on-agi-interventions>]

Eliezer tells me he thinks it pokes at some Rohin's questions

I'm surprised that you think this addresses (or even pokes at) my questions. As far as I can tell, most of the questions there are either about social dynamics, which I've been explicitly avoiding, and the

"technical" questions seem to treat "AGI" or "superintelligence" as a symbol; there don't seem to be any internal gears underlying that symbol. The closest anyone got to internal gears was mentioning iterated amplification as a way of bootstrapping known-safe things to solving hard problems, and that was very brief.

I am much more into the question "how difficult is technical alignment". It seems like answers to this question need to be in one of two categories: (1) claims about the space of minds that lead to intelligent behavior (probably weighted by simplicity, to account for the fact that we'll get the simple ones first), (2) claims about specific methods of building superintelligences. As far as I can tell the only thing in that doc which is close to an argument of this form is "superintelligent consequentialists would find ways to manipulate humans", which seems straightforwardly true (when they are misaligned). I suppose one might also count the assertion that "the speedup step of iterated amplification will introduce errors" as an argument of this form.

It could be that you are trying to convince me of some other beliefs that I wasn't asking about, perhaps in the hopes of conveying some missing mood, but I suspect that it is just that you aren't particularly clear on what my beliefs are / what I'm interested in. (Not unreasonable, given that I've been poking at your models, rather than the other way around.) I could try saying more about that, if you'd like.

[Tallinn][11:39] (Nov. 12)

FWIW, a voice from the audience: +1 to going back to sketching concrete scenarios. even though i learned a few things from the abstract discussion of goodness/corrigibility/etc myself (eg, that "corrigible" was meant to be defined at the limit of self-improvement till maturity, not just as a label for code that does not resist iterated development), the progress felt more tangible during the "scaled up muzero" discussion above.

[Yudkowsky][15:03] (Nov. 12)

anybody want to give me a prompt for a concrete question/scenario, ideally a concrete such prompt but I'll take whatever?

[Soares][15:34] (Nov. 12)

Not sure I count, but one I'd enjoy a concrete response to: "The leading AI lab vaguely thinks it's important that their systems are 'mere predictors', and wind up creating an AGI that is dangerous; how concretely does it wind up being a scary planning optimizer or whatever, that doesn't run through a scary abstract "waking up" step".

(asking for a friend; @Joe Carlsmith or whoever else finds this scenario unintuitive plz clarify with more detailed requests if interested)

23. November 13 conversation

23.1. GPT-n and goal-oriented aspects of human reasoning

[Shah][1:46]

I'm still interested in:

5. More concreteness on how optimization generalizes but corrigibility doesn't, in the case where the AI was trained by human judgment on weak-safe domains

Specifically, we can go back to the scaled-up MuZero example. Some (lightly edited) details we had established there:

Pretraining: playing all the videogames, predicting all the text and images, solving randomly generated computer puzzles, accomplishing sets of easily-labelable sensorymotor tasks using robots and webcams

Finetuning: The AI system is being trained to act well on the Internet, and it's shown some tweet / email / message that a user might have seen, and asked to reply to the tweet / email / message. User says whether the replies are good or not (perhaps via comparisons, a la Deep RL from Human Preferences). It would be more varied than that, but would not include "building nanosystems".

The AI system is not smart enough that exposing humans to text it generates is already a world-wrecking threat if the AI is hostile.

At that point we moved from concrete to abstract:

Abstract description: train on 'weak-safe' domains where the AI isn't smart enough to do damage, and the humans can label the data pretty well because the AI isn't smart enough to fool them

Abstract problem: Optimization generalizes and corrigibility fails

I would be interested in a more concrete description here. I'm not sure exactly what details I'm looking for -- on my ontology the question is something like "what algorithm is the AI system forced to learn; how does that lead to generalized optimization and failed corrigibility; why weren't there simple safer algorithms that were compatible with the training, or if there were such algorithms why didn't the AI system learn them". I don't

really see how to answer all of that without abstraction, but perhaps you'll have an answer anyway

(I am hoping to get some concrete detail on "how did it go from non-hostile to hostile", though I suppose you might confidently predict that it was already hostile after pretraining, conditional on it being an AGI at all. I can try devising a different concrete scenario if that's a blocker.)

[Yudkowsky][11:09]

I am hoping to get some concrete detail on "how did it go from non-hostile to hostile"

Mu Zero is intrinsically dangerous for reasons essentially isomorphic to the way that AIXI is intrinsically dangerous: It tries to remove humans from its environment when playing Reality for the same reasons it stomps a Goomba if it learns how to play Super Mario Bros 1, because it has some goal and the Goomba is in the way. It doesn't need to learn anything more to be that way, except for learning what a Goomba/human is within the current environment.

The question is more "What kind of patches might it learn for a weak environment if optimized by some hill-climbing optimization method and loss function not to stomp Goombas there, and how would those patches fail to generalize to not stomping humans?"

Agree or disagree so far?

[Shah][12:07]

Agree assuming that it is pursuing a misaligned goal, but I am also asking what misaligned goal it is pursuing (and depending on the answer, maybe also how it came to be pursuing that misaligned goal given the specified training setup).

In fact I think "what misaligned goal is it pursuing" is probably the more central question for me

[Yudkowsky][12:14]

well, obvious abstract guess is: something whose non-maximal "optimum" (that is, where the optimization ended up, given about how powerful the optimization was) coincided okayish with the higher regions of the fitness landscape (lower regions of the loss landscape) that could be reached at all, relative to its ancestral environment

I feel like it would be pretty hard to blindly guess, in advance, at my level of intelligence, without having seen any precedents, what the hell a Human would look like, as a derivation of "inclusive genetic fitness"

[Shah][12:15]

Yeah I agree with that in the abstract, but have had trouble giving compelling-to-me concrete examples

Yeah I also agree with that

[Yudkowsky][12:15]

I could try to make up some weird false specifics if that helps?

[Shah][12:16]

To be clear I am fine with "this is a case where we predictably can't have good concrete stories and this does not mean we are safe" (and indeed argued the same thing in a doc I linked here many messages ago)

But weird false specifics could still be interesting

Although let me think if it is actually valuable

Probably it is not going to change my mind very much on alignment difficulty, if it is "weird false specifics", so maybe this isn't the most productive line of discussion. I'd be "selfishly" interested in that "weird false specifics" seems good for me to generate novel thoughts about these sorts of scenarios, but that seems like a bad use of this Discord

I think given the premises that (1) superintelligence is coming soon, (2) it pursues a misaligned goal by default, and (3) we currently have no technical way of preventing this and no realistic-seeming avenues for generating such methods, I am very pessimistic. I think (2) and (3) are the parts that I don't believe and am interested in digging into, but perhaps "concrete stories" doesn't really work for this.

[Yudkowsky][12:26]

with any luck - though I'm not sure I actually expect that much luck - this would be something Redwood Research could tell us about, if they can [learn a nonviolence predicate](#) over GPT-3 outputs and then manage to successfully mutate the distribution enough that we can get to see what was actually inside the predicate instead of "nonviolence"

[Shah: ]

or, like, 10% of what was actually inside it

or enough that people have some specifics to work with when it comes to understanding how gradient descent learning a function over outcomes from human feedback relative to a distribution, doesn't just learn the actual function the human is using to generate the feedback (though, if this were learned exactly, it would still be fatal given superintelligence)

[Shah][12:33]

In this framing I do buy that you don't learn exactly the function that generates the feedback -- I have ~5 contrived specific examples where this is the case (i.e. you learn something that wasn't what the feedback function would have rewarded in a different distribution)

(I'm now thinking about what I actually want to say about this framing)

Actually, maybe I do think you might end up learning the function that generates the feedback. Not literally exactly, if for no other reason than rounding errors, but well enough that the inaccuracies don't matter much. The AGI presumably already knows and understands the concepts we use based on its pretraining, is it really so shocking if gradient descent hooks up those concepts in the right way? (GPT-3 on the other hand doesn't already know and understand the relevant concepts, so I wouldn't predict this of GPT-3.) I do feel though like this isn't really getting at my reason for (relative) optimism, and that reason is much more like "I don't really buy that AGI must be very coherent in a way that would prevent corrigibility from working" (which we could discuss if desired)

On the comment that learning the exact feedback function is still fatal -- I am unclear on why you are so pessimistic on having "human + AI" supervise "AI", in order to have the supervisor be smarter than the thing being supervised. (I think) I understand the pessimism that the learned function won't generalize correctly, but if you imagine that magically working, I'm not clear what additional reason prevents the "human + AI" supervising "AI" setup.

- I can see how you die if the AI ever becomes misaligned, i.e. there isn't a way to fix mistakes, but I don't see how you get the misaligned AI in the first place.
- I could also see things like "Just like a student can get away with plagiarism even when the teacher is smarter than the student, the AI knows more about its cognition than the human + AI system, and so will likely be incentivized to do bad things that it knows are bad but the human + AI system doesn't know is bad". But that sort of thing seems solvable with future research, e.g. debate, interpretability, red teaming all seem like feasible approaches.

[Yudkowsky][13:06]

what's a "human + AI"? can you give me a more concrete version of that scenario, either one where you expect it to work, or where you yourself have labeled the first point you expect it to fail and you want to know whether I see an earlier failure than that?

[Shah][13:09]

One concrete training algorithm would be debate, ideally with mechanisms that allow the AI systems to "look into each other's thoughts" and make credible statements about them, but we can skip that for now as it isn't very concrete

Would you like a training domain and data as well?

I don't like the fact that a smart AI system in this position could notice that it is playing against itself and decide not to participate in a zero-sum game, but I am not sure if that worry actually makes sense or not

(Debate can be thought of as simultaneously "human + first AI evaluate second AI" and "human + second AI evaluate first AI")

[Yudkowsky][13:12]

further concreteness, please! what pivotal act is it training for? what are the debate contents about?

[Shah][13:16]

You start with "easy" debates like mathematical theorem proving or fact-based questions, and ramp up until eventually the questions are roughly "what is the next thing to do in order to execute a pivotal act"

Intermediate questions might be things like "is it a good idea to have a minimum wage"

[Yudkowsky][13:17]

so, like, "email ATTTGAGCTTGCC... to the following address, mix the proteins you receive by FedEx in a water-saline solution at 2 degrees Celsius..." for the final stage?

[Shah][13:17]

Yup, that could be it

Humans are judging debates based on reasoning though, not just outcomes-after-executing-the-plan

[Yudkowsky][13:19]

okay. let's suppose you manage to prevent both AGIs from using logical decision theory to coordinate with each other. both AIs tell their humans that the other AI's plans are murderous. now what?

[Shah][13:19]

So assuming perfect generalization there should be some large implicit debate tree that justifies the plan in human-understandable form

[Yudkowsky][13:20]

yah, I flatly disbelieve that entire development scheme, so we should maybe back up.

people fiddled around with GPT-4 derivatives and never did get them to engage in lines of printed reasoning that would design interesting new stuff. now what?

Living Zero (a more architecturally complicated successor of Mu Zero) is getting better at designing complicated things over on its side while that's going on, whatever it is

[Shah][13:23]

Okay, so the worry is that this just won't scale, not that (assuming perfect generalization) it is unsafe? Or perhaps you also think it is unsafe but it's hard to engage with because you don't believe it will scale?

And the issue is that relying on reasoning confines you to a space of possible thoughts that doesn't include the kinds of thoughts required to develop new stuff (e.g. intuition)?

[Yudkowsky][13:25]

mostly I have found these alleged strategies to be too permanently abstract, never concretized, to count as admissible hypotheses. if you ask me to concretize them myself, I think that unelaborated giant transformer stacks trained on massive online text corpuses fail to learn smart-human-level engineering reasoning before the world ends. If that were not true, I would expect Paul-style schemes to blow up on the distillation step, but first failures first.

[Shah][13:26]

What additional concrete detail do you want?

It feels like I specified something that we could code up a stupidly inefficient version of now

[Yudkowsky][13:27]

Great. Describe the stupidly inefficient version?

[Shah][13:33]

In terms of what actually happens: Each episode, there is an initial question specified by the human. Agent A and agent B, which are copies of the same neural net, simultaneously produce statements ("answers"). They then have a conversation. At the end the human judge decides which answer is better, and rewards the appropriate agent. The agents are updated using some RL algorithm.

I can say stuff about why we might hope this works, or about tricks you have to play in order to get learning to happen at all, or other things

[Yudkowsky][13:35]

Are the agents also playing Starcraft or have they spent their whole lives inside the world of text?

[Shah][13:35]

For the stupidly inefficient version they could have spent their whole lives inside text

[Yudkowsky][13:37]

Okay. I don't think the pure-text versions of GPT-5 are being very good at designing nanosystems while Living Zero is ending the world.

[Shah][13:37]

In the stupidly inefficient version human feedback has to teach the agents facts about the real world

[Yudkowsky][13:37]

(It's called "Living Zero" because it does lifelong learning, in the backstory I've been trying to separately sketch out in a draft.)

[Shah][13:38]

Oh I definitely agree this is not competitive

So when you say this is too abstract, you mean that there isn't a story for how they incorporate e.g. physical real-world knowledge?

[Yudkowsky][13:39]

no, I mean that when I talk to Paul about this, I can't get Paul to say anything as concrete as the stuff you've already said

the reason why I don't expect the GPT-5s to be competitive with Living Zero is that gradient descent on feedforward transformer layers, in order how to learn science by competing to generate text that humans like, would have to pick up on some very deep latent patterns generating that text, and I don't think there's an incremental pathway there for gradient descent to follow - if gradient descent even follows incremental pathways as opposed to finding [lottery tickets](#), but that's a whole separate open question of artificial neuroscience.

in other words, humans play around with legos, and hominids play around with chipping flint handaxes, and mammals play around with spatial reasoning, and that's part of the incremental pathway to developing deep patterns for causal investigation and engineering, which then get projected into human text and picked up by humans reading text

it's just straightforwardly not clear to me that GPT-5 pretrained on human text corpuses, and then further posttrained by RL on human judgment of text outputs, ever runs across the deep patterns

where relatively small architectural changes might make the system no longer just a giant stack of transformers, even if that resulting system is named "GPT-5", and in this case, bets might be off, but also in this case, things will go wrong with it that go wrong with Living Zero, because it's now learning the more powerful and dangerous kind of work

[Shah][13:45]

That does seem like a disagreement, in that I think this process does eventually reach the "deep patterns", but I do agree it is unlikely to be competitive

[Yudkowsky][13:45]

I mean, if you take a feedforward stack of transformer layers the size of a galaxy and train it via gradient descent using all the available energy in the reachable universe, it might find something, sure

though this is by no means certain to be the case

[Shah][13:50]

It would be quite surprising to me if it took that much. It would be *especially* surprising to me if we couldn't figure out some alternative reasonably-simple training scheme like "imitate a human doing good reasoning" that still remained entirely in text that could reach the "deep patterns". (This is now no longer a discussion about whether the training scheme is aligned, not sure if we should continue it.)

I realize that this might be hard to do, but if you imagine that GPT-5 + human feedback finetuning does run across the deep patterns and could

in theory do the right stuff, and also generalization magically works, what's the next failure?

[Yudkowsky][13:56]

what sort of deep thing does a hill-climber run across in the layers, such that the deep thing is the most predictive thing it found for human text about science?

if you don't visualize this deep thing in any detail, then it can in one moment be powerful, and in another moment be safe. it can have all the properties that you want simultaneously. who's to say otherwise? the mysterious deep thing has no form within your mind.

if one were to name specifically "well, it ran across a little superintelligence with long-term goals that it realized it could achieve by predicting well in all the cases that an outer gradient descent loop would probably be updating on", that sure doesn't end well for you.

this perhaps is *not* the first thing that gradient descent runs across. it wasn't the first thing that natural selection ran across to build things that ran the savannah and made more of themselves. but what deep pattern that is *not* pleasantly and unfrighteningly formless would gradient descent run across instead?

[Shah][14:00]

(Tbc by "human feedback finetuning" I mean debate, and I suspect that "generalization magically works" will be meant to rule out the thing that you say next, but seems worth checking so let me write an answer)

the deep thing is the most predictive thing it found for human text about science?

Wait, the most predictive thing? I was imagining it as just a thing that is present in addition to all the other things. Like, I don't think I've learned a "deep thing" that is most useful for riding a bike. Probably I'm just misunderstanding what you mean here.

I don't think I can give a good answer here, but to give some answer, it has a belief that there is a universe "out there", that lots but not all of the text it reads is making claims about (some aspect of) the universe, those claims can be true or false, there are some claims that are known to be true, there are some ways to take assumed-true claims and generate new assumed-true claims, which includes claims about optimal actions for goals, as well as claims about how to build stuff, or what the effect of a specified machine is

[Yudkowsky][14:10]

hell of a lot of stuff for gradient descent to run across in a stack of transformer layers. clearly the lottery-ticket hypothesis must have been very incorrect, and there was an incremental trail of successively more complicated gears that got trained into the system.

btw by "claims" are you meaning to make the jump to English claims? I was reading them as giant inscrutable vectors encoding meaningful propositions, but maybe you meant something else there.

[Shah][14:11]

In fact I am skeptical of some strong versions of the lottery ticket hypothesis, though it's been a while since I read the paper and I don't remember exactly what the original hypothesis was

Giant inscrutable vectors encoding meaningful propositions

[Yudkowsky][14:13]

oh, I'm not particularly confident of the lottery-ticket hypothesis either, though I sure do find it grimly amusing that a species which hasn't already figured *that* out one way or another thinks it's going to have deep transparency into neural nets all wrapped up in time to survive. but, separate issue.

"How does gradient descent even work?" "Lol nobody knows, it just does."

but, separate issue

[Shah][14:16]

How does strong lottery ticket hypothesis explain GPT-3? Seems like that should already be enough to determine that there's an incremental trail of successively more complicated gears

[Yudkowsky][14:18]

could just be that in 175B parameters, combinatorially combined through possible execution pathways, there is some stuff that was pretty close to doing all the stuff that GPT-3 ended up doing.

anyways, for a human to come up with human text about science, the human has to brood and think for a bit about different possible hypotheses that could account for the data, notice places where those hypotheses break down, tweak the hypotheses in their mind to make the errors go away; they would engineer an internal mental construct towards the engineering goal of making good predictions. if you're looking at orbital mechanics and haven't invented calculus yet, you

invent calculus as a persistent mental tool that you can use to craft those internal mental constructs.

does the formless deep pattern of GPT-5 accomplish the same ends, by some mysterious means that is, formless, able to produce the same result, but not by any detailed means where if you visualized them you would be able to see how it was unsafe?

[Shah][14:24]

I expect that probably we will figure out some way to have adaptive computation time be a thing (it's been investigated for years now, but afaik hasn't worked very well), which will allow for this sort of thing to happen

In the stupidly inefficient version, you have a really really giant and deep neural net that does all of that in successive layers of the neural net. (And when it doesn't need to do that, those layers are noops.)

[Yudkowsky][14:26][14:32]

okay, so my question is, is there a little goal-oriented mind inside there that solves science problems the same way humans solve them, by engineering mental constructs that serve a goal of prediction, including backchaining for prediction goals and forward chaining from alternative hypotheses / internal tweaked states of the mental construct? or is there something else which solves the same problem, not how humans do it, without any internal goal orientation?

People who would not in the first place realize that humans solve prediction problems by internally engineering internal mental constructs in a goal-oriented way, would of course imagine themselves able to imagine a formless spirit which produces "predictions" without being "goal-oriented" because they lack an understanding of internal machinery and so can combine whatever surface properties and English words they want to yield a beautiful optimism

Or perhaps there is indeed some way to produce "predictions" without being "goal-oriented", which gradient descent on a great stack of transformer layers would surely run across; but you will pardon my grave lack of confidence that someone has in fact seen so much further than myself, when they don't seem to have appreciated in advance of my own questions why somebody who understood something about human internals would be skeptical of this.

If they're sort of visibly trying to come up with it on the spot after I ask the question, that's not such a great sign either.

This is not aimed particularly at you, but I hope the reader may understand something of why Eliezer Yudkowsky goes about sounding so gloomy all the time about other people's prospects for noticing what will

kill them, by themselves, without Eliezer constantly hovering over their shoulder every minute prompting them with almost all of the answer.

[Shah][14:31]

Just to check my understanding: if we're talking about, say, how humans might go about understanding neural nets, there's a goal of "have a theory that can retrodict existing observations and make new predictions", backchaining might say "come up with hypotheses that would explain double descent", forward chaining might say "look into bias and variance measurements"?

If so, yes, I think the AGI / GPT-5-that-is-an-AGI is doing something similar

[Yudkowsky][14:33]

your understanding sounds okay, though it might make more sense to talk about a domain that human beings understand better than artificial neuroscience, for purposes of illustrating how scientific thinking works, since human beings haven't actually gotten very far with artificial neuroscience.

[Shah][14:33]

Fair point re using a different domain

To be clear I do not in fact think that GPT-N is safe because it is trained with supervised learning and I am confused at the combination of views that GPT-N will be AGI and GPT-N will be safe because it's just doing predictions

Maybe there is marginal additional safety but you clearly can't say it is "definitely safe" without some additional knowledge that I have not seen so far

Going back to the original question, of what the next failure mode of debate would be assuming magical generalization, I think it's just not one that makes sense to ask on your worldview / ontology; "magical generalization" is the equivalent of "assume that the goal-oriented mind somehow doesn't do dangerous optimization towards its goal, yet nonetheless produces things that can only be produced by dangerous optimization towards a goal", and so it is assuming the entire problem away

[Yudkowsky][14:41]

well YES

from my perspective the whole field of mental endeavor as practiced by alignment optimists consists of ancient alchemists wondering if they can get collections of surface properties, like a metal as shiny as gold, as hard

as steel, and as self-healing as flesh, where optimism about such wonderfully combined properties can be infinite as long as you stay ignorant of underlying structures that produce some properties but not others

and, like, maybe you *can* get something as hard as steel, as shiny as gold, and resilient or self-healing in various ways, but you sure don't get it by ignorance of the internals

and not for a while

so if you need the magic sword in 2 years or the world ends, you're kinda dead

[Shah][14:46]

Potentially dumb question: when humans do science, why don't they then try to take over the world to do the best possible science? (If humans are doing dangerous goal-directed optimization when doing science, why doesn't that lead to catastrophe?)

You could of course say that they just aren't smart enough to do so, but it sure feels like (most) humans wouldn't want to do the best possible science even if they were smarter

I think this is similar to a question I asked before about plans being dangerous independent of their source, and the answer was that the source was misaligned

But in the description above you didn't say anything about the thing-doing-science being misaligned, so I am once again confused

[Yudkowsky][14:48]

boy, so many dumb answers to this dumb question:

- even relatively "smart" humans are not very smart compared to other humans, such that they don't have a "take over the world" option available.
- most humans who use Science were not smart enough to invent the underlying concept of Science for themselves from scratch; and Francis Bacon, who did, sure did want to take over the world with it.
- groups of humans with relatively more Engineering sure did take over large parts of the world relative to groups that had relatively less.
- Eliezer Yudkowsky clearly demonstrates that when you are smart *enough* you start trying to use Science and Engineering to take over your whole future lightcone, the other humans you're thinking of just aren't that smart, and, if they were, would inevitably converge towards Eliezer Yudkowsky, who is really a very typical example of a person that smart, even if he looks odd to you because you're not seeing the population of other [dath ilani](#)

I am genuinely not sure how to come up with a less dumb answer and it may require a more precise reformulation of the question

[Shah][14:50]

But like, in Eliezer's case, there is a different goal that is motivating him to use Science and Engineering for this purpose

It is not the prediction-goal that he instantiated in his mind as part of the method of doing Science

[Yudkowsky][14:52]

sure, and the mysterious formless thing within GPT-5 with "adaptive computation time" that broods and thinks, may be pursuing its prediction-subgoal for the sake of other goals, or be pursuing different subgoals of prediction separately without ever once having a goal of prediction, or have 66,666 different shards of desire across different kinds of predictive subproblems that were entrained by gradient descent which does more brute memorization and less Occam bias than natural selection

oh, are you asking why humans, when they do goal-oriented Science for the sake of their other goals, don't (universally always) stomp on their other goals while pursuing the Science part?

[Shah][14:54]

Well, that might also be interesting to hear the answer to -- I don't know how I'd answer that through an Eliezer-lens -- though it wasn't exactly what I was asking

[Yudkowsky][14:56]

basically the answer is "well, first of all, they do stomp on themselves to the extent that they're stupid; and to the extent that they're smart, pursuing X on the pathway to Y has a 'natural' structure for not stomping on Y which is simple and generalizes and obeys all the coherence theorems and can incorporate arbitrarily fine wiggles via epistemic modeling of those fine wiggles because those fine wiggles have a very compact encoding relative to the epistemic model, aka, predicting which forms of X lead to Y; and to the extent that group structures of humans can't do that simple thing coherently because of their cognitive and motivational partitioning, the group structures of humans are back to not being able to coherently pursue the final goal again"

[Shah][14:58]

(Going back to what I meant to ask) It seems to me like humans demonstrate that you can have a prediction goal without that being your final/terminal goal. So it seems like with AI you similarly need to talk about the final/terminal goal. But then we talked about GPT and debate and so on for a while, and then you explained how GPTs would have deep patterns that do dangerous optimization, where the deep patterns involved instantiating a prediction goal. Notably, you didn't say anything about a final/terminal goal. Do you see why I am confused?

[Yudkowsky][15:00]

so you can do prediction because it's on the way to some totally other final goal - the way that any tiny superintelligence or superhumanly-coherent agent, if an optimization method somehow managed to run across *that* early on, with an arbitrary goal, which also understood the larger picture, would make good predictions while it thought the outer loop was probably doing gradient descent updates, and bide its time to produce rather different "predictions" once it suspected the results were not going to be checked given what the inputs had looked like.

you can imagine a thing that does prediction the same way that humans optimize inclusive genetic fitness, by pursuing dozens of little goals that tend to cohere to good prediction in the ancestral environment

both of these could happen in order; you could get a thing that pursued 66 severed shards of prediction as a small mind, and which, when made larger, cohered into a utility function around the 66 severed shards that sum to something which is not good prediction and which you could pursue by transforming the universe, and then strategically made good predictions while it expected the results to go on being checked

[Shah][15:02]

OH you mean that the outer objective is prediction

[Yudkowsky][15:02]

?

[Shah][15:03]

I have for quite a while thought that you meant that Science involves internally setting a subgoal of "predict a confusing part of reality"

[Yudkowsky][15:03]

it... does?

I mean, that is true.

[Shah][15:04]

Okay wait. There are two things. One is that GPT-3 is trained with a loss function that one might call a prediction objective for human text. Two is that Science involves looking at a part of reality and figuring out how to predict it. These two things are totally different. I am now unsure which one(s) you were talking about in the conversation above

[Yudkowsky][15:06]

what I'm saying is that for GPT-5 to successfully do AGI-complete prediction of human text about Science, gradient descent must identify some formless thing that does Science internally in order to optimize the outer loss function for predicting human text about Science

just like, if it learns to predict human text about multiplication, it must have learned something internally that does multiplication

(afk, lunch/dinner)

[Shah][15:07]

Yeah, so you meant the first thing, and I misinterpreted as the second thing

(I will head to bed in this case -- I was meaning to do that soon anyway -- but I'll first summarize.)

[Yudkowsky][15:08]

I am concerned that there is still a misinterpretation going on, because the case I am describing is both things at once

there is an outer loss function that scores text predictions, and an internal process which for purposes of predicting what Science would say must actually somehow do the work of Science

[Shah][15:09]

Okay let me look back at the conversation

is there a little goal-oriented mind inside there that solves science problems the same way humans solve them, by engineering mental constructs that serve a goal of prediction, including backchaining for prediction goals and forward chaining from alternative hypotheses / internal tweaked states of the mental construct?

Here, is the word "prediction" meant to refer to the outer objective and/or predicting what English sentences about Science one might say, or is it referring to a subpart of the Process Of Science in which one aims to predict some aspect of reality (which is typically not in the form of English sentences)?

[Yudkowsky][15:20]

it's here referring to the inner Science problem

[Shah][15:21]

Okay I think my original understanding was correct in that case

from my perspective the whole field of mental endeavor as practiced by alignment optimists consists of ancient alchemists wondering if they can get collections of surface properties, like a metal as shiny as gold, as hard as steel, and as self-healing as flesh, where optimism about such wonderfully combined properties can be infinite as long as you stay ignorant of underlying structures that produce some properties but not others

I actually think something like this might be a crux for me, though obviously I wouldn't put it the way you're putting it. More like "are arguments about internal mechanisms more or less trustworthy than arguments about what you're selecting for" (limiting to arguments we actually have access to, of course in the limit of perfect knowledge internal mechanisms beats selection). But that is I think a discussion for another day.

[Yudkowsky][15:29]

I think the critical insight - though it has a format that basically nobody except me ever visibly invokes in those terms, and I worry maybe it can only be taught by a kind of life experience that's very hard to obtain - is the realization that *any* consistent reasonable story about underlying mechanisms will give you less optimistic forecasts than the ones you get by freely combining surface desiderata

[Shah] [1:38] (next day, Nov. 14)

(For the reader, I don't think that "arguments about what you're selecting for" is the same thing as "freely combining surface desiderata", though I do expect they look approximately the same to Eliezer)

Yeah, I think I do not in fact understand why that is true for any consistent reasonable story.

From my perspective, when I posit a hypothetical, you demonstrate that there is an underlying mechanism that produces strong capabilities that

generalize combined with real world knowledge. I agree that a powerful AI system that we build capable of executing a pivotal act will have strong capabilities that generalize and real world knowledge. I am happy to assume for the purposes of this discussion that it involves backchaining from a target and forward chaining from things that you currently know or have. I agree that such capabilities could be used to cause an existential catastrophe (at least in a unipolar world, multipolar case is more complicated, but we can stick with unipolar for now). None of my arguments so far are meant to factor through the route of "make it so that the AGI can't cause an existential catastrophe even if it wants to".

The main question according to me is why those capabilities are aimed towards achievement of a misaligned goal.

It feels like when I try to ask why we have misaligned goals, I often get answers that are of the form "look at the deep patterns underlying the strong capabilities that generalize, obviously given a misaligned goal they would generate the plan of killing the humans who are an obstacle towards achieving that goal". This of course doesn't work since it's a circular argument.

I can generate lots of arguments for why it would be aimed towards achievement of a misaligned goal, such as (1) only a tiny fraction of goals are aligned; the rest are misaligned, (2) the feedback we provide is unlikely to be the right goal and even small errors are fatal, (3) lots of misaligned goals are compatible with the feedback we provide even if the feedback is good, since the AGI might behave well until it can execute a treacherous turn, (4) the one example of strategically aware intelligence (i.e. humans) is misaligned relative to its creator. (I'm not saying I agree with these arguments, but I do understand them.)

Are these the arguments that make you think that you get misaligned goals by default? Or is it something about "deep patterns" that isn't captured by "strong capabilities that generalize, real-world knowledge, ability to cause an existential catastrophe if it wants to"?

24. Follow-ups

[Yudkowsky][15:59] (Feb. 21, 2022)

So I realize it's been a bit, but looking over this last conversation, I feel unhappy about the MIRI conversations sequence stopping exactly here, with an unanswered major question, after I ran out of energy last time. I shall attempt to answer it, at least at all. CC @rohin @RobBensinger .

[Shah: 😊] [Ngo: 😊] [Bensinger: 😊]

One basic large class of reasons has the form, "Outer optimization on a precise loss function doesn't get you inner consequentialism explicitly targeting that outer objective, just inner consequentialism targeting objectives which empirically happen to align with the outer objective given that environment and those capability levels; and at some point sufficiently powerful inner consequentialism starts to generalize far out-of-distribution, and, when it does, the consequentialist part generalizes much further than the empirical alignment with the outer objective function."

This, I hope, is by now recognizable to individuals of interest as an overly abstract description of what happened with humans, who one day started building Moon rockets without seeming to care very much about calculating and maximizing their personal inclusive genetic fitness while doing that. Their capabilities generalized much further out of the ancestral training distribution, than the empirical alignment of those capabilities on inclusive genetic fitness in the ancestral training distribution.

One basic large class of reasons has the form, "Because the real objective is something that cannot be precisely and accurately shown to the AGI and the differences are systematic and important."

Suppose you have a bunch of humans classifying videos of real events or text descriptions of real events or hypothetical fictional scenarios in text, as desirable or undesirable, and assigning them numerical ratings. Unless these humans are perfectly free of, among other things, all the standard and well-known cognitive biases about eg differently treating losses and gains, the value of this sensory signal is not "The value of our real CEV rating what is Good or Bad and how much" nor even "The value of a utility function we've got right now, run over the real events behind these videos". Instead it is in a systematic and real and visible way, "The result of running an error-prone human brain over this data to produce a rating on it."

This is not a mistake by the AGI, it's not something the AGI can narrow down by running more experiments, the *correct answer as defined* is what contains the alignment difficulty. If the AGI, or for that matter the outer optimization loop, *correctly generalizes* the function that is producing the human feedback, it will include the systematic sources of error in that feedback. If the AGI essays an experimental test of a manipulation that an ideal observer would see as "intended to produce error in humans" then the experimental result will be "Ah yes, this is correctly part of the objective function, the objective function I'm supposed to maximize sure does have this in it according to the sensory data I got about this objective."

People have fantasized about having the AGI learn something other than the true and accurate function producing its objective-describing data, as its actual objective, from the objective-describing data that it gets; I, of course, was the first person to imagine this and say it should be done, back in 2001 or so; unlike a lot of latecomers to this situation, I am skeptical of my own proposals and I know very well that I did not in fact

come up with any reliable-looking proposal for learning 'true' human values off systematically erroneous human feedback.

Difficulties here are fatal, because a true and accurate learning of what is producing the objective-describing signal, will correctly imply that higher values of this signal obtain as the humans are manipulated or as they are bypassed with physical interrupts for control of the feedback signal. In other words, even if you could do a bunch of training on an outer objective, and get inner optimization perfectly targeted on that, the fact that it was perfectly targeted would kill you.

[Bensinger][23:15] (Feb. 27, 2022 follow-up comment)

This is the last log in the [Late 2021 MIRI Conversations](#). We'll be concluding the sequence with a public [Ask Me Anything](#) (AMA) this Wednesday; you can start posting questions there now.

MIRI has found the Discord format useful, and we plan to continue using it going into 2022. This includes follow-up conversations between Eliezer and Rohin, and a forthcoming conversation between Eliezer and Scott Alexander of [Astral Codex Ten](#).

Some concluding thoughts from Richard Ngo:

[Ngo][6:20] (Nov. 12 follow-up comment)

Many thanks to Eliezer and Nate for their courteous and constructive discussion and moderation, and to Rob for putting the transcripts together.

This debate updated me about 15% of the way towards Eliezer's position, with Eliezer's arguments about the difficulties of coordinating to ensure alignment responsible for most of that shift. While I don't find Eliezer's core intuitions about intelligence too implausible, they don't seem compelling enough to do as much work as Eliezer argues they do. As in the Foom debate, I think that our object-level discussions were constrained by our different underlying attitudes towards high-level abstractions, which are hard to pin down (let alone resolve).

Given this, I think that the most productive mode of intellectual engagement with Eliezer's worldview going forward is probably not to continue debating it (since that would likely hit those same underlying disagreements), but rather to try to inhabit it deeply enough to rederive his conclusions and find new explanations of them which then lead to clearer object-level cruxes. I hope that these transcripts shed sufficient light for some readers to be able to do so.

Late 2021 MIRI Conversations: AMA / Discussion

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

With the release of [Rohin Shah and Eliezer Yudkowsky's conversation](#), the **Late 2021 MIRI Conversations** sequence is now complete.

This post is intended as a generalized comment section for discussing the whole sequence, now that it's finished. Feel free to:

- raise any topics that seem relevant
- signal-boost particular excerpts or comments that deserve more attention
- direct questions to participants

In particular, Eliezer Yudkowsky, Richard Ngo, Paul Christiano, Nate Soares, and Rohin Shah expressed active interest in receiving follow-up questions here. The Schelling time when they're likeliest to be answering questions is **Wednesday March 2**, though they may participate on other days too.