3rd discussion
(Mix of quotations and summaries; also hasn't yet been evaluated by Eliezer)

[Exact quotations are marked with quotation marks and italics below. -*Ed.*]

Richard, summarized by Richard:

> Your claims about recursive self-improvement (RSI) from your debate with Hanson are similar to your current claims about consequentialism, in that:
> - Both of them focus on one very powerful abstraction, without accounting much for the ways in which such abstractions become far messier when they interact with the real world.
> - Because of this oversight, in both cases you argued that there would be dramatic jumps from low levels of this abstraction to high levels, without much of an intermediate period during which we can learn and prepare.
>
> I claim that these are mistakes, which led to RSI being less applicable than you thought (specifically because of the deep learning revolution) and will also make your arguments about consequentialism less applicable than you think.

Eliezer, summarized by Richard:

> Due to the deep learning revolution, it turned out that there were ways to get powerful capabilities without RSI. But this isn't intrinsically a (strong) strike against the RSI abstraction, since I was never claiming that RSI would be the way that AI scales to human-level generality. And even if we expect another surprising revolution before AGI, that wouldn't be a good reason to doubt the consequentialism abstraction, unless we knew *particular* properties of that revolution (especially because consequentialism is significantly simpler than RSI). But most people have no experience operating genuinely useful, genuinely deep generalizations that extend to nonobvious things (as opposed to blathering in a non-expectation-constraining way) so they think that I'm doing the latter when I talk about consequentialism and expected utility.

Richard, summarized by Richard:

> You don't need to know particular properties of a revolution in order to expect that it will undermine existing abstractions. And if your abstractions are genuinely deep and useful, then they should have significant predictive power. What novel empirical predictions does expected utility theory make? Are predictive successes in economics good evidence for the importance of utility theory?

Eliezer, quoted and summarized by Richard:

> "*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*" - similar to how evolutionary

> psychology is a deep explanation of many traits of human minds, even though it's difficult to extract novel predictions from it. But "*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.*"

Richard, quoted and summarized by Richard:

> "*I expect deeper theories to make more and stronger predictions,*" and shed light in unexpected ways (as was the case for evolution). "*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.*"

Eliezer, quoted and summarized by Richard:

> "*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!* [...] *These are* predictions of the theory" - just not advance predictions.

Richard, quoted by Richard:

> "*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* [...] *(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.*"

Eliezer, quoted and summarized by Richard:

> "*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.*" When it comes to general optimization processes, "*we have only two major datapoints*[...]: *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.*" During this debate I've also been careful to include "*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*".

Richard, summarized by Richard:

> The connotations of “miracle” implied such negligible probability that I didn’t interpret this as practical advice. What probability do you assign to us being saved by what you call a miracle? More or less than 10%?

Eliezer, quoted and summarized by Richard:

> “*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.* [... I]*f 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.*” If they do react, we should ask: what kind of heuristic could have correctly led us to forecast the US's reaction to covid? The best I’ve come up with is ““*The government will react with a flabbergasting level of incompetence, doing exactly the wrong thing, in some unpredictable specific way.*" *It seems to me like the best observed case for government reactions* [...] *was the degree of cooperation between the USA and Soviet Union about avoiding nuclear exchanges.* [...I]*t's not unreasonable to take this as the upper bound of attainable cooperation*[.]”

Richard, quoted and summarized by Richard:

> Is that degree of cooperation “*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?*” How about a world where the only two actors involved in AGI development are the UK and US, or the US and China?

Eliezer, quoted and summarized by Richard:

> I don't expect that much change to happen at all, “*or even come remotely close to happening; I expect AGI to kill everyone before self-driving cars are commercialized.* [...] *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.*”

Richard, quoted and summarized by Richard:

> “*What’s the least impressive technology that your model strongly rules out happening before AGI kills*” everyone? And what's “*the most impressive technology you expect to be commercialised before AGI kills everyone*”?

Eliezer, quoted by Richard:

“[H]*ere'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*[... I]*t 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"*[.]

“[...] *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.* [...] *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*” (e.g. ”*mRNA vaccines, building houses, building cities*”).