EXHIBIT I

Private channel #the-research-meeting - 2022-06-20 (UTC)

Private June 20th 2022, 4:00 AM UTC

Greg Brockman, Ilya Sutskever, Jakub Pachocki and 2 others

IS Ilya Sutskever

an interesting discussion on broader questions around AGI alignment (h/t john): https://www.alignmentforum.org/posts/CoZhXrhpQxpy9xw9y/where-i-agree-and-disagree-with-eliezer 6/20/2022, 4:00 AM UTC

> Where I agree and disagree with Eliezer - AI Alignment Forum
> (Partially in response toAGI Ruin: A list of Lethalities.Written in the same rambling style. Not exhaustive.) …

WZ Wojciech Zaremba (OpenAI)

it seems that

"I think Eliezer is probably wrong about how useful AI systems will become, including for tasks like AI alignment, before it is catastrophically dangerous. I believe we are relatively quickly approaching AI systems that can meaningfully accelerate progress by generating ideas, recognizing problems for those ideas and, proposing modifications to proposals, *etc.* and that all of those things will become possible in a small way well before AI systems that can double the pace of AI research. By the time AI systems can double the pace of AI research, it seems like they can greatly accelerate the pace of alignment research. Eliezer is right that this doesn't make the problem go away (if humans don't solve alignment, then why think AIs will solve it?) but I think it does mean that arguments about how recursive self-improvement quickly kicks you into a lethal regime are wrong (since AI is accelerating the timetable for both alignment and capabilities).
"

becomes a critical part of our strategy 6/20/2022, 4:57 AM UTC

✔ 1

IS Ilya Sutskever

it is a response by paul to a document by eliezer explaining why he thinks AGI will kill everyone 6/20/2022, 4:01 AM UTC

it's very detailed and not the easiest to summarize, but it's a useful way to better understand the existing thinking on these questions 6/20/2022, 4:02 AM UTC

@Alec Radford mentioned a possible concern around 4, namely that it will have memorized a ton of data and therefore will be insanely good at regurgitation. what are our official thoughts on this? 6/20/2022, 6:51 PM UTC

GB Greg Brockman

so we do have an attempt at regurgitation prevention (for frequently-occurring strings). however, i am not sure if we know how well it's doing 6/20/2022, 7:25 PM UTC

Jakub Pachocki

JP

we have evals for code regurgitation, looking as expected I believe. I've definitely seen e.g. dr Seuss poems regurgitated verbatim, but not sure how well it does on longer form prose...

it's actually interesting, I bet Harry Potter is hidden somewhere in the model weights, but I doubt we can recover it with T=0 sampling (and definitely not T=1)... 6/20/2022, 11:06 PM UTC

JP

Jakub Pachocki

tbf memorizing short, extremely popular, poems, jokes etc seems pretty much unavoidable 6/20/2022, 11:07 PM UTC

✔ 1

SA

Sam Altman

also tbf, humans do that too 6/20/2022, 11:08 PM UTC

+ 1