Twitter Thread by Eli Tyre

Eli Tyre
@EpistemicHope

My catch all thread for this discussion of AI risk in relation to Critical Rationalism, to summarize what's happened so far and how to go forward, from here.

I started by simply stating that I thought that the arguments that I had heard so far don't hold up, and seeing if anyone was interested in going into it in depth with me.

https://t.co/c62D4oQccR

> CritRats!
>
> I think AI risk is a real existential concern, and I claim that the CritRat counterarguments that I've heard so far (keywords: universality, person, moral knowledge, education, etc.) don't hold up.
>
> Anyone want to hash this out with me?https://t.co/Sdm4SSfQZv
>
> — Eli Tyre (@EpistemicHope) December 26, 2020

So far, a few people have engaged pretty extensively with me, for instance, scheduling video calls to talk about some of the stuff, or long private chats.

(Links to some of those that are public at the bottom of the thread.)

But in addition to that, there has been a much more sprawling conversation happening on twitter, involving a much larger number of people.

Having talked to a number of people, I then offered a paraphrase of the basic counter that I was hearing from people of the Crit Rat persuasion.

https://t.co/qEFxP7ia8u

> ELI'S PARAPHRASE OF THE CRIT RAT STORY ABOUT AGI AND AI RISK
>
> There are two things that you might call "AI".

> The first is non-general AI, which is a program that follows some pre-set algorithm to solve a pre-set problem. This includes modern ML.
>
> — Eli Tyre (@EpistemicHope) January 5, 2021

Folks offered some nit-picks, as I requested, but unless I missed some, no one objected to this as a good high level summary of the argument for why AI risk is not a concern (or no more of a concern than that of "unaligned people").

I spent a few days and wrote up a counter-counter augment, stating why I thought the that story doesn't actually hold up.

https://t.co/3hXMyBHdXh

The very short version:

1. Criticism of goals is always in terms of other goals

2. There are multiple stable equilibria in the space of goal structures, because agents generally prefer to keep whatever terminal goals they have. And because of this, there is path-dependency in goal structures.

3. AGIs will start from different "seed goals" than humans, and therefore reach different goal equilibria than humans and human cultures do, even if AGIs are criticizing and improving their goals.

My hope is, that in outlining how _I_ think goal criticism works, folks who think I'm wrong can outline an alternative story for how it works instead, that doesn't lead to doom.

Multiple people requested that I write up the positive case for AI doom (not just a counter-counter argument).

So, after taking into consideration threads from the previous document, and my conversations with people, I wrote up a second document, in which I outline...

why I expect AGIs to be hostile to humans, starting from very general principles.

https://t.co/9HmQ6a2S0j

The basic argument is:

1. Conflict is common, and violence is the default solution to conflict. Non-violent solutions are found only when one of two conditions obtain for agents in conflict, either non-violence is less costly than violence, or the agents...

...intrinsically care about the well being of the other agents in conflict.

2. For sufficiently advanced AGIs, violence will not be cheaper than non-violence.

3. By default, there are strong reasons to think that AGIs won't intrinsically care about human beings.

Therefore, we should expect sufficiently advanced AGIs to be hostile to humans.

This second essay, is, in my opinion, somewhat crisper, and less hand-wavy, so it might be a better place to start. I'm not sure.

Some things that would be helpful / interesting for me, going forward in this conversation:

1) Presuming you disagree with me about the conclusion, I would like to know which specific logical link in the "on hostility" argument doesn't hold.

2) Alternatively, I am super interested in if anyone has an alternative account of goal criticism that doesn't entail multiple equilibria in goal-structure space, so that all agents converge to the same morality in the limit.

(An account that is detailed enough that we can work through examples together, and I can see how we get the convergence in the limit.)

3) If folks have counterarguments that don't fit neatly into either of those frames, that also sounds great.

However, I request that you first paraphrase my counter-counterargument to my satisfaction, before offering third order counter arguments.

That is, I want to make sure that we're all on the same page about what the argument I'm making IS, before trying to refute and/or defend it.

I would be exited if people wrote posts, for those things, and am likewise excited to meet with people on calls for 1, 2, or 3.

There are also a bunch of other threads about Bayesianism and Universality and the technical nature of an explanation and the foundation of epistemology, that are also weaving in and out here.

I'm currently treating those as separate threads until they prove themselves relevant to this particular discussion on AI risk.

I also don't know what's most interesting to other people, in this space. Feel free to drop comments saying what YOU'RE hoping for.

Some public, in-depth conversations:

With @ella_hoeppner (Sorry about the volume differential. I think I'm just too loud : /)

https://t.co/DUvMi6wr6r

With @DorfGinger

https://t.co/slxXeDsAlo