BUZZ CHRONICLES > SOCIETY
Saved by @Alex1Powell
See On Twitter

# Twitter Thread by Eli Tyre

**Eli Tyre**
@EpistemicHope

### 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

> In general, I am super up for short (1 to 10 hour) adversarial collaborations.
>
> If you think I'm wrong about something, and want to dig into the topic with me to find out what's up / prove me wrong, DM me.
>
> — Eli Tyre (@EpistemicHope) December 23, 2020

For instance, while I heartily agree with lots of what is said in this video, I don't think that the conclusion about how to prevent (the bad kind of) human extinction, with regard to AGI, follows.

https://t.co/nbXUsXvcmW

There are a number of reasons to think that AGI will be more dangerous than most people are, despite both people and AGIs being qualitatively the same sort of thing (explanatory knowledge-creating entities).

And, I maintain, that because of practical/quantitative (not fundamental/qualitative) differences, the development of AGI / TAI is very likely to destroy the world, by default.

(I'm not clear on exactly how much disagreement there is. In the video above, Deutsch says "Building an AGI with perverse emotions that lead it to immoral actions would be a crime."

I wouldn't usually put it in those words, but THAT is what the alignment problem is about:

We don't yet know how to reliably build AGI systems _without_ "perverse emotions." It seems like that might be pretty hard to avoid.)

But maybe I'm misunderstanding these arguments.

I would love to dig into this with someone who thinks that AI is not a serious existential risk for reasons related to the above, and together try and answer the question of how these AI is most likely to go.

My win conditions:

1. I change my mind about AI risk, in some way
2. I understand some new-to-me argument that I need to think about in depth
3. I viscerally "get" what I'm missing from the CritRat frame
4. There's a public refutation of the arguments that turn out to be flawed

Here's @reasonisfun vouching for me.

https://t.co/N3S5mER8Z9

> Crit rats!
>
> Eli\u2019s very nice to chat with (Bayesian/CFAR background) \u2014 curious, sharp and kind, would recommend: https://t.co/lNToaqOusr
>
> — Lulie \U0001f384 (@reasonisfun) September 2, 2020

I'm happy to talk to you even if your view is not fully representative of "all Critical Rationalists".

@DavidDeutschOxf @iamFilos @campeters4 @MatjazLeonardis @sashintweets @HermesofReason @adilzeshan @thenumber8008 @mansfield_pablo
@DorfGinger @ks445599 @jchalupa_ @RealtimeAI @JimiSommer @chuggfest
Feel free to share with whoever is most likely to be interested.

Feel free to DM me, if you're interested.