Ideas from Benedict Evan's weekly newsletter.

A significant part of the AI research world today, and especially the part that worries about some kind of

> existential risk and AGI, comes from a fairly tightly knit and incestuous set of subcultures in the San Francisco Bay Area, who've been saying these things for a decade.

This is very similar to crypto, which, again, came out of small groups of people who spent a long time talking to each other and telling each other that they were right (and also telling each other how clever they were). The internet in the early 1990s was a bit like this too, and all three came with some very fringe politics (there will be no laws on the internet, money is speech and 'fiat currency is evil, say), combined with a tendency towards a

> pretty basic lack of understanding of the complexity of the real work outside Berkeley. This is selection bias -
>
> ....ideas crazy enough to change the world (as the internet did!) sometimes have to come from 'crazy' people, and if they work, the crazy gets left behind.

The Microsoft engineer who wrote Microsoft Money was a *brilliant* engineer, and he refused to have a bank account.

This perspective is useful when listening to

> a certain kind of AI person talk about how this might all kill us all really soon: they really believe this, yes, but be careful not to confuse authority in talking about how the algorithms might be improved with authority in understanding how any of that might work in the real world.

One aspect of just how cultish and in-group parts of that field can be was hinted at in this Politico article: in between the bickering over who said what to Biden about AI policy (who cares?), Marc Andreessen told them that the doomers were a sex cult. And this New York Post piece, which, of course, one should take with a very large amount of salt, explains what he meant. POLITICO, POST

Mustafa Suleyman, one of the cofounders of DeepMind and now head of 'Microsoft AI', wrote an essay on AI safety with a strong and very specific criticism of the training approach taken by some other labs, especially Anthropic.

> His point is that their training techniques explicitly tell the model that it has feelings and opinions, and then the researchers 'observe' that the model 'seems' to be showing feelings and emotions: this is circular, with the model reflecting back what the researchers expect to see.

People have a very strong tendency to anthropomorphise almost anything (we see faces in clouds, which incidentally is a core weakness of the Turing Test), and

> if we presume the system is ‘conscious’ and build it with explicitly instructions to act as though it’s conscious, and then look at it and wonder if we can see consciousness, that way lies psychosis. LINK

Suleyman addressed Anthropic, but you can see a lot of the same anthropomorphism in OpenAI’s new framework for ‘misalignment’. These are bugs and engineering failures. When Word crashed and swallowed your work, we didn’t say that the software ‘decided’ to do that. LINK