Alignment For Foxes

1. Reflections on My Own Missing Mood
2. Parable: The Bomb that doesn't Explode

Reflections on My Own Missing Mood

Life on Earth is incredibly precious. Even a tiny p(DOOM) must be taken very seriously. Since we cannot predict the future, and since in a debate we should be epistemically humble, we should give our interlocutors the benefit of the doubt. Therefore, even if on my inside view p(DOOM) is small, I should act like it is a serious concern. If the world does end up going down in flames (or is converted into a giant ball of nanobots), I would not want to be the idiot who said "that's silly, never gonna happen."

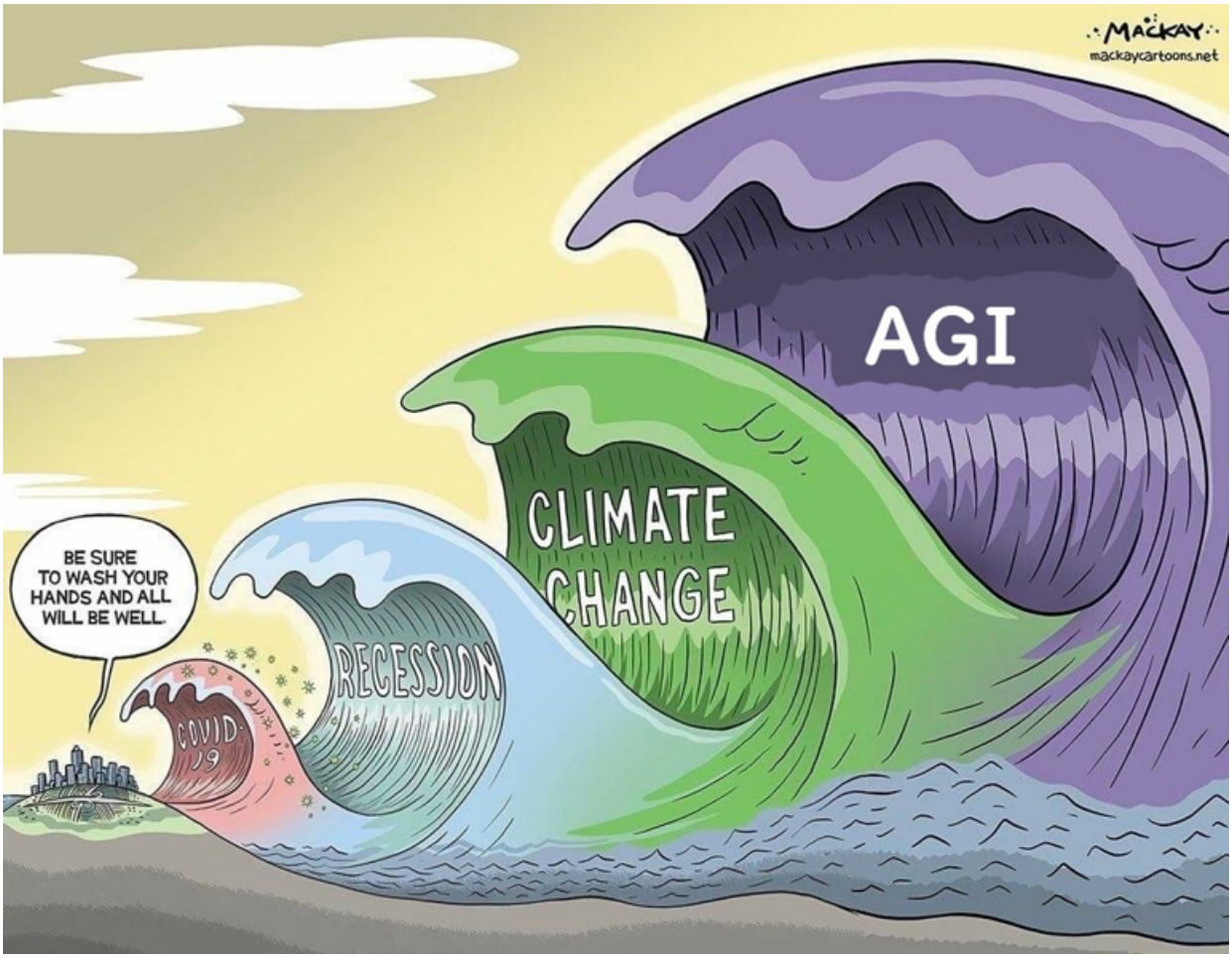
The problem is my emotional state can really only reflect my inside view and my intuitions. To account for the very real probability that I'm wrong and just an idiot, I can agree to consider, for the purposes of discussion, a p(DOOM) of say, 0.25, which is equivalent to billions of people dying in expected value. However, I can't make myself feel like the world is doomed if I don't actually think it is. This is the origin, I think, of missing moods.

For me personally, I have always had an unusually detached grim-o-meter. Even if an unstoppable asteroid were heading to Earth and certainly we all were going to die, I probably would not freak out. I'm just pretty far out on that spectrum of personality. Given that, and given that it makes me a bit reckless in my thinking, I'm very grateful that the world is full of people who *do* freak out (in appropriate and productive ways) when there is some kind of crisis. Civilization needs some people pointing to problems and fervently making plans to address them, and civilization also (at least occasionally) needs people to say "actually it's going to be okay." Ideally, public consensus will meet in the middle and somewhat close to the truth.

I've had three weeks to reflect on EY's post. That post, along with the general hyper-pessimism around alignment, does piss me off somewhat. I've thought about *why* this pessimistic mood frustrates me, and I've come up with a few reasons.

Crisis Fatigue

Here is a highly scientific chart of current world crises:



My younger sister is quite pessimistic about climate change. Actually scratch that: it's not climate change that she's worried about, it's biosphere collapse. I'm conflicted in my reaction to her worries. This is strikingly similar to my internal conflict over the alignment crisis. On the one hand, I'm very optimistic about new energy technologies saving the day. On the other hand, biosphere collapse is just about the most serious thing one could imagine. Even if p(Dead Earth) = 0.01, that would still be a legitimate case for crisis thinking.

With so many crises going on today, and with a media environment that incentivizes hyping up your pet crisis, everyone is a bit fatigued. Therefore, our standards are high for accommodating a new crisis into our mental space. To be taken seriously, a crisis must be 1) plausible, 2) not inevitable and 3) actually have massive consequences if it comes to pass. In my opinion, AGI and the biosphere threat both check all three boxes, but none of the other contender crises fit the bill. Given the seriousness of those two issues, the public would be greatly benefited if all the minor crises and non-crises that we keep hearing about could be demoted until things cool off a bit, please and thank you.

To be clear, I'm putting AGI in the "worth freaking out about" category. It's not because of crisis fatigue that I'm frustrated with the pessimistic mood around alignment. It's because...

Pessimism is not Productive

Given crisis fatigue, people need to be selective about what topics they engage with. If a contender crisis is not likely to happen, we should ignore it. If a contender crisis is likely to happen, but the consequences are not actually all that serious, we should ignore it. And (this is the kicker) if a contender crisis is inevitable and there is nothing we can do about it, we have no choice but to ignore it.

If we want the public (or just smart people involved in AI) to take the alignment problem seriously, we should give them the impression that p(DOOM | No Action) > 0.05 AND p(DOOM | Action) < 0.95. Otherwise there is no reason to act: better to focus on a more worthy crisis, or just enjoy the time we have left.

If you genuinely feel that p(DOOM | Action) > 0.95, I'm not sure what to say. But please please please, if p(DOOM | Action) < 0.95, do not give people the opposite impression! I want you to be truthful about your beliefs, but to the extent that your beliefs are fungible, please don't inflate your pessimism. It's not helpful.

This is why pessimists in the alignment debate frustrate me so much. I cannot manufacture an emotional reaction that I actually do not feel. Therefore, I need other people to take this problem seriously and create the momentum for it to be solved. When pessimists take the attitude that the crisis is virtually unsolvable, they doom us to a future where little action is taken.

Self-Fulfilling Prophecies

And genuinely freaking out about the alignment crisis can certainly make things worse. In particular, the suggestion of a 'pivotal act' is extremely dangerous and disastrous from a PR perspective.

(I'm keeping this section short because I think Andrew_Critch's post is better than anything I could write on the topic.)

We Need a Shotgun Approach

The alignment crisis could be solved on many, many levels: Political, Regulatory, Social, Economic, Corporate, Technical, Mathematical, Philosophical, Martial. If we are to solve this, we need all hands on deck.

It seems to me that, in the effort to prove to optimists (like me) that p(DOOM) is large, a dynamic emerged. An optimist would throw out one of many arguments: "Oh well it'll be okay because..."

- we'll just raise AI like a baby
- we'll just treat AI like a corporation
- it'll be regulated by the government
- social pressure will make sure that corporations only make aligned AIs
- we'll pressure corporations into taking these problems seriously
- AI will be aligned by default
- AIs will emerge in a competive, multi-polar world
- AIs won't immediately have a decisive strategic advantage
- there will be a slow takeoff and we'll have time to fix problems as they emerge

- we'll make sure that AIs emerge in a loving, humanistic environment
- we'll make sure that AIs don't see dominating/destroying the world as necessary or purposeful
- we'll design ways to detect when AIs are deceptive or engaging in power-seeking
- we'll airgap AIs and not let them out
- we'll just pull the plug
- we'll put a giant red stop button on the AI
- we'll make sure to teach them human values
- we'll merge with machines (Nueralink) to preserve a good future
- we'll build AIs without full agency (oracle or task-limited AI)
- it is possible to build AI without simplistic, numerical goals, and that would be safer
- we'll use an oracle AI to build a safe AI
- usually these kinds of things work themselves out

(This is definitely a complete list and there is nothing to add to it.)

In an attempt to make optimists take the problem more seriously, pessimists in this debate (and in particular EY) convinced themselves that all these paths are non-viable. For the purposes of demonstrating that p(DOOM | No Action) > 0.05, this was important. But now that the debate has shifted into a new mode, pessimists have retained an extremely dismissive attitude towards all these proposed solutions. It seems that this community has a more dismissive attitude in inverse proportion to how technical (math-y) the proposed solution is. I contend this is counterproductive both for bringing more attention to the problem and for actually solving it.

Moving Forward

There may not be much time left on the clock. My suggestion for pessimists is this: going forward, if you find yourself in a debate with an optimist on the alignment problem, ask them what they believe p(DOOM | No Action) is. If they say < 0.05, by all means, get into that debate. Otherwise, pick one of the different paths to addressing the problem, and hash that out. You should pick a path that both parties see as plausible. In particular, don't waste time dismissing solutions just because they are not in your personal toolbox. For example if regulatory policy is not something you nerd out about, don't inadvertently sow FUD about regulation as a solution just because you don't think it's going to successful. If you can raise p(AWESOME GALACTIC CIVILIZATION) by a fraction of a percent, that would be equivalent to trillions of lives in expected value. There is no better deal than that.

Parable: The Bomb that doesn't Explode

You're an engineer working for a military contractor. One day, your project manager comes to you and asks you to design a container for plastic explosive. It is to contain several kilograms of C4, enough to destroy a building. This is pretty dangerous, but you know that C4 is actually pretty safe. It can't be accidentally detonated by fire, impact or bullets. Only a detonator (another explosive) can trigger C4, so you don't include a detonator in your design, for safety.

Your PM reviews the design and gives you some feedback. "It would be a lot more useful if you put a blasting cap inside." You grimace. *More useful, yes. But a lot less safe.* Nevertheless, you do your job and install a blasting cap. The blasting cap has two electrical leads. If a voltage is applied across those leads, the C4 will explode, killing everyone around. To keep things safe, you snip off the the leads and put the blasting cap inside a pill-shaped plastic container, inside the larger container that contains the C4.

"Well, that's no good," your PM replies, "What if someone *does* want to put a voltage across the leads?" You grimace even more. *It is getting very difficult to make this design safe,* you think. But you have a solution: inside the pill-shaped plastic container, you install a Raspberry Pi. You install SELinux on the Pi and set it up to connect to WiFi -- but only secured networks. You program up a fancy web interface that allows the user to specify exactly what voltage they want to apply to the leads. The interface caps the voltage at a low level. It does not allow the user to apply a sufficient voltage to actually trigger the blasting cap -- or at least you hope that's how blasting caps work.

"Well, that's no good," your PM replies, "What if someone wants to apply a higher voltage?" You grimace again. *Then it'll explode! It'll kill everyone!* But this is not a workplace where you raise objections, so instead you just diligently do your job. If you don't make this dangerous thing, the company will hire someone else without your moral scruples, and certainly that person would make something really dangerous! So you alter the web interface, allowing the user to specify any voltage up to the limit of what the Pi can output. You add a big red warning screen, explaining how C4 is dangerous, and forcing the user to click "Yes, I'm really sure I want to apply this voltage." You also add a fancy cryptographic security key.

Your PM reviews the final design. She's a bit confused why this bomb is so damn complicated, but no matter, it serves its purpose well. The bomb goes on to kill someone whose name you cannot pronounce, thus defending the American people from great evil. (/s)

The moral of the story is, if you don't want a dangerous thing to get built, you have to actually not build it.