Supporting responsible AI: feedback

I am a researcher in the field of AI, and I'm glad to see the Australian government engage with the ramifications of this developing technology! I traveled from Australia to the United States 5 years ago now to work with folks here on the AI alignment problem. The motivating question is: how can we have powerful AI remain true to our values, and those of future generations?

The unfortunate truth is that we have no answer, and a long list of difficulties that make the problem look very tricky.

I write to you to flag two important considerations that I wish would inform policy choices as things unfold:

1. In the short term, AI companies are aiming to replicate human cognition in AI systems. They may largely succeed within 3-10 years.
2. In the long term, digital minds possess many advantages over our own, and could rob us of a future full of what we value. We (humanity) need to work to avoid that.

Some policy responses I think are warranted by this:

a. Instituting universal basic income, especially for labor replaced by AI.
b. Taking a collective moment to feel alarm, and notice the possibility that AI technology we are developing may replace us and what we value.
c. Shutting down frontier AI research that is not legibly safe. Specifically, research directed at replacing human thought with opaque systems.

If you have the time to read on, I will elaborate on these points.

The near-term: automating human thinking

If you look at the stated and implied beliefs of the leaders of AGI companies (here I'm thinking of DeepMind, OpenAI, and Anthropic), a few things stand out:

1. They believe "thinking as well as a human" is almost within reach of a deep-learning AI approach, perhaps just by training larger versions of existing models.
2. They assign substantial chance to achieving this goal within 5 years. Moreso within 10.
3. They expect their technology to be highly transformational.
4. They think the development of their technology risks human extinction.

Taken together this might seem like an absurd state of affairs! If you have the chance to interview people at these companies, it's worth asking them point-blank about these items.

In any case, these technology developers are trying to automate human cognition, replicating the thinking we can do inside AI systems. The LLM chatbots are showing us early signs of success, and think they will continue to succeed.

As this trend continues, we reach a point at which AI systems can ingest information and make decisions better than many human workers. We can now see large swathes of human cognitive labor replaced by AI systems.

(The bottleneck on this will be describing to the AI systems what tasks need to be performed and how. There will be some time between AI with the “raw capability” to act as a lawyer, and well-tuned AI lawyers that basically know what they’re doing and stay on-task. This makes the shift towards AI labor temporarily less abrupt.)

The longer-run: extinction risk

The end goal of the field of AI is machine intelligence that can go on to surpass our own. Unfortunately there are strong technical reasons to think this doesn’t end well for us, our children, and our future.

AI systems have no built-in access to human values. They aren’t made by copying things we value into computers, or making human brains that run faster. They are better thought of as alien intelligences, built the only way we know how out of lots of connected learning pieces. The pieces are hammered with crude tools that shape their *outer behaviour*, but not their *inner workings*.

For instance, controlling the outer behaviour of something like GPT-4 does not mean that you’ve solved the relevant challenges. Jailbreaks reveal that GPT-4 still has all the inner workings necessary to make nefarious plans, direct its thinking toward bioweapons, and etc.

So our tools remain very crude relative to the size and complexity of our AI systems. They don’t instill deep values and patterns of thought. They just give the systems the knowledge to behave like we ask, in a way that is visibly fragile.

Thoughts on what to do

In order to keep our future in our hands, we can’t keep building AI systems that are more complicated than we know how to understand or control. We can get away with it while they’re not very smart. But our AI companies have the explicit goal of making them smarter than us. This is a terrible idea while we have little clue about building deep benevolence into these systems.

What to do? I don’t have an adequate answer, but I know what good “first steps” would look like:

1. I think it's important that we take a collective moment of alarm, and orient toward the challenge.
2. We need a collective framework for how we relate to AI research efforts. We are currently not accounting for giant externalities in that work.
3. We need a re-negotiated relationship between world governments and AI labs. The current "do whatever you want" arrangement isn't adequately representing the interests of Australians or humankind.

I think world governments have an important role to play here, and there is a lot of room for leadership on all points. I would love to see the Australian government be a part of this effort, acting on AI to value our future generations, and all that could be.