Belief hierarchies that support the moratorium

Published on 11 May 2023 by Nik Samoylov

This is a supporting material used for our message testing work. Below are ***hypothetical (not validated)*** hierarchies of individuals' beliefs that support an AGI moratorium. We aim to understand the effectiveness of each belief structure, where the public stands on these beliefs, and what are the most effective ways to communicate along each belief hierarchy.

A logical case on lack of safety

Someone who is rationally convinced about the need for a moratorium will likely have the following **reasons**:

- **A1:** AI poses grave near-term dangers:
  - It can become or be made powerful in principle (i.e. be able to make big changes in the human environment).
  - Labs are able to advance these powerful AI capabilities quickly, in a matter of a few years.
  - At least one of:
    - It can be misused.
    - It can be critically misaligned.
- **A2:** Solutions are not currently available to counter these dangers:
  - Cannot simply switch off.
  - Do not currently have a way to encode complex human values in AI.
  - Do not currently have a bullet-proof way to prevent misuse.
- **A3:** There should be a stop/halt/pause on advancing these capabilities, at least until we have safety solutions.
- **A4:** The job of enforcing the stop should be with the governments because private actors are not sufficiently motivated on their own.
- **A5:** A moratorium is possible and can be effective.

Helpful, but not essential are beliefs regarding:

- Power concentration, job losses, dangers to democracy, etc. (ethics and economic concerns)
- Distrust for big tech companies
- Distrust for new technologies in general
- Human world is vulnerable and fragile
- AI does not need to be superhumanly smart in order to be very dangerous
- AI does not need to have creativity or consciousness in order to be very dangerous
- Current LLMs are indeed thinking, not rearranging previously memorized materials (i.e reject the "stochastic parrots" narrative)
- Advancement of AI capabilities is not inevitable
- AI labs make unsafe decisions (e.g. because of race dynamics between them)
- Anything that is smarter than people poses a grave danger
- The time of halt can be used for work on AI alignment and safety

On top of possessing these required and supporting beliefs:

- A person needs to walk through the chain of thought (or be walked through) in order for them to have a chance to come to the conclusion about the need for a moratorium.
- Objections need to be addressed.

A study on agreements with the constituent statements is available.

A Judeo-Christian spiritual case

- **B1:** Assessment of the situation needs to be close to:
  - People are assailing the role of God by creating a new reasoning being:
    - "Building a new Tower of Babel"
    - Association with cloning and bioengineering
  - People are building a new God
  - People are summoning Azazel or demons
- **B2:** Internal locus of control must be activated for prevention of evil:
  - Like: "Abhor what is evil; hold fast to what is good"
  - Not mere non-participation in evil: "For it is by grace you have been saved, through faith—and this is not from yourselves, it is the gift of God"
  - Not external locus of control: "God will not let it happen"
- **B3:** There should be a stop/halt/pause on advancing these capabilities
- **A4 + A5** same as above

Socio-economic and environmental reasoning

- **C1:** Further advancement of AI capabilities will likely lead to one or more of the following:
  - Concentration of power in the hands of big tech companies
  - Unacceptable proliferation of generated content (that will eclipse reality)
  - Detrimental economic consequences (for workers, companies, non-US governments)
  - Lock-in of unacceptable societal values (e.g. patriarchy)
  - Deterioration of the environment (e.g. data centers using processing powers)
- **C2:** There are no current solutions to these dangers
- **B3 + A4 + A5** same as above

An aesthetic rejection

- AGI is yucky or icky
- "I need to resist the emergence of AGI"
- It needs to be done in a grandiose way
- Moratorium is it

A case to avoid inhumane treatment of AGI

- Advanced AI may/can/will develop its own notion of suffering
- *Optional:* Because its cognitive powers are large, the suffering can be immense
- We can inflict suffering on it, either knowingly or unknowingly (and even lock in that suffering if it is done during training)
- We do not know how to avoid suffering
- The ethical thing to do is not to try to create a suffering machine
- **A4 + A5** same as above