Intro

The purpose of this document is to hash out a design of a survey to gauge people's priorities on AI risks.

It will introduce various concepts and see how much people wish to support exploration of that concept. Hopefully it will become clear as we go on.

Plan

Pick a place to host the survey
Refine document/questions
Post on Lesswrong and EA forum and relevant sub-reddits
Analyse and present statistics and then re-post on those locations

Questions

Intro questions

- Have you read SuperIntelligence
- Have you heard of SuperIntelligence
- What is the highest level of education you have achieved in an AI relevant profession (Computing/Neuroscience/Psychology)
- What is the highest level of education you have achieved
- How important do you think work into reducing the problems of AI risk are today.

Concepts

Questions for each concept

How strongly do you support researching this concept?

- I strongly oppose this research
- I only support this research with caveats
- I am ambivalent about this research
- I would like to talk about this research
- I would/am donating to help this research

- I would like to work on this subject
- I am working on this subject

If you are in favour of researching this concept is it because
- You think this concept is likely to be important
- You would prefer this outcome
- This outcome seems most dangerous
- Of a mixture of the above
- None of the above

If you are against research this concept is it because
- You think this concept is unlikely to be important
- You would not prefer this outcome
- This concept might lead to danger
- A mixture of the above
- None of the above

Satisficing

These systems do not have an in built goal they try and maximise. They might try and satisfy some abstract notion of goodness in a way analogous to reinforcement learning. The should be able to shift their ontology/paradigms as the notion of goodness is at a lower level.

Maximising

The agents are assumed to be able to maximise a goal related to the real world. E.g. the coherent extrapolated volition of humanity. It is currently unknown whether they can change their ontology/paradigms.

Proof-based Intelligence safety research

This is build mathematical models of agent based systems to try and figure out how to build them without having to do potentially dangerous tests in the real world.

Empirical Intelligence safety research

This would be building limited AI or IA systems and evaluating their failure modes and extrapolating them to more powerful systems.

Singleton safety research

It is possible that one Intelligence could dominate the entire world. It could mediate conflicts and make sure humanity does not wipe itself out. It could also make sure that evolutionary pressure does not make humans lose their humanity.

Multi-actor safety research

It might be that the way intelligence evolves that multiple entities will come to co-exist as peers over long periods of time. Research into this scenario might include trying to formulate political policy to minimise conflict between the actors. Or formulate developmental scenarios that will mean that the evolutionary pressures will not make the agents evolve to destroy humanity.

Artificial Intelligence safety research

Creating separate entities with their own motivation systems. May be easier to predict.

Intelligence Augmentation safety research

Work on safely improving human intelligence be it neural interfaces to adaptive computing (exo-brains) or improving brain processing.

Computer Security based research

One of the worries is that an advanced intelligence will be able to take over the whole internet and boost its intelligence greatly. This could be mitigated by shifting the paradigm of computing to be inherently more secure.