AI Existential Safety Research Definition

“AI existential safety research” refers to:

1. Research that analyzes the most probable ways in which AI technology could cause an existential catastrophe (that is: a catastrophe that permanently and drastically curtails humanity’s potential, such as by causing human extinction), and which types of research could minimize existential risk (the risk of such catastrophes). Examples include:

    1. Outlining a set of technical problems and arguments that their solutions would reduce existential risk from AI, or arguing that existing such sets are misguided.

    2. Concretely specifying properties of AI systems that significantly increase or decrease their probability of causing an existential catastrophe, and providing ways to measure such properties.

2. Technical research which could, if successful, assist humanity in reducing the existential risk posed by highly impactful AI technology to extremely low levels. Examples include:

    1. Research on interpretability and verification of machine learning systems, to the extent that it facilitates analysis of whether the future behavior of the system in a potentially different distribution of situations could cause existential catastrophes.

    2. Research on ensuring that AI systems have objectives that do not incentivize existentially risky behavior, such as deceiving human overseers or amassing large amounts of resources.

    3. Research on developing formalisms that help analyze advanced AI systems, to the extent that this analysis is relevant for predicting and mitigating existential catastrophes such systems could cause.

    4. Research on mitigating cybersecurity threats to the integrity of advanced AI technology.

    5. Solving problems identified as important by research as described in point 1, or developing benchmarks to make it easier for the AI community to work on such problems.

The following are examples of research directions that do not automatically count as AI existential safety research, unless they are carried out as part of a coherent plan for generalizing and applying them to minimize existential risk:

1. The mitigation of non-existential catastrophes, e.g. ensuring that autonomous vehicles avoid collisions, or that recidivism prediction systems do not discriminate based on race.

2. Increasing the general competence of AI systems, e.g. improving generative modelling, or creating agents that can optimize objectives in partially observable environments.