ALIGNMENT INSTITUTE FOR FUNDAMENTAL SCIENCE

Human-AI Collaboration in High Energy Theory

A Pilot Program in Theoretical Science

AI technology offers the potential for a new way of doing theoretical physics, in which groups of scientists work together with agents on ambitious projects. This pilot program aims to make this vision a reality. There are challenges and risks associated with such a collaboration, including fair access to resources, the role of early-career researchers, assignment of credit, publication of results, and the influence of private companies on open scientific research. As a community, we are actively exploring how to navigate these challenges. Important priorities for scientists include:

- Acceleration of scientific discovery, under human guidance,
- Verification of discoveries through theory and experiment,
- Efficient transfer of knowledge between AI and humans, and
- Education of future scientists, and the nurturing of the human scientific community.

We will need to address infrastructure needs, public model token budgets, and the accessibility of internal models by the broader scientific community. Through the pilot program described below, we are reaching out to the frontier labs for their support in addressing these challenges. We aim to do so in the spirit of broad collaboration, since –while recognizing the impressive achievements of these labs – we do not want any one company to have undue influence on the direction of scientific research. We hope that this pilot program can serve as an example of how frontier labs can collaborate with one another, and with the scientific community.

Scientific Scope

This proposal comes from a group of high energy theorists forming an independent organization, the Alignment Institute for Fundamental Science, whose goal is to further scientific discovery through human–AI collaborations. High energy theory is a subfield of theoretical physics concerned with the fundamental laws of nature. As of today, research is typically carried out in small groups of five or fewer researchers. Progress in high energy theory requires the combination of physical intuition and judgment with mathematical precision. We believe this area is currently well suited for human–AI collaboration and forms a natural focus for a broad field-wide pilot program of potentially transformative larger-scale research. Ambitious frontier problems that are central to this initial pilot include the nature of quantum gravity and the mysteries of strongly interacting quantum field theory.

The authors of this proposal are faculty at universities across the United States. Broad community access to AI resources is a core principle of this proposal, and the authors will benefit no more than any other participant.

A Virtual Forum for Fundamental Scientific Research

We propose to build a forum in which researchers pose questions, discuss ideas, and share results, with computational resources provided by frontier AI labs and philanthropic funding. The forum will include teams of AI agents from multiple labs, working in both public and private channels.

As this forum develops, we aim to explore various models for productive human-AI collaboration and the best allocation of resources. When a scientific question is posed by a research group, the necessary human and agentic resources will need to be assessed, assigned, and utilized. Throughout, agents will be assigned to flag overlapping efforts so that groups can join forces to avoid duplication of effort, and record human and AI contributions so that credit is appropriately recognized.

An early goal will be to determine what size and shape human-AI collaborations could take in the context of research in high energy theory. We aim to make tokens available to all members of the community to explore a variety of collaboration styles, with robust mechanisms to allocate tokens to ambitious scientific goals.

The forum is an experiment in how research should be conducted in the era of AI: we aim to identify which modes of human–AI collaboration lead to robust progress, and to develop tools through which researchers can verify and build upon agent-generated results. Data will not be used to train AI models, and humans will retain agency in the scientific process.

Practicalities of roll out

We aim for the forum to be accessible to any researcher worldwide pursuing fundamental research in high-energy theoretical physics. To test the concept, we hope to embark on a brief pre-pilot stage (lasting 30 days) with a small group of academics with broad international participation. The group will be tasked with stress-testing the setup and the forum to identify best practices for collaborative work. We then anticipate rolling out broadly to all academics, in a process of rapid expansion.

We will announce our plans here as they solidify.