Karlsruher Institut für Technologie

Institute for Technology Assessment and Systems Analysis

Expert assessment tender

Existential risk of AI: technical conditions

Discussions about possible existential risks associated with artificial intelligence (AI) are increasingly finding their way into scientific and political debates. The focus is on speculative scenarios that describe AI as potentially uncontrollable, potentially constantly self-improving, potentially misanthropic and a threat to the survival of humanity. For example, the Center for AI Safety (CAIS) has collected signatures from AI experts, the wider scientific community, and public figures for the following statement: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" (https://www.safe.ai/work/statement-on-ai-risk).

The aim of the expert assessment is to provide an overview of the current state of the art and trends. Accordingly, the focus of the report should not be on the speculative scenarios themselves, but on examining the conditions that need to be met for such scenarios to occur, and the extent to which these conditions have already been met. Accordingly, a careful distinction must be made between verified knowledge, well-founded assumptions and conjecture.

Please describe in your proposal how you intend to address the following tasks. You are welcome to include further research questions in your proposal – if you do so, please also state how you intend to approach these additional questions. Interim results must be presented and discussed with the client about halfway through the agreed period. The overall report as well as each individual chapter must contain a summary. It is essential that the report is understandable for non-computer scientists. It is also expected that the report will be published by the authors in a generally accessible and citable format as soon as possible after acceptance by ITAS.

1. State of research. Presentation and evaluation of the state of research, with particular attention to the following questions:

a. What capabilities/characteristics are cited in the literature as reasons for classifying AI as an existential risk, and how is this justified? Which of these capabilities do current AI applications already have, and to what extent?

b. What methods/procedures are available to demonstrate these capabilities/characteristics? What are the advantages and disadvantages of these methods/procedures?

c. Is there any point in testing individual capabilities/characteristics at all? How can you test for a potentially dangerous interaction of capabilities/characteristics?

d. What external conditions must be met for these capabilities/characteristics to lead to a loss of control? (access to resources, interconnectivity, ...)

2. State of research on Instrumental convergence. Instrumental convergence refers to the assumption that all intelligent agents, whether human or machine, develop instrumental intermediate goals to achieve their objectives. In the context of possible existential risks of artificial intelligence, the intermediate goals of self-protection and self-preservation, utility function or goal-content integrity, self-improvement and resource acquisition are often mentioned. Describe and evaluate the current state of research, paying particular attention to the following questions:

a. What is the status of the instrumental convergence thesis? Is it an assumption? Is it a conclusion that necessarily follows from the structure of advanced AI systems? Is it ...?

b. Is there any evidence or indication that the "instrumental convergence" thesis is correct?

c. To what extent does the current state of technology already allow the development of instrumental intermediate goals through AI, in particular those listed above?

3. Technical (counter-)measures. What technical measures are proposed to ensure that AI does not develop any of the capabilities/characteristics discussed under 1 and 2 or that such capabilities/characteristics could be controlled? Please assess the underlying rationale as well as the feasibility of the proposed measures.

Context

The expert assessment is intended to contribute to a better understanding of the existential risks of AI. It is part of the project “Systemic and Existential Risks of Artificial Intelligence”, which is funded by the Federal Ministry of Education and Research (BMBF) (funding reference 01IS23075). The project is being carried out by the Institute for Technology Assessment and Systems Analysis (ITAS) at the Karlsruhe Institute of Technology (KIT). The project's research aims to better identify, assess and avoid or mitigate existential risks of AI and to derive insights for its governance.

As the assessment is to be produced in an interdisciplinary project context, the presentation of the expert assessment should be comprehensible to an interdisciplinary audience. It is expected that the assessment will be published by the authors (one or more co-authors) in a citable format in a timely manner after approval by ITAS.

ITAS is the point of contact for all scientific questions around the project and responsible for reviewing and approving the final assessment. Willingness to engage in intensive discussions and close cooperation with ITAS is a prerequisite.

Renumeration and deadlines

The maximum renumeration for an expert assessment is EUR 70,000.

- The deadline for submitting proposals is March 31, 2025.
- Work on the expert assessment is intended to begin on May 01, 2025.
- The expert assessment must be submitted to ITAS by September 30, 2025.

Notes on the preparation of the proposal

The proposal can be written in German or English. ITAS will review and scientifically evaluate the proposals and award the expert assessment. In order for ITAS to be able to evaluate the quality of the proposals, qualitative criteria must be considered when preparing the proposal. These criteria will be given equal weight in the evaluation:

- The proposal must demonstrate and document the particular expertise of the specific scientific personnel employed in the requested subject area in a detailed, clear, well-founded and transparent manner. In particular, the relevant scientific and research experience and/or other outstanding competencies (including acknowledgements and

successes) in the subject area must be listed, both in terms of breadth and depth. Generally, this is to be demonstrated by presenting past projects with responsible accomplishment, activities relevant to the topic and (scientific) consulting services, as well as relevant publications.

- The overall quality of the content and form of the proposal will also be considered and evaluated. A clear structure is required. The planned effort and approach for preparing the assessment must be clarified and justified in a detailed and comprehensible manner. Aspects listed in the call ought to be considered and addressed (as completely as possible).
- The description of the intended methodological approach for achieving the scientific expertise and work results relevant to the assessment will also be assessed. The chosen methodology and its particular suitability for the purpose of the assessment must be presented clearly and justified convincingly. The relation between the respective work packages, allocated time, and delivered content must also be transparent, clear, and justified.

  Lastly, the price of the respective proposals is also considered in the evaluation.

Please note the mandatory information that needs to be included in the proposal (see below). Please send your proposal as an electronic version to the e-mail address provided under 'Contact'. In our experience, detailed proposals often require revisions, e.g. with respect to formalities or calculations. If we shortlist your proposal after reviewing it, we will ask you to make the necessary revisions and then to send a signed written proposal to ITAS (P.O. Box 3640, 76021 Karlsruhe, Germany).

If you are awarded the expert assessment, a contract between ITAS and you will be drawn up and signed.

Contact

Reinhard Heil

reinhard.heil@kit.edu

Notes on mandatory information

In order to comply with the formal regulations of the KIT for proposals, please use the following wording for your proposal:

Proposal to the Karlsruhe Institute of Technology (KIT),

Institute for Technology Assessment and Systems Analysis (ITAS)

The following information must be included in your proposal:

- Name and exact address (no P.O. Box) of the proposing institution or person; for providers who work at a university or comparable public institution, but propose as a private individual, the private address is required.
- Function, title, first name and surname of the provider or authorised signatory (representing the institution, e.g. the chancellor in the case of universities/colleges)
- Exact title of the assessment
- If applicable, the person responsible for the assessment
- Date of the proposal
- Processing period: from ... to ...
- Date of submission of the assessment. Please note that the final version of the assessment will be delivered as an electronic version (PDF), which also contains the original files of the tables and figures.
- Cost calculation including a separate VAT rate or a declaration that you are exempt from VAT. For personnel costs, the underlying time expenditure and estimated rates should be stated. The total price is treated as a fixed cost price.
- The proposal and further documents can be submitted electronically as PDFs.
- A short CV of the persons working on the project and, if applicable, a short introduction of the providing institution should be included as an attachment.