Systemic Risks of Interacting AI

Description of expert assessment

("Leistungsbeschreibung")

As part of the project "Systemic Risks of Artificial Intelligence" we are looking for experts to assess the state of the art of interacting AI and their potential systemic risks.

The understanding of systemic risks of AI seems to be in its beginnings. A generally recognized definition is still lacking. Threats to the functions of a system resulting from complex interactions between system elements are often seen as an important characteristic. Sectoral systems or industries (e.g., the financial system), the social system, or global systems are hold as relevant systems. There appear to be different types of systemic risks, which can differ, for example, in terms of their (un-)predictability during development or application of AI systems or whether damage can be attributed to a single actor.

To our understanding, systemic risks of AI can arise, for example, from the wide reach of AI providers or applications affecting large parts of society. They can also result from the propagation of potential harms from AI models in downstream applications or from the interaction between machines and/or humans. Furthermore, systemic risks may emerge from inappropriate stimuli or failures in collective behaviour leading to undesirable outcomes for society. However, the full range of systemic risks from AI may still be unknown, as they are largely unexplored and the technologies and applications of AI are developing rapidly.

Background

The release of ChatGPT has spurred a hype around Large Language Models (LLM) and AI in general. Since its inception, multiple models, products and applications followed the trend. The vast resources and investments that are mobilized for data center infrastructure suggest that this trend will not subside anytime soon. Recently, the attention of major AI players has shifted towards inference focused models with reasoning capabilities and agentic AI. The latter go hand in hand with a plethora of possible future application scenarios and imagined futures of fully-automated workflows and infrastructures.

AI agents are expected to independently assist in or autonomously perform specific tasks they were previously trained on. In this role they add to existing processes by replacing conventional algorithms or other technical solutions in more complex tasks and problems with multiple steps. In some cases, they are even expected to replace humans and their work altogether. The newest iteration of these agents is based on LLMs and Deep Learning techniques. However, the idea and technology behind the automation of tasks is older than LLM-based AI agents. Previous iterations were mostly based on variations of machine learning, expert systems and/or rule-based statistical algorithms in the course of the wider trend of digitization. Even multiple autonomous agents were used in so-called multi agent systems (MAS).

In this sense, agentic AI brings with it known use cases, potential benefits and risks that are comparable to previous waves of automation, but also new ones based on their inner workings and large amounts of processed data. Some of the more prominent risks are hallucinations, their opaque decision making and brittleness. In our project we are particularly interested in systemic risks. Interacting AI possibly amplifies the systemic risks present in isolated AI agents and their socio-technical context. Therefore, they include but are not limited to unintended feedback loops from misaligned models, inherent bias, privacy issues, collusion, interoperability issues, loss of control or specification gaming. Interacting AI are potentially more likely to affect more than a single context and their unintended effect could ripple through several domains, companies, infrastructures or society at large.

For the report we are interested in the state of the art of interacting AI, with a particular focus on systemic risks they currently pose and can be expected to exhibit in the reasonably foreseeable future. We are also interested in previous risks, failures and accidents of interacting (more or less) autonomous technologies and what can be learned from them. The report should however, focus on the specificities and genuine risks of the current iteration and its developmental trajectories.

Guiding questions

1) Collect and comprehensively present the types of interacting AI (including agentic AI, MAS, expert systems, algorithms and the layering of AI in hierarchical control structures and AI agents).
2) Map the types of systemic risks (e.g. material damage, societal harm or the violation of fundamental rights etc.) and assess their possible impact, if they should manifest.
3) How can these systemic risks and their potential impact be avoided, prevented and mitigated?

4) What can be learned from previous experiences regarding systemic risk causes and mitigation for algorithm-based systems of automated interaction and multi-agent systems. For instance, the role of high frequency trading in a flash crash, autonomous vehicle crashes or accidents with autonomous weapon systems. How do they differ? Can something similar be expected of agentic AI and to what extend?
5) How do the systemic risks of interacting AI change with the context they are deployed in or interact within? For instance, several interacting chatbots might be turned off, when they malfunction without considerable harm to the operation of the companies that employed them, but comparable interactions among autonomous agents in critical infrastructures might spell disaster (e.g. blackouts in the energy industry).

Context

The expert assessment is intended to contribute to a better understanding of the systemic 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 systemic 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.

Available financial means and deadlines

The maximum amount available for the expert assessment in the project is EUR 70,000.

- The deadline for submitting proposals is March 31, 2025.
- Work on the expert assessment is intended to begin on April 31, 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

Alexandros Gazos

alexandros.gazos@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 (doc or docx format), which also contains the original files of the tables and figures in the possible MS Office formats.
- 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.

We are looking forward to your proposal!