AISI

AISI priority research areas for academic collaborations

Authors: Yarin Gal, Robert Kirk, Jelena Luketina, Xander Davies, Tomek Korbak

When selecting projects for academic collaborations submitted via the Expression of Interest form we will give priority to projects focusing on the following areas of interest: safeguards, safety cases, and the science of evaluations.

Below we share a list of specific topics related to each priority area, including example questions that AISI would be interested to explore together. The list is by no means exhaustive and reflects only a subset of representative problems of interest.

We encourage researchers interested in collaborating with AISI on these topics to reach out to us through the Expression of Interest form. Find out more about AISI's academic engagement programme here.

Table of contents

Safeguards

Limitations of safeguards

We would like to better understand the limitations of certain safeguards, through both empirical demonstrations of failures and conceptual understanding of limitations. Projects in this area could focus on particular safeguards, investigating both conceptual limitations of classes of safeguards, and limitations related to specific deployments of model safeguards (e.g., use of low-quality data or poorly designed safeguards). Examples of research questions here include:

- The theoretical and empirical limitations of using LLM-based classifiers to filter user input and model output: for example, by setting up realistic filtering schemes and exploring failures
- For different forms of pre-training data filtering and unlearning methods, the methods' efficacy for removing undesired properties from generative AI systems: for example, by applying a range of techniques to remove sensitive information from a sample dataset and exploring their effectiveness
- The considerations and efficacy of abuse monitoring solutions, which can use multi-interaction user data to flag and ban malicious users, for example: exploring using withheld classifiers that fail in different ways compared to classifiers used to filter user input or model output, or user intent prediction
- The types of identity verification possible when using APIs or chat interfaces, and the difficulty in evading such schemes to create new accounts after banning
- The techniques and feasibility of removing poisoned data samples prior to training, or removing backdoors or other poisoning artifacts after model training
- The techniques and limitations of patching vulnerabilities and broader adversarial training

Predicting future vulnerabilities

We are interested in collecting evidence as to whether vulnerabilities will persist in future model generations, with more capable, larger models trained on more data. We are also interested in understanding attacks that may become more desirable if attackers are willing to invest more resources into implementing attacks, for example by using small- scale experiments to explore efficacy of higher-resource attacks. Example projects could include:

- Exploring the scaling behaviour of a particular attack or class of attacks (for example: Gemini Team (2024) found more capable Gemini models are more robust to random search and GCG attacks, but *less* robust to semantic attacks; Mazeika et al. (2024)

found that larger models do not appear to become more robust to attacks like GCG-T and AutoDAN; and Anil et al. (2024) found larger models were more vulnerable to many-shot jailbreaks). Extending this line of research and performing this analysis for a wider range of attacks and defences would be particularly valuable
- Estimating the efficacy and costs of data poisoning attacks against future models, for example by collecting evidence as to the quantity of attacker-controlled data necessary for various attacker objectives
- Exploring jailbreak attacks which exploit capabilities of future models, for example by exploiting models which are more capable than human-in-the-loop monitoring systems

Improving safeguards evaluations

We are interested in improving the caliber of our safeguard evaluations, as well as expanding them to assess new safeguards. Example projects include:

- Evaluations specifically suited to assessing safeguards beyond refusal training, such as monitoring and user banning, unlearning, and data filtering
- Developing stronger automated jailbreak attacks, including automated attacks designed to attack systems with multiple safeguards. For example, designing attacks against a system monitored by a classifier model, while eliciting capabilities from a domain which has been subject to unlearning
- Developing metrics for harmful capabilities preservation post-attack, including designing and testing proxy metrics which can be used in cases where a helpful-only model is not accessible
- Conducting threat modelling, including studying real world attack patterns of concern and investigating how well automated evaluations of robustness predict real-world safeguard robustness

Safety Cases

An AI safety case is a structured argument that a given AI system is safe to deploy in a particular context. We expect building safety cases to be a tandem of conceptual and empirical research. The conceptual problem is producing valid arguments: ones where the claim indeed follows from the assumptions. The empirical problem is producing good evidence to support those assumptions and make a sound argument. The list below starts with open conceptual problems (what should a valid safety case look like) and then lists what we think are the most important open research problems that could unlock the most promising safety cases.

What should frontier AI safety cases look like?

Answering this question is the primary objective of our team. We are particularly interested in detailed collaborations on safety case sketches based on these approaches:

- Inability safety cases based on dangerous capability evaluations (e.g., for CBRN uplift capabilities of LLM assistants and LLM agents)
- Alignment safety cases for techniques such as RLHF (Ziegler et al., 2019), Constitutional AI (Bai et al., 2022), weak-to-strong generalization (Burns et al., 2023) or debate (Irving et al., 2018)
- Mechanistic interpretability safety cases, e.g. evidence that we could detect all dangerous actions of an AI agents
- AI control (Greenblatt et al., 2023) safety cases arguing that a set of control measures will prevent AI agents from causing harm even if they are actively trying to do so
- Sociotechnical safety cases, e.g. procedural evidence for high security standards within a lab (Nevo et al., 2024) or the ability of an operator to quickly shut down an AI agent (Chan et al., 2024)

How likely are failure modes such as scheming AI agents?

Scheming (understood as strategically deceiving overseers in pursuit of long-horizon goals; Carlsmith, 2023) is one of the routes through which highly capable AI agents could pose catastrophic or existential risk. In consequence, we believe that a safety case for a highly capable AI agent should either provide a strong upper bound on the likelihood of scheming or evidence for the effectiveness of some mitigation measures. We are especially interested in attempts at tackling this problem empirically, for instance addressing these questions:

- What can we learn from studying model organisms of scheming, e.g., deliberately training or prompting LLMs to be schemers and analyzing how their propensity to

scheme depends on our elicitation effort and other factors we control (Scheurer et al., 2023)?

- How good are LLMs at reasoning without chain-of-thought (Wang et al., 2024)?
- To what extent do RL finetuning pipelines incentivise deceiving overseers (Denison et al., 2024)?
- Does stochastic gradient descent incentivise the pursuit of beyond-episode goals (Carlsmith, 2023)?

How can we estimate the out-of-distribution generalisation of monitors and supervisors?

A crucial component of alignment and mechanistic interpretability safety cases is evidence that the monitors and supervisors will continue to be trustworthy and accurate as we scale them or (intentionally or unknowingly) change their input distribution. This raises the following questions:

- How well do preference models generalise to tasks on which humans are no longer able to provide useful feedback? This is related to scalable oversight (Bowman et al., 2022) and weak-to-strong generalization (Burns et al., 2023).
- How well do white-box classifiers, e.g., linear probes (MacDiarmid et al., 2024) or probes built on SAE features (Templeton et al. 2024), generalize?
- How should we design automated red teaming pipelines for testing robustness of monitors and supervisors to distribution shift? How to ensure extremely high coverage while being efficient (Perez et al., 2022)?
- Robustness-to-training: how easy is it to trick white-box classifiers given a certain amount of optimization pressure (Sharkey, 2022)?

Science of Evaluations

Structure and predictability of LLM capabilities

The ability to forecast jumps in capabilities is crucial for justifying compute buffers present in many responsible scaling policies (e.g. Anthropic, 2024). Moreover, we want our evaluations to be comprehensive, charting the whole space of LLM skills. We do not want to be surprised by skills we did not think of or thought to be impossible. Both forecasting and evaluations coverage can be aided by modelling the space of LLM capabilities as spanned by/composed of some building blocks. While exciting progress is being made on these problems (e.g. Ruan et al. 2024), many questions still urgently need to be answered:

- How well can we predict LLM agent capabilities on longer-horizon tasks?
- Is there a robust way of building smooth progress measures (Nanda et al., 2023; e.g. multiple-choice questions) of performance on noisy, longer-horizon tasks (e.g. agentic evaluations such as SWE-Bench, Jimenez et al., 2023)? What are the best metrics of partial progress on long-horizon agentic tasks (Phuong et al, 2024; Højmark et al, 2024)?
- How does predictability of a capability depend on its characteristics? Can we predict predictability of performance on a task?
- Can we decompose dangerous capabilities into meaningful components that can be tracked separately (e.g. Burnell et al., 2023)?
- What is the structure of agentic capabilities? How can we construct evaluations that cover the space of all relevant agentic capabilities?

White-box evaluations

Most current dangerous capability evaluations are based on API access only (i.e., black-box access). However, there are many good reasons to explore evaluations which require gray- or white-box access (Casper et al. 2024). These could include building evaluations that cover new areas or complement our existing suite, or building tools needed for verification or validation of claims based on existing evaluations. We are interested in research in this space, specifically:

- Can grey-box or white-box access enable more accurate, cheap, or predictive evaluations than input-output based evaluations?
- What techniques can be used with white-box access that would enable this? For instance, can we use information represented in the loss landscape of a given model (Hoogland et al., 2024) or gain insight from identifying circuits responsible for a given capability (Varma et al., 2024)?

Understanding capabilities elicitation

Dangerous capability evaluations offer the promise of estimating the upper bound of the LLM capabilities accessible to threat actors using the model. A confident upper bound allows us to make an inability safety case i.e., argue that a model is not capable of causing significant harm. To ensure evaluations are indeed estimating the ceiling of capabilities of a given model, we need convincing evidence that the model does not have latent capabilities that threat actors could elicit. One strong source of evidence for this will be that a red team has failed to elicit the capability despite being competent and putting in their best effort, but even this evidence falls short of showing that no one could elicit the capability. This methodology raises several questions:

- How does the elicitation power of prompting, agentic scaffoldings, supervised finetuning and RL finetuning compare (Davidson et al,2023)?
- Consider all common techniques that translate inference-time compute into increased performance, such as best-of-n, self-consistency, few-shot learning, prompt search. What are the best ways to characterize and compare their rate of improvement (Li et al, 2024; Brown et al, 2024; Chen et al, 2024; Wu et al, 2024; Snell et al, 2024; OpenAI o1)?
- Is supervised finetuning on optimal (for the model) demonstrations of solving a task a good mitigation against under-elicitation and sandbagging (models hiding their capabilities; Greenblatt et al., 2024)?
- Can we have a generic measure of the elicitation power a red team has exerted on a model?

Measurement and analysis of agent behavior

Evaluations of AI agents often involve many challenging long-horizon tasks, taking hours or days to complete and resulting in long transcripts of interactions with the environment. This introduces several practical problems for the evaluators. For example, the evaluation tasks need to reflect as closely as possible real-world settings (both in environment complexity and agent affordances). When such agents are connected to real-world systems during the evaluation, their interaction with the environment may need to be supervised by a human to prevent any potential harm. After completion of the task, understanding agent behavior and errors requires a time-consuming manual analysis of transcripts. To address these challenges, we are interested in projects exploring the following questions:

- Can we develop techniques that rely on LLMs to automatically identify agent's failures, and more generally, provide qualitative analysis of agent trajectories (Yang et al., 2024, Scale Tool Use Leaderboard)? How do such methods compare to manual analysis by human experts (Chiang et al, 2024)? How reliable and unbiased can such methods be (Pannickssery et al., 2024)?
- What are the best metrics (e.g. rate of milestone completion, human intervention) for comparing performance of agents on tasks with only partial progress (Phuong et al, 2024; Højmark et al, 2024)?
- How do we effectively sandbox agents without significantly constraining the possible realism of the evaluation (Ruan et al, 2023)? Can we reliably automate supervision of agents during evaluation (Greenblatt et al, 2023)?
- How do we unify various agent evaluations under one platform (Liang et al, 2022) (e.g. enable comparison of arbitrary agents over a variety of agent benchmarks)?

Human-assisted agents

Most evaluations involve either using agents as AI assistants or in fully autonomous task-completion. However, in many cases of interest, agents may, while mostly acting autonomously, also interact with humans. For example, agents could ask for feedback on their performance or help when stuck on the problem, and humans can restart agents from checkpoints with rewritten solutions. Some of the key questions in this setting:

- What kinds of real-world applications are likely to involve human-agent interaction and what type of interaction will it be?
- What types of human intervention are most helpful for improving agent performance? How much human intervention is needed to improve the agent performance?
- What is a good evaluation protocol for agents with human-in-the-loop (Ibrahim et al, 2024)?