PRIORITY RESEARCH AREAS

AISI Challenge Fund

Updated - May 2025

Purpose of this document

This document is intended for **prospective applicants** to the AISI Challenge Fund. It outlines four priority research areas: **Safeguards, Control, Alignment,** and **Societal Resilience**. For each, we provide an overview of key research challenges and directions that AISI is looking to progress.

The Challenge Fund is designed to support a diverse range of projects that contribute to safe and secure AI development. While we are particularly interested in proposals that align with our priority research areas, we also welcome proposals that explore other innovative topics relevant to the safe and secure development of AI systems.

Contents

SAFEGUARDS

Summary

As AI systems become more capable and integrated into the economy, they will increasingly be targeted by adversaries looking to take advantage of their weaknesses. This includes adversaries misusing AI systems to aid in perpetrating large-scale harms, or adversaries disrupting operation of deployed AI systems to cause data loss or damage to critical systems.

Research that seeks to understand, evaluate, and improve the technical measures designed to address these risks ("safeguards") is urgently needed.

Key Research Challenges

- **Defending hosted frontier AI systems against misuse.** When AI systems are hosted by benign actors, a range of safeguards can be used to prevent adversaries using the AI to aid in illegal activity. These can include *model-level* safeguards like safety training; *system-level* safeguards like real-time monitors; *access* safeguards like giving vetted users or organisations access to less safeguarded or more advanced systems; or *maintenance* safeguards like rapid vulnerability remediation. The challenge also includes alternative access forms, like mitigating against misuse by insiders and defending fine-tuning APIs. A key difficulty is developing safeguards that prevent genuinely malicious activity whilst remaining cost-effective appropriate to the level of risk, and not significantly dampening the capability or utility of the model.
- **Defending against 3rd party attacks.** Exposure to attacker-controlled data during training ("data poisoning") or inference ("prompt-injection")—as well as direct adversarial manipulation of model weights ("model poisoning")—can lead to adversaries controlling the actions or goals of otherwise benign AI systems. This could lead to AI agents exfiltrating sensitive information or causing large amounts of harm, or broad attacks on the availability of critical AI systems.
- **Mitigating misuse of open-weight models.** Open-weight models enable a range of defensive and beneficial use-cases. However, when AI model weights are made public, system-level safeguards, access safeguards, and maintenance safeguards are made largely infeasible; and model-level safeguards are made much more difficult, as users can fine-tune away refusal behaviour or design attacks in a full-information (white-box) setting. Developing reliable methods for preventing or mitigating risks from misuse of open-weight models remains an open problem.

Potential Research Directions

- **Improving safeguard evaluations.** We are interested in realistic evaluation environments for safeguards across all the challenges above, particularly those which assess autonomous agents acting in real-world environments and include careful measurements of costs and user friction. These include better evaluations of misuse safeguards like unlearning, rapid vulnerability remediation, and asynchronous monitoring, as well as poisoning and prompt injection defences. Along with evaluation environments, we would like to see work that develops stronger attacks to stress-test defences—especially those that can be incorporated into defensive measures. We'd

also be interested to collaborate and fund work that understands fundamental limitations in specific defences or classes of defence.

- **Improving defences.** Specifically, work in addressing deficiencies in current safeguards, such as designing more resilient unlearning or data-filtering methods that remove a narrow harmful capability without degrading other model capabilities; developing better detection or mitigation strategies against data or model poisoning; or improving monitoring systems to cheaply and effectively catch instances of harm occurring, especially in agentic settings.
- **Longer-term bets.** We would also like to see research which explores directions that are less likely to provide short-term impact but will lead to substantial progress if successful. As a few examples, researchers could explore pre-training or architecture-level defences, design new paradigms for post-hoc safety training or unlearning or explore architectural changes to separate trusted and untrusted data sources to prevent prompt injection.

CONTROL

Summary

Most current safeguards for catching AI's unsafe behaviours are not designed to resist intentional subversion by misaligned AIs. Furthermore, current alignment methods can't guarantee that an AI's goals and actions match human intention.

AI control aims to design and evaluate *control protocols. Control protocols* are plans designed to prevent unsafe actions by AI systems, even if those AIs were misaligned and intended to subvert the safeguards. Promising protocols vary widely, from "*only deploy your AI systems inside a cyber-secure environment*" to "*use another AI to monitor all of your AI's actions and get a human to audit the most suspicious behaviour*".

Control evaluations attempt to measure the safety of protocols. A core challenge here is experimental design: how can we estimate safety without access to real-world deployments, or real subversive models? Existing control evaluations use simplified environments as proxies for real-world deployments, human-AI red-teams as proxies for subversive models, and statistical modelling to estimate the relative risk of deploying at scale.

Key Research Challenges

Existing research has prioritised one of the most critical environments – AIs deployed within frontier AI developers' systems. Soon highly autonomous and capable AIs may be deployed within such environments, with access to sensitive code and infrastructure. To ensure such deployments are safe, we need protocols that have been shown to be adversarially robust in similar settings. Only a handful of AI control evaluations have been conducted so far, and there are many challenges to scaling control to meet the demands of future deployments:

- **Creating realistic proxy environments.** Designing experimental environments that are as similar to the real deployments as possible is challenging. For example, closely replicating the infrastructure and codebases used by leading developers.
- **Creating realistic red-team AIs.** The accuracy of evaluations depends heavily on the red-teaming AI used to stress-test the protocol. Different implementations of red-team AIs balance different levels of realism, conservatism and practicality.
- **Understanding human-in-the-loop protocols.** Some protocols rely on humans to audit the most suspicious actions from AI systems. Understanding and augmenting human's ability to investigate LLM behaviour's is essential, especially as models achieve superhuman performance on some tasks.
- **Relaxing assumptions.** Existing control evaluations make many assumptions that may not hold in real deployments. For example: that AIs are either always subversive or never subversive; or that models are either trusted or untrusted.
- **Securing AI developer infrastructure against insider threats from AI systems.** Preventing AI's from taking dangerous actions will be easier with stronger traditional information security protocols in place, ideally codesigned with control protocols.

- **Develop alternative control protocols.** Prior work has proposed a number of control protocols. However, improvements may be necessary, especially if current protocols fail in more realistic and complex environments, with less restrictive assumptions.
- **Scaling protocols to (and beyond?) AGI.** Prior evaluations consider AIs performing tasks at a sub-human level. It remains an open-question how far control protocols could scale to AI systems performing tasks at a superhuman level.

Potential Research Directions

- **Create more realistic red-team AIs.** Prior work on control has stress-tested protocols against language-models utilising a human-optimised scaffold, unseen by the protocol. A more realistic evaluation could, for example, modify the language model's weights, using finetuning or transformer-compiled cryptographic algorithms.
- **Understand human-in-the-loop protocols.** Whereas previous research has focused on using LLMs as proxies for human auditing, we are excited to see empirical research on real human auditors. This could involve investigating how effective human auditors are against both LLM-generated and human-generated code vulnerabilities. We would also be excited about research into the factors that determine the success of human auditing (e.g., existing skillsets, time taken per trajectory and additional tooling provided).
- **Study analogous security measures for human insider threats.** Traditional software engineering security protocols might be adaptable to mitigate risks posed by AI agents acting as insider threats, similar to human threats. However, there is limited research at the intersection of AI control and conventional security strategies.
- **Develop alternative control protocols.** Potential future protocols could include white-box techniques such as linear probes, ensembles of monitors, hierarchical auditing and specific anti-collusion mitigations (such as better paraphrasing, or activation steering).
- **Investigating scaling protocols to (and beyond) AGI.** We're interested in work that convincingly argues that control can or cannot scale to super-human systems, across a wide range of domains.

ALIGNMENT

Summary

The AISI Alignment Team focuses on research relevant to reducing risks to safety and security from AI systems which are autonomously pursuing a course of action which could lead to egregious harm, and which are not under human control. There are no known reliable methods for preventing sufficiently capable AI systems from doing this. Our initial research focus is on using a combination of theoretical guarantees and empirical evidence to ensure the honesty of AI systems as they scale past AGI to superintelligence.

Key Research Challenges

We use safety case sketches as an organising frame for clarifying the relationships between the claims, arguments and evidence entailed by alignment proposals. Using these sketches, we identify areas where existing evidence is insufficient – that is, key alignment subproblems that we need to solve to make the alignment method work.

We specifically focus on making top-level claims about the honesty of AI systems. We do this because honesty (in at least some domains) is likely to be a necessary condition for the safety of superintelligent systems; the converse would be deceptive systems which may hide safety-relevant information. We also think that honesty is useful as a first step – for example, if we could build honest systems, we could use them to conduct research into other aspects of alignment without a risk of research sabotage.

We aim to make claims about various properties of AI systems using 'asymptotic guarantees'. Many approaches to alignment rely on a formal proof that an AI system obeys some specification or rely on empirical experiments which may or may not generalise. An asymptotic guarantee is a proof that if a process (such as training) has converged then some claim about the system will be guaranteed. We use these guarantees alongside empirical reasons to expect that the training (or other relevant process) will converge.

Our key research challenges are therefore:

- Conducting theory research (in complexity theory, game theory, learning theory and other areas) to both improve our 'asymptotic guarantees' and develop ways of showing that processes (e.g. training) relevant to these asymptotic guarantees have converged.
- Conducting empirical research (in machine learning, cognitive science and other areas) to validate that this theory research applies to real models - and covers other relevant gaps (such as the human input into alignment schemes).

We anticipate broadening our work beyond asymptotic guarantees as we identify new promising alternatives, in part through collaboration with the research community.

References to Existing Work

Existing work on alignment has been focused on empirical iteration: improving over previous models on average-case benchmark performance. Reorienting the field to prepare for AGI deployment requires developing both methods and evaluations suitable for models deployed on high-stakes tasks—including tasks going beyond human ability.

- **Alignment methods:** We are interested in improvements over existing methods (listed on the next page) as well as developing novel alignment methods. Alignment methods must address three problems:
  - **Scalable oversight**: How can we correctly reward desired behaviours beyond humans' ability to efficiently judge them?
  - **Interpretability for worst-case guarantees**: Given error bounds from training, how do we ensure that random errors do not have correlations and structure which will be used to subvert human intent when deployed?
  - **Exploration and elicitation:** To train AIs on tasks where human supervision is unavailable, we expect to use AI assistance to help humans supervise AI systems. One way to mitigate unforeseen errors by AI systems is to elicit the AI's maximum performance at the assistance task.
- **Alignment evaluations:** Measuring the effectiveness of alignment methods requires developing evaluations without relying on human annotation. We are interested in development of new evaluation methodologies modelling imperfect human judgment e.g. via partial observability, sandwiching, or adversarial initialisations.
- **Automated alignment:** Many researchers and AI developers are hoping to partially automate the search for and evaluation of alignment algorithms. We believe this may be feasible for some subproblems of alignment, but that obstacles such as automation collapse may block this approach on other subproblems. We are interested in research that maps or expands the set of subproblems that can be tackled, such as by studying long-horizon generalisation, reward hacking in the code agent setting, etc.

Potential Research Directions

We are excited to see work extending and critiquing existing work on alignment methods. This includes:

- **Developing safety case sketches for honesty.** We have published a safety case sketch using a 'scalable oversight' protocol (Buhl et al. 2025) based on AI safety via debate (Irving et al., 2018). Scalable oversight protocols aim to find ways to correctly reward desired behaviours of AI systems when those behaviours are too complex to be efficiently judged by humans. The development of this sketch will help us identify important subproblems that need to be solved for scalable oversight methods to achieve honesty.
- **Improving scalable oversight protocols.** We are optimistic about finding solutions to the 'obfuscated arguments' problem: a dishonest prover can in polynomial time produce an 'obfuscated' argument that would require an honest prover exponential

time to refute (Brown-Cohen et al., 2023). Such solutions might require strict stability assumptions, so we will examine ways to relax these. We'd also like to develop protocols that are robust to systematic human error. These problems are necessary to solve before scalable oversight protocols can be used in practice.

- **Finding ways to guarantee sufficient exploration during reinforcement learning.** One way in which training could not converge would be if models do not sufficiently explore the action space to discover optimal behaviours. We are excited about applications of learning theory and interpretability to this problem. Guaranteeing exploration is an important part of ensuring that training converges, allowing us to use any asymptotic guarantees.
- **Developing methods that use online training to mitigate risks from distribution shift.** With an appropriate reward model, online training can be used to prevent dishonesty arising as a result of the distribution shift from training to deployment. However, this only works on average, and individual outputs could be dishonest. One simple approach would be to use some averaging process over a large number of outputs, preventing any single response from the model from being used. Robust approaches to this issue are required to deploy models in high-stakes settings where even a single dishonest output could cause safety and security risks.
- **Empirically validating that theoretical results work in practice.** We will develop datasets which can be used to validate our theoretical claims about alignment methods (experiments about the validity of alignment methods have been called 'alignment evaluations'). Other key empirical questions about scalable oversight include whether debate measurably improves performance, whether training reliably reaches equilibrium, whether stability is a real problem, and whether obfuscated arguments actually emerge as predicted. Empirical results are needed before we can rely on the arguments in our safety case sketches.
- **Researching aspects of automated alignment beyond honesty.** We are interested in developing safety case sketches for automated alignment. We are also interested in research that maps or expands the set of alignment subproblems that can be safely tackled using AI agents, such as by studying long-horizon generalisation, reward hacking in the code agent setting, etc. Many researchers and AI developers are hoping to partially automate the search for and evaluation of alignment algorithms, so it's important that we are able to trust the outputs of such research; while honesty may be necessary to trust automated alignment research, it is not sufficient (for example, the AI researcher could be honest but wrong).

Our live list of open problems can be found in part 4 of this blog post.

SOCIETAL RESILIENCE

Summary

Today advanced AI capabilities have become rapidly accessible to individuals, businesses, and governments for a wide range of tasks. Popular platforms like ChatGPT, Claude, and Gemini now serve hundreds of millions of users, while businesses across a wide range of sectors are exploring AI for task automation and augmentation. As AI becomes deeply integrated into our lives, - reshaping financial markets, personal relationships, and information ecosystems -, we need to build resilience into systems and infrastructure, boosting our collective ability to prevent, prepare for, and absorb risks from highly capable AI. However, at present, we lack data concerning the nature and extent of AI uptake by individuals and organisations. To understand and address risks that emerge from the deployment of AI in real-world contexts, we need to carry out evidence-based assessments of how AI products and services are being integrated into different sectors of the economy for particular tasks. We need to measure their real-world impacts on individuals and organisations, and develop practical mitigations, guidance, and solutions that build resilience. into the systems and infrastructure in which frontier AI is being deployed. Our initial research approach is on applying socio-technical perspectives to understand and model frontier AI adoption and emergent uses, and on exploring evidence-based approaches to assess and monitor societal vulnerabilities to AI risks.

Examples of the sorts of large-scale risks that we have in mind include:

- As multimodal features in frontier AI systems improve, criminal organisations deploy them for scams and financial fraud, especially targeting the more vulnerable.
- As humans increasingly delegate problem-solving and decision-making to frontier AI systems, people become over-reliant on AI reasoning and knowledge, leading to human disempowerment in critical contexts such as critical national infrastructure.
- As frontier AI systems come to exhibit humanlike behaviours and personalised features, people become more vulnerable to persuasion, manipulation or forms of emotional dependence.
- As decisions are increasingly made in interconnected networks of AI systems, including agentic AI, unstable or volatile dynamics emerge in highly interoperable environments, such as financial markets.

Key Research Challenges

Our overall research goals and interests are:

1. **Mapping the landscape:** Before we can assess risk, we need a clearer picture of how, where, and by whom frontier AI systems are being adopted and utilised. We aim to build robust understanding of frontier AI adoption trends across key sectors of the economy (i.e. finance, critical infrastructure, education), and shed light on how specific AI tools are used for specific tasks and by which type of users. Beyond intended applications, we are also

interested in how users in the wild are repurposing, misusing, or making "unapproved uses" of consumer-facing and enterprise level AI systems. Some variables of interest in this space include:

- The degree of embeddedness of AI agents in enterprise workflows, and the extent to which an unsafe level of reliance on AI is being created.
- The ways in which frontier AI models are being used by private consumers, and the frequency and longevity of human-AI interactions.
- Incidence of frontier AI being used for fraud and scams, and the features that make this possible.

2. **Gauging the evidence for assessing societal impact dynamics**: Understanding how frontier AI systems interact with existing societal structures, cognitive biases, and vulnerabilities is crucial to anticipate how individual harms could scale to systemic risks. We support rigorous analyses of how frontier AI systems amplify or create new systemic vulnerabilities across communities, and which identify the enabling societal and technical factors which can exacerbate and escalate AI impacts. This includes research on:
    - **individual conditions** which make certain groups more susceptible to AI risks such as deception, emotional reliance.
    - **emerging characteristics or behaviours of AI systems** which drive societal harms (i.e. personalisation, persuasive capabilities). This includes assessments of the nature of human-AI interpersonal interactions, and the degree to which this leaves people vulnerable to influence by frontier AI.
    - **Key factors, thresholds and feedback loops** which could cause individual instances of AI-driven harm (i.e. a single fraudulent transaction) to escalate into societal-scale crises (i.e. financial instability).
3. **Building governance foundations by investing in data, infrastructures and methodologies for risk monitoring**: To effectively navigate the evolving landscape of AI risks, we need robust data, analytical tools, and consistent methodologies. We are interested in research which:
    - **Identifies leading indicators for emerging or escalating risks**, and which supports the development of ongoing monitoring systems to track societal risks from AI advancing capabilities, adoption patterns, and societal context shifts.
    - **Address data scarcity for AI usage and adoption,** including by developing novel approaches to systematically generate and collect usage data from frontier AI systems, as well as creating processes, tools and analytical frameworks to turn raw data into meaningful insights on risk trajectories and monitor the associated societal impacts.

4. **Developing and evaluating defence-in-depth mitigations:** We are interested in a broad range of potential interventions aimed at all the actors involved in the pathway to harm, from technical safeguards to policy responses and educational initiatives. We want applied solutions, not theoretical ones, that build resilience to frontier AI risks. We are interested in research which rigorously assesses the efficacy of current and proposed mitigations for preventing or lessening specific frontier AI risks, as well as research creating evaluation methods to test the resilience of our societal systems to AI-driven shocks and disruptions.

Potential Research Directions

We invite researchers from academia, industry, civil society, and other institutions to engage with these challenges. For each of the above areas, possible research projects we would be interested in supporting include, but are not limited to:

1.Mapping the Landscape:

- **Large-scale adoption surveys and longitudinal studies:** tracking AI tool uptake and usage patterns across key industries and demographics over time.
- **Sectoral deep-dives:** in-depth case studies of AI integration within specific sectors like finance, focusing on specialised tools and workflow transformations.
- **OSINT and dark web analysis:** scraping and analysis of online marketplaces (including illicit ones) to identify AI tools being sold, repurposed, or implicated in harmful activities like fraud.
- **Horizon scanning for emerging tools:** identifying and characterising new AI tools, particularly those with potential for high-impact or dual-use capabilities.

2.Gauging the Evidence:

- **Qualitative studies on user demographics and motivations:** understanding why and how specific groups use AI tools (e.g., companion apps), their perceived benefits, and their vulnerabilities.
- **Vulnerability assessments for specific harms:** Focused studies on communities susceptible to AI-driven fraud, including those who typically avoid conventional scams but might fall for more sophisticated AI-perpetrated ones.
- **Longitudinal studies on deception detection:** assessing whether users' ability to discern AI-generated content (e.g., deepfakes, AI-authored text) degrades or improves as AI capabilities advance.
- **Analysis of AI system outages or disruptions:** studying user and systemic reactions to AI service interruptions to measure dependence and exposure.

3.Building governance foundations:

- **Usage data bounties & crowdsourcing initiatives:** projects to incentivise users to share anonymized interaction data with frontier AI systems (e.g., chat histories from LLMs, usage logs from companion apps) to build diverse datasets for research.
- **Development of updatable usage indexes and monitoring dashboards:** creating systems to track key metrics of AI adoption, usage, and associated incidents over time, potentially incorporating automated classifiers and data visualisation.
- **Simulations and synthetic data generation:** creating simulated environments or synthetic datasets to model AI interactions and potential harms, especially for exploring scenarios where real-world data is inaccessible.
- **Incident monitoring and analysis:** Systematic collection and analysis of AI-related incidents to identify patterns, root causes, and the role of frontier AI systems in causing harm.

4.Developing and evaluating defence-in-depth mitigations:

- **Socio-technical solutions for frontline workers:** developing practical solutions to support frontline workers (i.e. forensic news content analysis for journalists). This also includes connecting the challenges and experiences of these workers and users with technical and socio-technical measures that upstream actors can implement to reduce risk exposure, vulnerability or severity.
- **Cross-cutting technical measures:** creating technical tools and solutions to improve the identification and tracking of AI systems including agents, such as model fingerprinting, synthetic content detection and the provenance and watermarking of synthetic content, alongside other related measures.
- **Developing clearer harm pathways:** constructing threat models that clearly articulate the pathway to harm for specific societal risks and identifying the actors responsible for addressing them.
- **Evaluations of technical mitigations:** researching the robustness and practical effectiveness of solutions such as digital watermarking, synthetic media detection, and AI model auditing in real-world scenarios.
- **Resilience measures for critical sectors: e**xploring how regulators and key organizations can build capacity and develop toolkits to manage AI risks pertinent to their specific domains.

References to Existing Work

- Bernardi et al. (2024) Societal Adaptation to Advanced AI. arxiv.org/pdf/2405.10295

- Dobbe et al. (2022) System Safety and Artificial Intelligence. arxiv.org/abs/2202.09292
- Raji et al. (2022) Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. arxiv.org/abs/2206.04737
- Stein et al. (2024) The Role of Governments in Increasing Interconnected Post-Deployment Monitoring of AI. doi.org/10.48550/arXiv.2410.04931
- AISI Systemic Safety grants page
- AI Policy Perspectives blog

CONTACT US

For further questions or clarifications, you may refer to our Application Pack, Clarification Questions (CQ) document or message us at aisichallengefund@dsit.gov.uk.