AI SAFETY INSTITUTE

# Examples of systemic AI safety projects

Systemic AI safety draws upon sociotechnical AI research, which is a broader endeavour that considers the impact of AI on people and society (Weidinger et al. 2023). Systemic AI safety is also related to safety science within engineering, which studies how to make systems and infrastructure safer (Dobbe, 2022). We hope to bring together researchers from these and other communities to tackle systemic risks from AI.

For this grants programme, we are focused on systems-focused approaches to AI safety. We distinguish systemic AI safety from interventions that focus on AI models themselves. To make this distinction clear, consider the problem of AI-generated misinformation:

- A **model-focused** approach might attempt to find fine-tuning regimes that improve the factuality of AI model outputs (this is out of scope for this call).
- A **systems-focused** approach might consider how to build user trust in legitimate digital content, even where AI outputs are often unreliable (this is in scope for this call).

We are excited about impactful, evidence-based work that addresses both ongoing and anticipated risks to societal infrastructure and systems.

We recognise that future risks from AI remain largely unknown. We are open to a range of plausible assumptions about how AI technologies will develop and be deployed in the next 2-5 years (we are less interested in highly advanced capabilities that may take much longer to develop). For example, over the near term

- The uptake of AI models across different sectors of the economy is likely to grow
- AI models will become capable at taking actions on behalf of the user
- The capacity of AI models to generate audio and video content will improve
- AI models may become more personalised to the tastes and beliefs of the user
- More situations will arise where AI models interact with each other

Below, we provide examples of potential systemic AI safety problems to help you better understand what we are looking for. We include examples of both cross-cutting and sector-specific problems.

We hope that the ideas below will serve as helpful starting points, but they are not intended to provide an exhaustive list of topics in systemic AI safety. We are sure there are many other important problems to address - if you have one in mind, then please do apply.

## Examples of cross-cutting interventions

1. **Developing tools to monitor systemic AI risks.** A comprehensive approach to systemic AI safety needs to monitor and track the ways that AI is being deployed in society. Questions might include:
    a. Which societal systems are most likely affected by advanced AI, in what way might they be affected, and how should interventions be prioritised (Avin et al., 2018)?
    b. How can we ensure that information is shared between those who identify risks and those who are in a position to address them (e.g., via incident reporting or engagement with civil society) to create effective and responsive governance, whether centrally or decentralised?
    c. What can be learned from simulations, scenario analysis, and stress testing, drawing from fields such as climate science, epidemiology, and financial risk management?
    d. What are the implications of market structure on systemic risks of Advanced AI?

2. **Designing markets for AI risks.** Ideal markets would accurately align incentives to societal values and spur the responsible deployment of AI models. Questions might include:
    a. How can market mechanisms be developed to quantify systemic risks and safety?
    b. How can we improve the way the insurance industry prices risks for both ongoing harms that impact specific communities and catastrophes that are infrequent but massive in magnitude?

3. **Building infrastructure for AI agents.** Increased delegation to 'AI agents' — models capable of performing complex tasks on digital platforms with limited supervision — may exacerbate or introduce new societal risks. These risks could manifest in interactions between the human and an AI agent, between an AI agent and the world, between two AI agents ("cooperative AI"), or as emergent properties from an ecosystem of AI agents. Questions might include:

a. Which new threats can we expect to see with agentic AI models that cannot be addressed with pre-deployment measures?
b. How can infrastructure like agent identifiers, real-time monitoring, and activity logging be implemented (Chan et al. 2024)?
c. As AI models become more advanced, how do capabilities like situational awareness and self-modification affect the ways that agents interact with one another?
d. How can we implement and empirically test theoretical work on AI collusion for different societal sectors (e.g. Foxabott et al. 2024)?

4. **Building governance tools for systemic safety**, including solutions that are both technical and institutional in nature. Questions might include:
    a. What is the technical architecture needed for the various bodies, domestically and internationally to monitor and respond to systemic risk across sectors?
    b. Which technical tools, in hardware or software, could be designed and deployed to better help risk-owners monitor or respond to risk? for example, could privacy-preserving technology enable real-time monitoring and response without compromising privacy or adoption (Aarne et al., 2024)?
    c. How could circuit-breaking algorithms or similar technical governance tools look like & be implemented?

5. **Mapping systemic over- and under-reliance on AI.** AI models are increasingly being integrated into critical infrastructure such as communications and finance. Systemic overreliance occurs when many users or infrastructure companies excessively rely on AI and start acting upon incorrect outputs, or are unprepared for scenarios where AI systems suddenly become unavailable. Systemic under reliance on AI solutions could also occur. Questions might include:
    **a.** What are the degrees and modalities of acceptable reliance with respect to the substitutability of the AI models and the criticality of the tasks performed?
    **b.** How can we develop robust systems-informed safety and security standards for AI use in these contexts?
    **c.** What do context-specific failsafe mechanisms and contingency plans for AI failures look like?
    **d.** How do existing power dynamics shape human-AI interaction and reliance?

# Examples of sector-specific interventions

1. **Democracy and Media.** Research has indicated that AI could be used to manipulate public opinion, interfere with democratic processes, or undermine trust in institutions, though evidence about current impact is limited . On the other hand, AI could also help make democratic participation more accessible and augment deliberative processes. Research directions may include measuring and identifying methods to enhance trust in democratic processes with increasing AI-generated information and creating ways for AI models to be used thoughtfully to improve democratic inputs.

2. **Education.** We are seeing the early phases of adoption of generative AI in education to personalise learning, broaden student capabilities, and automate teacher tasks, but also note emerging concerns about potential costs to educational outcomes and disruption to assessment methods. Potential research directions could include building a system for gathering evidence around frontier AI in education or examining how teaching methods should adapt with the adoption of increasingly capable AI models.

3. **Economy and the labour market.** Often, technological change is not exogenous but rather responds to shifting skill supplies and profit opportunities. Technological adoption — and the way it impacts the labour market — also depends on organisational structure in firms. Could AI be leveraged to aid worker retraining and upskilling? What are possible interventions that could encourage task complementarity over substitution, thus making workers across the income range more (and not less) valuable (e.g., Acemoglu and Autor, 2011, Eloundou et al., 2023)?

4. **Health, biosafety, and biosecurity.** Frontier AI is enabling significant advances in biological R&D, and generative AI is also being trialled in healthcare settings. Potential projects in this space could include: identifying key decision makers and risk owners in healthcare AI adoption, and assessing how this system of responsibilities might respond to more capable and autonomous AI systems; investigation of how advancements in biological R&D shift the balance across defensive (e.g., diagnostics, vaccines, and therapeutic design) and offensive capabilities, and identify promising interventions that could accelerate safety-enhancing technologies; or developing specific defensive technologies, such as AI tools for nucleic acid synthesis screening or early pathogen identification.

5. **Finance and insurance.** Greater adoption of AI-powered tools and agents in financial services could exacerbate harms such as collusion, unfair pricing, and insider trading, potentially even without human intent. AI tools could also amplify the scale and efficacy of market manipulation, and a lack of understanding and robustness could lead to large-scale coordinated failures. Projects could look at developing better methods to assure AI tools in finance, simulate potential failures, and monitor markets in real-time. More ambitiously, risk-oriented financial services, primarily insurance, could potentially incentivise distributed resilience to AI-related risks if they were better characterised and priced.

6. **Legal.** AI tools could increase access to legal information and increase productivity in the legal profession. However, given the known failure modes of these systems, there can be real challenges if legal users do not sufficiently understand AI tools or if there is unequal adoption. AI risks also pose challenges to legal decision-making, as synthetic media could undermine trust in evidence, and AI-assisted decision-making could challenge notions of accountability and liability. Interventions could focus on the education of legal practitioners and judges, better tooling for assessing the reliability of evidence that may be AI-generated, and methods for co-design of legal AI tools that increase understanding, access, and fairness.

7. **Emergency services.** AI tools are already seeing adoption in emergency response, from real-time monitoring for wildfires to AI assistants for police call handlers. The potential benefit, in terms of faster and more effective response, is significant, but so are risks of misallocation of emergency resources, the reinforcement of societal biases, coordinated failures at times of crises, and AI-assisted attacks on emergency services as an amplifier to other malicious harms. Interventions in this space could focus on data provenance and governance, and on training and institutional/incentive design to identify and prevent failures before they occur.

8. **Transportation.** While there has been a gradual adoption of autonomous vehicles over the past few years, AI has already been incorporated into intelligent transport systems including traffic prediction and road maintenance, and employed extensively for public, air, and water transport planning. How can we assess the contribution of AI adoption to the risk of concurrent failure across transport systems? Are there systemic biases in AI transport systems with downstream social effects on urban planning, access to services, etc.? How can these be mitigated?

9. **Food, water, and energy.** AI models are already seeing early adoption in these domains, from AI-assisted climate prediction that informs AI-assisted planning, AI-assisted real-time management of flows, AI-based decision support tools for

farmers, and autonomous vehicles and drones used in agriculture and infrastructure maintenance. Concerns have been raised about the technology incentivising further centralisation of production and control, with increased systemic risks in case of failures or adversarial disruption. How could such risks be monitored and alleviated?

10. **Communications, information technology, and operational technology.** AI models could be used to identify and exploit vulnerabilities in digital and physical infrastructure, leading to large-scale cyberattacks or making cyberattacks more accessible to a wider range of actors. Research could focus on developing AI-powered tools, such as automated threat detection and response systems, that exploit the defender's information asymmetries and ensure that advancements in cyber-capable AI actually improve (and not degrade) systemic safety.

# Example of problem statement

- Main Question: How can the degree of reliance on AI models be classified and monitored? Which tools might support adequate reliance on legal and medical information generated by AI models?
- Agentic AI models are becoming more and more integrated in society (>90% of Fortune 500 use GPT-4 according to OpenAI, new agentic AI models like GPT-4o, AI Software Engineers etc. are being built on-top)
- Previous research suggests that excessive reliance on AI models might lead to severe incidents and cascade risks (see AI incident database and autonomous vehicle incidents).
- However, AI models provide a wide range of automation benefits, like cost reductions, increased access, higher quality service etc. (Bomassani et al. 2021: Benefits and risks of foundation models)
- Especially in the legal and medical context, AI chatbots providing advice are on the rise. In the next 2-5 years, these chatbots will likely be integrated more broadly, and more deeply. Deeper integration includes increasingly critical legal and medical decisions being made based on AI suggestions. Their advice might lead to actions with severe individual consequences.
- However, on a systemic level, possible large-scale harms and cascading risks are unclear.
- Public bodies and industry associations are lacking:
    - a) information on the degree of use of AI models for medical and legal information
    - b) standardised metrics to understand the degree of appropriate reliance and quality management processes to ensure appropriate reliance,

    - c) scenarios about the consequences of the use of AI models on the fundamental functioning of legal and medical sectors
- Clarity on each of a)-c) is important to ensure adequate governance and reliance
- Existing research is not fully addressing these problems.
    - i): Previous work focused on capability benchmarks of AI models (e.g. MedQA, LegalBench), while measures of propensity in sociotechnical contexts and usage data are lacking.
    - ii): Usage monitoring is not standardised. There are existing datasets including legal and medical advice interactions (e.g. IntheWildChat), however these are limited.
    - iii): There have been structural monitoring indicators proposed for other sectors (e.g. structural monitoring indicators for the information space as part of the Digital Service Act in the EU). However, for the legal and medical setting, these are lacking.
    - iv) Scenario modelling exercises for advanced AI has remained qualitative, without specific quantitative modelling like in climate science (see Undheim & Armad 2024)
- Each of these subproblems apply to the legal and medical context, but also generalise to other sectors: The lack of usage monitoring standards, structural indicators and quantitative scenarios is profound in most critical infrastructure related to AI