ILINA 5.0. JRF Areas of Interest (Governance Track)

For applicants interested in AI governance, we are particularly keen to support projects in the following seven areas:

African AI policymaking

As AI adoption becomes more widespread across Africa, policymakers are increasingly recognising the need for AI policy within their jurisdictions. This workstream focuses on projects that seek to guide African policymakers on how to govern AI in ways aligned with AI safety objectives. Suitable projects include work touching on generating and translating evidence (such as evaluation results and primary data on the impacts of AI in African countries) into policy recommendations for African countries. Other viable projects include those examining what African AI policy should look like in the wake of other policies such as US policy, and geopolitical considerations that might constrain African AI policymaking.

EU AI Law and Global South AI Safety

The EU AI Act (Regulation 2024/1689), together with instruments such as the General-Purpose AI (GPAI) Code of Practice, represents the first comprehensive regulatory framework for AI. Due to the EU's first-mover advantage, this regulatory regime may influence AI governance approaches beyond the EU, including in Global South countries, in a manner similar to the Brussels Effect. This workstream focuses on how EU AI law is likely to affect Global South countries as the regime continues to evolve. Suitable projects explore questions such as: for countries that choose to borrow from EU AI Law, what should localisation look like in light of differing capacity and safety goals? What other ways can Global South countries actively shape the effect of EU AI law to advance their own safety and security goals? Where collaboration is relevant, which mechanisms or avenues should be pursued, and what frameworks should guide such collaboration?

Whistleblower Protection in AI Safety

Whistleblowing plays a critical role in accountability across industries. Established whistleblower protection frameworks were designed for harms like financial fraud or workplace safety violations, but AI safety presents distinct challenges. These challenges include emerging capabilities whose dangers are contested, speculative long-term risks that do not yet fall under existing regulations, and systemic harms from deployed systems. As a result, legal standards like “reasonable belief” or "foreseeable and material risk” become ambiguous when applied to AI safety, presenting questions such as: how should we assess materiality for low-probability catastrophic risks, or what counts as reasonable belief when experts disagree? We are interested in projects that analyse how whistleblower protection frameworks apply to AI safety, whether through interpretation of specific standards, comparative analysis across jurisdictions or sectors, or proposals for alternative frameworks. Strong proposals will identify a specific unresolved question, explain why existing research has not adequately addressed it, and demonstrate clear implications for who receives protection and under what circumstances.

Sociotechnical AI misuse evaluations

This category focuses on how AI systems can be misused—either intentionally or accidentally—due to their dual capabilities. Fellows would ideally have a background in the social sciences and basic knowledge of or interest in quantitative methods. We are interested in projects that include evaluations focused on: persuasion and manipulation, biological and chemical hazards, surveillance and censorship, as well as general misuse risks in critical sectors such as healthcare. This work is applied and hands-on, requiring fellows to design, implement, and execute evaluations.

Threat modeling for Global South Contexts

Threat modeling is the process of identifying and mapping out risk pathways in order to mitigate them. In this workstream, fellows would focus on conceptual work identifying models specifically relevant to Global South contexts and, particularly, harms that would not occur without frontier AI systems. This is primarily theoretical research requiring a structured, methodical approach. Fellows would conduct extensive literature reviews and likely collect qualitative data through expert interviews.

AI Agents

AI Agents that can independently plan and execute tasks with limited human involvement are considered the next major step in the development of AI. Leading companies have already begun to release agents that can serve as software developers, research assistants and even personal assistants. Current agents are still more limited than the most helpful, sophisticated and general purpose AI Agents that companies believe will transform society. Defining AI Agents and understanding their distinct features has been a core part of the initial work in this area. A broad range of existing areas of law could be useful in managing the impacts of AI Agents e.g. agency law, tort law and contract law. This workstream tackles both questions about how the law is likely to apply to AI agents and how it should be adapted. While there is research examining potential technical guardrails to mitigate the risks of AI Agents, we still need to figure out several broader policy questions: What goals are these guardrails meant to achieve? How should these goals be balanced? What should the regulation of AI agent development look like? What role should different actors play in mitigating risks from AI Agents? How should implementing technical guardrails be done? What existing laws make it easier or difficult to operationalize technical safety practices?

Legal doctrinal questions relating to AI Safety

When considering how to appropriately anticipate, measure, and govern AI risks, it is common practice to try and establish (i) whether AI is any different from already existing technologies that are regulated, and (ii) if it is different, to what degree. Answering these questions frames our understanding of how far existing legal frameworks can be stretched to apply to regulating AI, and when new legal doctrines or institutions are necessary. Questions exploring the application of existing legal frameworks, procedures, and doctrines when it comes to large scale harms caused by AI fit within this category of work. This workstream can also include work that interprets the application and impact of new laws on AI design and governance. We are also interested in research specifically addressing how we attribute responsibility for harms caused by AI under liability law. How do we make determinations of liability for AI harms when these harms are often difficult to predict and mitigate? What mitigation measures are technically feasible and sensible to expect from AI developers when harm is caused?