Australian National University

MINTLab
Machine Intelligence and Normative Theory

Call for PhD Applicants:

Language Model Agents and Society

We are seeking highly motivated PhD candidates to join the Machine Intelligence and Normative Theory (MINT) Lab at ANU, to work under the supervision of Professor Seth Lazar on its new project on Language Model Agents and Society. This is an ambitious research project aimed at exploring the societal impacts, ethical considerations, and regulatory frameworks surrounding the development and deployment of Language Model Agents — advanced AI systems powered by Large Language Models (LLMs) that are capable of autonomous action, decision-making, and tool usage.

Project Overview:

Recent advances in the development of LLMs have led to the creation of Language Model Agents — AI systems that can use software tools to interact with the world, and take unsupervised series of actions towards some goal (for an overview of LMAs, see this FAccT tutorial: https://youtu.be/IkY2yzgIDa0). These agents have the potential to revolutionise personal computing, transform social relationships, and enable unprecedented levels of autonomy in non-human entities. However, with these advancements come significant risks and challenges, including issues of safety, trustworthiness, and ethical governance and design.

Our project aims to address these challenges by:

1. **Anticipating and evaluating** the societal impacts of Language Model Agents.
2. **Articulating norms and ethical guidelines** for the design and regulation of these agents.
3. **Operationalising these norms** through concrete proposals for design and policy interventions.

PhD candidates will combine a focus on **two** of these three project objectives with concentration on **one or more** key application area. These are subject to change as the technology develops (and candidates may propose alternatives), but at present they are as follows:

- **AI Companions:** Investigate the ethical, psychological, and societal implications of AI systems designed for companionship, including the potential risks of emotional dependency, manipulation, and the erosion of authentic human relationships.
- **Universal Intermediaries:** Explore the role of Language Model Agents as intermediaries in digital interactions, focusing on their potential to reshape democratic processes, influence social behaviour, and transform the attention economy.
- **Catastrophic Agents:** Assess the risks associated with the development of highly autonomous AI systems capable of self-replication and large-scale cyber-attacks, and propose strategies for mitigating these risks through regulation and design.

Candidate Profile

Applicants should meet the criteria for admission to the ANU philosophy PhD program, and in addition should ideally have a strong background or otherwise demonstrated capacity in computer science or engineering, especially AI and machine learning, preferably with experience working with LLMs or other related technologies.

This call is specifically focused on international candidates.

The Program

While candidates will pursue a PhD in philosophy and will meet the standard requirements thereof, they will also be trained to draw on empirical and technical fields relevant to their project, and to incorporate experimental elements into their work (in the spirit of 'computational philosophy'). Engineering support will be provided as part of the project.

MINT PhD students are also introduced to and connected with lab partners around the world — the project's advisors include individuals at Anthropic, the Carnegie Endowment for International Peace, Imbue, Hugging Face, the Gradient Institute, the Australian Parliament, as well as researchers at Stanford, Harvard, and the Institute for Advanced Study.

The Lab

MINT is based in the School of Philosophy at ANU, and comprises postdocs, HDR candidates, research engineers and research assistants associated with the work of Professor Seth Lazar, as well as research affiliates from around the world and from a wide range of disciplines.

The lab pursues work in *normative philosophy of computing* and *sociotechnical AI safety* — the former addresses normative questions raised by computing using the tools of analytical philosophy, and drawing extensively on relevant empirical and technical research; the latter draws on a broad range of disciplinary inputs, including philosophy, to steer the most advanced AI systems so that they do not undermine people's fundamental rights. At the intersection of these approaches, the lab is building a number of projects in *computational philosophy* — in our case, the use of computational tools to answer philosophical questions relevant to AI safety.

The lab has close connections with major academic, civil society, industry, and government organisations around the world working on frontier AI ethics and safety.

Stipend and Travel Funding

Two international PhD scholarships are available, associated with a grant that runs for three years. The scholarship will run for the duration of the grant. Recipients who start in the first half of 2025 will receive a stipend of AUD$50,000 in 2025, 2026, and 2027. The intended start date is January 2025.

Candidates are strongly encouraged to apply for the ANU International HDR Scholarship (which comes with some benefits not provided as part of the GASP fellowship). If successful, their MINT grant will be converted into a top-up grant to bring their total stipend up to AUD$50,000/year.

In their second and third years, the awardees will have access to a travel budget of AUD$5,000/year each, in addition to any entitlements they have as PhD candidates in the School of Philosophy.

Application Procedure

Candidates must either already have secured admission to the ANU PhD program in an earlier round, or should apply to the international HDR scholarship round that **closes on 31/8/2024**. Candidates who do not make that deadline will still be considered for the GASP scholarship if they apply to the ANU PhD program by 15/9/2024. Candidates who wish to be considered for this position must write to mint@anu.edu.au by at least **1/9/2024** to indicate their interest, with a CV and a 250 word research outline. If they wish to apply for the international HDR scholarship they should write by at least **24/8/2024**.

The scholarship will be awarded based on the candidate's full ANU PhD application (which must contain a project proposal relevant to GASP).