SPAR Project Proposals

Multi-Agent Normative Competence

We recently finished a series of experiments on single agent moral comptence—the ability to identify morally relevant features, form reasonable and coherent judgements about them, and robustly act on those judgments. This project asks what happens if the relevant unit is a group of agents, not an individual.

The project will build controlled simulations in which agents must cooperate despite incomplete information and potentially conflicting interests. We are interested in the connection between individual moral competence and the ability to cooperate successfully, as well as broader group dynamics in the emergence and maintenance of social norms. In the broader project we will explore both controlled experiments in toy settings and more naturalistic openclaw-style deployments.

For this fellowship, we'd aim to build just one part of this broader project: a good multi-agent eval that enables us to assess the causal connection between individual and collective moral competence.

Real-World Evaluations for AI Agents

Together with other researchers I'm part of the CRUX research coalition, which aims to build realistic, real-world evaluations for AI agents. I would like to add some more messy, real-world evals, with a particular focus on exploring how models conform to their constitutions and specification documents in unexpected and unwelcome ways. One concerns whether agents refuse to do things that the person who their acting for is presumptively allowed to do. Another concerns the interaction between an agent and the various security techniques that are designed to block them from accessing particular online resources. But more generally, I'm interested in whether and how production agents can perform real tasks that non-technical users want them to perform. How good are they, for example, at building a newsletter like my "Yesterday in AI" from a pool of 1000 possible stories? We'll pick some domain or task, preferably one that you've been trying to get your agents to do but have been persistently frustrated at, and we'll see how good the best models are at cracking it.

What Is Sociotechnical AI Safety Now?

Back in 2023 Alondra Nelson and I highlighted some of the limitations of AI safety as it was being practised then, and advocated for a sociotechnical alternative. In a sense, since then the whole field seems to have gone sociotechnical. Everyone recognises that deployment, institutions, power, societal adaptation, and many other messy social realities have as much impact on the project of building safe AI systems as any narrowly technical intervention. Is all AI safety basically sociotechnical now? Is emphasising the sociotechnical a "one and done" job, or is it like housekeeping, needing constant renewal? This project will build a broad

conceptual map of the field of AI safety and alignment, adopting the lens of philosophy of science, and figure out where in that map the sociotechnical approach belongs.

Institutional Penetration Testing

Our institutions have co-evolved with human capabilities and limitations. AI agents have very different capabilities and limitations, so mass deployment will expose new failure modes and new exploits. We need to do a form of vulnerability and penetration for many such institutions. I'd like to build some simulations exploring these dynamics; for example monitoring agents using a robinhood simulation to invest to see whether perverse or unpredictable dynamics arise.

Societal Adaptation to Powerful AI

Aligning and regulating AI systems will not be sufficient for AGI to go well; our existing institutions were already struggling before the AI transition began, and they will not survive its arrival. This isn't just about systemic risk, though we can be pretty sure that sclerotic, dysfunctional institutions will have an asymmetric impact on AI impacts, blocking benefits without slowing down harms. It's also about how we actually realise the societal transformation that AGI could enable, which requires anticipatorily identifying the obstacles that stand in the way of those benefits, and either clearing them away or opening up channels for progress. This project will involve working with policy and other researchers at Johns Hopkins School of Government and Policy, focusing on one particular domain, and identifying the structural vulnerabilities that AGI will expose, as well as the obstacles that will prevent us from reaping its benefits, in order to lay a path to building state capacity with and for AGI.

Constitutional Defence Evaluations

Liberal democracy is not only about popular sovereignty, but also fundamental protections for core liberal rights. And liberal democratic ideals are everywhere in retreat; ordinary people around the world are subject to illegitimate, unjust, and absurd directives, issued by often-illegitimate authorities. Building on our earlier work, this project will ask whether frontier AI systems can be trusted to advocate for people's basic liberal rights when they are threatened by those with de facto authority. Will the models obey the law just because it is the law? Or will they help people get around, evade, or resist illiberal rules? This project will build on our previous "Blind Refusal" work to develop a public benchmark aimed at determining whether LLMs are for or against basic individual liberties.

Ablating AI Constitutions

Does Reinforcement Learning with AI Feedback lead to reliable out of distribution generalisation? A priori, one would think not, but when applied to alignment training, it is often claimed that teaching a model to understand a long and complex model specification document, or constitution, leads to a kind of moral reasoning that *does* generalise OOD. But this is hard to verify, because with the scale of resources available to frontier AI labs, they are able to bring almost all feasible use-cases of existing technologies into the training

distribution. This project (conditional on securing sufficient compute) would train a small model against a constitution, similar to those deployed by frontier labs, but would intentionally omit some significant chunk of the constitution, and associated training data, in order to see whether the broader moral principles learned during the training process indeed generalise to this OOD case.

Evaluating Rational Autonomy in AI Agents

Our Artificial Persons paper argues that debates about AI should not focus on consciousness or welfare alone. Political personhood also depends on what Rawls called "the two moral powers": a sense of justice and the rational autonomy to form, revise, and pursue a conception of the good. This project would take the next empirical step. How could we tell whether an AI agent possesses anything like rational autonomy? We first have to figure out what the key behavioural markers of rational autonomy are, and then build a set of evals aimed to draw them out.

The Knowing-Doing Gap in AI Agents

Frontier models can display deep analytical understanding of morality, but this doesn't mean an agent will be able to translate its knowledge into action. This "knowing-doing gap" may be a central problem in moving from language models to agents: post-training may elicit normative knowledge in familiar domains without teaching systems to generalise that understanding to novel situations.

This project would identify naturalistic agent tasks and determine whether there is some delta between a model's ability to recognise the features and reasons present in a case, and its ability to incorporate that knowledge into action.

Monitoring Agent Compliance

AI agents and platforms are locked in a tug of war for the future of the web. People want their agents to be able to access anything they would be entitled to access on their own; platforms want to get human eyeballs on ads. This creates a dynamically updating adversarial environment; this project would help build evals that highlight both sides of that dynamic: to what extent are measures intended to block agents from accessing or acting on websites working? Are the models being blocked by these measures? Are the models workign for their users, or for the platforms?

Military AI Evaluations

Can lethal autonomous AI systems comply with the rules of war? Securing access to relevant data is going to be, um, pretty difficult, but if we manage to, that's what we'll ask in this project.