Entrepreneur First AI Safety Hackathon - some subject ideas

A Socio-Technical Problem

The issue of AI safety is multidisciplinary, and the solution *must* be holistic. AI ethics, AI alignment, and AI governance must work hand in hand:

- **AI ethics asks which values we can incorporate in these complex systems**
- **AI alignment asks how to control autonomous systems, whatever the value incorporated in those systems**
- **AI governance asks how to adopt the solutions at societal levels.**

Different explanations lead to the conclusion that AI could be an existential risk:

- Natural Selection Favors AIs over Humans
- How harmful AIs could appear - Yoshua Bengio
- Is Power-Seeking AI an Existential Risk?
- AGI Ruin: A List of Lethalities - AI Alignment Forum
- The alignment problem from a deep learning perspective
- Other scenarios are given in Threat Model Literature Review, by DeepMind.

In all these scenarios, international coordination and governance would have a significant influence, which is why we talk about AI governance in the following document, with a focus on what technical contributions can bring to this field.

Governance of AI

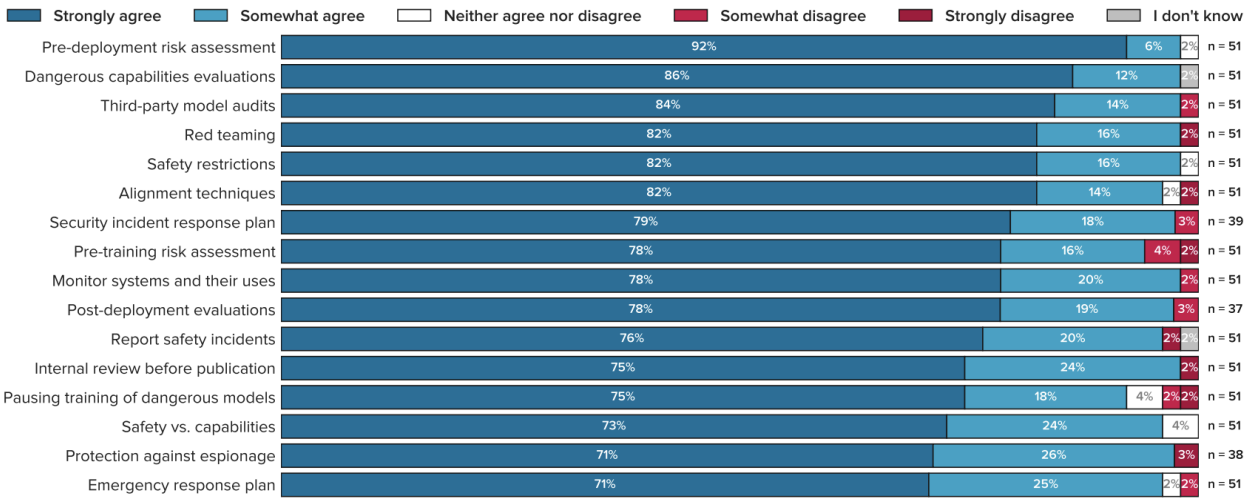
The unregulated advancement of artificial intelligence is not inevitable. Just as aviation is now regulated and requires an audit process to build and use a new airplane for public use, just as nuclear regulations are extremely cautious, and it is unthinkable to build bridges without large safety margins, we can choose to adopt ambitious safety standards for this disruptive technology.

In a world where the safety of AI is not a consensus, even if alignment solutions are found, they must be applied. **There are two main objectives in AI governance: 1. To give time and means to find technical solutions, and 2. To increase the likelihood that these solutions are used, thanks to coordination.**

These safety ambitions require the establishment of a large AI security industry (a neologism inspired by the cybersecurity industry).

AI Safety Objectives

A paper by the GovAI organization that lists 50 policy proposals: Towards best practices in AGI safety and governance: A survey of expert opinion, is a very important paper. Another one is: Existing Policy Proposals Targeting Present and Future Harms. Each of the measures proposed in these papers could be a pretext for creating a specialized organization to address these issues.



Towards best practices in AGI safety and governance: A survey of expert opinion.

The following measures all got >70% "Strongly agree" and at most "3%" "strong disagree" (note that not everyone answered each question, although most of these 14 did have all 51 responses):

1. Pre-deployment risk assessment. AGI labs should take extensive measures to identify, analyze, and evaluate risks from powerful models before deploying them.
2. Dangerous capability evaluations. AGI labs should run evaluations to assess their models' dangerous capabilities (e.g. misuse potential, ability to manipulate, and power-seeking behavior).
3. Third-party model audits. AGI labs should commission third-party model audits before deploying powerful models.
4. Safety restrictions. AGI labs should establish appropriate safety restrictions for powerful models after deployment (e.g. restrictions on who can use the model, how they can use the model, and whether the model can access the internet).
5. Red teaming. AGI labs should commission external red teams before deploying powerful models.
6. Monitor systems and their uses. AGI labs should closely monitor deployed systems, including how they are used and what impact they have on society.
7. Alignment techniques. AGI labs should implement state-of-the-art safety and alignment techniques.
8. Security incident response plan. AGI labs should have a plan for how they respond to security incidents (e.g. cyberattacks).

9. Post-deployment evaluations. AGI labs should continually evaluate models for dangerous capabilities after deployment, taking into account new information about the model's capabilities and how it is being used.
10. Report safety incidents. AGI labs should report accidents and near misses to appropriate state actors and other AGI labs (e.g. via an AI incident database).
11. Safety vs capabilities. A significant fraction of employees of AGI labs should work on enhancing model safety and alignment rather than capabilities.
12. Internal review before publication. Before publishing research, AGI labs should conduct an internal review to assess potential harms.
13. Pre-training risk assessment. AGI labs should conduct a risk assessment before training powerful models.
14. Emergency response plan. AGI labs should have and practice implementing an emergency response plan. This might include switching off systems, overriding their outputs, or restricting access.

Contributing to AI governance with technical skills

Here are governance agencies/companies which require technical skills :

- **An audit agency**: Require the mandatory formation of an expert team (red team) by third parties; these would need level x access for y months before market launch in order to verify dangerous failure modes. The "red teamers" are external auditors who uncover vulnerabilities or flaws in AI system security.
    - This would mean conducting studies in line with Evaluating Dangerous Capabilities and the paper Model Evaluation for Extreme Risks.
    - Representative work examples are Red-Teaming the Stable Diffusion Safety Filter, the GPT-4 System Card | OpenAI, or the Update on ARC's Recent Eval Efforts.
- **Interagency oversight commission**: Require the examination of major AI models before releasing them into the wild. When Apple wants to release a new iPhone, it must first submit a prototype and documentation to the FCC. This encourages companies to pay more attention to security.
    - Improve the definitions of AGI, following Risto Uuk et al's paper (FLI) on how to operationally define a General Purpose AI. The t-AGI framework by Richard Ngo is also an appreciated refinement.
- **External incident investigation team :** When an airplane crashes, an independent team investigates to find out what happened and how to prevent it from happening again. We still don't know why Bing chat repeatedly threatened users. We want to prevent future AIs from threatening people.

Below are some – not necessarily detailed, other subject ideas for projects that could make progress on AI safety.

- **Removing bias from Language Models.**
    - Classifying LLM output to avoid manipulation/bias/harms

- **Mitigations against prompt injection**
- **Backdoor auditing** - Pair-up with Hugging face to assess the presence of Trojans in open Source LLMs. Backdoors are patterns that allow neural networks to be manipulated. The classic example is a stop sign on which patterns have been placed: the autonomous car's neural network has been trained to react by accelerating when it sees these patterns, which could allow malicious individuals to cause accidents. It is becoming increasingly easy to upload pre-trained networks (foundational models) to the web for everyone to use. Setting up verification mechanisms that allow such networks to be audited before distribution is a major problem in AI safety.
- **Policy development** - Creating a leading institute which would propose policy development work—i.e. developing concrete proposals about what key actors (e.g. governments, AI labs) should do to make AI go well.
- **Compute governance** - creating a Startup pour to implement the software to facilitate GPU regulation as proposed in the paper What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring, this requires technical knowledge of hardware in Deep Learning. Investigating questions related to AI hardware, e.g. the technical feasibility of tamper-proof monitoring/ verification of AI training runs. There is a large literature on compute governance works.
- **Corporate governance** - Create a consultant organization implementing to audit the leading AI labs and Open Source model developers.
- **Consensus building**, by creating a non-profit organization who would coordinate the creation of "the IPCC of AI Risks".
- **InfoSec** - Improving information security at organizations working on AGI development and their suppliers (more)
- **Observatory and forecasting** - Forecasting on questions related to the development of advanced AI (more)

More resources

A few actors in the field of AI safety
Here are some actors that would have been successful during this hackathon!

- Conjecture - a London startup doing AI safety research
- RedWood research - an organization that does research in theoretical interpretability.
- Arc Evals – an AI Audit organization, which audited GPT-4
- Apollo Research - another AI Audit organization, delving deeper in interpretability