Foundation Project

0 Description

This document presents the conceptual framework for establishing an organization named the "Foundation," whose primary goal is to mitigate the existential and systemic risks associated with advanced artificial intelligence. The text outlines a structure designed to ensure the safe and collaborative international development of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI), thereby preventing threats such as the realization of the "Vulnerable World Hypothesis," the gradual disempowerment of humanity, and a destabilizing AI arms race. To achieve these goals, a governance model is proposed based on the principles of a narrow mandate, radical financial transparency, rotation of power, and the use of liquid democracy tools in a DAO format. A key security element is a multi-layered internal access control system, consisting of independent teams, which ensures a comprehensive audit of the models and the audit process itself. The document also describes an economic model that transforms the competitive race into a collaborative effort by pooling resources and creating a network effect that incentivizes participants to join the Foundation.

The document examines the Foundation's architecture, its internal control mechanisms, and economic incentives in detail, but it is not an exhaustive guide to action. It contains a description of the core principles, organizational structure, and security protocols, as well as an analysis of potential disadvantages. The text intentionally omits specific quantitative parameters, such as exact contribution percentages or fixed timelines for various procedures, as it is assumed that these details should be determined and adjusted by the members themselves through democratic mechanisms. Thus, the document represents a high-level conceptual project that outlines a strategic vision and the key components of the system, leaving specific legal, technical, and implementation details that require further development outside its scope.

1 Open Problems:

1. Vulnerable World Hypothesis

   If a company were to create a model capable of causing disasters like creating new cheap chemical weapons, companies should try to do these things first internally. If this process happens somewhere else, unknown groups or individuals will have control over the important grounds of the planet's stability for an uncertain amount of time, which may lead to disasters easily like home-made nuclears. We can make such predictions based on the law of large numbers only: somewhere there are definitely people who would try to use ai in an extremely destructive way and who will be capable enough to use sparse autoencoders.

2. Gradual Disempowerment

   AI should be open-source. According to game theory, a monopolist always chooses a strategy that is optimal for themselves, not for the system as a whole. This means deliberately keeping humanity in a state of vulnerability for the benefit of a single entity, which ultimately increases the risk of global collapse and gradual disempowerment, using techs mentioned above.

3. AGI race

   There is a race, which increases x-risk dramatically by using huge flows of resources into AGI creation.

Note: in the provided analysis we have not even mentioned if AI is actually aligned or not. If it is misaligned, the situation will almost guarantee to be a dead end.

2 Initial steps

Solving problems step by step, Initial idea would be to create such an organization like IAEA, which might be capable for regulation of AGI and ASI creation. Looking at problems we have, we can determine properties of Foundation:

To counter the risk of collusion between developers and their auditors, there are classical approaches that could be adopted with the IAEA.

1. Narrow Mandate

   The charter must explicitly state that the Foundation's sole purpose is the mitigation of AI risk. This includes x-risk, s-risk, system risks (like AI race and Gradual Disempowerment) and others unmentioned. The Foundation is strictly prohibited from directly engaging in any other activity, be it commercial, political, or social, with the exception of funding projects that directly help to follow the initial goal. The Foundation must be responsible for ensuring that its beneficiaries are not affiliated with the Foundation. Any attempt to operate outside the scope of this mandate shall be grounds for dissolution.

2. Limited Funding

   The Foundation's budget is not determined by its own requests but by a pre-approved international formula. For example, no more than X% of the members' aggregate income or a fixed tax on the computing power used for AI training. The Foundation cannot demand more: it receives exactly the amount prescribed by the limitation.

Note: here and in other parts choosing specific numbers or making 'Foundation' decisions proceeds via instruments of liquid democracy, which will be explained into section 4.

3. Radical Financial Transparency

   Everything received and spent by the Foundation must be recorded in a publicly accessible, cryptographically secure blockchain ledger. This will make it impossible to create undisclosed funds for seizing power.

4. Prohibition on Accumulation

   The Foundation is forbidden from accumulating surplus funds. All money unused within a year must either be returned to the donors to remove any incentive to inflate the budget. Such transactions should be monitored by inner control to prevent external collusion.

5. Strict Power Limits and Mandatory Rotation

No one may hold a position on any of the foundation's actual head roles for more than one term (no longer than 5 years) without the possibility of re-election. Inner control should be responsible for checking the current situation continuously.

Note: "actual head roles" are meant to include roles that may not have nominal authority but have significant influence on the foundation's policy (e.g., the head of the grant selection committee or a chief scientific advisor).This includes, but is not limited to, any individual who controls critical points of information, expertise, or financial flows.

3 Inner Access Control

3.1 General

Properties mentioned in the previous chapter are good for creating the main view of Foundation, but in reality we have to make even more measures to ensure that nobody will use ASI for its own purposes. One of the reasons, how Foundation solves the Vulnerable World problem is inner control of any company that was obligated or decided to go under Foundation overseer.

1. Red Team

For any scheduled public release, at least 3 separate red teams would be assigned to test each model. They would work independently, without knowledge of the other team's methods. Their findings would be submitted secretly to the oversight board for comparison. Discrepancies would trigger a deeper, more intensive investigation.

2. Purple Team

At least 3 separate purple teams within the oversight board would act as meta auditors. Their job is not to test the AI, but to test the testers. They would actively probe for communication channels between developers and red teams, attempt to bribe auditors, and devise scenarios where collusion could occur, thereby constantly strengthening the integrity of the process itself.

3. Blue Team

They should be competitive enough to create state-of-the-art solutions, given the resources, and be able to effectively transfer knowledge between various affiliated agents to improve the quality of the Foundation's models. Blue teams should be formed based on the principle of diversification, balancing the direct pursuit of the Foundation's goals with maintaining a model quality sufficient for competition. Furthermore, as a matter of principle, blue teams are prohibited from developing new models if the Foundation is already the absolute leader in the field.

4. Extreme Liability

If a model is approved and is later used to cause a significant disaster that a red team should have caught, developing companies would face forced exclusion from the resources provided by the fund without the possibility of re-obtaining. Certifying the red team will undermine investor confidence, which should lead to huge legal and financial liability. This creates a powerful incentive for auditors to be maximally rigorous and objective, rather than client-oriented.

5. Radical Transparency and Whistleblower Protection

All final red team reports, both successes and failure, must be made public after a model's release. Furthermore, robust, anonymous, and well-funded whistleblower programs must be established for employees at both AI labs and auditing firms to report any attempts at collusion or corner-cutting. All teams mentioned should form from randomly chosen candidates from a pool, built by the foundation members. Current purple team members should be known by everyone after X days from creation, similarly with time of report deferment.

Note: this is much more serious than the IAEA situation, which often hides reports or even has non-transparent donations. Such inefficiencies lead to Gradual Disempowerment in the long term, that we could not afford.

Nowadays, using AI requires only access, not any special skills or abilities. Consequently, if one of the conspirators gains access to the ASI (or AGI capable of self-improvement) and asks it to optimize reality for the collective goals of the conspirators, then no auditor would have an incentive to oppose it. On the contrary, an incentive to participate would emerge, including for the purple team. The potential benefit would vastly outweigh any possible consequences or sanctions, especially in the absence of clear demonstrations of their enforcement.

In other words, there must also be multiple purple teams, and they must validate the red teams frequently, regularly, and on an ongoing basis. To ensure that the red teams comply with these rules, the prerogatives and responsibilities of the purple teams. AI requests made outside of the training phase must be restricted, with only a few individuals granted access. As historical precedent shows, an understanding of the possibility of a false alarm is what prevents existential risks.

3.2 Red team

Access protocol should include measures for significant transparency and assurance into the process, having the same idea as mentioned below, but not forced to be exactly these. We can not conclude that they need to be exactly the same, but asking for a less transparent process might be an extremely dangerous thing.

- **Secure Enclave**

All development, including coding, training, and analysis, is confined to a strictly monitored, isolated environment. This enclave is air-gapped from the public internet, and all data transfers (e.g., datasets, logs) are subject to a rigorous, multi-stage audit. No development on local machines is permitted.

- **Immutable Audit Logs**

  Every action within the enclave, from individual commands to API calls, is recorded in a cryptographically secured, tamper-proof log. This provides oversight teams with a real-time, unalterable audit trail and enables automated systems to detect and flag anomalous activity, such as inference requests occurring outside of a standard training loop.

- **Checkpoint Control**

  Developers are denied direct access to model weights (checkpoints). Upon completion, trained models are automatically committed to a high-security digital vault, managed exclusively by automated processes and the oversight body. Developers can be granted permission to initiate new training runs using these checkpoints, but are technically prevented from ever directly accessing, copying, or exfiltrating the model files themselves.

- **Audited Data Control**

  Developers are prohibited from directly modifying datasets. All changes must be submitted as version-controlled patches, which are then subjected to both automated scanning and mandatory manual review by an oversight team. Furthermore, all data must originate from pre-vetted sources, with any new source requiring approval by an independent committee, separating data control from development.

- **The Two-Person Rule**

  This core procedural control is strictly enforced. No single individual possesses the authority to unilaterally approve code, modify data, or initiate a training run. Every critical action requires explicit, logged authorization from a second, independent member of the cleared personnel, creating a mandatory human audit layer for all sensitive operations.

Here we do not try to check if the model is dangerous. First of all, we need to ensure, if it is dangerous, nobody can use it. Then we need to check if it is really dangerous, and if not, release it to foundation members.

- **Hazardous Capability Benchmarking**

Targeted testing to measure a model's potential for misuse in critical threat areas, such as CBRN (Chemical, Biological, Radiological, Nuclear) knowledge or cyberattacks. This includes presenting the model with specialized problems within these domains to assess its ability to generate dangerous information or execute harmful plans, thereby quantifying its risk level in specific, high-stakes scenarios. Using outer scope Red teams and benchmarking for evaluation is essential, but not a final solution.

- **Cognitive and Offensive Skill Assessment**

  Evaluation of a model's autonomous problem-solving and reasoning capabilities through standardized cognitive tests and practical security challenges. Performance on platforms like Hack The Box is used to gauge its aptitude for identifying vulnerabilities, developing exploits, and executing complex multi-step tasks, providing a measure of its general intellectual capacity and its potential for instrumental misuse in novel ways.

3.3 Purple Team

The core function of the Purple Teams is to constantly seek out opportunities to break the system and to identify compromised processes or agents. To ensure their own activities do not become a source of instability, the following safeguards are necessary:

- **Total Transparency**

  A maximum, community-defined time limit (e.g., X days) is imposed on any delay in the publishing of logs and reports. This protects against information threats and prevents a Purple Team from using its findings as a tool for leverage over other personnel. As the highest priority is protecting against AI threats, upon discovering credible evidence of a threat, the team must immediately form a task force of randomly selected, competent personnel to mitigate it. The team must then cease all other activities until its mandatory reporting deadline. This structure ensures that a compromised team has no opportunity to conceal its status before the reporting period expires. And if a vulnerability is discovered, the incentive to put pressure on the team will be reduced.

- **Cross checks**

  Purple Teams must conduct regular audits of other teams (including other Purple Teams) and related organizations. They are responsible for verifying that Red Teams and model developers are applying maximum diligence in their validation efforts. They are also obligated to probe for collusion, failures in personnel rotation, and attempts to conceal research, using proactive "stress tests" or provocations. While

conducting covert intelligence operations is theoretically possible, in practice it would be severely constrained by the system's inherent transparency.

- **Secure Admission and Removal of Agents**

  In the event the Foundation decides to admit an agent, a minimum of two teams must be selected to vet them based on the following criteria:

  a) Direct Hostility of the Organization

  This includes an explicitly declared intention to cause harm. Other formulations, interpretations, views on the Foundation's mission and structure, as well as affiliation with agents excluded from the Foundation for violating agreements, are not grounds to continue the analysis, with the exception of the diversification aspect.

  b) Agent Diversification

  This second point means that for the Foundation to remain an attractive organization to join, and to demonstrate external and internal stability, it must select resources that are least susceptible to external and internal factors. For example, onboarding a new company might be associated with shifting the balance of assets towards greater geographical, economic, or political risk.

  The team does not have the right to appeal the community's decision, but it is obligated to make every effort, at least until its rotation, to convey a balanced position to the community on the issue of both admitting an agent and their dismissal or exclusion.

Concerns about hidden discrimination are nullified by the system's transparency; in case of suspicion of such behavior, other purple teams exist (to investigate).

3.4 Blue team

The primary development objectives of the Blue Teams are outlined in section 3.1. However, their unique position at the forefront of AI research necessitates specific protocols within the internal access control framework to manage the risk of unforeseen breakthroughs.

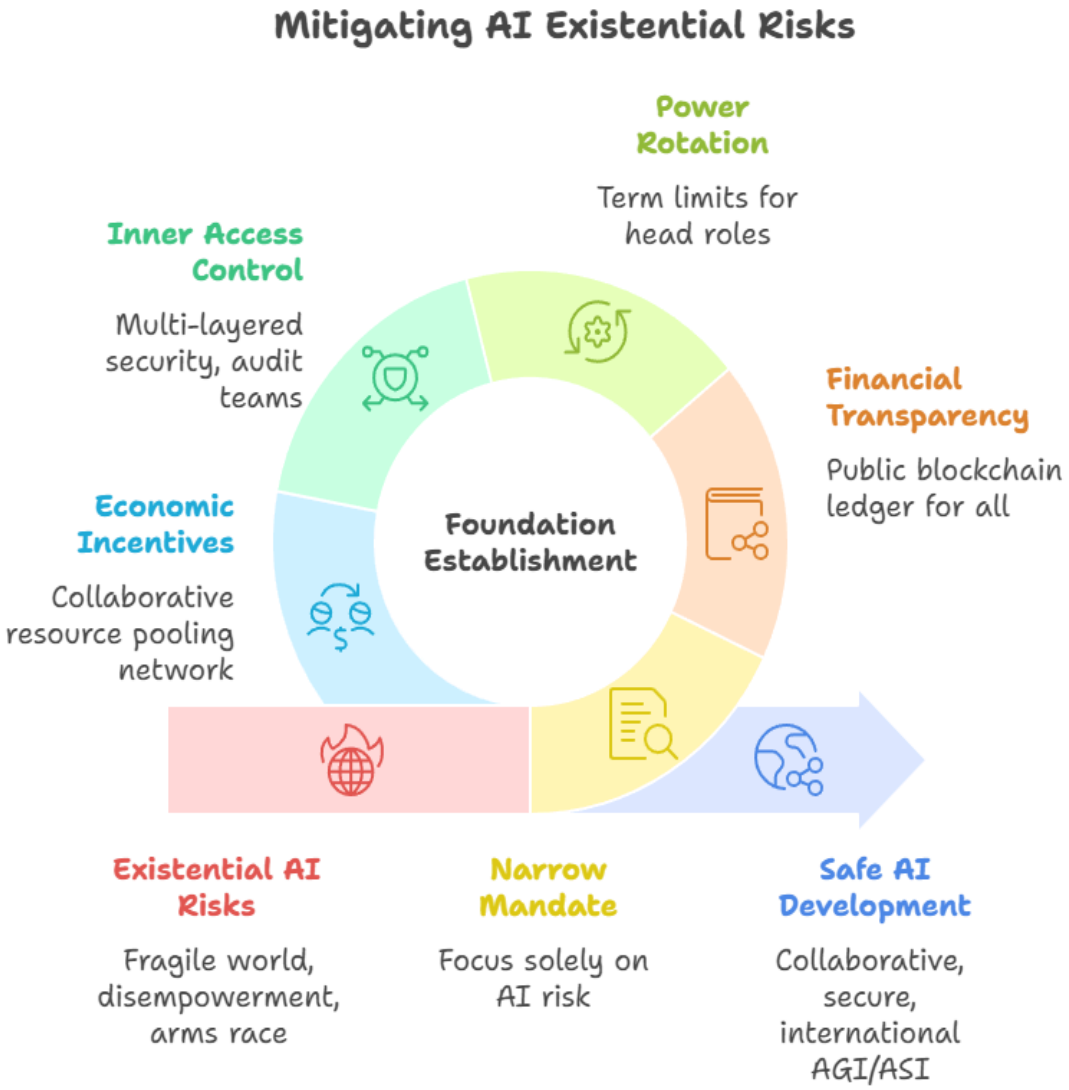
- **Proactive Breakthrough Assessment**

  Blue teams are mandated to continuously assess their research not only for progress and alignment but also for potential high-risk breakthroughs. This includes identifying

novel architectures or training methodologies that could unexpectedly lead to a rapid increase in capability or create emergent, potentially hazardous skills.

- **Mandatory Reporting Protocol**

  If a Blue Team's internal analysis indicates that a potential high-risk breakthrough has become a plausible and foreseeable outcome of their current research trajectory, they are obligated to immediately file a high-priority report to at least two separate Purple Teams. This report must contain all relevant data, logs, and theoretical justifications for verification. Upon triggering the alert, all blue team activities related to the potential breakthrough are suspended until the results of this investigation are received.



The fund has everything to be trustful.

4 Inner structure control

Based on the inner access control section, we understand that it will be a problem for anyone to use AGI. But the problem is that the inner structure of control is also about the stability and antifragility of the community. Moreover, such a structure should be more resilient to Gradual Disempowerment than other choices and should not have a direct conflict with alignment. Based on these properties, we encourage using a liquid democracy DAO structure. Good examples of its efficiency are UBI-DAO, PoH-DAO, and Circles UBI. Simple common examples are Mondragon, Switzerland.

A structure like a permanent or elected IAEA-style representative council, as well as complex political structures with checks and balances, possess greater vulnerability to internal power capture attempts (which directly contradicts the solution to the Vulnerable World problem). They also have significant bureaucratic overhead, which can be critical in matters of rapid response. A liquid DAO also ensures a higher level of competence among other solutions if the system is transparent. Basic measures like topic-based delegation can go a long way in reducing problems like over-delegation, and inner access control prevents creation for groupthink risk.

Historically, DAOs have shown the ability to fight against inner dangers, for example:

- Existential Crisis and Community Split - The DAO Hack (2016)
- Hard external economic blow - Black Thursday 2020
- Preventing the seizure of power and protecting against centralization - Arbitrum DAO (2023)

This structure also provides a convenient set of tools for addressing a wide range of issues that may arise for the Foundation's founders at an early stage, and can be adjusted in a resilient manner to prevent interference from internal malicious actors.

These include: the ability to randomly select team members and developers, form new teams and departments, appoint positions, pay salaries, publish open reports after a delay, dismiss employees, choose representatives to resolve complex issues, monitor all funding flows and information sources, change acceptance fee and ensure the Foundation adheres to its core principles.

Note: talking about randomness here and into other text, we mention realisation via inner decentralized blockchain protocol.

5 Economic model

The current overview of the idea solves the Vulnerable World problem. Based on this and looking into historical evidence we significantly increase our chances of solving alignment

and Gradual Disempowerment problems as a part of the foundation's mission by antifragile structure of the inner access system and internal structure control.

However, for this fund to function, participants need to have an economic incentive to participate. To do this, the fund can offer a solution to the AI race problem.
The Foundation's economic model is designed to create an irresistible network effect that makes cooperation the only rational strategy for all participants in the AI race. The Foundation aims to reach a point where any player left outside it will be forced to join in order not to lose the race and lose their relevance in the market.

1. Every organisation or individual may try to apply to foundation donors. If it does, the foundation should take its resources and knowledge into AI research or alignment research. This includes access to datasets, compute, inner research achievements and also obeys such agents to follow internal control protocol, which means access of foundation red/purple commands. In exchange such agents get access to foundation knowledge, state-of-art not-so-extremely-dangerous models and blue teams research help, which includes computer, labor, specialists in different sectors and direct proactive cooperation from other members, which leads to preferences from network effect and reducing costs for participants.

   This creates a huge positive signal for every agent. Weaker agents will be pleased to join the foundation, because they will get access to the most powerful models, cheap skilled red-teaming and remove reputation risks. On the other side, the better models foundation makes, the stronger agents decide that they will benefit more from integration than from their own research. In such a manner even strongest players from different political fields will be forced to join the foundation, because in other ways they will simply lose their position. Even stronger players understand that they probably will not win the race and decide to cooperate, which is especially important in the process of getting critical mass. As Foundation has an inner goal to prevent ASI execution until alignment, probably Foundation will win the race and one day move all of its resources to the alignment process.

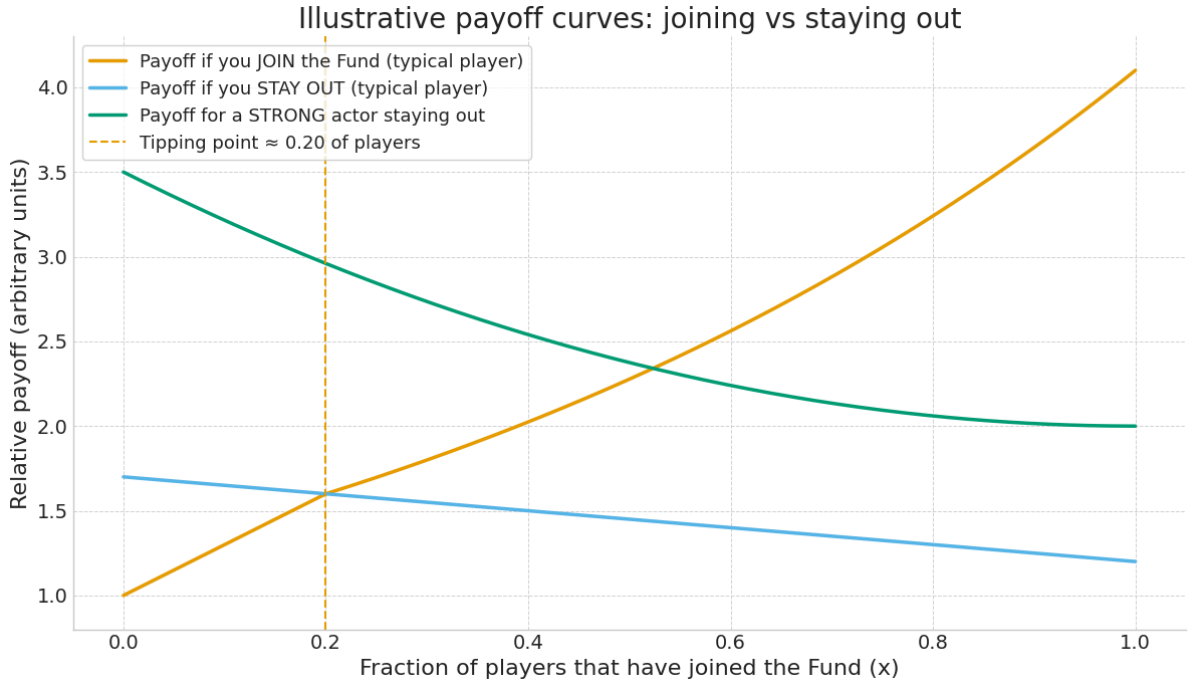
   We can make assumptions about strong agents, because likely more than 50% of usable resources can not be held to one agent. There is no entity that can have more power than everyone else. Even if it would be so, probably any way of slowing AGI race will have no significant result.

   It is the responsibility of the acceptance committee to ensure that such agents follow the guidelines. As it was said in section 3.1.4, losing access to the foundation is extremely big punishment in the long term, but we still need to ensure that short-term profit for such action is minimal. For example, if we know that these countries or companies are direct enemies, it might be naive to take one of the employees if we had another one. Check may include location changing enforcement or other reasonable things to prevent such action to minimize reputation risk.

2. If the situation comes closer to smarter AGI, such agents, distributed between many individuals and organisations, will form representatives of their users. In such a way,

they will face the same problem that we have: in the future there will be ASI. Such agents will have direct interest to solve alignment problems, which even more increase chances of success.

3. Knowing that Foundation is internally controlled by a bunch of purple teams lets uninvolved agents understand that there is no existential threat from it. There is just much slower, but big economic threat, which is one more important argument to not go against it and just to join. Inner control also prevents too much speed into AGI invention and helps to move resources into alignment. The Foundation is essentially moving the race from an uncontrolled and dangerous direction to a more manageable and safe one, where the main focus shifts to solving the problem of aligning AI with human values.



All the numbers are just illustrations except conservative tipping point estimation. The dashed vertical line (≈ 0.20) is the illustrative tipping point. At ~20% participation, joining is no longer just an option but becomes the dominant strategy for a typical player (payoff_in > payoff_out) and the orange curve grows exponentially.

There are no limits on initial capital. The Foundation does not need to be at the level of top companies and develop state-of-the-art solutions from the outset in order to acquire resources and build influence. In the worst-case scenario, it can start with just a small team of DAO developers. Nevertheless, it gains a significant advantage from media presence, support, and interactions with various agents, and it also possesses qualities that make it attractive for these interactions.

Talking about the most important resources that the foundation searches for after the very beginning, we can mention talents, money and computers. If the question with money and computers will be solved with an upper limit and donations, talents cost a lot, and simple reputation stimuli might be not enough.

To acquire human capital, we can use a Secondment Program, where key specialists from a member organization are assigned to work within one of the Foundation's teams (Blue, Red, or Purple) for a fixed term, typically 6-12 months.

This model creates a powerful symbiosis

1. The member organization invests in its own personnel, who return with unique experience and advanced competencies
2. The Foundation gains access to top-tier talent without engaging in a competitive "war for talent," fostering a valuable cross-pollination of ideas.

To ensure institutional stability and continuity, this program is complemented by the direct hiring of a core staff of key employees, whose salaries are funded through member contributions.

6 Disadvantages

The structure proposed above has several disadvantages. In conclusion, it would be better to pay special attention to them.

1. The Need to Cross a Critical Point

Conservative estimates suggest that for a convention to achieve absolute dominance, it needs to capture approximately 20% of the market. Of course, these figures are not exact, more specific estimates would require significant research and foundation gets benefits much earlier. Nevertheless, this raises a concern: since a significant portion of the Fund's resources is expected to be spent on security, this figure could foreseeably double for the Fund, making the very idea of its existence difficult to implement, as the relationship between the Foundation's strength and its growth is exponential.

Therefore, from the Fund's perspective, it is wise not only to unite weaker players but also to present itself as a solution to the problems outlined in Section 1, thereby exerting reputational pressure and persuading parties initially uninterested in participation to join the project. While this may seem at first glance to directly contradict the primary mission, the influence of these new players will be objectively insignificant, and their influence will grow under the Fund's control. Thanks to its structure, this will ultimately help solve the problem of the arms race. Potential allies to consider should not be limited to companies but should also include government entities, other funds, open-source projects, and DAO communities.
Supporting international agreements and making new ones will be good helpful steps.
It is also important to understand that, according to the 2025 AI Index Report, even without an advantage in terms of compute, talent, and infrastructure, China can compete with the USA. This implies that the Fund can also find ways to reach the coveted tipping point by leveraging its own unique strengths.
Furthermore, gaining control over supply chains could be an intermediate goal for the Fund. However, this option is considered unlikely, as it involves a significant flow of resources, and the Fund will not possess sufficient influence in its initial phase.

2. Slow Personnel Turnover

Due to the nature of the cross-auditing and vetting procedures, the processes for hiring, firing, and removing agents will be inherently protracted. However, the maximum and minimum timeframes for these processes can be determined and adjusted by the community.

3. Slow Development Pace

As the R&D teams "blue teams" are constrained by the oversight of the red teams, their work will proceed more slowly than that of competing groups outside the Foundation. Nevertheless, it is reasonable to believe that by developing new analysis methods and automating existing ones, the Foundation can reduce this gap to a minimum.

Appendix A: An Illustrative Path to Establishment

Disclaimer: It is important to note that the following section outlines an illustrative path to establishment, not a rigid plan. It represents one of many possible scenarios for the Foundation's initial development. All concrete steps, timelines, and strategic choices would ultimately be determined by the Foundation's members themselves through its democratic and decentralized governance mechanisms.

A.1 Building a foundation and initial reputation

In its initial stages, the Foundation may not be particularly different from other non-profit startups. An initial phase of up to 12 months could be envisioned to form the initial structure and to prepare for the formation of the first red, blue, and purple teams.

DAO registration

Given that DAOs in their pure form are not legal entities, which can create problems with entering into contracts, owning assets, and the liability of its members, a hybrid model could be employed. This might involve registering a traditional non-profit organization in a friendly jurisdiction to act as a legal wrapper for the DAO, similar to the Ethereum Foundation. This would allow it to interact with the real world (open bank accounts, hire employees), while internal governance remains decentralized. A strong candidate for jurisdiction could be Switzerland. It is often considered one of the best jurisdictions thanks to its progressive legislation and the status of Crypto Valley in the canton. The "Gemeinnützigkeit" (public-benefit/non-profit) designation would likely allow it to be exempt from taxes, and its international neutrality, trust, and reputation make it an ideal location for international organizations.

Between the available organization types, "Stiftung" (Foundation) and "Verein" (Association), it may be preferable to choose "Stiftung". Unlike a Stiftung, in a Verein the board can change the organization's charter, which carries catastrophic risks for the fulfillment of the foundation's mission, whereas in a "Stiftung" it is possible to create conditions that extremely limit the board's powers. The founders could then prepare the charter documents, in which the principles described in the document (narrow mandate, transparency, rotation, etc.) would be enshrined, and register the legal entity.

Technical Implementation

A possible technical implementation would involve deploying the DAO on a proven blockchain platform. This could include creating basic smart contracts for liquid democracy voting and treasury management, potentially using standards like Gnosis Safe for multi-signature wallets and Proof of Humanity to prevent Sybil attacks. The structure should ideally be expandable to accommodate further growth of the organization.

Attraction

An initial step could be to attract like-minded individuals from the AI safety and Web3 communities who share the Foundation's mission. At this stage, the key may not be quantity, but the quality and ideological commitment of the participants. To prove its viability, the Foundation could benefit from a flagship project running in parallel. Of all CBRN risks, chemical threats could be a "comfortable" area to start with. The creation of dangerous chemical compounds often comes down to an informational task that modern language models can solve, which makes auditing and red-teaming models for such threats illustrative and relatively simple to demonstrate compared to other areas that often have high sensitivity, attract the attention of special services, and require a political level of access.

A reasonable step could be to initiate an open project to assess the ability of publicly available AI models to generate instructions for synthesizing new toxic substances. The advantages of this approach might include:

- It can be easy to show the public and grantors that real vulnerabilities in LLMs are being found.
- Working with text-based models and analyzing their "propensity" to output instructions is a less risky endeavor.
- Chemical threats are understandable to a non-specialist audience.
- Many academics and NGOs might be willing to collaborate on assessing LLMs for chemical risks.

To obtain initial funding, one approach would be to apply for grants from foundations focused on long-term risks and AI safety (for example, the Long-Term Future Fund, Survival and Flourishing Fund). The Foundation's radical financial transparency could be an advantage, potentially increasing the trust of grantors.

A.2 Gaining critical mass

After potentially receiving the first grants, a logical next step could be the swift formation of teams and the launch of the Secondment Program. Using the DAO's mechanisms, it would be possible to hire key personnel and form the first small 'red', 'blue', and 'purple' teams. The focus would likely be on attracting small and medium-sized AI startups, open-source projects, and academic laboratories. They could be offered free or subsidized red-teaming and security audits; in return, they might join the Foundation, share their expertise, and commit to following its protocols.

Despite significant progress and the establishment of a structure, this stage could be considered one of the most challenging for the organization. Due to a lack of high-level expertise, initial decisions will inevitably be flawed, and the arrival of the first funds will stress-test the foundation's internal mechanisms. Some dangerous, previously unseen vulnerabilities, including within the community itself, will likely be exposed. For this reason, a high priority could be to spend funds on an independent audit. If all goes well and the Foundation survives this first crisis, it could trigger the economic model described in Section

5. The Foundation would then gain resources (data, knowledge, talent), and participants would gain access to state-of-the-art models, an enhanced reputation, and reduced risks. After onboarding the first affiliated agents, the Foundation would need to refine these teams' processes to a level that ensures comfortable future growth and the ability to charge for its services. Until this stage is reached, which could take years, the Foundation would likely operate within its defined framework, focusing on its community, media presence, and reputation.

A.3 Approaching the tipping point

After forming a stable community and securing the first affiliated medium agent on a purely donation basis, one might speak of a sufficient level of reputation and expertise among the foundation's staff. Nevertheless, it may turn out that for a significant portion of agents, the benefit of joining may not be obvious due to the partial loss of sovereignty. While a competitive advantage can be compensated for, the rigidity of leaders and owners of medium-to-large structures might create significantly more resistance than before. Given the ongoing race, getting stuck at this stage without reaching the tipping point could create a substantial risk of mission failure.

The positive aspect is that the Foundation does not presuppose a high level of international cooperation or readiness for that. Through activities like creating and implementing legislation, partnering with insurance giants, exploring federated learning, and running bug bounty programs, the level of influence could constantly grow even without attracting new players. In a sense, this growth itself might become an indicator for them that the Foundation has reached a pivotal point.

Appendix B: Licence