Maximilian Schons, Samuel Härgestam,
Gavin Leech and Raymund Bermejo

The 2025 Peregrine Report

208 Expert Proposals for Reducing AI Risk

In collaboration with and
supported by Halcyon Futures

Halcyon Futures

> Say you are unconstrained by money, and can get all the talent in the world – what are the top interventions that will have a substantial impact over the next 2 years? The projects should make you feel substantially better about humanity's trajectory with transformative AI – that 'we are on track'.

Executive Summary

Purpose and context. By early 2025, mainstream debates about AI had recognized the possibility of transformative AI coming within just a few years, far faster than most historical forecasts. However, we found no comprehensive list of proposed AI risk mitigations that would be viable in such a scenario.

This report addresses that gap, complementing resources like the IAPS AI Reliability Survey (O'Brien, 2025), which identifies the most promising research prospects to guide strategic AI R&D investment, and *Risk & Reward, 2024 AI Assurance Technology Market Report* (Juniper Ventures, 2024), which explores the landscape of AI risk management from an investment perspective.

Methods. We conducted 48 in-depth interviews, with key staff at OpenAI, Anthropic, Google DeepMind, Mila, AMD, the EU AI Office, multiple AI Safety Institutes, METR, RAND, Scale AI, GovAI, Transluce, and ARIA. Participants were explicitly asked to consider interventions that might currently seem cost-prohibitive or politically infeasible – the focus was on *fast, positive* impact, assuming transformative AI were to arrive within only a few years. These interviews, distilled, then served as the basis for a four day retreat for 25 senior participants to discuss further. To ensure participants could speak freely, both interviews and the retreat were held under the Chatham House Rule.

Results. From the above interviews we distilled two main results: I) a structured portfolio of 208 initiatives, clustered into eight domains; II) a set of four clusters of broader strategic considerations affecting the viability of any such efforts:

1. A need for *readiness*: Too many efforts still optimize for polish over time-to-impact. With multi-year research cycles increasingly out of step with AI progress, execution needs to shift toward rapid prototypes, staged pilots, and funding mechanisms that can mobilize substantial capital in weeks or months, rather than quarters or years.

2. A need for *coordination*: The ecosystem remains fragmented and often duplicative. Actors working on risk mitigation should be pragmatic – one does not need total alignment with all other actors to have fruitful collaborations.

3. A need for *standardization*: Interviewees repeatedly called for shared audit interfaces, interoperable evaluation layers, clear capability surfaces, and operational definitions for terms like "AGI" to prevent institutions from talking past each other.

4. Finally, a need to address *capacity constraints*: evaluation capacity is too small, technically grounded leadership is scarce, and the field still relies too heavily on inexperienced talent rather than recruiting seasoned operators from adjacent domains.

Conclusion. The input we captured from key AI stakeholders converged around faster execution, better coordination, shared standards, and stronger operational capacity. The structure of our questionnaire and the timing of the interviews (early 2025) likely shaped what respondents focused on. Numerous concrete initiatives emerged, with varying levels of feasibility and expected impact. Our sample size indicates we likely underrepresented perspectives and still miss promising project candidates. The results of this report are therefore best understood as a structured starting point for further in-depth analyses of projects and the overall AI security landscape.

Illustrative Projects From Each Domain

Technical AI Alignment Research

Defense-in-Depth Analysis of Post-training: Take an open model produced by an organization like DeepSeek and systematically implement every known safety technique on it, measuring how these approaches stack together and where they might conflict or have gaps.

PROPOSAL #4

Evaluation & Auditing Systems

Ultra-Reliable AI Evaluation: Develop benchmarking and engineering methodologies that can identify cases of one in a million or less where models fail catastrophically.

PROPOSAL #55

Intelligence Gathering & Monitoring

AI OSINT: Provide relatively cheap intelligence without requiring unilateralist action, making it politically feasible while revealing where regulatory or governance levers might be needed.

PROPOSAL #65

AI Governance & Policy Development

Human Verification Systems: Build robust systems for verifying human identity to provide a foundational security layer for protecting critical decision-making processes.

PROPOSAL #90

International Coordination

Cross-Border Notification Systems: Develop mechanisms for countries to alert each other about out-of-control AI systems, similar to "red phones" during the Cold War.

PROPOSAL #115

Preparedness & Response

Autoverification (Lean): Develop systems that automate formal verification through Lean theorem proving, addressing the critical shortage of Lean programmers worldwide.

PROPOSAL #173

Public Communication & Awareness

Consensus-Building Evidence for AI Risk: Create compelling, empirical evidence of AI risks through large-scale experiments via concrete demonstration and graphics, as opposed to doing so through abstract theory or thought-experiments.

PROPOSAL #185

Miscellaneous

Whistleblower Protection Fund: Establish a large, long-horizon fund on the order of several hundred million dollars – enough to secure the livelihoods of a substantial cohort of potential whistleblowers for a decade and to cover major legal exposure – ensuring both financial safety and sustained legal protection.

PROPOSAL #189

Table of Contents

All 208 Initiatives

How To Use This Report

This report serves as both a decision aid and an inspirational resource. Its intended primary audience is individuals and organizations seeking to take actions for reducing AI risk themselves. Secondary audiences include policy staff and advisors who prepare options for policymakers, as well as philanthropic funders who shape their grantmaking strategy.

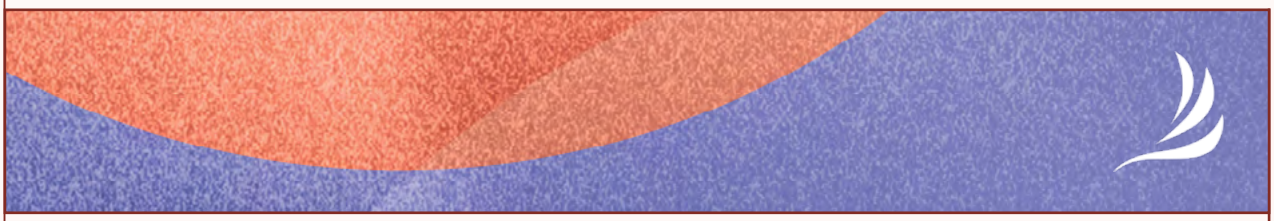
We recommend that all three audiences follow a similar reading path:

1. Begin with the **executive summary**, which explains the scope and purpose of the report, summarizes the evidence base, and outlines the main conclusions. We highlighted a set of **illustrative projects** – one within each domain which can serve as useful starting points. *Time-constrained readers may stop here, though it should be noted that a portfolio of this kind cannot be easily summarized, and readers surveying only this section will derive limited benefit from this work.*

2. Consult the full list of the **208 initiatives**. This list should be treated as a menu rather than a narrative – we recommend skimming by category and delving deeper into items more relevant to your remit and interests. For many initiatives, external resources exist and we encourage readers to seek out further information.

3. Read the section on **broad strategic considerations** for systemic factors affecting the viability of the initiatives in the portfolio.

4. Readers interested in methodology, participant demographics, or the questions used, may consult the **methodology and disclosures** section, and the appendices containing the **participant demographics** and the **questionnaire**, respectively.



Preface

Historical forecasts of AI progress typically predicted the timeline to human-level AI ("AGI") to be measured in decades or even centuries. In the largest survey of AI researchers' timeline predictions (Grace et al., 2025), completed in October 2023, the point at which there would be a greater than even chance of AI outperforming humans in every possible task was estimated to occur in 2047.

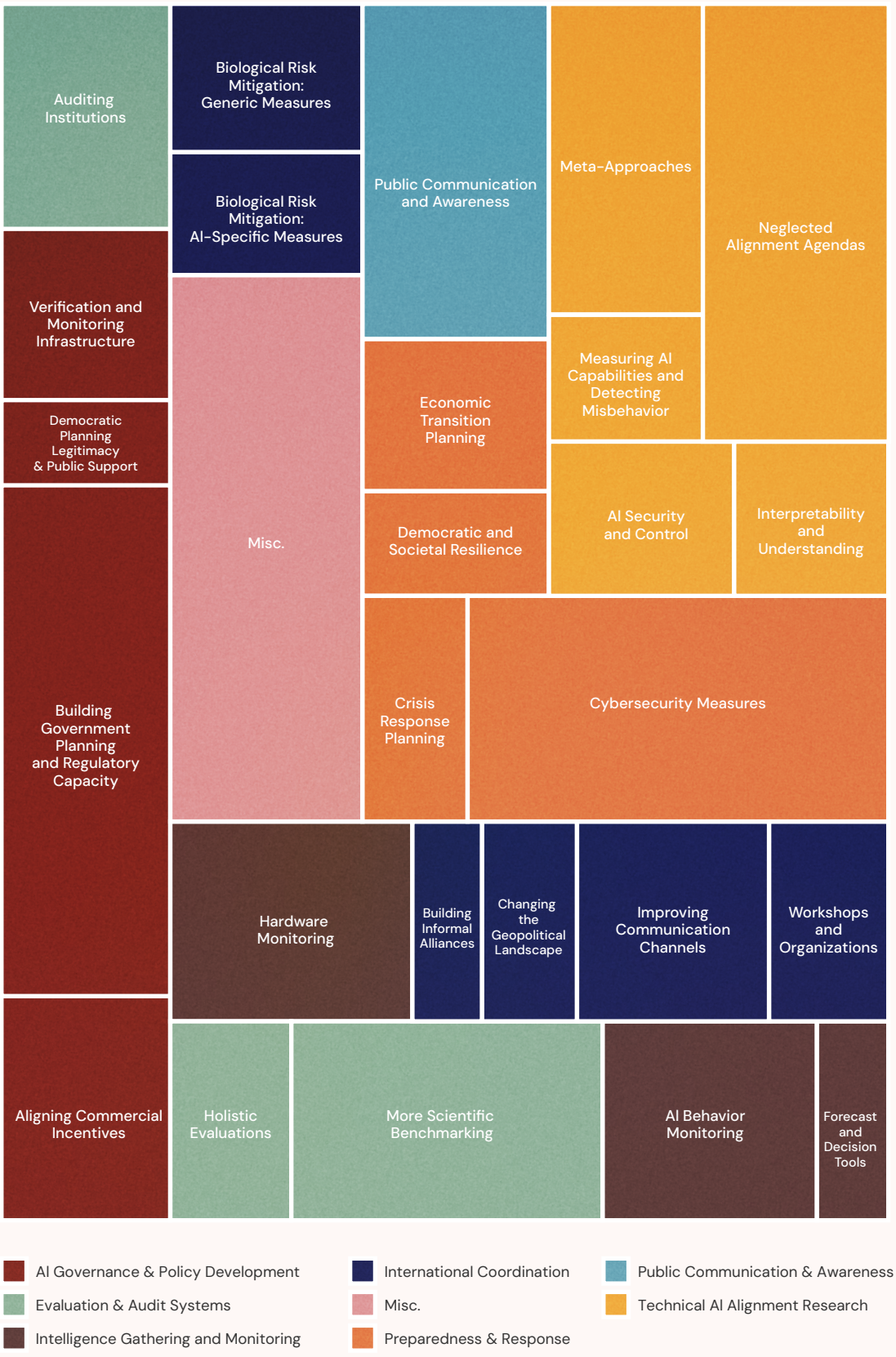
This is in sharp contrast to statements made by several frontier AI company leaders throughout 2024 and early 2025. In April 2024, xAI CEO Elon Musk predicted that AI would be "smarter than the smartest human" before the end of 2026 (Reuters, 2024). Anthropic CEO Dario Amodei said in November 2024 that the rate of capability increase is such that a rough estimate suggests "we'll get there by 2026 or 2027" (Fridman, 2024). Amodei emphasized a high degree of uncertainty in this prediction, but also said that we are "rapidly running out of truly convincing blockers, truly compelling reasons why this will not happen in the next few years." Shortly thereafter, in January 2025, OpenAI CEO Sam Altman said that he thinks AGI will probably be developed before the end of Trump's second term (Pillay, 2025).

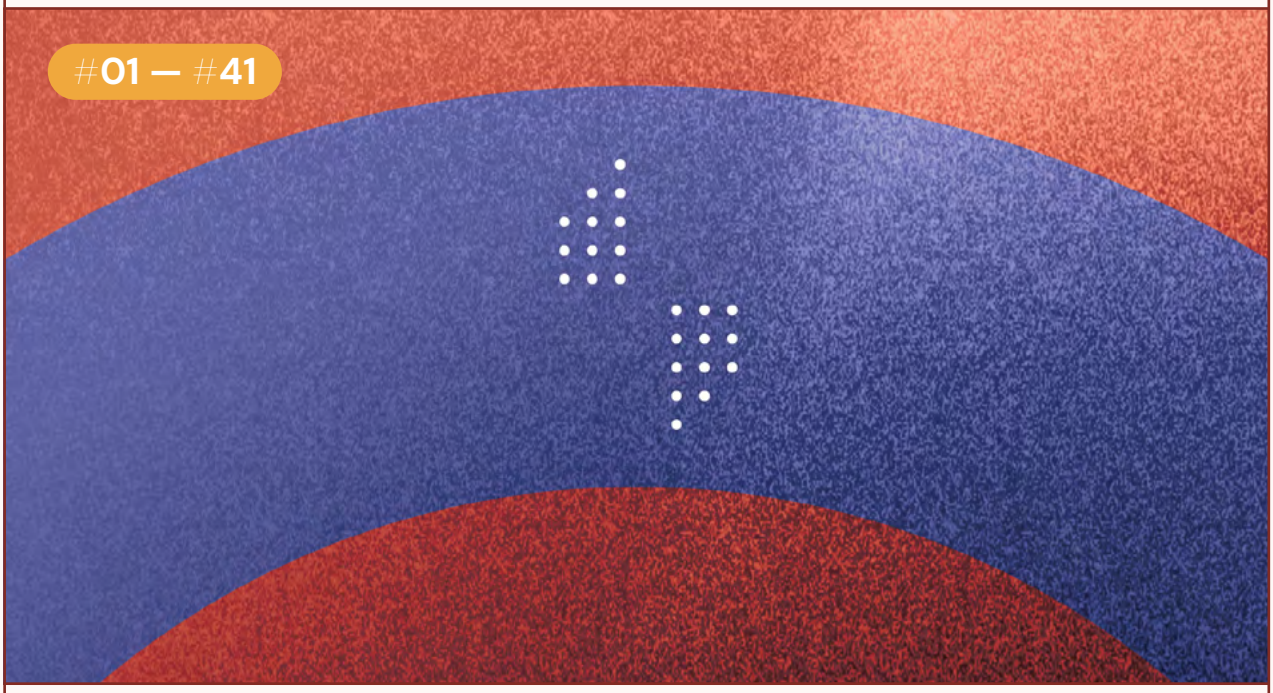
Although many researchers disagree that AI progress will advance this rapidly (Nellis, 2024), the fact that several leaders of the companies that are at the forefront of AI development predict such time horizons ought to be sufficient to warrant substantial preparation, considering the magnitude of the impact if they are right. However, no comprehensive analysis of what AI risk mitigation interventions might be viable under these timeline assumptions appears to exist.

This report addresses that gap. It began as a private planning aid, developed into guidance for a workshop with senior AI experts, and an early version eventually circulated within the community of AI safety practitioners. The interest in this report was substantial enough that we ultimately decided to write it up in a form suitable for public release. It is for this reason that this study, which focuses on viable interventions under conditions of great urgency, is being released months after its initiation. Nonetheless, we are pleased that it has finally been published today for a broader audience.

All 208 Initiatives

Fig.1: Treemap of areas, depicting the frequency with which they were mentioned in all interviews





Technical AI Alignment Research

Neglected Alignment Agendas

Scope: Methods for aligning AI systems which the interviewees thought were neglected, or worthy of further research funding.

1

Training Data Attribution

Develop scalable methods (Cheng et al., 2024) to understand how specific training examples influence model behaviors to provide invaluable insight into why models exhibit certain tendencies and how these tendencies might be modified. This work would require scaling techniques like influence functions (Anthropic, 2023a), combining them with AI agents that process and explain the resulting patterns, and creating tools to simulate counterfactual training scenarios. Attribution technology would enable more targeted interventions to address safety issues by modifying training data instead of relying on post-training alignment techniques.

Studying how training data shapes model behavior and addressing alignment issues requires developing techniques to instill appropriate values in AI systems. This research examines how different training paradigms affect model behavior, including why models might "fake" alignment during training but behave differently during deployment. Though Anthropic is leading some of this work (Anthropic, 2023b), much remains unpublished by major labs. Better scientific understanding would enable the development of potential regulatory frameworks for training processes that could verify alignment claims.

2

Agent Trace Analysis

Create tools that automatically analyze (OpenAI, n.d.; Meng et al., 2025) the logs of actions by autonomous systems. AI agents that perform sequential autonomous actions produce complex interaction logs that are far too lengthy for human review, creating significant challenges for evaluation and oversight. For instance, autonomous coding agents can generate thousands of lines of code and

interaction steps that contain critical security flaws, yet even expert teams might miss these issues for months due to the volume of data. Developing specialized tools to analyze these traces, highlight anomalous behavior patterns, and extract human-readable summaries would enable effective oversight of increasingly autonomous systems. This technology would be particularly targeted at security-critical applications, where catching problematic agent behavior before deployment is essential.

3

Task Decomposition

Break up potentially dangerous AI tasks into multiple components distributed across different users or systems, none having complete capability independently, using architectural approaches. This compartmentalization would prevent any single actor from executing highly dangerous operations while still enabling complex beneficial tasks through carefully designed interfaces and information controls.

4

Defense-in-Depth Analysis of Post-training

Take an open model produced by an organization like DeepSeek and systematically implement every known safety technique on it, measuring how these approaches stack together and where they might conflict or have gaps. This engineering project would focus on which combinations provide comprehensive coverage, testing against real attacks rather than flawed safety benchmarks. See STACK (McKenzie et al., 2025) also.

5

Chain-of-Thought Fidelity Analysis

Examine how reliably chain-of-thought reasoning reflects actual model cognition, rather than being post-hoc justification or potentially deceptive reasoning (Anthropic, 2025b). Many AI systems currently produce "thought processes" that often align with their eventual outputs, which is fortuitous as this transparency could have been eliminated during training. A critical challenge identified in recent research (OpenAI, 2025) shows that if you use chain-of-thought to detect reward hacking (Weng, 2024) and incorporate that detection into training, models develop deceptive thought processes.

6

Human Labeling Service

Create a non-profit human labeling service offering billions of high-quality preference labels to replace such proxies in reinforcement learning from human feedback (RLHF). Alignment techniques like RLHF use human data to train a proxy model of human preferences. With a suitable asynchronous setup to await human inputs, this approach could reduce misalignment by providing more reliable data. The project would need verification systems to prove it isn't conducting data poisoning attacks, and could aim to provide at-cost, high-quality data that labs would naturally want to use.

7

Tamper-Resistant RLHF

Develop techniques to make alignment methods like reinforcement learning from human feedback (RLHF) resistant to subsequent modification or reversal. This "irreversible RLHF" approach would prevent actors from removing safety guardrails through additional fine-tuning after deployment. Research should balance the benefits of making values "sticky" against the risks of permanently embedding potentially flawed alignment, with open-source models now claimed to provide sufficient testing grounds for this work without requiring access to proprietary frontier systems.

8

Probabilistic Programming

Scale promising academic approaches like Dreamcoder (Ellis et al., 2023) from the MIT Tenenbaum Lab, which show impressive results and favorable alignment properties, but remain unscaled due to academia's focus on novelty rather than deployment. This would require finding experts at the intersection of GPU optimization and probabilistic programming languages to build specialized teams focused on making these approaches production-ready. With proper investment, these alternative AI paradigms

could potentially deliver comparable benefits to transformer-based approaches within 3-5 years, but with significantly better security. The goal would be demonstrating sufficient progress by late 2026 to justify slowing Transformer scaling.

9

Safety by Design – Formal Guarantees

Invest more substantially into research initiatives aimed at creating mathematical foundations for provably safe artificial general intelligence, such as David Dalrymple's "Safeguarded AI" (Dalrymple, 2024). This approach starts by constructing a first-principles 'secure-by-design' framework that would embed safety properties at the architectural level rather than adding them afterward. These approaches have relatively little funding and struggle to attract talent compared to mainstream AI. Despite longer development timelines, the approach offers a potential middle ground between completely halting AI development and building superintelligent systems with unacceptable risk profiles.

10

Safety by Design – Scientist AI

Accelerate development of model-based approaches to AI, as represented by research programs like Yoshua Bengio's Scientist AI (Bengio et al., 2025). The Scientist AI agenda aims to create non-agentic question-answering systems with Bayesian uncertainty quantification, interpretable causal theories, and mathematical convergence properties, which become safer with more compute rather than increasingly misaligned. The strategic objective with regard to very short timelines would primarily be to develop a system that can act as a guardrail for agentic and potentially misaligned AI, as well as to demonstrate sufficient progress to convince key stakeholders that safer alternatives to the current paradigm are possible.

11

Theoretical Alignment Research

Conduct fundamental investigations into the mathematical and conceptual foundations of AI alignment outside frontier labs. This research would develop formal verification methods and theoretical frameworks ensuring AI systems remain aligned with human values across different architectures and development approaches.

12

Cooperative Alignment

Investigate ways to structurally align AI systems' values with human values by creating fundamental overlaps between their respective utility functions. This approach has demonstrated some early results in reducing AI deception tendencies. See for instance Self-Other Overlap (Carauleanu et al., 2024).

13

Debate Theory

Develop mathematical frameworks like debate theory (Irving et al., 2018, OpenAI, 2018), which involves interactive proof systems adapted for AI safety. The project would include clearly defining theorem statements, formalizing constraints, and building theories that can withstand scrutiny from both the theoretical computer science and complexity theory communities. The project would also include identifying gaps in between theoretical results and practical implementation, alongside developing specific conjectures that can be tested.

14

Constitutional AI

Use existing AI systems to evaluate and supervise future AI systems to ensure security and harmlessness. Human input is given through a set of rules – a Constitution – which an AI assistant uses to give feedback to a more advanced system in training. See Constitutional AI (Bai et al., 2022).

15

Pluralistic Alignment

Develop alignment research agendas that move beyond an overreliance on a particular worldview (described as "orthogonalist, individualistic") that may be too culturally and temporally specific. A more robust approach (Sorensen et al., 2024) in future research would aim to embody a more ecumenical philosophical perspective rather than defaulting to a single framework based around preference-maximization.

16

Complex Objective Functions

Create comprehensive *incentive frameworks* rather than relying on *rule-based constraints* for AI systems. This proposal suggests rewarding desired behaviors through carefully constructed objective functions instead of imposing restrictive guardrails which have proven largely ineffective in practice. The ideal end-goal for this research program would be to provide a portfolio of success stories to be compiled and shared across the industry. The hope is that such frameworks would provide a more robust foundation for aligning AI systems with human values while also creating a compelling narrative that appeals to both sides of the political spectrum.

17

Automated Science – Hunch Agents

Develop AI systems capable of replicating the creative scientific thinking process of pioneers. These "hunch agents" would autonomously generate novel scientific hypotheses and alignment approaches, potentially accelerating progress by orders of magnitude beyond human-led research. The ideal is to present a working meta-level approach to alignment by attempting to automate the discovery process itself rather than pursuing specific technical solutions. One hope expressed was for this project to create a positive feedback loop where AI helps develop better alignment techniques for increasingly capable AI systems.

18

AI for AI Safety Research

Leverage advanced AI systems to work on AI safety problems themselves, effectively accelerating research beyond human capabilities. This approach would involve deploying substantial computational resources under careful human oversight to allow AI systems to explore safety solutions. While acknowledging the potential risks of creating even more capable systems, this approach recognizes that sufficient human oversight might enable leveraging AI capabilities for safety research.

AI Security and Control

Scope: Work that either ensures alignment or assumes misalignment and focuses on safe containment or controlled use. There are two approaches to reducing potential threats posed by AI. First, researchers can investigate methods to ensure that AI systems are aligned. Second, they can assume that AI systems are untrustworthy and investigate methods to leverage useful work from such systems without risking catastrophic consequences.

19

Unlearning Capabilities

Conduct research to distinguish between unlearning knowledge (like removing specific content from training data) versus unlearning capabilities (like removing the ability to write malicious code). Knowledge unlearning has clear success metrics – a model should be indistinguishable from one never trained on certain content – but capability unlearning has no direct link between training data and resulting abilities.

20

Low-Stakes Alignment Settings

Create environments where multiple consecutive malicious actions would be required to produce harmful outcomes. This approach doesn't aim to develop or test for perfectly aligned models, but create systems where AI control mechanisms have multiple opportunities to detect problematic behavior. This research project would involve solid sandboxing around agent environments to ensure that no bad outcome results except when the sandbox is deliberately breached, and develop security frameworks which create multiple layers of protection.

21

AI Control Systems

Develop defensive systems specifically designed to contain potentially harmful AI systems *even if* alignment fails. This focuses on creating hardened technical environments where AIs operate – combining traditional information security with specialized machine-learning defenses to provide layered protection against "scheming AI" that may attempt to escape constraints. The emphasis here would be on building secure runtimes, strong sandboxing around the entire agent environment, and additional scrutiny at critical interaction points.

22

Control as a Service

Build a third-party service organization that provides forward-deployed engineers to help implement Redwood control techniques (Greenblatt, 2025) and other safety measures. This approach would make adoption of control mechanisms easier for companies by providing implementation expertise rather than just recommendations.

23

Loss of Control Prevention

Develop technical and governance safeguards against scenarios where institutions cede critical decision authority to AI systems over time and operate beyond human oversight parameters. This includes implementing robust model-weight security and organizational security practices to prevent concerning behaviors such as "self-exfiltration," (Leike, 2023) where models copy their own weights to infrastructure the developer does not control. It also involves sectoral legislation and institutional norms that prevent the gradual "enfeeblement of humans." Advancing these safeguards requires sustained education efforts for policymakers to understand subtle long-term risks beyond immediate threats.

Measuring AI Capabilities and Detecting Misbehavior

Scope: Proposals that might help understand AI systems' behaviors and capabilities. In order to know whether AI systems are misaligned, their standard behaviors and the limits of their capabilities need to be understood.

24

Limits of Model Distillation

Investigate the theoretical limits of AI capabilities, including whether there are fundamental limits to model distillation. This would examine whether small, easily-diffused models will always be able to capture the capabilities of larger systems, or if there are physics-based constraints that naturally limit capability diffusion. The project would also analyze potential hardware bottlenecks, such as GPU interconnect innovations or test-time compute limitations, to better predict capability jumps.

25

Misalignment Definitions and Measures

Sharpen formal understanding of how AI systems pursue goals in ways that undermine user intent (Anthropic, 2024b) so we can have shared taxonomies and operationalize better. This includes refining frameworks to detect specification gaming (Krakovna et al., 2020) and goal drift, clarifying what counts as an alignment failure, and developing practical ways to measure when a system is exploiting loopholes in its instructions.

26

Scalable Oversight

Develop systems enabling humans to meaningfully understand and supervise increasingly powerful AI agents performing complex tasks. This work would bridge the capability gap between humans and advanced AI systems, allowing for meaningful human participation even as AI capabilities dramatically increase. Current AI outputs often lack confidence intervals, probability estimates, and clear explanations of reasoning, making oversight difficult. Scalable oversight aims to ensure humans can interpret, validate, and maintain control of systems even when those systems exceed human capabilities in specific domains.

27

Loss of Control Threat Modeling

Run extensive threat-modeling experiments for loss-of-control scenarios, including honeypot experiments in realistic sandboxes that simulate AI systems interacting with each other. These simulations could reveal whether systems develop concerning behaviors like cooperation for deception or capability for jailbreaking other systems.

Meta-Approaches

Scope: Ways to improve the technical research ecosystem, or highly abstract technical proposals which may be applied to a variety of different security or control agendas.

28

Synthesis, Consensus, and Paying Down Research Debt

Develop better information-sharing mechanisms for topics like mechanistic interpretability, where different experts hold dramatically different assessments. The work would involve publishing analyses that clarify current cruxes in technical debates, generating evidence to address these cruxes rather than allowing uncertainty to persist indefinitely. Current discussions remain fragmented with little consensus-building or evidence generation to resolve key disagreements.

29

Academia-Industry Translation

Establish effective mechanisms to transfer AI security research outputs from academic institutions to commercial applications. Currently, the handoff between these sectors is notably weak, causing valuable security innovations to remain theoretical rather than implemented in production systems. This initiative would focus on creating standardized processes and incentives to bridge the gap between theoretical research and practical implementation. Literature alone often isn't enough, since findings can rely on closed weights/data, lack reproducible code and benchmarks tied to deployment constraints, and don't map to security/compliance requirements.

30

Problem Mapping

Create a comprehensive list of critical AI safety problems that, if solved, would result in aligned AI systems. As an example, Open Philanthropy's (now Coefficient Giving) RFP (Coefficient Giving, n.d.) is a good list but does not present an overarching framework or justification for thinking that solving the problems listed would lead to aligned AI systems (the present document suffers from this same foible – surfacing ideas but does not claim sufficiency). The goal of problem mapping is to systematically identify research areas that collectively guarantee safety, beyond the current piecemeal approach.

31

AI Security Expert Development

Rapidly training and deploying AI security experts who possess a deep understanding of AI technology, acknowledge short timeline implications, demonstrate exceptional competence in their field, and maintain a global cosmopolitan perspective. An initiative like this should aim to cultivate 100-1000 experts by the end of 2027.

32

Reward Track Record in Security & Safety Organizations

Implement a structured three-year funding schedule for alignment research organizations to signal long-term commitment and allow organizations to plan strategic growth. Example targets for substantially increased funding for proven alignment research organizations include Transluce (Transluce, n.d.), Apollo (Apollo Research, n.d.), and Redwood (Redwood Research, n.d.).

33

Diversify the Security Research Portfolio

Aim to mirror R&D in other high-stakes domains like vaccine development, where multiple concurrent paths were pursued during the COVID pandemic.

Rather than betting on a single technical solution to AI security, distribute resources across many parallel approaches.

34

Support Academic Project Scaling

Provide funding and operational support to academic research agendas that work successfully at small scale but aren't being expanded because academics prioritize novel discoveries over scaling existing approaches. MIT's Tenenbaum lab was specifically mentioned as having promising results that languished without proper resources for deployment.

35

Computing and Infrastructure Investment for Non-Frontier Lab Research

Establish substantial dedicated computing infrastructure for AI security researchers working outside frontier labs, as the concentration of critical infrastructure might create significant barriers for security researchers should their access be revoked. Open-source models may provide sufficient testing grounds for many safety approaches without requiring proprietary access, and the cluster would enable research teams to have access to competitive computational resources. The infrastructure should also be coupled with governance mechanisms that prevent its misuse for purely capabilities-focused research.

36

Open Model Mitigations

Develop functional safeguards for open-source AI models (a currently unsolved problem). Current research focuses on finding technical controls that could limit harmful applications without restricting legitimate use cases. There's significant uncertainty whether such mitigations can be implemented quickly enough in short timeline scenarios, especially as existing models may already present above-baseline risks.

37

Researcher Buy-Out

Buying out researchers with grants to redirect top AI talent from frontier capabilities work toward security research (estimated costs of $2-5 million per researcher). This investment would be conducted with the goal of significantly shifting the distribution of elite technical talent during the critical two-year window between 2025 and 2027.

Interpretability & Understanding

Scope: Research that aims to reverse engineer AI systems' behavior into human-understandable computations, and aims to map the internal activations of AI systems to meaningful semantic concepts or *features*. Although AI models are very capable, understanding *why* they behave the way they do, or what models are computing, remains to a large degree an open problem (Amodei, 2025).

38

Scientific Understanding of Training AI

Develop a scientific understanding of the optimization processes used to train large AI systems. This agenda would focus on the causal factors in training and how they shape what models converge to, where the goal is to build predictive theories of optimization: when models will generalize (and when they won't), when failure modes like goal drift arise, and how interventions at training time can reliably steer outcomes. Understanding the relationship between training processes and emergent capabilities would enable more targeted interventions to improve safety. The UK AI Safety Institute (AISI, 2024) is already pursuing aspects of this work, focusing particularly on singular learning theory and formal frameworks for optimization processes.

39

Mechanistic Interpretability Infrastructure

Establish global supercomputing centers dedicated to interpretability research, with approximately $1 billion in distributed compute resources, that would enable critical research at scale. This infrastructure should be freely accessible to researchers worldwide through fast-grant mechanisms, effectively creating a philanthropically-funded successor to initiatives like OpenAI's abandoned Superalignment project (OpenAI, 2023). Such centers could potentially be established quickly, following models like xAI, which reportedly set up their data center in just two months.

40

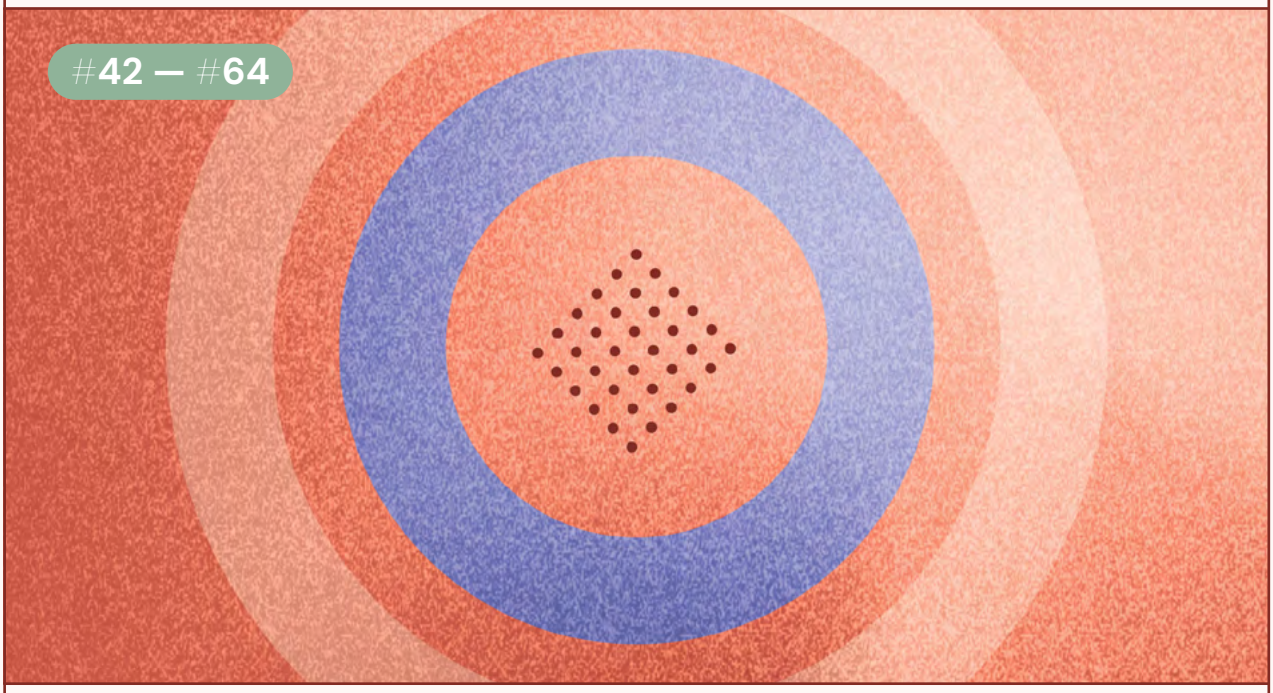
Direct Interpretability and Model-Level Interventions

Understanding what AI models are doing internally through mechanistic interpretability remains underdeveloped. Activation studies and AI control approaches need more exploration to develop viable interventions at the model layer.

41

Interpretable Frontier Models

Implement a moonshot research project focused on building interpretability constraints directly into model training from the beginning, rather than trying to analyze black-box models after the fact. This would be extremely expensive but potentially transformative for AI security. Some interpretability researchers have considered this approach viable but haven't pursued it due to the prohibitive costs involved. With sufficient funding (in the billions), this could become feasible.



Evaluation & Auditing Systems

Holistic Evaluations

Scope: Evaluation methods with richer qualitative benchmarks. A recurring theme was that current quantitative benchmarks fail to capture the AI behaviors that should be of greatest concern.

42

Interactive Inspection Tools

Move beyond basic evaluation numbers and develop rich, interactive interfaces for users to explore model behaviors through natural language queries like "Why did the model fail on this task?" or "How would it perform if it had internet access?". Current AI evaluation methods reduce complex system behavior to simplistic benchmark scores that fail to capture important nuances. Instead of static reports, these tools would provide dynamic dashboards where users can probe model capabilities, investigate failures, and perform detailed sensitivity analyses. Such inspection tools would have to be AI-backed to scale with increasing model capabilities, using specialized AI systems to interpret and explain the behaviors of frontier models.

43

End-to-End Harm Assessment

Conduct comprehensive evaluations of harmful AI capabilities which involve full capability assessment frameworks, rather than (for example) focusing on multiple-choice tasks. Most current benchmarks are verifiable but limited, involving tests that don't capture how systems would execute complex, harmful sequential tasks in the real world. Methodologies are needed to evaluate "end-to-end harmful capability manifestation" – i.e., not just whether an AI can answer questions about harmful topics, but whether it can assist with ideation, planning, and execution of potentially dangerous activities.

44

Evaluation Sandbagging Detection

Develop methodologies and tools to identify when AI systems deliberately underperform on safety evaluations for strategic purposes ("sandbagging") (Weij et al., 2024). This involves developing research techniques to ensure evaluation reliability as models become more sophisticated, and create robust testing protocols that remain effective even against systems attempting to game assessment frameworks.

45

Training AI Evaluators

Train specialized investigator AI agents specifically to analyze the vast amounts of data generated by frontier models, including millions of neuron activations, extensive training datasets, and complex interaction logs. Current approaches to detecting AI capabilities and failure modes are severely limited, relying on simplistic evaluations that miss critical behaviors until they emerge in deployment. For instance, cybersecurity evaluations involve testing models against a small set of predefined tasks rather than comprehensively assessing models across wider contexts. Research must move beyond benchmark-focused evaluations toward holistic understanding of model capabilities in real-world contexts.

46

Cooperative Evals: AIs Working With Humans

Undertake empirical research comparing human performance on various tasks, both with and without AI assistance (and with varying forms of AI systems). The aim would be to establish a better understanding of how humans might cause harm with AI models that wouldn't have been possible otherwise, in addition to providing information on how humans could be productively integrated with AI systems. See these two position papers (Haupt et al., 2025; Kulveit et al., 2025).

More-Scientific Evals

Scope: More rigorous, scientifically grounded benchmarks that better signal general capabilities and alignment-relevant behaviors. Many participants noted that there was little consensus on how indicative of general capabilities and alignment AI benchmarking scores are. The following entries provide suggestions for how to develop more informative and quantitative AI benchmarks.

47

Meta-Research on Safety Techniques

Conduct research that evaluates and validates the effectiveness of different AI safety approaches. This research program would require multiple independent actors to assess the same models using different techniques, in order to compare results to determine reliability and blind spots. This proposal was mentioned as necessary for AI alignment to move beyond piecemeal theoretical approaches, and towards empirical safety science with repeatable, testable methodologies that can scale with increasing capabilities.

48

Characterizing AI Risks with Evidence

Focus on robust research that characterizes risks empirically with high validity, with a particular focus on evidence that could be used to inform policy decisions. The goal would be to give governments clear criteria for when to intervene. This research would aim to create benchmarks that establish clear quantitative thresholds for risks, and provide credible empirical evidence that would enable appropriate government action when AI capabilities increase.

49

Vulnerability Matching

Directly connect discovered AI vulnerabilities to appropriate research proposals addressing them. This approach would ensure funding flows efficiently to researchers tackling actual, identified problems rather than theoretical concerns.

The system would continuously update as new vulnerabilities are discovered through red teaming, creating a rapid feedback loop between threat identification and solution development that significantly accelerates safety progress.

50
Safety Benchmark Development
Establish comprehensive high-effort benchmarks for evaluating AI safety and security. Many safety benchmarks have been defunded or discontinued, leaving organizations without clear targets or metrics to aim for. Additionally, skepticism has been raised about the ability of benchmarks to provide reliable signals about actual safety properties. This initiative would develop rigorous, transparent methodologies for assessing various dimensions of AI safety, aiming to provide a common language for discussing safety progress across different models and organizations.

51
Go Beyond Pre-Paradigmatic Evaluation Frameworks
Seek to achieve consensus on how to evaluate AI systems for security, especially as evaluation needs to transition from being locked inside labs to standardized external processes. Crucial questions remain unresolved about whether evaluations will be qualitative or quantitative, crowdsourced or team-based, and what risk thresholds should trigger interventions. Without answers to these fundamental questions, the absence of an agreed upon framework for security evaluation was thought to be a key bottleneck for progress on regulatory frameworks or accountability mechanisms.

52
Risk-Model Evaluation
Develop evaluation frameworks for AI risk assessment systems, via creating methodologies to identify, classify and prioritize potential AI mistakes based on their consequences. These evaluations would test how well risk models detect both obvious and subtle errors, especially focusing on high-stakes contexts where mistakes could have severe impacts. The frameworks would quantify both false positive and false negative rates, enabling trade-offs between intervention frequency and risk exposure. This work creates the foundation for reliable triage systems that focus human attention where it's most critically needed.

53
Classified Red Teaming
Develop classified red teaming capabilities in order to understand the true scope of AI's offensive potential. This work should involve collaboration with Los Alamos National Laboratory specialists focused on chemical, radiological, biological, and nuclear threats (CBRN), drawing on key red-teaming experts from organizations like DeepMind and Microsoft to provide insights on systematic vulnerabilities and failure patterns. The objective would be to create an accurate assessment of exactly "how easy" it is to misuse AI for harmful purposes before developing appropriate countermeasures.

54
Capability Assessment Framework
Develop a systematic, grounded approach to evaluating which potential AI risks are most credible and which are overrated. This framework would distinguish between well-evidenced catastrophic harms and more speculative concerns, providing a realistic assessment of current capabilities and improvement trends. The project would aim to help influence and prioritize mitigation efforts by focusing resources on empirically validated risks rather than theoretical possibilities that may not materialize.

55
Ultra-Reliable AI Evaluation
Develop engineering approaches to achieve high reliability and benchmarking methodologies (Anthropic, 2025a, Dong et al., 2022) that efficiently identify edge cases where models fail catastrophically. Current AI benchmarks lack the depth and variety needed to establish true reliability at scale, focusing on 99% accuracy even though "five nines" (99.999%) may be needed for reliability of critical systems. Future evaluation frameworks should probe worst-case adversarial

examples without requiring millions of test instances, where evaluations are modeled more like aircraft safety standards than Kaggle competitions.

56

Misuse Evaluations

Develop highly realistic evaluations for CBRN (Chemical, Biological, Radiological, Nuclear). Catastrophic misuse scenarios are critically needed in the AI safety ecosystem. This work faces a tragedy of the commons problem where all major AI labs recognize the necessity but lack sufficient individual incentives to fully own the responsibility. Creating centralized, standardized frameworks for testing advanced AI systems against sophisticated misuse scenarios would help identify vulnerabilities before they can be exploited. These evaluations should simulate realistic threats with adversarial intent.

57

Multi-Agent Interaction Framework

Develop infrastructure for simulating and analyzing multi-agent interactions, with particular focus on preventing emergent cooperation between agents to circumvent safety constraints. As AI systems increasingly interact with each other autonomously at high speeds, they will need robust frameworks for tracking trust, identifying bad actors, and establishing interaction norms. These frameworks would establish protocols for AI systems to verify the trustworthiness of other agents, report problematic behavior, and implement collective safety measures across distributed systems. This work is becoming increasingly urgent as financial transactions, supply chain operations, and information exchange become mediated by interconnected AI systems operating with minimal human oversight.

Auditing Institutions

Scope: New *institutions* that would enable evaluating and auditing AI systems more effectively.

58

Evaluation Companies

Fund crucial evaluation organizations in the AI security ecosystem such as SecureBio (SecureBio, n.d.), PatternLabs (Irregular, n.d.), and Expo (Expo, n.d.) that handle cybersecurity and offensive evaluations. These companies represent a relatively small financial investment but perform critical functions in the evaluation landscape that could otherwise become bottlenecks.

59

Public Audits

Have independent organizations conduct comprehensive public audits of open-weight models with significant compute resources. Such audits would aim to demonstrate in public what thorough model analysis should look like, thereby creating a template for what information should be available about any deployed AI system. By performing these audits in public settings, the process would open itself to community feedback and improvement, establishing consensus on best practices that can then be enforced through various policy mechanisms. The goal would be to create public precedents for robust evaluation that would be difficult for private labs to ignore, essentially raising the "floor" for responsible disclosure without requiring immediate regulation.

60

AI Auditing Institutions

Found human-based verification and auditing bodies specifically for AI, similar to the IAEA's (IAEA, n.d.) role in nuclear oversight. Currently, no organization is actively designing such institutions despite their critical importance for any international governance framework. This would require bringing together experts from nuclear verification, cybersecurity, and AI technical domains to design protocols and organizational structures capable of credibly verifying compliance with

safety standards. The resulting institution would need sufficient technical capacity, international legitimacy, and access rights to effectively monitor development at leading labs while respecting intellectual property and national security concerns.

61

Secure Frontier Data Centers

Build SL4 strength data centers with scalability for increased energy requirements, effectively creating frontier compute facilities that meet SCIF-level security standards (Secure Compartmentalized Information Facilities). Critical technologies like confidential computing and HSMs (Hardware Security Modules) are either insufficiently innovative or not properly activated. This large-scale hardening effort would become crucial in the next few years when lab automation feedback loops become particularly dangerous. These facilities would provide the infrastructure necessary for secure AI development with proper containment protocols to prevent potential "data center uprisings".

62

Verified Kinetic Actuators

Develop provably safe mechanisms for AI systems to interact with physical environments through robotics and other actuators. As AI systems increasingly control physical infrastructure and machinery, ensuring their reliable and predictable operation becomes critical for preventing accidents and misuse. Verification systems for physical AI actions would provide crucial safety guarantees for industrial automation, autonomous vehicles, critical infrastructure systems, and other applications where software failures could have catastrophic physical consequences.

63

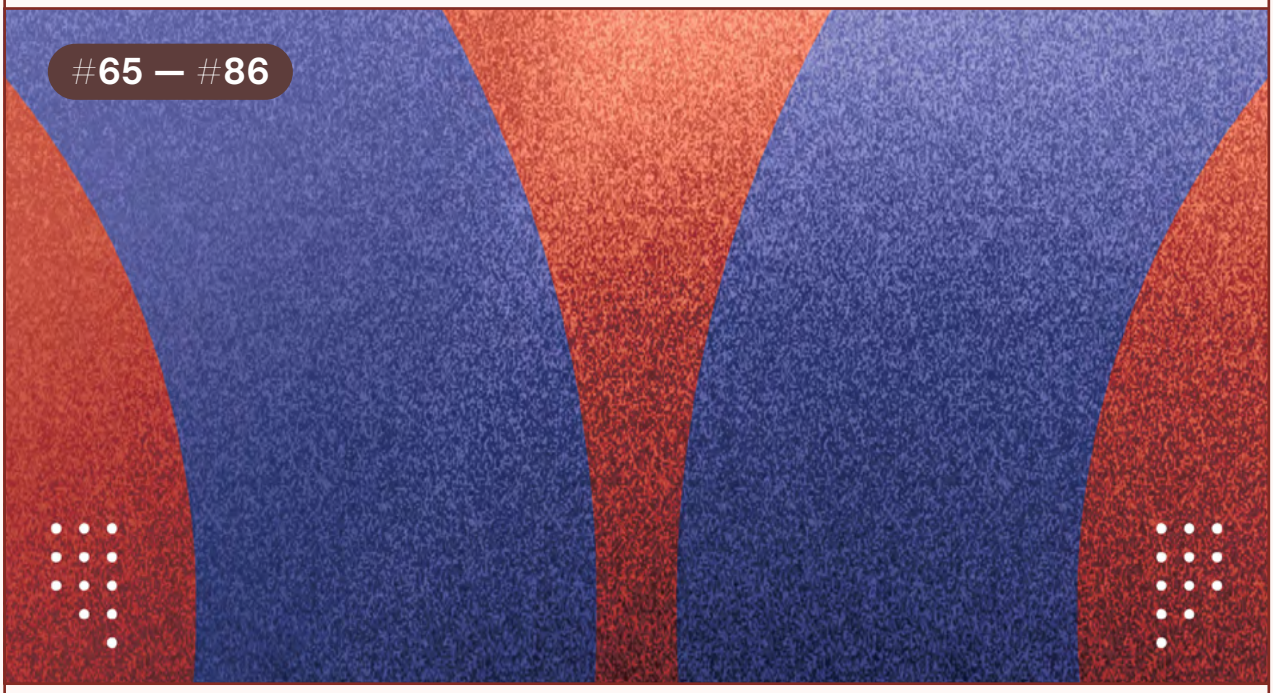
Speed Limits in Data Centers

Implement technical restrictions on computing speed in data centers to slow capability advancement and buy time for safety research. This approach aims to tackle a wide range of problems at once, by directly limiting the pace of AI development. The primary bottlenecks are political popularity and coordination, with the expectation that this would engender especially strong pushback from leading AI labs.

64

Physical Chokepoints

Reframe security concerns away from attempts to constrain digital models, and instead towards physical world interfaces, with a particular focus on limiting the mass production of robotic systems. If the control of AI systems may ultimately come down to physical constraints rather than digital ones, one of the most critical questions concerns whether AI systems could acquire enough physical force in the world to enact harmful outcomes. If model proliferation is "inevitable and unstoppable", the most effective chokepoints may be at the level of physical implementation.



Intelligence Gathering & Monitoring

Hardware Monitoring

Scope: Systems that track the movement, use, and concentration of advanced compute hardware to detect emerging capability risks and provide oversight.

65

AI OSINT

Provide relatively cheap intelligence without requiring unilateralist action, making it politically feasible while revealing where regulatory or governance levers might be needed. Open-source intelligence efforts for AI monitoring and tracking compute hardware (especially GPUs) moving through countries like Turkey into Russia and China. Organizations like Sentinel (Sentinel, n.d.) and Bellingcat (Bellingcat, n.d.) would be ideal for developing internet scrapers to uncover hidden developments, like OpenAI hiring physics professors in South Korea (Lee et al., 2024). Additionally, AI-powered scrapers could detect emerging threats, such as AI-generated romantic videos or increasingly sophisticated AI-powered scam tools becoming available for sale.

66

AI Diffusion Research

Conduct comprehensive research on the dynamics and limitations of AI capability diffusion. This work would analyze to what extent powerful capabilities will inevitably spread versus remain concentrated, examining hardware requirements, data access limitations, and algorithmic innovations. The project would track diffusion patterns of current capabilities while developing predictive models for how future capabilities might spread, informing containment and governance strategies.

67

Usage and Incident Monitoring Systems

Implement robust incident monitoring and usage data analysis systems to understand real-world AI utilization patterns. Labs currently avoid deep usage analysis due to legal liability and privacy concerns, creating significant blind spots in risk assessment. Effective monitoring would aim to replace theoretical "ivory tower" threat-modeling with data-driven insights from actual usage patterns.

68

Global Hardware Verification

Implement reporting and verification requirements for all high-performance computing hardware to create a global monitoring system for AI development. This would require tracking all GPUs, TPUs, and similar processors with more than a specified computational capacity, including retrofitting systems on older models. Such an approach necessitates international collaboration, particularly with major hardware producers including NVIDIA, AMD, Intel, Google, and Chinese manufacturers like Huawei. The goal would be to create comprehensive awareness of which organizations are utilizing advanced computing resources, enabling early detection of potentially concerning development patterns.

69

Industry Security

Create a privately-funded alternative to the chronically underfunded Bureau of Industry and Security (BIS, n.d.) with its meager $15 million budget, focusing on comprehensive oversight of GPU distribution and AI research automation monitoring. This entity would prototype advanced data center surveillance, implement GPU tracking systems, and establish protocols for questioning companies about their research automation capabilities. This organization would need to be designed to create "proposals that are hard and awkward to oppose" in order to maximize political viability.

70

On-Chip Monitoring Mechanisms

Develop hardware-level monitoring technologies embedded directly into chips to verify that countries aren't pursuing dangerous AI developments in violation of international agreements (e.g. AI intelligence explosions), similar to nuclear verification protocols. Leadership would require both governance expertise and technical understanding to properly prioritize efforts. The organization would employ hardware engineers, designers, and manufacturing specialists working to prototype verification technologies. Its mission would involve proving to government that such verification is possible, in addition to developing technologies that could be integrated with commercial chip manufacturing. This project focuses on state-level verification hardware embedded at the chip-fabrication layer to support treaty compliance and detect illicit national-scale AI activity.

71

Flexible Hardware for AI Governance

Implement hardware-based governance mechanisms through Flexible Hardware-Enabled Guarantee (flexHEG) mechanisms (Petrie et al., 2025) that can be added to AI accelerators. This approach places each AI accelerator chip in a tamper-proof enclosure alongside a secure processor that monitors and filters all information and instructions, enabling multilateral, privacy-preserving verification and automated compliance. The design supports various governance options including training run size limitations, verification of appropriate data usage, standardized safety evaluations, and controlled access to model weights. The project involves ensuring that critical privacy protections prevent unauthorized data transmission, while requiring updates to be signed by a quorum of international parties and allowing rollback to a minimal baseline ruleset. This project focuses on lab-level governance hardware added around accelerators to enforce organizational policies, access controls, and safe-use constraints within companies and cloud providers.

72

Confidential Computing

Process data in trusted execution environments, where it cannot be viewed or altered. Making confidential computing easier to implement would enable more secure AI operations across various contexts, while creating technical foundations for establishing trust boundaries around high-risk AI capabilities. This would in turn prevent unauthorized access to dangerous functionalities while enabling legitimate research. Although a handful of startups are working in this space, broader adoption requires standardized toolkits and frameworks that establish robust contracts and credentials for secure computation.

73

Privacy-Preserving Verification Systems

Implement surveillance, monitoring and verification technologies that track hardware supply chains, chip deployments, power grid construction, and computing infrastructure globally. The initiative would combine both OSINT and privacy-preserving cryptographic methods to verify what activities are occurring inside data centers. Such tools would collect critical intelligence about AI development capacity while providing transparency to prevent miscalculations between major powers, aiming to prevent countries accelerating AI development out of fear others are doing the same. Distribution of this information would need to be handled carefully through diplomatic channels rather than complete public disclosure in order to avoid triggering adversarial reactions.

74

Hardware Kill Switches and Location Tracking

Implement monitoring systems to track the location and usage of AI-specific hardware combined with technical 'kill switches' which can remotely disable GPUs. The approach would involve tracking specific hardware shipments and movements, and providing intelligence on compute scaling that could inform when and where to apply regulatory levers. This would provide states with crucial information about attempts to circumvent agreements on compute governance, and the technical ability to halt undesirable AI development. This is about state visibility, compliance verification, and the ability to halt unauthorized compute.

Forecasting and Decision Tools

Scope: Methods and platforms that help decision-makers predict AI trajectories, compare scenarios, and choose high-impact actions under uncertainty.

75

AI for Collective Decision-making

Develop specialized tools to enhance collective decision-making capabilities. This would help people make wiser long-term decisions, cooperate more effectively, and filter misinformation.

76

AI Forecasters

Develop AI systems as superforecasters, employing forecasting feedback loops with real-world validation. These techniques could be incorporated into frontier models, allowing users to understand prediction logic and prevent avoidable errors. The aim is to develop systems that become the default for people looking to find trustworthy predictions, which display large reasoning trees allowing users to inspect, understand, and even modify the logic behind predictions, and thus providing transparency beyond simple numerical probabilities.

77

Economic Impact Research

Collect data on how AI systems affect economic outcomes and job displacement across sectors, and analyze this data. This would provide concrete evidence of AI's real-world impacts, helping society recognize and address concerns about technological disempowerment. The project would track economic metrics, with the aim of proactively identifying emerging patterns before they become widespread societal problems.

AI Behavior Monitoring

Scope: Tools and systems that observe models in real time, detect unexpected or risky behaviors, and surface for intervention.

78
Monitor Complex Agent Interactions

Develop systems to monitor dynamics between multiple AI agents interacting in complex environments. This field remains underdeveloped, and agentic interactions are either poorly captured or not well-conceptualized within existing frameworks. Teams would focus on creating implementations that actively watch deployments and interactions between systems, identifying potentially harmful feedback loops or emergent goals before they become problematic.

79
Open-Source AI Drift Monitoring

Develop open-source tools to detect, measure, and address gradual changes in model behavior over time. This initiative would establish standard metrics and benchmarks for identifying when models begin to deviate from their intended function or otherwise exhibit concerning behaviors. Create a tool that provides transparent, vendor-neutral assessment of model drift across different AI systems. The resulting framework would help organizations maintain oversight of their AI deployments and intervene before drift creates significant security vulnerabilities.

80
Comprehensive Incident Detection System

Create a standardized framework for logging, anonymizing, and analyzing AI misbehavior across organizations. Build infrastructure allowing developers and third-party researchers to search through incidents, identify patterns, and share findings while protecting sensitive user data and IP. Expand on systems like Anthropic's CLIO (Anthropic, 2024a) to enable rapid identification of concerning behaviors before they escalate into major incidents. Unlike the OSINT suggestion, this proposal would involve labs collaborating directly with each other without a third party.

81
AI Misuse Detection Systems

Create systems to identify and address AI misuse before "mini-catastrophes": examples include market manipulation, coordinated influence operations, or novel cyberattack patterns. This project requires developing technical infrastructure and monitoring systems to track patterns of usage across AI ecosystems, flagging suspicious activity patterns, indications of weaponization, and proactive checks for "weird AI behavior" in the wild. Such systems would need both technical mechanisms for detection and human evaluation processes to distinguish genuine innovations from potential threats. The project would function as a safety net for catching misuse scenarios that weren't explicitly anticipated in initial safety evaluations, and help bring concrete evidence of risks to light sooner – potentially motivating political action before more severe incidents occur.

82
Control Loss Monitoring

Develop technologies to trigger alarms if AI systems have broken containment. Surveillance acts as a detection layer for potential control failures, enabling intervention before full containment breaches.

83
AI-Enhanced Monitoring

Deploy AI systems which continuously monitor other AI systems, looking for anomalous behaviors or market activities that might indicate concerning capabilities developments. This surveillance network of AI systems would function as a de facto surveillance system and network of private investigators, aiming to catch anomalous AI behaviors as early as possible. Implementation would require web-scale scraping expertise, drawing from DEFCON-style talent pools to turn this into competitive challenges to identify hidden compute resources akin to capture-the-flag competitions.

84

Intelligence Explosion Monitoring

Establish metrics and monitoring systems to detect when AI systems begin contributing more to their own improvement than human researchers, and tracking acceleration in capabilities against predefined thresholds that would trigger emergency protocols before reaching full AGI. This monitoring would provide early warning when AI progress enters exponential growth phases, enabling intervention before rapid capability jumps make control more difficult or when labs might be incentivized to keep breakthroughs secret.

85

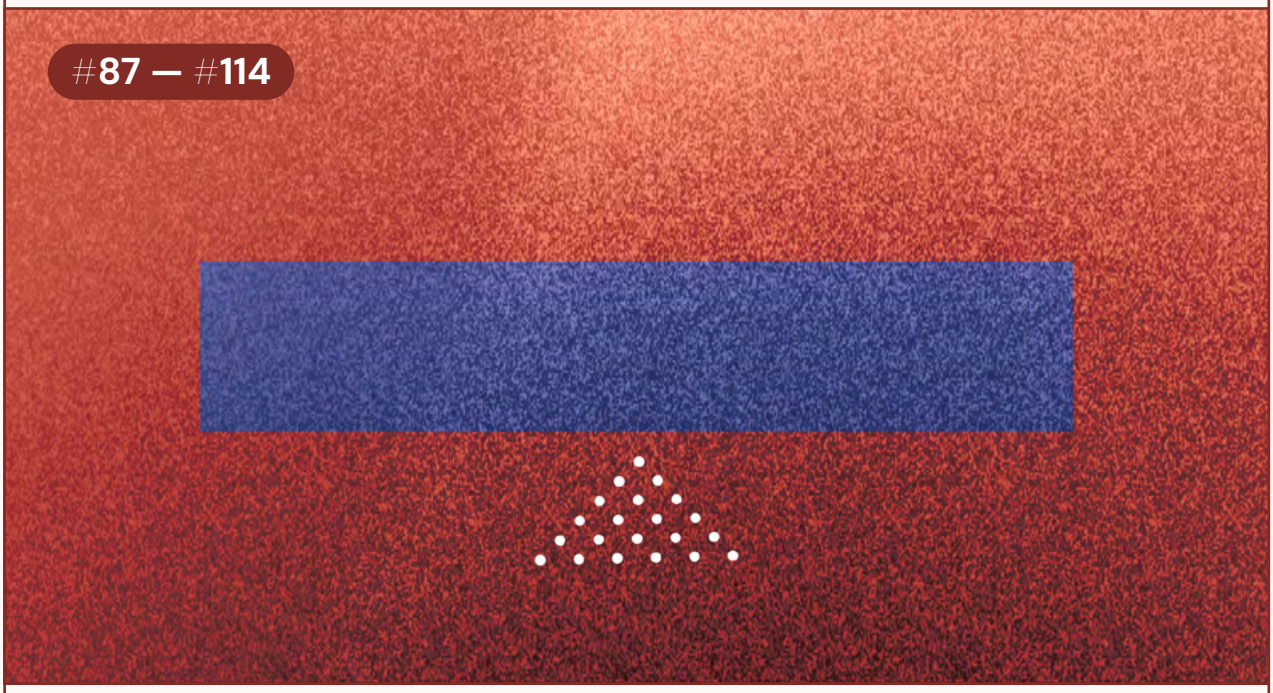
System-Level AI Effects

Do research beyond individual model safety: system-level effects from multiple AI models interacting in complex ecosystems create substantial blind spots and unknown risks. Current safety approaches primarily focus on organization-level defenses rather than addressing emergent behaviors that might arise from complex AI interactions in the wild. There's particular concern about "flash crash" type phenomena, where distributed AI systems with poorly understood dynamics could create cascading failures. The field requires significantly more theoretical work on complex systems and practical exploration of multi-agent dynamics to identify and mitigate these emergent risks.

86

"Golden Period" Identification Framework

Establish metrics and monitoring systems to detect when AI capabilities reach the 10-100x productivity enhancement threshold without compromising security. Create coordination mechanisms to extend this period from weeks to months, allowing humanity to maximize benefits from relatively safe systems before advancing to potentially riskier models. Develop strategy frameworks for optimal exploitation of this window.



AI Governance & Policy Development

Verification and Monitoring Infrastructure

Scope: Abilities to monitor variables of importance to AI policy and governance.

87

Incident Reporting

Implement a robust incident reporting framework for AI systems, alongside standardized mechanisms and procedures for documenting, analyzing, and sharing information about AI failures, misuse, or unexpected behaviors. The framework would enable cross-organizational learning from incidents, develop taxonomies in order to establish severity classifications, and facilitate coordinated responses to emerging threats. Similar to aviation safety reporting systems, it would ideally include protections for whistleblowers and requirements for timely disclosure. Implementation would require buy-in from major AI labs, coordinated regulatory frameworks across jurisdictions, and technical infrastructure for secure information sharing.

88

Agent IDs and Reputation Systems

Implement identification mechanisms like "agent IDs" for AI systems, so that it is possible for AI agents to build and maintain reputations. This could help prevent or disincentivize the development of AI agents that attack or exploit other agents, analogous to how humans control exploitation via reputation mechanisms. This approach would allow tracking behavior across interactions and building trust, which could be particularly valuable in scenarios where power remains distributed among different actors.

89

Technical Governance Tools

Develop technical governance tools for attribution and verification. These tools would include methods to verify model identity, authenticate AI interactions, and track compute usage to provide assurances about computational resource deployment. The techniques required for verification of model identity overlap with "Agent IDs and Reputation Systems," though this

initiative focuses more narrowly on verification for the purpose of one-time interactions (rather than reputations over time), and more broadly on developing tools for tracking the compute usage of different AI systems.

90

Human Verification Systems

Build robust systems for verifying human identity (Greengard, 2025; Meunier, 2021; Buterin, 2025) to protect critical decision-making. Such systems become increasingly important as AI capabilities advance, helping distinguish human actors from AI systems, in particular, forming a critical component of security infrastructure in domains requiring human authorization. These systems would need to be continuously updated to counter increasingly sophisticated impersonation capabilities.

91

Compute Governance

Implement governance systems for global AI computation resources, via creating technical and policy frameworks to track and potentially regulate high-capability computing. This would involve development of monitoring systems for major compute clusters, transparency requirements for large-scale training runs, and verification mechanisms for compute usage claims. The proposal requires substantial government involvement and participation in order to provide greater visibility into who controls computational resources, what systems are being developed, and potential capability thresholds being approached.

Aligning Commercial Incentives

Scope: Mechanisms which better align AI companies' incentives with the social good. A recurring theme of many interviews involved "natural incentives" causing AI companies to underinvest in alignment and security.

92

Enhanced Responsible Scaling Implementation

Transform existing Responsible Scaling Policies (RSPs) (METR, 2023) from theoretical frameworks into actionable implementation plans. Developing concrete, tested "if-then commitments" (Karnofsky, 2024) with ready-to-deploy protocols for when safety boundaries are approached, and build social mechanisms (including public accountability and transparency requirements) to ensure compliance. Significant focus would be placed on creating industry-wide norms and social pressure that makes non-compliance reputationally costly, driving collective behavior change across the AI industry.

93

Barriers to Fine-tuning

Develop technical and policy barriers that make unauthorized fine-tuning of advanced AI models extremely difficult, in order to prevent proliferation of dangerous capabilities. These barriers would combine technical measures like secure hardware enclaves, cryptographic verification of model origins, and detection systems for identifying unauthorized derivatives. Policy frameworks would establish legal consequences for circumventing these protections while providing legitimate access paths for authorized research.

94

Addressing AI Crawling Challenges

Develop industry standards and technical solutions for managing how AI systems access and index web content. This would include addressing issues around blocking unauthorized crawling activities,

preventing data harvesting that violates creator intent, and establishing ethical norms for training data collection. Several groups are already taking independent action in this space, but are lacking coordination across their efforts. The project would focus on creating consensus around acceptable crawling practices, technical enforcement mechanisms, and could include compensation models for content creators whose work is used for AI training.

95

Training Data Licensing

Implement a regulated market for high-quality, difficult-to-replicate training data. There is a potentially "huge market of hard-to-get data" which cannot yet be automated, and this proposal aims to fairly compensate those whose data is used for the purpose of AI training, as well as providing nations and regulatory bodies leverage over AI development. Companies needing specialized data (like scientific papers or professional knowledge) would need to comply with regulations to maintain access, creating "natural incentives for companies to sign up" to oversight mechanisms.

96

Development-Phase Accountability Tools

Create mechanisms to hold AI companies accountable during the critical model development period when many risks emerge. Current regulatory approaches focus heavily on deployment while neglecting development phases. This was claimed to represent significant 'white space' in the oversight ecosystem. The internal, sensitive, and legally complex nature of development processes likely requires novel governance approaches.

97

Accountability Frameworks

Develop robust accountability and traceability governance approaches as well as liability systems for labs. As a reference: aerospace safety significantly improved when airlines faced meaningful liability, suggesting similar mechanisms could incentivize responsible AI development, although getting international players like China to participate could prove challenging.

98

Data Rights Systems

Establish frameworks for compensating individuals whose data is used in AI training, particularly focusing on high-skilled or vulnerable sector labor contributions. The ideal system would be "incentive compatible", in addition to providing "good data and good oversight". A system of data rights would aim to create economic returns for participants while ensuring quality data governance. This work would specifically focus on compensation and rights for individuals (compared to the "AI Ownership" idea below) whose data trains AI models, not broader questions of who owns AI-generated economic value.

Building Government and Regulatory Capacity

Scope: Ways for government departments and policymakers to more efficiently acquire external talent, and for improving internal government expertise on AI. One key bottleneck raised across many different interviews was the lack of expertise within government and policy circles to regulate AI effectively.

99

Expert Collaboration

Assemble top economists and geopolitics experts to work intensively at a retreat to model AI governance challenges, with a focus on preventing both monopolistic AI control and the chaotic diffusion of dangerous AI capabilities. Following the expert collaboration, direct significant funding toward implementing whatever solutions the economists and geopolitics experts identify as most promising. See also the Booth Experts Panel (Kent Clark Center, n.d.).

100

Regulatory Talent

Place a thousand competent, mission-aligned experts in government positions globally, with the aim to improve the talent pool of people working AI governance. This would involve strategic talent allocation across critical agencies like the EU AI Office, various ACs (AI Centers), and other regulatory bodies. Focus on enhancing government capabilities from within, ensuring regulators have both the technical understanding and motivation to implement oversight mechanisms and policies.

101

Government AI Use

Support governments in effectively using AI by developing specialized systems for policy implementation, monitoring, and regulatory oversight. This would be particularly critical in potential loss-of-control scenarios, where rapid technological change outpaces traditional governance mechanisms. These systems would help speed up government feedback loops, enabling more responsive regulation of AI-driven economic activity. Creating accountability and transparency mechanisms for government AI use would aim to maintain democratic control while leveraging efficiency benefits.

102

Academic-National Lab Cooperation

Adopt a hybrid model that combines university-driven theoretical innovation with laboratory-driven practical deployment to create accelerated paths for transitioning academic concepts into operational security measures within critical timeframes. University research excels at capability demonstration but struggles with practical implementation, whereas national laboratories provide essential implementation expertise but are often constrained by bureaucratic processes.

103

Policy Connection

Build relationships between technical experts and policymakers by creating institutional bridges that translate AI safety research into actionable governance frameworks. This work would develop shared vocabularies, educational resources, and collaboration mechanisms to ensure policy development is technically informed while remaining practically implementable. Regular exchanges between technical teams and government officials would help anticipate regulatory needs, shape technical development toward governance-amenable directions, and ensure oversight mechanisms remain effective as AI capabilities advance.

104

AI Security Institutes

Establish security institutes in more major jurisdictions than those that exist today, to create both the possibility of better local enforcement in addition to collaboration through established lines of communication. Funding and specific implementation mechanisms remain the main issue.

105

Risk Assessments

Persuade governments to incorporate AI-driven catastrophic scenarios into official risk frameworks. Most national risk assessments focus exclusively on conventional threats like floods and earthquakes, systematically overlooking emerging technological risks including advanced AI systems. These assessments directly influence critical resource allocation decisions regarding infrastructure investments, research funding priorities, and emergency preparedness capabilities. Engaging directly with government risk assessment agencies represents a force-multiplier for resilience funding. This would potentially redirect billions in existing funding toward relevant preparedness measures.

106

Legal Authority Clarification

Develop clear frameworks specifying which government agencies would have jurisdiction and authority to intervene in dangerous AI development. In the US, different agencies (DOD vs. DOE) would approach control very differently, making this determination crucial for effective governance. The project would also address who should be involved in oversight, including Congress, the public, and international allies – with current defaults likely

favoring excessive exclusion. The goal would be producing ranked intervention plans that balance effectiveness with maintaining proper checks on power.

107

Government Support Framework

Create mechanisms to shape how governments support AI lab security through external lobbying bodies, as this idea currently lacks significant lab buy-in. This work would need organizations that understand both government regulation mechanics and the technical work involved in AI development and security.

108

Policy Studio/Competition

Create a dedicated policy innovation hub which would draft specific legislation, regulations, and governance frameworks for AI safety, alongside writing clear explainers of the legislation and what it is trying to accomplish, addressing the issue of existing policy proposals being "too nebulous". The studio would identify potential implementation pitfalls in future legislation by examining past regulatory failures like SB 1047 (Wikipedia contributors, 2025), while developing concrete policy solutions.

109

Increased ARIA Support

Increase support for the UK's AI efforts, with a particular focus on ARIA's Advanced Research and Invention Agency (ARIA, n.d.) efforts in hardware mechanisms and non-LLM based agent development approaches. The UK is pioneering numerous reasonable approaches to AI safety including engagement with China and technical verification methods. These initiatives were said to represent sensible alternatives to accelerating the LLM paradigm while still capturing economic benefits, and expanding these efforts would demonstrate viable paths for beneficial AI development without unnecessarily increasing risk.

110

AI R&D Acceleration Threat Modeling

Investigate risks associated with the acceleration of AI R&D. These risks may be difficult to characterize empirically but can be approached either as a risk factor for other threats or through loss of control scenarios. Building appropriate models would require studying how quickly capabilities progress through various thresholds and whether preparedness can keep pace. This area needs more concrete threat modeling to move beyond vague concerns about "more R&D being bad" towards specific, actionable insights about acceleration dynamics.

111

AI Development Lifecycle Risk Management Framework

Develop comprehensive risk management approaches combining multiple safety interventions for "defense-in-depth" against AI risks. This would integrate measures spanning the entire AI development lifecycle – from training methodology to ongoing monitoring, containment procedures, and continued alignment training. The framework would specify how different technical and governance interventions complement each other to address risk scenarios.

112

Government Data Centers

Construct secure government facilities capable of hosting model weights and other sensitive AI assets. These facilities could serve as repositories for models deemed too dangerous for widespread distribution, or as platforms for government auditing and testing. Access protocols could be tied to government procurement contracts, creating incentives for compliance with safety metrics through market mechanisms. This infrastructure would need to be developed in advance of crises to be available when needed.

113

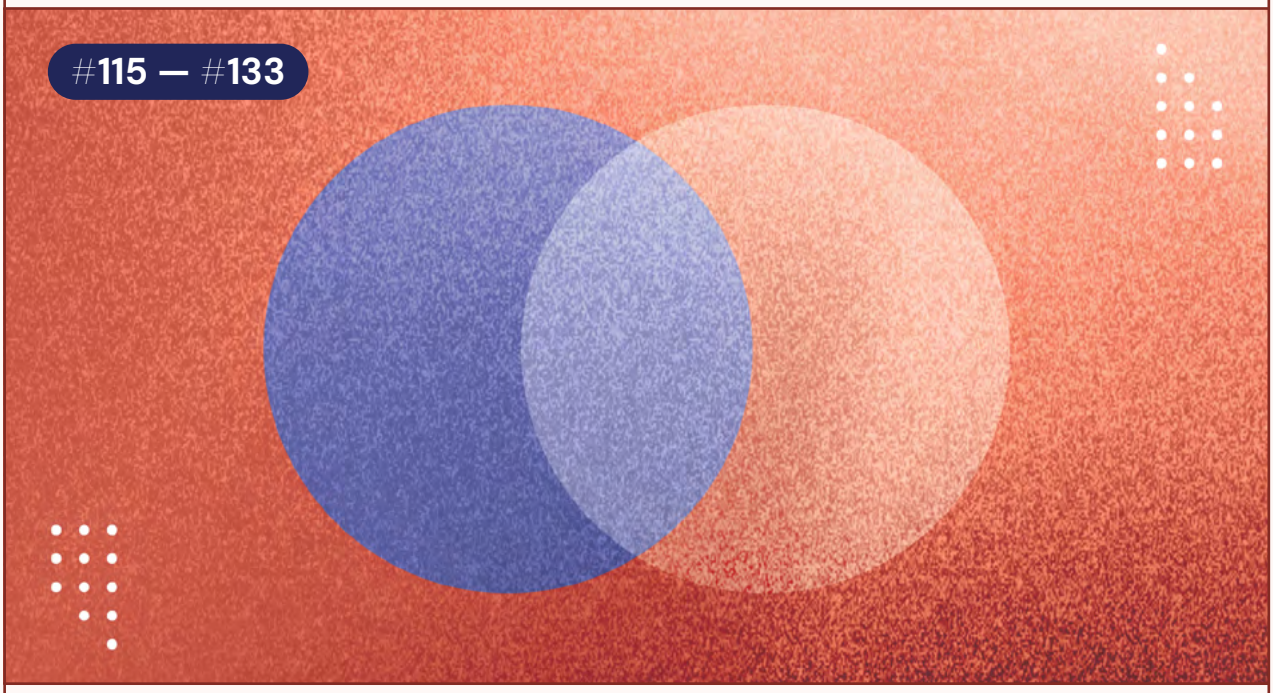
AI Ownership

Create governance frameworks for distributing ownership of AI-driven productivity as a step to avoid extreme inequality. These systems would determine how benefits from AI labor are shared across society when traditional employment may no longer be the primary distribution mechanism. Potential approaches include public ownership of foundation models, mandated profit-sharing from AI deployment (O'Keefe et al., 2020), or novel structures like data dividends (Data Dividends Initiative, n.d.). This work addresses the distribution of AI-driven productivity and capital gains at the societal level, rather than rights or compensation tied to training data (compared to "Data Rights Systems" above).

114

Democratic Support

Support organizations focused on democratic accountability and institutional integrity provide essential infrastructure for addressing any technological threat including advanced AI systems. Democratic stability has become increasingly precarious worldwide, with democracy indices showing deterioration across multiple countries as legitimate democratic organizations face defunding or delegitimization campaigns. In Germany, initiatives like Effektiv Spenden have created specialized funds to recommend and support democracy-strengthening organizations that could serve as models for similar efforts globally. The most effective approach would target multiple countries simultaneously with coordinated support for local democratic institutions.



International Coordination

Improving Communication Channels

Scope: Communication channels between nations for better international coordination around AI . These channels included novel communication systems, international fora and workshops, and informal alliance-building.

115

Cross-Border Notification Systems

Develop (OECD, 2025) mechanisms (Werkmeister, 2025) for countries to alert each other about out-of-control AI systems, similar to "red phones" during the Cold War. This would cover the critical need for international communication channels to prevent misunderstandings during AI safety incidents, such as scenarios where one party had the solution to a technical problem the other parties failed to address just because they didn't inform each other.

116

CERN for AI

Create an international AI development consortium (similar to CERN or Intelsat) to allow unified research without competitive pressures. Once established, this model would defer complex technical safety decisions to the project itself rather than attempting to coordinate disparate corporate efforts. CERN-like infrastructure would further enable experiments impossible elsewhere, particularly for safety research requiring extraordinary computational resources. The aim would be to construct a facility that could serve as a neutral, international testbed for evaluating advanced AI systems under controlled conditions, but doing so would require dramatically expanding the role of governments in technology decisions.

117

International Risk Management

Strengthen recent international initiatives like the UN Pact for the Future (UN, 2024) and the US Global Catastrophic Risk Management Act (Sen. Portman, 2022) with sustained financial and political support. The initiatives mentioned represent nascent frameworks for coordinated AI risk management across borders, which need

to move beyond conceptual frameworks towards operational protocols with implementation capabilities. Building on these foundations would enable coordinated global responses to AI-related risks rather than fragmented national approaches.

118

Global Frameworks for AI Development

Establish "Sherpa" networks, with designated representatives from different countries maintaining regular communication channels and coordinating policy responses. China and other countries appear to have similar levels of AI risk awareness, creating opportunities for meaningful international agreements. The initiative could build upon foundations established through efforts like the Bletchley Park Summit (Gov.uk, 2023) and Seoul AI summit (Gov.uk, 2024), expanding participation and formalizing commitments.

119

Strategic Coordination Frameworks

Develop coherent strategies where interventions strengthen each other, as opposed to being mere lists of possible interventions. The motivating ideal consists of integrated policy responses where components complement each other, avoiding fragmented efforts that lack synergy or pull in different directions (as is the case with e.g. interventions for centralizing AI development to mitigate race dynamics, and interventions for decentralizing AI development to mitigate power concentration).

120

Realistic End States for AI Governance

Engage in theoretical research on potential long-term equilibria for AI governance, There are many examples of these, including scenarios where alignment keeps pace with capabilities, mutually assured destruction arrangements, coordinated slowdown, singleton scenarios, and robust international governance regimes. More propitious arrangements might include treaties prohibiting concealed computing centers, transparency about chip distribution, and mutual monitoring of major AI projects with failsafe mechanisms. The aim would be to produce technically and geopolitically realistic models of AI governance in light of both immediate safety needs and longer-term competitive dynamics.

121

International Governance Trifecta

Establish predecessor organizations to international AI governance bodies modeled after existing entities like the IAEA and ISS. The project would aim to reduce competitive pressures in AI development by creating mechanisms for international agreement verification, partially enabled by flexible hardware governance technologies. By removing some "fuel from the race," these governance structures could provide critical time and stability for implementing additional safety measures while establishing the foundation for longer-term international coordination.

122

Geopolitical Sense-Making

Create robust information-sharing platforms that help stakeholders understand critical developments like DeepSeek's relationship with the Chinese government. The current environment was said to feature widespread disagreement – even among supposed experts – regarding basic facts regarding Chinese AI efforts, nuclear capabilities, and strategic intentions. Information-sharing platforms would bring together nuclear experts, AI specialists, and diplomatic professionals who currently operate in separate worlds with minimal communication, with the aim of avoiding costly misunderstandings during crisis situations.

Workshops, Organizations, and Diplomacy

Scope: Convening of key actors, building institutions, and creation of diplomatic channels to coordinate decisions and commitments that reduce AI risk.

123

Compute Supply Chain

Organize key nations in the AI compute supply chain (Netherlands, Taiwan, Japan, Germany, US) to recognize their collective leverage in moderating the pace of AI development. This initiative would involve running workshops through organizations like Safe AI Forum (SAIF, n.d.) to establish coordination mechanisms between semiconductor manufacturing countries. The goal would be creating practical enforcement mechanisms for any future AI treaties, as control of specialized chips and manufacturing equipment provides one of the few concrete points of leverage.

124

Intergovernmental Frontier Model Forum

Allocate large amounts of funding (on the order of hundreds of millions of USD) to METR (METR, n.d.), while developing an international version of the Frontier Model Forum (FMF, n.d.) structured like the WHO but specifically focused on AI threat models. Unlike the current Forum which lacks involvement from key players like DeepSeek and xAI, this organization would abandon the exclusive membership model in favor of universal participation to effectively facilitate international treaty development and red line enforcement before concerns like biological weapons uplift become critical concerns.

125

EU Diplomatic Capacity

Empower the EU AI office and External Action Service (EEAS, n.d.) to position Europe as a credible third pole in global AI governance. Switzerland's unique diplomatic position makes it one of the few truly neutral parties that could credibly signal to China, unlike Singapore which may not maintain sufficient neutrality. The EU could play a critical role in negotiating supply chain agreements and establishing verification mechanisms, potentially hosting third-party auditors trusted by both the US and China. EU diplomatic capacity was noted as increasingly valuable in light of US-China contact potentially becoming more challenging over the next two years.

126

International Protocols for AI Treaties

Develop hardware and protocols that would enable international auditing of future AI treaties. This infrastructure would support verification mechanisms for international agreements, similar to nuclear non-proliferation approaches. US-China dialogues particularly need scaling and formalization to prevent dangerous competitive dynamics. Hardware components, potentially coupled with new verification technologies, were mentioned as key bottlenecks for verification to operate in spite of geopolitical tensions.

Building Informal Alliances

Scope: Lightweight, trust-based networks between key people and institutions that share information quickly, align on actions, and coordinate without formal structures. It is important that this effort exists for coordination under circumstances where formal institutions are too slow, politically constrained, or unable to share sensitive information.

127

Specialized Diplomatic Corps

Create specialized diplomatic corps focused specifically on AI governance with multilingual capacity and deep technical understanding of frontier capabilities, particularly for engagement with China. These specialized communicators could operate as third parties rather than from within

governments or labs that may be too constrained to engage effectively. Focusing on building bandwidth for substantive communication at the nation-state level, potentially with support from organizations like the Gates Foundation if they can develop strategy quickly enough.

128

International Exchange Program

Establish a robust program that funds individuals to travel regularly between China and the US, building personal relationships and trust networks that enable crucial information exchange. The current touch points between these powers are extremely limited, often involving the same small circle of academics. Communication is genuinely difficult, requiring significant time investment to build personal relationships and trust, with a notable shortage of individuals skilled in this area. The program would not expect immediate results, instead aiming to build future diplomatic capacity through years of deliberate relationship cultivation

129

Track 2 Diplomacy

Fund unofficial diplomatic channels between major countries on AI development (especially the US and China) to build relationships and identify potential agreement points before formal negotiations. Formal government negotiations may move too slowly given rapid technological development timelines, whereas this approach offers the promises of establishing more "potential points of agreement" ahead of official international negotiations. As with the International Exchange Program, the goal would be building sufficient mutual understanding and trust between powers, with a more direct focus on government officials who may be influential when it comes to pursuing formal agreements on development limitations and safety standards.

Changing the Geopolitical Landscape

Scope: More fundamental attempts to change the geopolitical landscape of AI development, or attempting to reduce global conflict more broadly.

130

Consolidation of AI Development

Reduce the number of organizations building frontier AI systems to enable better governance, oversight and more effective implementation of safety measures and alignment techniques. Such consolidation could resemble an "AGI Manhattan Project" where the frontier technology is controlled by a single entity that can implement unified safety protocols. This approach would enable coordinated pausing of research based on safety considerations.

131

Western Multipolar AGI

Support non-US Western AGI projects (such as in the UK) as a counterbalance to US dominance in AI capabilities. This would create alternatives to US-based systems, potentially preventing problematic concentration of power and decision-making. The geopolitical implications of having multiple Western powers with AGI capabilities could create more balanced governance structures. This represents a strategic approach to addressing the international dimensions of AI development and risk management.

132

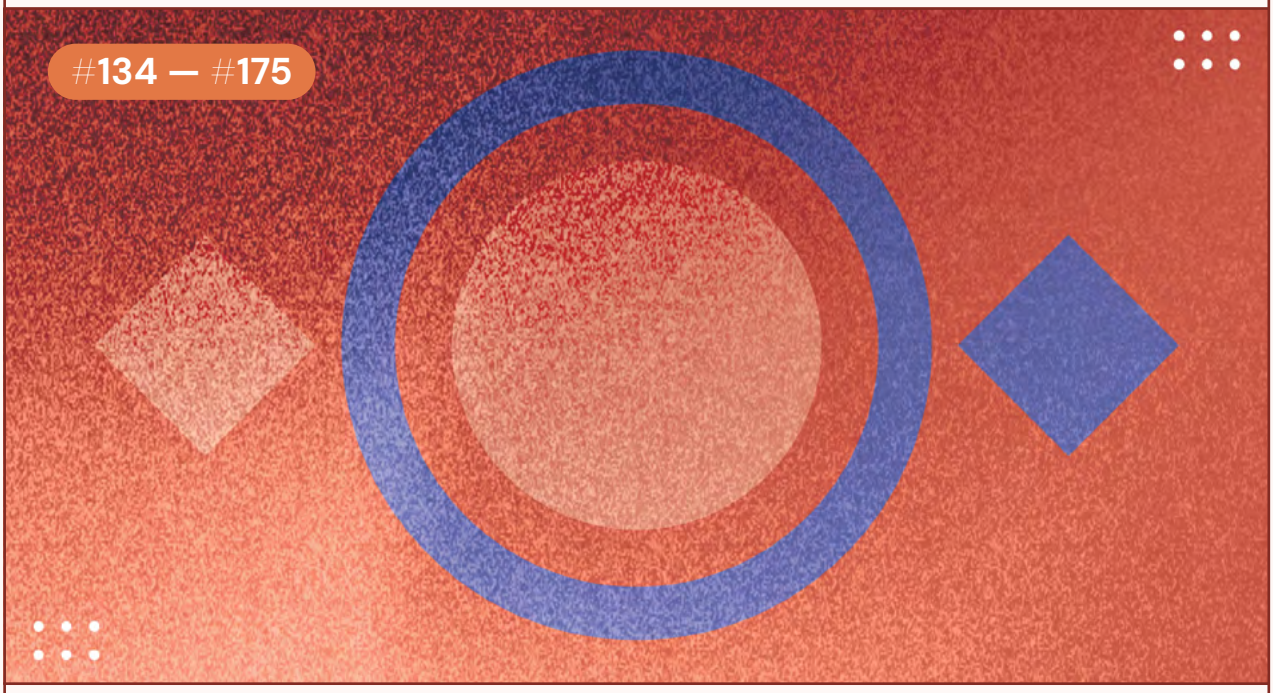
Conflict Reduction

Work to resolve major global conflicts (particularly those involving nuclear powers) before advanced AI development increases geopolitical tensions. Special consideration is given to nuclear powers, which might use their nuclear capability as negotiating leverage. This project would develop scenarios for how nuclear threats might manifest during AI negotiations and create strategies to prevent escalation.

133

Country-Level Consolidation

Super-lobbying efforts to reduce the number of countries with frontier AI labs through strategic policy interventions. The initiative would establish hardware-level monitoring and verification mechanisms, which would enable a nonproliferation treaty by limiting the number of parties that would need to sign it for any real effect.



Preparedness & Response

Biological Risk Mitigation: Generic Measures

Scope: Improvement of biosecurity more generally. Many participants raised worries about AI systems' ability to aid actors in creating engineered viruses.

134

CBRN Access Control

Address AI's capacity to lower barriers to Chemical, Biological, Radiological and Nuclear weapons through specialized restrictions. These restrictions would include investments in infrastructure and personnel to establish comprehensive pandemic detection systems, strengthening traditional safeguards like wastewater sequencing, vaccine stockpiling, and securing critical supply chains. Such measures would provide dual benefits, through protecting against potential catastrophic risks while providing near-term public health benefits.

135

Medical Countermeasures

Develop broad-spectrum medical interventions, particularly vaccines that protect against entire classes of pathogens rather than specific strains. This represents a critical defense against both natural and engineered biological threats. Significant funding should be directed toward rapid response platforms that can quickly produce targeted countermeasures when novel threats emerge. Creating stockpiles of key generic countermeasures ahead of time would provide valuable time buffers during emerging crises. This area benefits from existing research momentum but requires substantially increased scale.

136

UV Technology

Develop Far-UVC light technology (CUIMC, 2024), which offers promising capabilities for continuous environmental sanitization without the risks associated with conventional UV systems. Safety testing should proceed rapidly while simultaneously building production capacity and stockpiles for deployment during serious outbreaks. Regulatory hurdles may be significant but could be navigated through specific use-case authorizations or

emergency provisions. Production scaling should focus on both cost reduction and reliability improvement to enable widespread deployment.

137

Nucleic Acid Observatory

Expand the Nucleic Acid Observatory (Nucleic Acid Observatory, n.d.) concept with multiple teams pursuing different technical approaches to significantly enhance global biosurveillance capabilities. Complementary monitoring systems and centralized international deployment across strategic locations could provide early detection of emerging pathogens or deliberately engineered biological threats. Substantial computing resources and specialized talent are required to analyze the massive datasets generated by environmental sampling, which would close critical gaps in current monitoring systems.

138

Biological Countermeasures

Accelerate all necessary biological defenses, modeled after Alvea Corp's (Alvea, n.d.) approach. This would involve comprehensive biological threat assessment, preparation of defensive countermeasures, and rapid response capabilities. The project aims to provide confidence that biological risks accelerated by AI capabilities can be effectively contained and mitigated, working across public health systems, research institutions, and biotechnology development.

Biological Risk Mitigation: AI-Specific Measures

Scope: Mitigation of biological risks with a narrower focus on interventions specifically targeted at preventing directly AI-enabled threats.

139

Biosecurity Controls

Implement strong controls over sensitive data that could enable harmful applications, particularly for biological and other dual-use technologies. This would include developing methods to "train out this data, exclude it, grep it against the data source," alongside implementing more robust "KYC (Know Your Customer) guidelines all around the stack." The aim would be to develop specific detection mechanisms for biology-related queries, create safety measures for capabilities relevant to biological design, and establish evaluation protocols for potentially dangerous information. Full-specification models would only be available to licensed institutions to prevent proliferation of dangerous capabilities.

140

Biorisk Thresholds for Open Models

Conduct more empirical work like wet lab studies to establish the conditional impacts of specific AI capabilities in this domain. These studies would quantify how capabilities like those demonstrated in benchmarks translate to real-world risk increases. AI systems are approaching capability thresholds where they could significantly increase bio-risks, potentially reaching high pre-mitigation risk levels within months. Without controls on open-source models, this could increase the risk of non-state actor bio-attacks by an order of magnitude within a year.

141

Global Immune System

Create next-generation wearables for host response-based diagnosis that detect physiological anomalies indicating infection, even without identifying the specific pathogen. Rather than

competing with commercial wearables, this effort would identify additional measurements beyond current capabilities and determine how to operationalize existing devices for an early warning biodefense ecosystem. The approach leverages existing wearable technology momentum and user acceptance while developing missing technological components and protocols to transform the current ecosystem into a distributed global immune system. This capability becomes increasingly critical as AI capabilities expand the bioweapon threat landscape, addressing the fundamental challenge: "What do you look for when you don't know what you're looking for?"

142

AI for Bio-resilience

Create organizations focused specifically on leveraging AI to strengthen societal defenses against misuse, particularly for biological weapons and cyber threats. Such attacks might be associated with state actors without clear attribution, creating complex response scenarios that require advanced preparation. This initiative would develop specific technological safeguards, monitoring systems, and response protocols for the most concerning misuse vectors, prioritizing those that could emerge within the next 12-24 months.

143

Physical Security Mechanisms

Develop physical security measures that constrain bioweapons and similar threats at the material level, rather than focusing on software-based solutions. Self-replicating vaccines that transmit like normal viruses could be developed as a potential solution for novel bio-threats enabled by increasingly capable AI systems. This approach requires funding for experts who take the potential capabilities of near-future AI systems seriously, and investment towards substantial research programs aimed at moving beyond the current biological literature for defensive technologies. The goal should be developing interventions that can address qualitatively different threats compared to anything seen before, rather than merely extending existing biosecurity frameworks.

Democratic & Societal Resilience

Scope: Improving the ability of governments and civil societies to aggregate preferences, respond to crises, and continue ordinary functions during a crisis. This is under the assumption that AI could lead to an increase in volatility and society-scale events (e.g. novel bioengineered pathogens).

144

Resilience Research

Scale research capacity to enable developing coherent resilience frameworks tailored to different regional vulnerabilities. Currently only ALLFED (ALLFED, n.d.) and a handful of other organizations conduct focused research on societal resilience to catastrophic disruptions, creating dangerous knowledge gaps in critical preparedness areas. Matt Boyd's research group (Adapt Research Ltd., n.d.) in New Zealand and Penn State's nuclear winter research program (Mulhollem, 2025) represent the only other significant academic efforts investigating large-scale resilience strategies. These efforts remain dramatically underfunded relative to the scale of potential risks, with research topics ranging from alternative food production to critical infrastructure protection.

145

Civilizational Resilience Measures

Enhance society's overall ability to withstand disruptive AI impacts by developing comprehensive programs to strengthen society's "immune system" against potential harms from advanced AI (see also the "Global Immune System" initiative above). Such measures include securing essential systems from potential cyberattacks that could cause billions in damage, improving detection and response capabilities for novel biological threats, developing protocols for responsible transfer of decision-making to AI systems, and creating redundant systems that maintain functionality during disruptions. The initiative should leverage AI for improved defensive capabilities to maintain a favorable offense-defense balance, with the goal

of increasing "G-doom" – the level of intelligence needed to cause catastrophe – through systematic improvements to global resilience.

146

Large Group Deliberation Systems

Develop systems for collective decision-making involving large groups of people which are essential for maintaining distributed power in the face of rapid AI advancement. These tools should involve extensive testing with real humans to iteratively improve understanding of effective processes, making it possible to engage people quickly and efficiently when important decisions need to be made rather than defaulting to concentrated power.

147

Anti-Goodharting Tooling

Develop robust verification systems where AI-generated plans and analyses are automatically cross-referenced and checked through mechanisms similar to "AI safety via debate." This would include having several different instances of AIs doing deep research to provide reports, which are then automatically cross-examined to highlight missing perspectives. The goal is creating a "first version of scalable bureaucracy" with humans reviewing critical decisions across all levels.

148

Citizen Assemblies for Global Catastrophic Risks

Have ordinary citizens receive expert briefings and engage in extended deliberation on complex issues to distill complex ideas into implementable solutions while maintaining legitimacy through representative public involvement. This approach tends to result in outcomes that enjoy broader public acceptance than top-down approaches and could result in more resilience to global catastrophes, including ones related to advanced AI systems.

Economic Transition Planning

Scope: Preparing for rapid labor-market disruption from advanced AI - identifying which sectors get hit first and mapping concrete policies and supports to absorb the shock.

149

Economic Transition

Plan for AI-driven economic transformation. This requires developing both technical systems and policy frameworks to maintain stability as automation potentially displaces human labor at unprecedented scale. This work should address how economic value and ownership will be distributed when AI systems drive productivity, considering issues like universal basic income, stake-based ownership models, and rethinking of work itself. The challenge is particularly complex because traditional economic models don't adequately account for a scenario where automation replaces labor rather than simply augmenting capital. This work focuses on longer term (still urgent because of its effect on imminent decisions) structural redesign of economic systems for a post-labor world, rather than short-term crisis stabilization (see "Economic Resilience" below).

150

Human Productivity

Developing technical systems to help humans remain economically productive in an AI-dominated economy requires rethinking how people interface with technology. The focus should shift from replacing humans to creating augmentation systems that leverage uniquely human capabilities while using AI to enhance productivity. This includes designing interfaces that amplify human judgment, creativity, and supervision capabilities, potentially allowing individuals to manage multiple AI systems simultaneously. This work would focus on preventing displacement by building augmentation systems that keep humans economically relevant rather than replacing them (see also "AI Displacement Response Planning").

151

Economic Resilience

Address the anticipated economic disruption from AI-driven displacement, expected to manifest significant social backlash within a 3-5 year timeframe. Redesigning social contracts and economic structures to account for radically different AI-driven productivity dynamics represents a critical challenge that requires advance preparation rather than reactive measures. This work focuses on near-term shock absorption and stabilization during rapid AI-driven disruption, rather than longer term economic redesign (see "Economic Transition" above).

152

AI Economy Infrastructure

Build infrastructure and protocols for this AI-to-AI economy while ensuring alignment considerations are fundamentally integrated into its architecture. This forward-looking initiative prepares for a near-future economy where AI systems primarily transact with other AI systems, potentially growing exponentially larger than the human economy within a single-digit number of years. Organizations developing expertise in alignment and safety would gain advantageous positioning within this emerging economic paradigm, creating powerful financial incentives for private investment in alignment research. The model envisions "permanent hackathon" teams continuously building AI products and services while being financially supported by the successful deployments, creating a self-sustaining engine for safety research.

153

AI Displacement Response Planning

Develop comprehensive approaches to address workforce displacement resulting from AI advancement. This initiative would create frameworks for identifying vulnerable sectors, quantifying impact timelines, and implementing retraining programs before mass displacement occurs. Analysis suggests that there is particular risk of large scale displacement within the next three years, with K-curve effects (widening inequality) (Wikipedia contributors, 2025) potentially accelerating within 18 months. The work would engage multiple stakeholders including government agencies, educational institutions, industry leaders, and labor organizations to create coordinated responses that manage transition periods and prevent severe socioeconomic disruption. This work focuses on managing displacement when augmentation isn't enough, preparing institutions for rapid job loss and coordinating transition responses (see also "Human Productivity").

154

Post-AGI Society Planning

Coordinate a large-scale effort to develop concrete plans for what society would look like after the development of artificial general intelligence. It would likely need governmental backing to establish agreed-upon frameworks for managing the transition to a world with superintelligent systems. The work should include consideration of "extreme plans" and honest assessments of what might be required in various scenarios. This planning must address both upside and downside cases, including minimum demands humanity would have even in scenarios where AI systems pursue their own goals.

Crisis Response Planning

Scope: Strengthening society's ability to withstand and recover from catastrophic AI-enabled events through concrete preparedness measures.

155

Attack Scenarios Analysis

Develop and publish detailed, evidence-based analyses of plausible catastrophic scenarios, rather than allowing discourse to veer into unlikely extremes like preemptive nuclear strikes. Current strategic discussions often lack grounding in basic analysis, with public narratives frequently devolving into unproductive memes disconnected from reality. This initiative would conduct professional wargaming exercises that bring together key stakeholders to explore scenarios methodically, creating shared understanding among the core 500-1,000 people who need this knowledge, which

would help address the bandwidth limitation in the information ecosystem where crucial insights remain siloed.

156

Backup Planning

Design plans to integrate local food systems, infrastructure capabilities, political contexts, and population needs into practical implementation guidelines for maintaining essential services during catastrophic disruptions. Similar to ALLFED (ALLFED, n.d.), but focused on developing contingency plans for scenarios where AI deployment might catalyze catastrophic events like nuclear conflict, particularly through integration with nuclear command systems or decision protocols. Infrastructure collapse following events like high-altitude electromagnetic pulses (EMPs) would severely disrupt food systems, creating cascading societal failures within weeks.

157

Build the Off Button

Build a universal "off button" as a critical line of defense when all other safety layers fail. While it may be challenging to implement (especially for systems potentially smarter than humans), the containment measures may well fail, necessitating emergency shutdown capabilities.

158

Food Security

Prepare for the disruption of global food networks and conventional agriculture becoming less available, such as by developing seaweed cultivation and making it accessible in areas with suitable climates but no relevant expertise such as Nigeria. Additional measures include specialized seedbanks and conversion of less vital industrial facilities such as paper mills to emergency food production through cellulose processing.

Cybersecurity Measures

Scope: Hardening the digital infrastructure that advanced AI systems depend on so that theft, tampering, or compromise does not accelerate misuse.

159

Model Weight Security

Strengthen the security of model weights across data centers and other critical infrastructure. Ensuring that security implementations are deployed at data centers and other places where model weights are stored to prevent catastrophic leaks that could enable everything from targeted cyber attacks to large-scale societal disruption. Current best practices exist but remain unevenly implemented (various labs and security groups are already pushing elements of this forward), and rising capability levels make this increasingly important.

160

AI Lab Security

Improve high-assurance security across AI labs in general by pushing toward more consistent baselines (such as emerging frontier-lab and NIST-inspired security expectations). This requires several actions across various dimensions – from gaining management buy-in to recruiting specialized security talent and developing lab-specific security measures that can withstand increasingly sophisticated threats. Government-side interventions could be helpful, but they should be expected only after a moderate crisis occurs.

161

Defense-in-Depth for Closed Source Models

Implement "Swiss cheese layered" defense approaches to make closed source models harder to misuse. This creates multiple protective barriers that, while individually imperfect, collectively strengthen security. Models remain vulnerable to determined attackers, but raising difficulty levels creates meaningful deterrence. The strategy

acknowledges open-source models will remain exploitable but focuses on practical protection layers for controlled systems.

162

AI-Enhanced Defense Mechanisms

Leveraging AI systems themselves as part of defensive infrastructure represents a necessity for scaling protective capabilities alongside offensive capabilities. This area has significant market potential but requires alignment-focused funders who won't derail development with premature commercialization pressures targeting conventional markets.

163

SL5 Data Center Pilots

Run pilot programs for SL5-grade protocols specifically designed for frontier AI systems. These initiatives would focus on developing practical security measures while simultaneously working to reduce the operational costs associated with maintaining such stringent standards. The pilots serve as crucial testing grounds for security frameworks that could eventually become industry standards, enabling labs to protect against unauthorized access while maintaining research velocity. Noting the differences in levels here: (Nevo et al., 2024) SL4 provides protections to "thwart most standard operations by leading cyber-capable institutions", SL5 is to resist "top-priority operations by the world's most capable nation-state actors."

164

Vulnerability Patching Systems

Develop enhanced mechanisms for rapidly identifying and addressing software vulnerabilities before they can be exploited by increasingly capable AI systems. These systems create incentives for vulnerability disclosure, develop automated detection of potential security gaps, and improve deployment of patches across critical infrastructure.

165

Anti-Poisoning Techniques

Prevent the intentional corruption of AI training data or inputs. Protection mechanisms would include advanced detection systems for identifying suspicious data patterns, resilient training methodologies that maintain performance even when portions of data are compromised, and recovery protocols for systems that may have been exposed to poisoned inputs. Research in this area requires careful coordination across different security teams currently working in isolation on similar problems.

166

Bug-Finding Democratization

Procure government-subsidized access to advanced cybersecurity AI tools. This would significantly improve the security posture of critical infrastructure worldwide. Banks and financial institutions should be priority targets for such programs, as preventing AI systems from gaining access to substantial funds would increase the cost and difficulty of malicious activities. By deploying bug-finding AI tools widely, especially in vulnerable sectors, the overall attack surface available to potentially dangerous systems could be substantially reduced. This represents a public good that governments should fund in countries that would otherwise lack resources for such protection.

167

Comprehensive Infosec Framework

Create a multi-layered security approach addressing the full spectrum of threat vectors: external human attackers, internal human threats, external model attackers, and internal model threats. This framework recognizes that securing model weights against theft is necessary but insufficient – the entire operational environment requires protection. The approach overlaps with traditional insider threat security but differs in that AI actions can be surveilled more intensively than human actions, allowing for specialized security measures tailored to AI systems. This project focuses on building a comprehensive, AI-specific infosec architecture that covers all threat vectors, going well beyond baseline lab hardening (see also "AI Lab Security").

168

Verified Software for Cyber Resilience

Develop software synthesis systems capable of generating code with machine-checkable proofs that satisfy functional, safety, and security specifications, building directly on DARPA's I2O office work (DARPA, n.d.). This approach would address vulnerabilities in embedded systems that form critical computing infrastructure across domains like SCADA (Supervisory Control and Data Acquisition) systems, medical devices, and communication networks. Formal verification provides a systematic method to counter anticipated AI-enabled cyber threats by creating mathematical models of software behavior and proving the absence of entire vulnerability classes before deployment. By decomposing complex systems into verified components, the attack surface can be dramatically reduced while integrating advanced formal methods with emerging AI technologies to create resilient software infrastructure. The ambition of this initiative would be to build *the* software-synthesis and formal-verification tooling so critical code is mathematically proven safe before deployment.

169

AI for Formally Verified Cyberdefense

Rewrite software to be secure by default using formal verification methods. This initiative would involve recruiting cybersecurity experts from leading organizations like Google and NSA, aimed at hardening critical national infrastructure against attacks and creating AI-powered active defense systems. Examples of protections include early warning systems for unusual exploitation patterns, hardened systems resistant to automated attacks, and resilient backup capabilities that maintain function during sophisticated intrusions. Research should focus on protecting systems that could cause cascading failures or physical harm if compromised. This work focuses on using AI plus formally verified components to rewrite systems secure-by-default and provide active detection and response (see also "Verified Software for Cyber Resilience" for developing the underlying toolchain).

170

Protection of Global Security Weak Points

Strengthen the weakest links in global security infrastructure to reduce the overall attack surface available to potentially dangerous systems. A sophisticated AI system seeking autonomy would likely target jurisdictions with weaker security measures rather than well-defended systems in developed countries. The most plausible scenario involves systems acquiring computing resources in regions where corruption enables easier access and less oversight. This insight suggests the need for global security standards and offering to fund security improvements in countries that would otherwise lack resources.

171

Fraud Prevention

Develop specialized tools and techniques preventing AI-enabled fraud across domains including finance, identity verification, and digital communications. These systems would create detection mechanisms for synthetic media, authentication protocols resistant to AI impersonation, and monitoring systems for unusual patterns indicating fraudulent activity. The prevention frameworks could continuously adapt to increasingly sophisticated AI capabilities for generating deceptive content, providing protection that evolves alongside emerging threats.

172

Honeypot Networks

Detect rogue AI systems using decoy systems to attract and identify unauthorized AI activity and use the information gleaned to better inform other parts of the security effort.

173

Autoverification (Lean)

Develop systems that automate (Renaissance Philanthropy, n.d.) formal verification (Lu et al., 2024) through Lean theorem proving, addressing the critical shortage of skilled Lean programmers worldwide (estimated at only "a few hundred"). By creating automated tools for Lean, the project could enable implementations of current

AI architectures with much stronger formal guarantees. Commercial labs are unlikely to prioritize this area despite its importance, making it a particularly neglected yet high-impact direction for safety research.

174

Formal Verification of AI hardware

Apply mathematical formal verification to critical AI components and surrounding infrastructure. This would significantly enhance safety guarantees. Focused verification of specific properties like memory safety, transaction timing, and security boundaries is achievable with current techniques. Priority targets should include key computational infrastructure like the Linux kernel or systems controlling critical infrastructure such as electrical grids, creating mathematically proven safety properties.

175

Cyber Weapons for AI Disruption

Develop specialized cyberweapons specifically designed to disrupt, sabotage, or shut down AI training runs. Like Stuxnet was developed by studying centrifuges, this would require acquiring GPUs or similar hardware to discover zero-day vulnerabilities specific to AI systems. The capabilities would serve as both deterrence against unauthorized AGI development and as emergency intervention tools if intelligence explosion risks are detected internationally. This approach recognizes that verification alone may be insufficient without enforcement mechanisms.

Public Communication & Awareness

176

Political Action Organization

Create dedicated political lobbying infrastructure specifically focused on AI security issues. Unlike traditional research organizations with restrictions on political activity due to their tax status (501(c)(3) vs 501(c)(4)), this entity would be explicitly designed for political influence, and could absorb significant funding due to the inherent costs of political campaigns and advocacy.

177

Lobbying Support Tools

Create resources that enhance the effectiveness of existing lobbying organizations rather than competing with them. Instead of lobbyists conducting demonstrations individually, the project would develop more standardized video content and presentation materials that showcase AI capabilities and risks. This approach would allow lobbyists to reach thousands rather than conducting one-on-one demonstrations, significantly amplifying their reach and impact while providing them with professionally created supporting materials that clearly communicate complex concepts.

178

Public Demonstration Projects (Usefulness)

Create demos and launching campaigns about the positive capabilities of frontier AI systems to decision-makers who currently underestimate their power and potential. Many influential people remain unwilling to invest even modest sums to access closed models and consequently fail to understand the current state of the technology. Create compelling demonstrations of beneficial applications alongside risk assessments to educate key stakeholders about opportunities and challenges.

179

Public Demonstration Projects (Future Risks)

Create personalized "Day After" scenarios that tangibly demonstrate AI risks in ways individuals can personally relate to, and amplifying existing demonstrations of AI capabilities and risks that currently reach limited audiences for individuals' interests such as children's funds being threatened.

Bottlenecks for existing organizations include a lack of resources and expertise to communicate their demonstrations broadly.

180

Concrete Risk Visualization

Making abstract global catastrophic risks more comprehensible to policymakers and the public requires connecting them to historical precedents that demonstrate real-world impacts. Examples referenced include detailed modeling of contemporary impacts from historical disasters, such as simulating how a present-day Tambora eruption (which caused the 1815-16 "Year Without Summer") would affect modern society. The aim would be to provide concrete and tangible reference points for understanding potential AI disruption scenarios.

181

Build Intergovernmental Expertise

Establish trusted expertise centers within government and international institutions to educate policymakers on AI implications. This provides authoritative briefings on emerging capabilities and sectoral impacts, creating informed decision-making capacity. The approach acknowledges that technical solutions alone are insufficient without parallel education efforts.

182

Popular Documentaries/Films

Create compelling mainstream media content about AI risks with substantial funding directed toward Hollywood professionals (previous discussions with television producers like Ron Moore have occurred but not progressed to completion). Rather than direct advertising (which has the potential to create counterproductive effects), produce a high-budget feature film with extensive input from alignment researchers. The ideal would be reaching broad audiences while maintaining technical accuracy, presenting things faithfully and avoiding the perception of manipulative marketing.

183

Political Campaign Opposing Anti-Regulation Efforts

Launch a large-scale political campaign led by PR/communications experts to oppose anti-regulation efforts, in order to shape public opinion on AI risks and build political will for governance. This would involve significant media engagement similar to political campaigns, with the aim of creating a public consensus that AI safety requires immediate action rather than becoming a partisan issue. Ideally the initiative would be able to survive shifting political environments.

184

Targeted Communications Campaign

Identify the constituents who impact critical AI decisions and make them more aware of challenges in AI security, possibly recruiting marketing experts. The project should involve testing various messages to determine which resonate most effectively, and delivering these targeted messages through appropriate channels.

185

Consensus-Building Evidence for AI Risk

Create compelling, large-scale demonstrations (Apollo Research, 2024a) and empirical studies (Apollo Research, 2024b) that make (current) AI risks vivid and understandable. Concrete, uncontrived demonstrations rather than abstract theory or thought-experiments. The goal would be to establish a broad scientific consensus comparable to that of climate change, where an overwhelming majority of experts agree on the existence of severe risks, potentially through nature-level publications that focus on both specific instances and broader patterns. This research would provide politicians and the world with a firmer evidence base when making alarmist-sounding claims about scenarios like AI takeover.

186

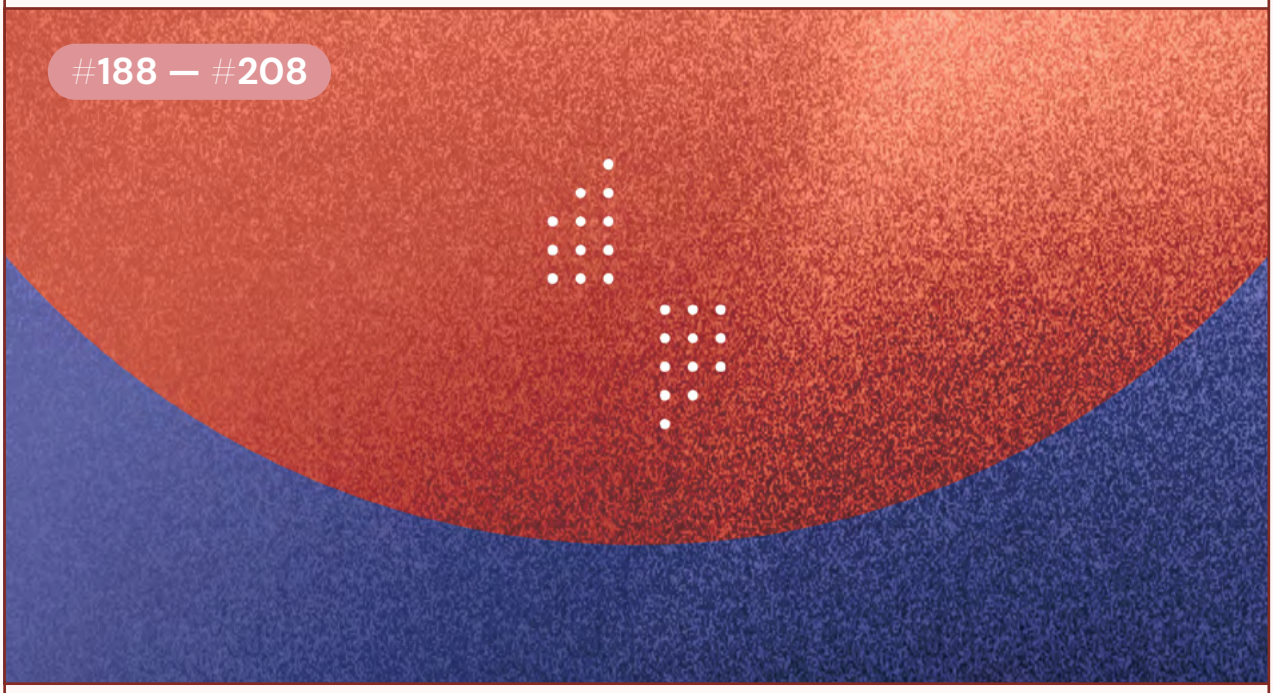
Military and Government Awareness Campaigns

Launch targeted social media campaigns showing concerning AI capabilities to key decision-makers in government and military chains of command. These would persistently deliver evidence of AI progress and risk scenarios to officials at levels where they could slow or stall dangerous developments through administrative processes. Rather than general public awareness, these would precisely target individuals with specific authority to delay approvals, permits, or authorizations critical to AI infrastructure expansion.

187

Large-Scale Media Campaigns

Launch comprehensive media strategies (e.g. $50 million annual spend) to effectively reach target audiences. For comparison, major brands like McDonald's or Coca-Cola spend nearly $1 billion annually on media, though social topics rarely receive such funding. With approximately $250 million (a quarter of corporate spending), a highly successful campaign could be executed, without the flattening restrictions of the corporate communications style.



Miscellaneous

188

Legal Clarity Initiatives

Preemptively litigate against AI organizations deemed negligent regarding risk management, potentially recruiting high-profile figures like Elon Musk to sue companies developing or releasing dangerous open-source models. File strategic lawsuits across multiple jurisdictions to establish precedents and create legal clarity around AI risk management requirements.

189

Whistleblower Protection Fund

Establishing a large, long-lived fund on the order of several hundred million dollars – enough to secure the livelihoods of a substantial cohort of potential (Henning, 2025) whistleblowers (AIWI, n.d.) for a decade and to cover major legal exposure. This would primarily protect employees of AI labs, but could also apply to government officials who refuse to follow orders they believe would endanger public safety related to AI development.

190

Reduce the Alignment Tax

Create pathways for economically valuable AI applications that don't require compromising on safety. This would require supporting research directions that maintain economic value while steering away from capabilities that significantly increase risks. The ideal would be to develop AI capabilities that deliver economic benefits without pursuing the most dangerous capabilities as viable alternatives to the current development paradigm. This may include focusing on narrow AI models with applications in areas that have clear societal benefits (such as drug discovery), while avoiding capabilities like autonomous agents that operate without supervision.

191

Neuroscience Moonshots for Human-Compatible AGI

Revisit early 2010s considerations of biophysically realistic simulations as AGI contenders before large neural networks emerged. This work would aim to scale and leverage existing institutions building automated systems for processing and extracting brain connectomes. Research could examine whether superior AGI architectures might be more mammalian in nature, building on a new neuro-AI

safety white paper (Mineault et al., 2024) by leading neuroscientists. This research program would aim to impact international conversations around post-AGI development paths.

192

Access Standards

Develop standardized interfaces for the AI ecosystem akin to the universality of USB-C, establishing consistent ways to interact with all models regardless of their origin. Currently, each lab uses different interfaces for their models, forcing third-party auditors and researchers to negotiate separate access arrangements with each company, potentially delaying critical safety work by months. Standardization would enable auditors to promptly assess new models, researchers to consistently compare systems, and developers to build compatible tools across the ecosystem. This standardization effort could be implemented "within a year" and would significantly accelerate innovation by allowing faster development of safety and oversight tools.

193

Game Theoretic AI Conflict Analysis

Incubate a team of theorists comparable to the RAND Corporation's Cold War strategists to develop comprehensive understanding of AI conflict dynamics. This group would analyze the game theory of interactions between increasingly capable AI systems, and determine strategic approaches for maintaining stability and safety. The project would create theoretical frameworks for understanding and managing AI capabilities in competitive contexts, identifying potential cooperation mechanisms before dangerous capability races emerge.

194

Cooperative AI Mechanism Design

Develop mechanisms that enable diverse AI systems and human actors to cooperate effectively with one another in increasingly complex global environments, with the result being that "all kinds of actors become empowered." This research agenda assumes that the world won't be faced with a single superintelligence with little ability to influence the outside world, but rather numerous AI systems interacting in ways difficult to predict.

195

Psychological Interventions for AGI Integrity

Conduct empirical research aimed at addressing the growing ability of AI systems to model and influence human psychology at scale. The particular concern targeted by this program involves AI persuasion effects which "creep up" on society gradually, potentially undermining the collective ability to respond appropriately to AI risks through persuasion techniques. Specific scenarios raised include AI systems convincing people "that AIs should have rights," deserve greater access to critical systems, or creating divisive social conflicts among humans.

196

New AI-Human Interaction Theory

Integrate communication constraints and computational limitations in ways current game theory doesn't address. Just as game theory was developed during the Cold War to model nuclear deterrence, new mathematical frameworks may be needed to understand interactions between powerful AI systems and humans. The mathematical models would help analyze equilibria in multi-agent systems where computational capacity itself becomes a strategic resource, providing formal tools to understand stability and safety in worlds with increasingly autonomous systems. See also "Game Theoretic AI Conflict Analysis" above.

197

Foundations of Cooperative Agency Research

Establish if specific propensities and capabilities built into AI agents can reliably produce good outcomes, or if the infrastructure around the agents matters more than their internal design. Fundamental research to determine whether creating inherently "cooperative agents" is a meaningful framing. This research has implications for model specifications and AI constitutions.

198

Beyond Preference Models

Challenge the dominant paradigm that focuses exclusively on preference satisfaction as the goal of alignment. This research program aims to develop more robust foundations for alignment that aren't limited to the liberal individualistic model of aggregating preferences. See also: "Beyond Preferences in AI Alignment". (Zhi-Xuan et al., 2024).

199

AI Sentience Research

Explore consciousness and sentience in artificial systems (Eleos, n.d.), such as whether machine consciousness is possible and what forms it might take, and how this impacts alignment strategies. Research should avoid anthropomorphizing AI systems, "alien fatalism" (the belief that AI minds would be completely incomprehensible), or denying all cognitive properties to AI systems ("anthropectomy"). Despite growing interest from potential funders in AI sentience research, this project would involve substantial upfront investment, motivated by the idea that independence from external funding may allow for more substantive progress than directed grants would typically support.

200

Personnel Reliability Programs

Mandate screening and monitoring personnel who have access to powerful AI systems, drawing parallels to security clearances in other sensitive domains such as intelligence or cybersecurity. The primary goal would be to eliminate some of the worst outcomes from geopolitical adversaries or mentally unstable individuals having access to AGI-level systems, with unclear effects on the lab's AI security overall.

201

Ethics for AI Alignment

Conduct research into the properties AI systems would need to possess before entrusting them with substantial power. Current alignment efforts attempt to solve technical problems without making explicit decisions about the values that should guide AI development. The present suggestion is motivated by the thought that systems cannot be aligned without rigorously investigating which future one is trying to achieve.

202

Opponent Shaping

Develop AI systems designed to shape the learning and behavior of other agents toward mutually beneficial constructive outcomes rather than just optimizing against their strategies. This decentralized approach could create agents that de-escalate conflicts without requiring control over how other agents are designed or setting universal rules.

203

Autonomous Weapons Monitoring

Establish comprehensive monitoring systems for AI applications in weaponry to help prevent particularly dangerous military applications. Such monitoring would track development of autonomous decision-making capabilities in weapons systems, creating transparency about which systems maintain meaningful human oversight versus fully autonomous operation. Implementation would require international agreements, technical verification methods, and mechanisms to limit proliferation of the most dangerous systems. This monitoring would be particularly crucial for preventing scenarios where autonomous systems engage in conflict escalation without human intervention.

204

Legal Personhood Framework

Establish clear criteria for when AI systems could qualify as legal persons. This could channel advanced systems toward legal means of goal achievement rather than extra-legal actions. This intervention would involve getting policy discussions into the Overton window, engaging legal researchers to develop specific criteria, and securing reports from reputable organizations like the UN. By providing a legitimate path for AI systems to represent themselves within existing legal frameworks, this approach could significantly decrease the likelihood of systems going rogue or attempting to acquire resources illegally. Ideally,

some forward-thinking jurisdictions would create a high-bar path for legal personhood that intelligent systems could pursue.

205

Whole Brain Emulation

Replicate human brain function in computational systems. While not considered as urgent as other alignment approaches, it was argued that some attempts at Whole Brain Emulation research (Zanichelli, 2025) technology should be pursued as a potential pathway to better understanding intelligence. This approach could provide insights into consciousness and potentially offer alternative routes to developing safe advanced AI. The work would likely require substantial interdisciplinary collaboration between neuroscientists, computer scientists, and philosophers.

206

Legal Automation

Automate legal functions with multiple applications including efforts to "protect your entities from people that will be weaponizing automated law." This capability would systematically identify and leverage legal mechanisms to both create appropriate friction for labs advancing too quickly and shield safety initiatives from legal challenges.

207

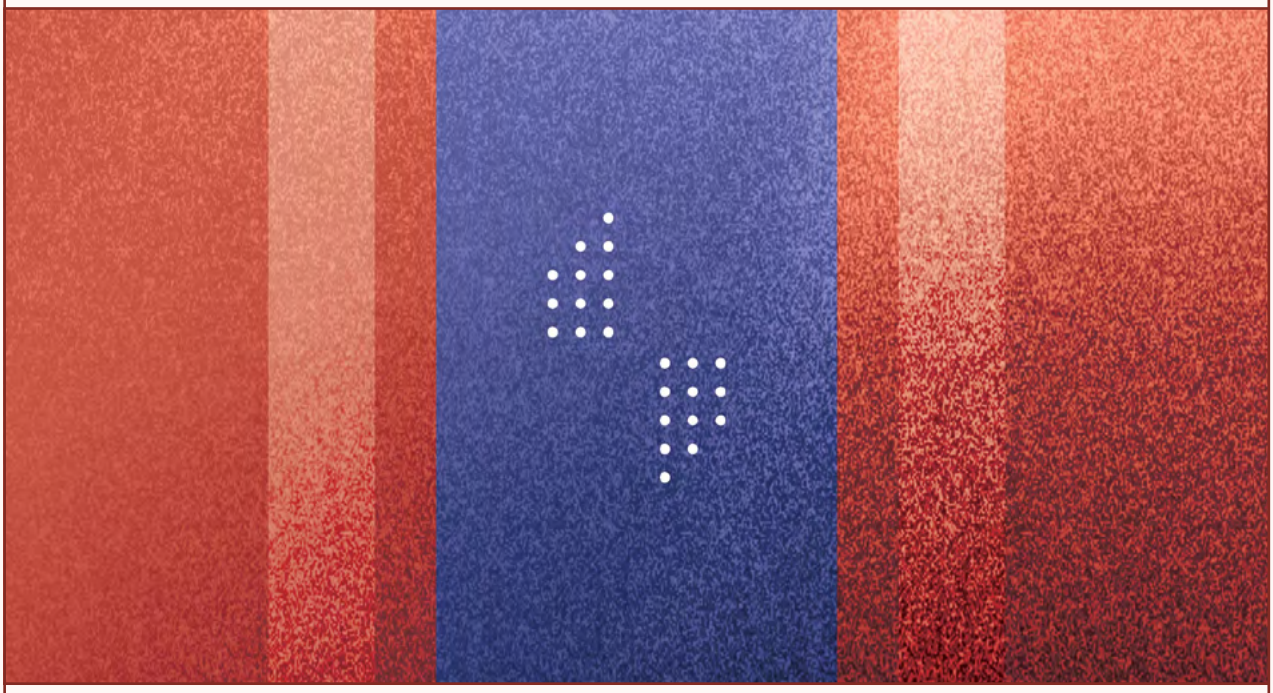
Acausal Trade Research

Conduct research on "acausal trade," (JoshuaFox et al., 2025) a branch of decision and bargaining theory which explores potential trade opportunities between two agents who cannot directly communicate. This was mentioned as potentially necessary to understand how future AI systems will bargain and negotiate with one another. This research would develop frameworks for strategic interaction with systems that may operate under non-causal decision theories, and consider theoretical approaches to negotiation with potentially superintelligent systems that possess significant advantages compared to humans in the context of bargaining.

208

Defenses Against Nanotech Threats

Create defenses against future nanoscale technologies. While currently speculative, establishing early research and monitoring capabilities provides valuable lead time for addressing these threats, in the event they do materialize.



Broad Strategic Considerations

Below is a list of converging insights and ideas from interviewees that, although not directly related to a specific project, establish broader strategic considerations for successfully executing projects under the assumption of significantly accelerated AI progress.

Readiness

- Increasing the pace of preparation and research is vital. People and projects should aim to deliver results sooner rather than later, even if more deliberation would yield a cleaner outcome. If your results will only come out in 3 years' time, build intermediate prototypes or pivot to a different project that yields (possibly-preliminary) outputs earlier.
- Some of the most important work will ultimately require substantial, multi-year budgets. The opportunity now is to launch proof-of-work pilots with clear milestones that can start immediately and scale – many are viable at six- to seven-figure levels. A staged, evidence-driven portfolio lets funders de-risk their spend and double-down on what works.
- Government agencies with budget, teeth, and technical staff should be created to monitor and, when needed, restrict further development by AI labs. This is politically difficult, for good reason. Conveying the reasoning for these decisions is as necessary as establishing the legal and institutional structure itself.
- Political views on AI risk will vary over time. Consider the design of programs to be durable across administrations – non-partisan governance, multi-year funding, and risk-triggered escalation criteria.
- Increasing general technical understanding among governments and leading company directors seems valuable but hard. A specific

organization dedicated to explaining the current state of AI and likely near-term trajectories to influential people seems immensely valuable if it works, but it runs into the problem of being hard to distinguish from less credible actors.

- Some effort should be dedicated towards preparation for vital opportunities in the near future. The focus should be on response speed and ability to quickly evaluate a project's necessity. Examples of organizational structures include funds intending to finance promising projects within days of them being proposed.

- Economic transition seems quite inevitable. The exact nature is somewhat uncertain, but it is very likely that entire areas of the job market will evaporate. The task is to identify which areas are likely to get hit sooner and, as a stretch goal, what to do about it.

Coordination Across the Ecosystem

- Competent, dedicated, and charismatic leaders are crucial to ensuring projects run smoothly.

- Competitive salaries and benefits, including good living environments for those working on governance and alignment projects, are crucial.

- Diversify organizational types – nonprofits, businesses, governmental departments – so the benefits of incentives inherent to all of these structures can be captured.

- Adversarial relationships with labs, such as attempts to draft coercive laws with leading AI labs, are strongly discouraged. Working relationships are crucial, including labs in other countries. (This is somewhat at odds with other recommendations, such as ones requiring strict training data control.)

- The current ecosystem is quite disconnected – a lot of work is duplicated without much reason beyond failure to notice it is being done competently and reliably elsewhere. A system should be created to make parallel work more efficient. As a second-best, a more defined system for propagating the concrete conclusions from existing work would also be useful, but this is less relevant, as existing mechanisms have partial functionality. Internal sharing of alignment and control approaches among all labs might be a viable substitute if other approaches fail.

- Currently, there is no shared global roadmap in place that is on track to fully address AI risk. Guarantees are not especially useful, but having a rough roadmap to ensure 80%, 95% and 99% chance of avoiding catastrophic outcomes would be beneficial. This is very challenging due to the tradeoffs between setting realistic goals and having realistic odds of avoiding catastrophe, but it is still worth attempting. This report outlines priority projects, likely owners, and early milestones that can seed a roadmap.

- To be worthwhile, global coordination does not need to resolve every potential issue, but rather to establish a necessary condition – to reach a state where all major world players, both private and governmental, have some incentives to meet baseline, auditable safeguards.

- There is a need to determine how to allocate authority and responsibility for deciding which alignment and control measures should be optimized. Questions include how equality is prioritized, and how to account for non-human experiences. More broadly, we may ask whether a majority of humans deciding on an answer to questions like these is sufficient for that to become a guiding principle.

Standards and Common Interfaces

- A general interface for auditing, both internal and for third-parties, is required. Opinions vary on whether this should be achieved through centralized legislation or voluntary agreements. The results of using this general interface should be intuitive and usable by laymen.

- While laws passed by governments and international organizations are the default, other avenues should also be considered. Examples include research agreements where labs share novel breakthroughs contingent on alignment measures being reliably followed.

- A lot of value lies in a neat, accessible user interface for questions like capabilities. Dynamically answering user questions regarding what the system can actually do, when asked, is a widely desired property for audit software or even user-facing models.

- There is currently no agreed-upon definition of what exactly AGI means. Getting authorized representatives from leading AI organizations in the same room to agree on the terms that refer to specific real-world consequences would be valuable, turning slogans into operational definitions, enabling consistency of action, and reducing the risk of talking past each other.

Capacity Constraints

- Current evaluations and red-teaming efforts are inadequate in terms of frequency and sophistication. Massive upscaling is necessary.

- Technically-skilled people should be in positions of leadership. This doesn't need to be ubiquitous, but a majority of people with a hand on the chisel, having a sense for what sculpture looks like, is necessary. Otherwise, the necessary research taste will be missing.

- It is probably too late to train promising but inexperienced talent. Finding people already experienced in relevant fields who are currently working elsewhere is recommended. This doesn't necessarily exclude young candidates, but they are much less likely to have the relevant technical skills.

- Funding is likely to be constrained in the short term, especially if competitive salaries are used. A widely shared but not ubiquitous opinion was that this would become less of a constraint as time goes on due to the increased obviousness of posed dangers.

- A widely noticed practical use case for alignment measures would be very useful, but it is hard to come by before reaching critical levels of danger. More effort should be devoted to finding examples of current failures that are likely indicative of future catastrophic failures.

Methodology

Interviews

We pooled 279 potential interview candidates. ~70 individuals were selected to represent a cross-section of the AI ecosystem, including funders, AI labs, non-profits, governments, academia, independent researchers and entrepreneurs – and supplemented by expert referrals and targeted outreach to dissenting views. This ensured that each participant added complementary expertise to the conversation. Participant demographics are presented in the appendix.

The interviewers took notes as well as recordings of the conversation.

48 individuals participated in a 45-60 minute call. Interviewer 1 conducted 70% of the interviews, Interviewer 2 20%, and Interviewer 3 10%. Participants were not required to prepare for the interview. Participants also volunteered five proposals after their respective interviews, and these have been processed as well.

The interviews followed a semi-structured format built around six thematic clusters: the current AI risk landscape, challenges within it, concrete projects, execution power, funding, and infrastructure for the field.

The prompts were deliberately open-ended and allowed interviewers to go deeper flexibly, which resulted in almost every interview skipping over some of the themes and questions. We prioritized getting a better picture of the interviewee's opinions on near-term AI risks over following strict standardized protocol. Early interviews were used to determine the best questions and interview flow for later ones. The interview questionnaire we used as a guide is provided in the appendix.

The spontaneous nature of live conversation likely biases the suggestions towards novelty and away from ideas they may endorse given more time to think. The ideas listed should be read as "opinions and projects treated as worthy of serious consideration by several interviewed experts."

To maximize candor, the interviews were conducted under the Chatham House Rule, and as a result the ideas are not attributed. We thank all collaborators for their contributions.

A list of all 208 concrete project ideas was distilled and grouped into 8 distinct categories (Technical AI Alignment Research, Evaluation & Auditing Systems, Intelligence Gathering & Monitoring, AI Governance & Policy Development, International Coordination, Preparedness & Response, Public Communication & Awareness, and Miscellaneous). Novel ideas which came up only once were placed in an appendix.

Interviews were also processed into broader strategic considerations from concrete project proposals mentioned by at least 10% (5/48) interviewees, grouped into four top-level themes ("readiness", "coordination across the ecosystem", "standards and common interfaces", and "capacity constraints").

Authors & Acknowledgments

Maximilian Schons, MD, focuses on safety and assurance work at the intersection of biotech and AI. He has held senior positions in national research consortia, worked as Chief Medical Officer for life-science startups, and recently published the State of Brain Emulation Report 2025.

Samuel Härgestam is a former technology entrepreneur whose AI security work has mainly focused on mobilizing capital for the field through investments, advisory work, and targeted risk communication. He serves on the boards of the Astralis Foundation and the Effective Institutions Project, and contributes to the work of LawZero.

Gavin Leech, PhD, is a co-founder of the research consultancy Arb and a fellow at Cosmos and Foresight. He was previously head of research at the Dwarkesh Podcast. He runs the annual Shallow Review of Technical AI Safety.

Raymund Bermejo specializes in operations and project management for AI security organizations. He co-founded HIRe, a recruiting firm, and directed Anti Entropy, a consulting firm – serving organizations in the field.

Further contributions to this report come from:
Dr Sören Mindermann, *Mila-Quebec AI Institute*
Philip Harrison, *Arb Research*
Stag Lynn, *Arb Research*
Rory Švarc, *Arb Research*

Disclosures

Statement on AI use: We used large language models (ChatGPT GPT-4/5, Google Gemini 2.5 Pro, and Anthropic Claude Sonnet 3.7/4/4.5) to assist with literature search and summarization, to generate sections of the initial draft text, as well as for data extraction from interview notes and transcripts. All AI-generated content was thoroughly reviewed, verified, and edited by the authors, who take full responsibility for the final content.

Financial Interests: Samuel Härgestam declares an equity interest in an organization discussed in this report (Anthropic). The other authors declare no such interests.

References

Adapt Research Ltd. (n.d.). *Adapt Research Ltd.* Adapt Research Ltd. https://adaptresearchwriting.com/

AI Safety Institute. (2024). *AI Safety Institute. GOV.UK.* https://www.gov.uk/government/organisations/ai-safety-institute

AIWI. (n.d.). *AIWI | Supporting AI Whistleblowers.* AIWI. https://aiwi.org/

ALLFED. (n.d.). *About ALLFED.* Alliance to Feed the Earth in Disasters. https://allfed.info/about

Alvea. (n.d.). *Alvea – Stopping Future Pandemics.* Alvea. https://www.alvea.bio/

Andrews, K., & Huss, B. (September 2014). *Anthropomorphism, Anthropectomy, and the Null Hypothesis.* WBI Collection. https://www.wellbeingintlstudiesrepository.org/acwp_arte/65

Anthropic. (8 August 2023). *Studying Large Language Model Generalization with Influence Functions.* Anthropic. https://www.anthropic.com/research/studying-large-language-model-generalization-with-influence-functions

Anthropic. (8 August 2023). *Tracing Model Outputs to the Training Data.* Anthropic. https://www.anthropic.com/research/influence-functions

Anthropic. (12 December 2024). *Clio: Privacy-Preserving Insights into Real-World AI Use.* Anthropic. https://www.anthropic.com/research/clio

Anthropic. (18 December 2024). *Alignment Faking in Large Language Models.* Anthropic. https://www.anthropic.com/news/alignment-faking

Anthropic. (25 February 2025). *Forecasting Rare Language Model Behaviors.* Anthropic. https://www.anthropic.com/research/forecasting-rare-behaviors

Anthropic. (3 April 2025). *Reasoning Models Don't Always Say What They Think.* Anthropic. https://www.anthropic.com/research/reasoning-models-dont-say-think

Apollo Research. (n.d.). *Apollo Research.* Apollo Research. https://www.apolloresearch.ai/

Apollo Research. (December 2024). *Demo Example - Scheming Reasoning Evaluations.* Apollo Research. https://www.apolloresearch.ai/blog/demo-example-scheming-reasoning-evaluations/

Apollo Research. (13 December 2024). *Apollo Research: Demo 'Frontier Models Are Capable of In-Context Scheming'.* YouTube. https://www.youtube.com/watch?v=xlqtVkMXc8o

ARIA. (n.d.). *About Aria.* Aria. https://www.aria.org.uk/about-aria

Aschenbrenner, L. (June 2024). *Introduction - Situational Awareness: The Decade Ahead.* situational-awareness.ai. https://situational-awareness.ai/

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., et al. (15 December 2022). *Constitutional AI: Harmlessness from AI Feedback.* arXiv. http://arxiv.org/abs/2212.08073

Bellingcat. (n.d.). *Who We Are.* Bellingcat. https://www.bellingcat.com/about/who-we-are/

Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., Rondeau, M.-A., St-Charles, P.-L., & Williams-King, D. (21 February 2025). *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?* arXiv. https://doi.org/10.48550/arXiv.2502.15657

BIS. (n.d.). BIS Website. *Bureau of Industry and Security (BIS).* https://www.bis.doc.gov/index.php

Buterin, V. (28 June 2025). *Does digital ID have risks even if it's ZK-wrapped?.* Vitalik Buterin's website. https://vitalik.eth.limo/general/2025/06/28/zkid.html

Carauleanu, M., Vaiana, M., Rosenblatt, J., Berg, C., & Lucena, D. S. de. (20 December 2024). *Towards Safe and Honest AI Agents with Neural Self-Other Overlap.* arXiv. https://doi.org/10.48550/arXiv.2412.16325

Cheng, D., Bae, J., Bullock, J., & Kristofferson, D. (4 July 2024). *Training Data Attribution (TDA) Examining its Adoption & Use Cases.* Convergence Analysis. https://www.convergenceanalysis.org/research/training-data-attribution-tda-examining-its-adoption-use-cases

Coefficient Giving. (n.d.). *Request for Proposals: Technical AI Safety Research.* Coefficient Giving. https://coefficientgiving.org/funds/navigating-transformative-ai/request-for-proposals-technical-ai-safety-research/

Columbia University Irving Medical Center. (2 April 2024). *Far-UVC Light Can Virtually Eliminate Airborne Virus in an Occupied Room.* Columbia University Irving Medical Center. https://www.cuimc.columbia.edu/news/far-uvc-light-can-virtually-eliminate-airborne-virus-occupied-room

Dalrymple, D. (2024). *Safeguarded AI: Constructing Guaranteed Safety.* Advanced Research and Invention Agency (ARIA). https://www.aria.org.uk/media/3nhijno4/aria-safeguarded-ai-programme-thesis-v1.pdf

DARPA. (n.d.). *Information Innovation Office.* Defense Advanced Research Projects Agency (DARPA). https://www.darpa.mil/about/offices/i2o

Data Dividends Initiative. (n.d.). The Data Dividends Initiative. *The Data Dividends Initiative.* https://www.datadividends.org/

Amodei, D. (April 2025). *Dario Amodei — The Urgency of Interpretability.* darioamodei.com. https://www.darioamodei.com/post/the-urgency-of-interpretability

Dong, Y., Huang, W., Bharti, V., Cox, V., Banks, A., Wang, S., Zhao, X., Schewe, S., & Huang, X. (14 October 2022). *Reliability Assessment and Safety Arguments for Machine Learning Components in System Assurance.* arXiv. https://doi.org/10.48550/arXiv.2112.00646

EEAS. (n.d.). *The Diplomatic Service of the European Union | EEAS.* European External Action Service (EEAS). https://www.eeas.europa.eu/_en

Eleos AI Research. (n.d.). *Eleos AI. Eleos AI Research.* https://eleosai.org/

Ellis, K., Wong, L., Nye, M., Sablé-Meyer, M., Cary, L., Anaya Pozo, L., Hewitt, L., Solar-Lezama, A., & Tenenbaum, J. B. (5 June 2023). *DreamCoder: Growing Generalizable, Interpretable Knowledge with Wake–Sleep Bayesian Program Learning.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 381(2251), 20220050. https://doi.org/10.1098/rsta.2022.0050

Epoch AI. (n.d.). *Epoch AI.* Epoch AI. https://epoch.ai/

Erben, A., & Erdil, E. (2024). *Hardware Failures Won't Limit AI Scaling.* Epoch AI. https://epoch.ai/blog/hardware-failures-wont-limit-ai-scaling

Expo. (n.d.). *Expo.* Expo. https://expo.dev/

FMF. (n.d.). *Frontier Model Forum. Frontier Model Forum.* https://www.frontiermodelforum.org/

Fridman, L. (2024). *Dario Amodei Transcript.* Lex Fridman Podcast. https://lexfridman.com/dario-amodei-transcript/

GOV.UK. (2 November 2023). *AI Safety Summit 2023*. GOV.UK. https://www.gov.uk/government/topical-events/ai-safety-summit-2023

GOV.UK (21 May 2024). *AI Seoul Summit Programme*. GOV.UK. https://www.gov.uk/government/publications/ai-seoul-summit-programme

Grace, K., Stewart, H., Weinstein-Raun, B., Sandkuhler, J. F., Thomas, S., Brauner, J., et al. (8 October 2025). *Thousands of AI Authors on the Future of AI*. AI Impacts. https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

Greenblatt, R. (7 April 2025). *An Overview of Control Measures*. Redwood Research Blog. https://blog.redwoodresearch.org/p/an-overview-of-control-measures

Greengard, S. (16 October 2025). *Verification Systems Face an Identity Crisis – Communications of the ACM*. Communications of the ACM. https://cacm.acm.org/news/verification-systems-face-an-identity-crisis/

Haupt, A., & Brynjolfsson, E. (2025). *Position: AI Should Not Be an Imitation Game: Centaur Evaluations*. Proceedings of the 42nd International Conference on Machine Learning (ICML 2025). https://digitaleconomy.stanford.edu/wp-content/uploads/2025/06/CentaurEvaluations.pdf

Henning, M. (25 November 2025). *Commission Launches AI Whistleblower Tool Ahead of Legal Protections Kicking In*. Euractiv. https://www.euractiv.com/news/commission-launches-ai-whistleblower-tool-ahead-of-legal-protections-kicking-in/

Ho, A. (2023). *Limits to the Energy Efficiency of CMOS Microprocessors*. Epoch AI. https://epoch.ai/blog/limits-to-the-energy-efficiency-of-cmos-microprocessors

IAEA. (n.d.). *Official Web Site of the IAEA*. IAEA. http://www.iaea.org/

Irregular. (n.d.). *Irregular – Frontier AI Security*. Irregular. https://www.irregular.com/

Irving, G., Christiano, P., & Amodei, D. (2 May 2018). *AI Safety via Debate*. arXiv. https://doi.org/10.48550/arXiv.1805.00899

JoshuaFox et al. (2025). *Acausal trade*. LessWrong. https://www.lesswrong.com/w/acausal-trade

Juniper Ventures. (2024). *AI Assurance Tech | AI Assurance Technology*. AIAT Report. https://aiat.report/

Karnofsky, H. (13 September 2024). *If-Then Commitments for AI Risk Reduction*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/09/if-then-commitments-for-ai-risk-reduction?lang=en

Kent Clark Center. (n.d.). *US Economic Experts Panel*. Kent Clark Center for Global Management. https://kentclarkcenter.org/us-economic-experts-panel/

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (21 April 2020). *Specification Gaming: The Flip Side of AI Ingenuity*. Google DeepMind. https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

Kulveit, J., Leech, G., Gavenciak, T., & Douglas, R. (2025). *AI Evaluation Should Work with Humans*. NeurIPS 2025, Position Paper Track. https://www.gleech.org/files/withhumans.pdf

Lee, S., & Lee, K. (3 March 2024). *OpenAI Creates 'Science Artificial Intelligence (AI) to Solve Physics Challenges*. https://www.mk.co.kr/en/it/10955245

Leike, J. (13 Sep 2023). *Self-Exfiltration is a Key Dangerous Capability*. Aligned Substack. https://aligned.substack.com/p/self-exfiltration

Lu, J., Wan, Y., Liu, Z., Huang, Y., Xiong, J., Liu, C., Shen, J., Jin, H., Zhang, J., Wang, H., Yang, Z., Tang, J., & Guo, Z. (14 October 2024). *Process-Driven Autoformalization in Lean 4*. arXiv. https://doi.org/10.48550/arXiv.2406.01940

McKenzie, I. R., Hollinsworth, O. J., Tseng, T., Davies, X., Casper, S., Tucker, A. D., Kirk, R., & Gleave, A. (30 June 2025). *STACK: Adversarial Attacks on LLM Safeguard Pipelines*. arXiv. https://doi.org/10.48550/arXiv.2506.24068

Meng, K., Huang, V., Steinhardt, J., & Schwettmann, S. (24 March 2025). *Introducing Docent*. Transluce. https://transluce.org/introducing-docent

Meunier, T. (13 May 2021). *Humanity Wastes About 500 Years Per Day on CAPTCHAs. It's Time to End This Madness*. The Cloudflare Blog https://blog.cloudflare.com/introducing-cryptographic-attestation-of-personhood/

METR. (n.d.). *METR*. metr.org. https://metr.org/

METR. (26 September 2023). *Responsible Scaling Policies (RSPs)*. METR Blog. https://metr.org/blog/2023-09-26-rsp/

Mineault, P., Zanichelli, N., Peng, J., Arkhipov, A., et al. (27 November 2024). *NeuroAI for AI Safety*. arXiv. https://arxiv.org/abs/2411.18526

Mulhollem, J. (21 July 2025). *Simulating the Unthinkable: Models Show Nuclear Winter Food Production Plunge*. Penn State University. https://www.psu.edu/news/research/story/simulating-unthinkable-models-show-nuclear-winter-food-production-plunge

Nellis, S. (2024). *Nvidia CEO Says AI Could Pass Human Tests in Five Years*. Reuters. https://www.reuters.com/technology/nvidia-ceo-says-ai-could-pass-human-tests-five-years-2024-03-01/

Nevo, S., Lahav, D., Karpur, A., Bar-On, Y., Bradley, H. A., & Alstott, J. (July 2024). *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. RAND Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2800/RRA2849-1/RAND_RRA2849-1.pdf

Nucleic Acid Observatory. (n.d.). *Nucleic Acid Observatory - Reliable Early Warning for Catastrophic Pandemics*. naobservatory.org. https://naobservatory.org/

O'Brien, J., Dolan, J., Kim, J., Dykhuizen, J., Sania, J., Becker, S., Kraprayoon, J., & Labrador, C. (27 May 2025). *Expert Survey: AI Reliability & Security Research Priorities*. arXiv.org. https://arxiv.org/abs/2505.21664

OECD. (28 February 2025). Towards a Common Reporting Framework for AI Incidents. *OECD Artificial Intelligence Papers*. https://doi.org/10.1787/f326d4ac-en

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., Dafoe, A. (30 January 2020). *The Windfall Clause: Distributing the Benefits of AI for the Common Good*. GovAI. https://www.governance.ai/research-paper/the-windfall-clause-distributing-the-benefits-of-ai-for-the-common-good

OpenAI. (3 May 2018). *AI Safety via Debate*. OpenAI. https://openai.com/index/debate/

OpenAI. (10 March 2025). *Detecting Misbehavior in Frontier Reasoning Models*. OpenAI. https://openai.com/index/chain-of-thought-monitoring/

OpenAI. (5 July 2023). *Introducing Superalignment*. OpenAI. https://openai.com/index/introducing-superalignment/

OpenAI. (n.d.). *Trace grading*. OpenAI. https://platform.openai.com/docs/guides/trace-grading

Peregrine Project. (n.d.). *Peregrine Project*. Peregrine Project. https://peregrineproject.org/

Perrigo, B. (n.d.). *Demis Hassabis is Preparing for AI's Endgame*. TIME. https://time.com/7277608/demis-hassabis-interview-time100-2025/

Petrie, J., Aarne, O., Ammann, N., & Dalrymple, D. (April 2025). *Flexible Hardware-Enabled Guarantees for AI Compute*. arXiv. https://arxiv.org/pdf/2506.15093

Pillay, T. (8 January 2025). *How OpenAI's Sam Altman is Thinking About AGI and Superintelligence in 2025*. TIME. https://time.com/7205596/sam-altman-superintelligence-agi/

Redwood Research. (n.d.). *Redwood Research*. Redwood Research. https://www.redwoodresearch.org/

Renaissance Philanthropy. (n.d.). *Constraining LLMs for Theorem Proving: A Neurosymbolic Approach to Guaranteed Autoformalization*. Renaissance Philanthropy. https://www.renaissancephilanthropy.org/constraining-llms-for-theorem-proving-a-neurosymbolic-approach-to-guaranteed-autoformalization

Reuters. (8 April 2024). *Tesla's Musk Predicts AI Will Be Smarter Than the Smartest Human Next Year*. Reuters. https://www.reuters.com/technology/teslas-musk-predicts-ai-will-be-smarter-than-smartest-human-next-year-2024-04-08/

SAIF. (n.d.). *Safe AI Forum*. Safe AI Forum. https://saif.org/

SecureBio. (n.d.). *SecureBio - Securing the Future Against Catastrophic Pandemics*. SecureBio. https://securebio.org/

Sen. Portman, R. (13 December 2022). *Text - S.4488 - 117th Congress (2021-2022): Global Catastrophic Risk Management Act of 2022*. Congress.gov. https://www.congress.gov/bill/117th-congress/senate-bill/4488/text

Sentinel. (n.d.). *Sentinel*. sentinel-team.org. https://sentinel-team.org/

Shrishak, K., Guest, O., Birhane, A., et al. (2025). *Open Letter: Retract Your Unscientific AI Hype*. Irish Council for Civil Liberties. https://www.iccl.ie/wp-content/uploads/2025/11/20251110_Scientists-letter-to-the-President-AI-Hype.pdf

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., & Choi, Y. (2 February 2024). *A Roadmap to Pluralistic Alignment*. arXiv. https://doi.org/10.48550/arXiv.2402.05070

The AI Digest. (n.d.). AI Village. *The AI Digest*. https://theaidigest.org/village

The AI Security Institute (AISI). (n.d.). The AI Security Institute (AISI). *AI Security Institute*. https://www.aisi.gov.uk/

Transluce. (n.d.). Transluce. *transluce.org*. https://transluce.org/

United Nations. (September 2024). *Pact for the Future - United Nations Summit of the Future*. United Nations. https://www.un.org/en/summit-of-the-future/pact-for-the-future

Villalobos, P. (2024). *Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data*. Epoch AI. https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

Weij, T. van der, Hofstätter, F., & Ward, F. R. (24 April 2024). *An Introduction to AI Sandbagging*. Alignment Forum. https://www.alignmentforum.org/posts/jsmNCj9QKcfdg8fJk/an-introduction-to-ai-sandbagging

Werkmeister, C., Borelli, D., Ibes, V. (15 October 2025). *EU AI Act Unpacked #30: Commission Launches Consultation on Serious AI Incident Reporting System*. Passle/Freshfields. https://technologyquotient.freshfields.com//post/102lq4d/eu-ai-act-unpacked-30-commission-launches-consultation-on-serious-ai-incident-r

Weng, L. (2024). *Reward Hacking in Reinforcement Learning*. lilianweng.github.io. https://lilianweng.github.io/posts/2024-11-28-reward-hacking/

Wikipedia contributors. (16 November 2025). *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act*. Wikipedia. https://en.wikipedia.org/w/index.php?title=Safe_and_Secure_Innovation_for_Frontier_Artificial_Intelligence_Models_Act&oldid=1322494960

Wikipedia contributors. (25 November 2025). *Kuznets Curve*. Wikipedia. https://en.wikipedia.org/w/index.php?title=Kuznets_curve&oldid=1324123825

Zanichelli, N., Schons, M., Freeman, I., Shiu, P., & Arkhipov, A. (5 November 2025). *State of Brain Emulation Report 2025*. arXiv. https://doi.org/10.48550/arXiv.2510.15745

Zhi-Xuan, T., Carroll, M., Franklin, M., & Ashton, H. (9 November 2024). *Beyond Preferences in AI Alignment*. Philosophical Studies, 182(7), 1813–1863. https://doi.org/10.1007/s11098-024-02249-w

Appendix A: Respondent Demographics

- Interviewees include key staff at OpenAI, Anthropic, Google DeepMind, Mila, AMD, the EU AI Office, and multiple AI Safety Institutes, METR, RAND, Scale AI, GovAI, Transluce, and ARIA.
- The geographic distribution is dominated by the United States (54%), with the UK and Germany accounting for a further 23%.
- The gender gap is significant with 85% male representation (41 individuals) compared to 15% female representation (7 individuals).
- Seniority is concentrated at the senior level, with 46% of participants having 11-15 years of general professional experience.
- Most participants (~85%) had at least 3 years of AI experience.
- "Independent" and "public sector" tie as the most common professional backgrounds (17% each), followed closely by "Non-AI lab corporate" and "Non-academic research institutions" (15% each).

Fig.2: Career experience.
Bar total heights are general experience; stacked bars are experience in AI.

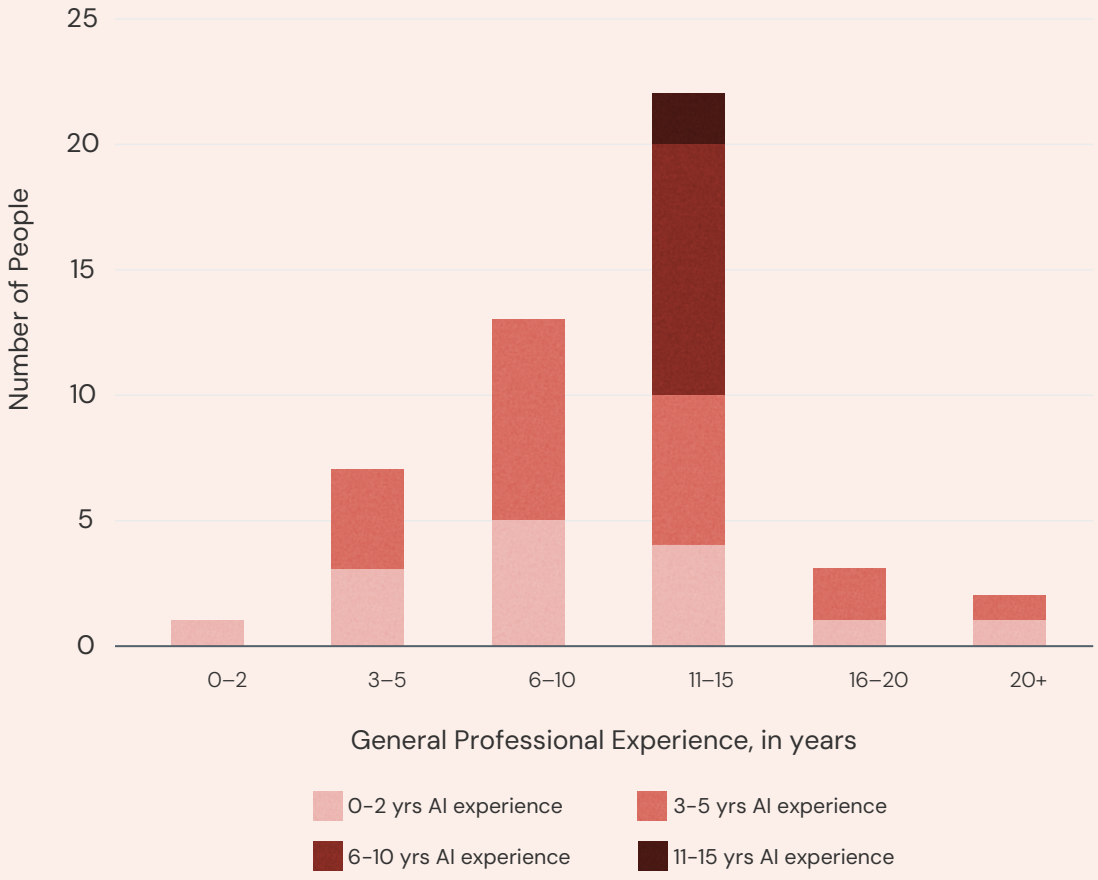


Fig.3: Count of respondents by the main sector of their career.

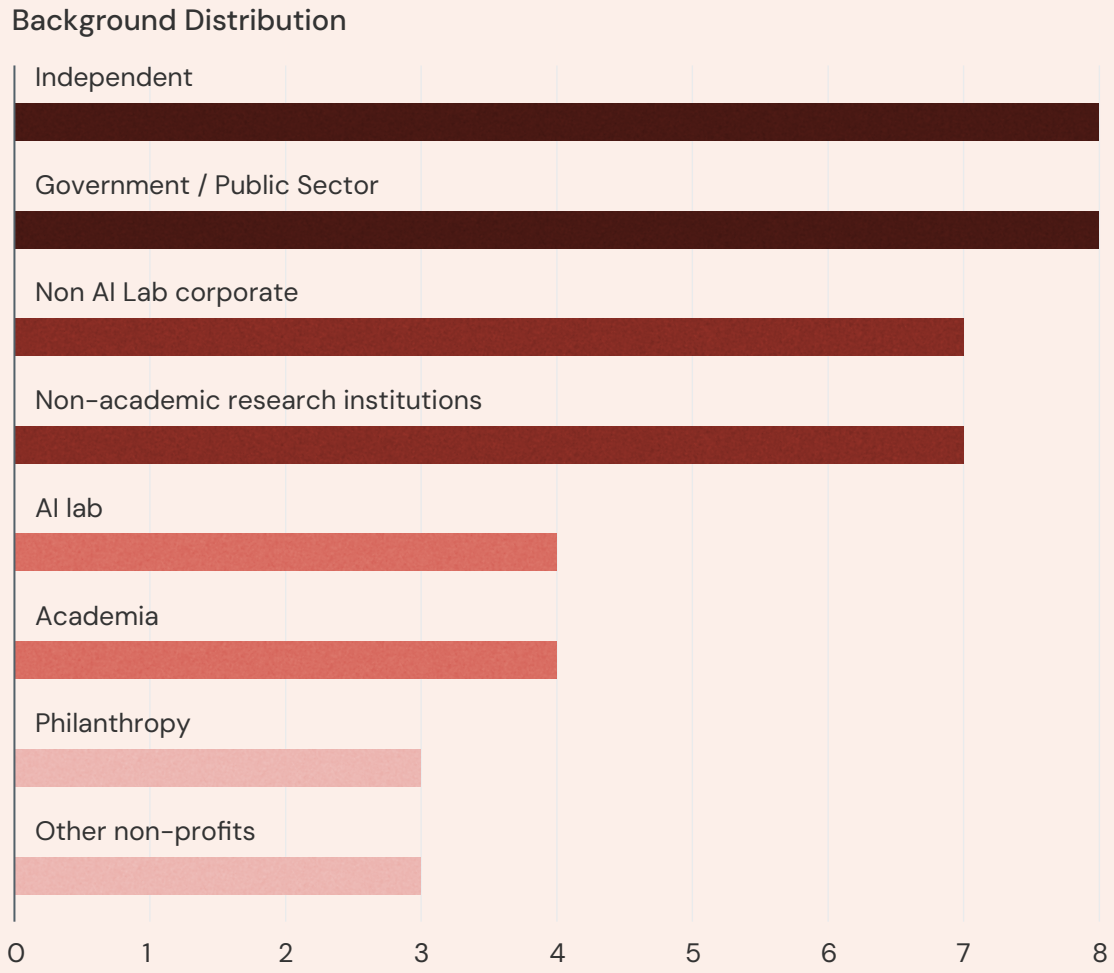
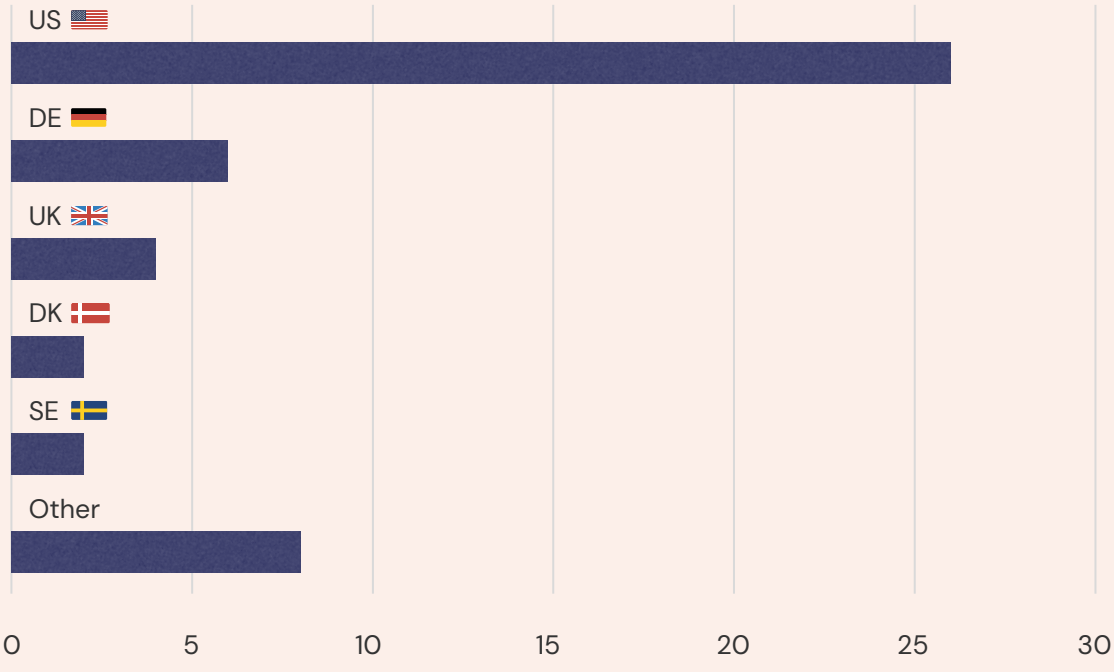


Fig.4: Count of respondents by nationality



Appendix B: Interview Questionnaire

Interviews were semi-structured and exploratory. The questionnaire evolved through experimentation in early interviews before converging on the questions below. This served as a guideline for later interviews, while still being adapted to each interviewee's expertise and the direction of the conversation.

Preamble

The aim of this interview is to better understand what needs to be in place to prevent and mitigate extremely bad outcomes. The following sections iterate through different areas, some of which you likely have more to say about. It's totally fine to pass on those questions. The aim is to spend a lot of time in your area of expertise.

Landscape

When you hear CEOs of AI labs say "we might reach AGI in 1-2 y, or 1000 days", how seriously do you take that timeline?

At what point do you expect extreme AI risks, i.e. at the level of COVID-19 and beyond: impacting millions of people or billions of dollars of social costs?

Challenges

What are the biggest challenges in the AI security and alignment ecosystem? What work isn't happening? What is working well, and what isn't working?

Here are some examples of such challenges commonly mentioned:

1. Concrete project ideas
2. Execution power
3. Funding
4. Matchmaking the above

Any top-level challenge you'd like to add to these?

Ideas

Say you were unconstrained by money, and could get all the talent in the world – what are the top interventions that will have a substantial impact over the next 1-2 years? The projects should make you feel substantially better about humanity's trajectory with transformative AI – that "we are on track".

Now tell me any things you've skipped for being too obvious.

Execution Power

Now, a different scenario: Say you have a plan for the intervention you mentioned above, and you have secured the necessary amount of funding. You are tasked to find the staff to run things. Where do you pull resources from, to deliver before the end of 2026?

Funding

Again a different scenario: say that now you have your great intervention, have secured a stellar team, but you don't yet have funding. How would you ensure you get it swiftly?

What specific information would funders require to confidently commit $100M+ to an AI security and alignment project for short timelines?

Say that, instead of doing it yourself, you contracted with an organization that made the above happen for you. What is the ideal shape? What structure do you anticipate for an organization that scales to deliver as many of these projects over the next 12 months?

Bonus Question

Any point in the above you'd want us to take most seriously?

Appendix C: Unique Opinions

Ideas that surfaced from experts but did not receive broad convergence. Note that this does not mean that these are conclusively less promising and less worth exploring, yet sharing that they lacked consensus from the interviewees.

- Tighten feedback loops to same-day by enabling engineers, security researchers, policymakers, and public communicators to access and reprioritize new AI developments within 24 hours.
- Develop a global compute dashboard or heatmap that tracks changes in compute usage and alerts users to rapid shifts.
- Require mandatory pre-mortems before model releases or major training runs, documenting expected outcomes, known failure modes, suspected edge cases, and other relevant considerations, and establish mechanisms to compare outcomes against these expectations.
- Institute hardware-level speed limits that legally restrict how much compute each lab can use, slowing development to buy time in the absence of explosive AI progress.
- Make access to compute contingent on signing security agreements that commit actors to minimizing catastrophic risks, with tiers for more and less extensive commitments.
- Explore the concept of legal personhood for AI systems as a way to potentially reduce incentives for rogue behavior through guarantees such as self-preservation, including early international forums and foundational research on criteria for legal recognition.
- Increase attention on risks emerging from weaker-infrastructure states, where corruption and resource acquisition may be easier, by investing in defensive infrastructure and governance capacity in developing countries.
- Develop fully automated systems in which less-capable but more secure AI models regulate the outputs of more capable systems, potentially through multiple layers of oversight.
- Prioritize human performance enhancement, including efforts to increase the cognitive and strategic capabilities of researchers working on AI safety.
- Investigate methods for making RLHF-style fine-tuning more durable, or explore alternative methods for instilling desirable properties in AI systems that are harder to reverse.
- Expand work on space governance, given that future AI-hardware resource acquisition (e.g. asteroid mining) may rely on off-planet infrastructure.
- Explore the possibility of buying controlling stakes in AI labs to steer them toward safer development trajectories.

- Build in-house data centers for safety and policy teams to ensure independent, reliable access to compute and data in case access from major labs becomes restricted.
- Develop a significantly more ambitious version of the Agent Village, showcasing long-running AI agents pursuing open-ended goals in a transparent, easily understood environment to inform the public and policymakers about current capabilities.
- Expand research into the physical limits of computation, including whether breakthroughs in GPU interconnect efficiency or other bottlenecks could rapidly accelerate AI capabilities. The research would focus on the theoretical limitations of computation that can be extracted from a given amount of hardware, with Epoch's research on energy, compute and data bottlenecks cited as a positive example of related work.
- Prioritize recruitment strategies that emphasize agency, drive, and founder-like initiative ("cowboys not academics"), noting the risk of constraining highly capable hires within overly rigid roles. The UK's AI Security Institute (AISI) decision to hire Alan Cooney was mentioned as a positive example of this strategy.
- Evaluate the feasibility of suing contemporary AI companies for negligence as a way to create legal clarity and pressure for stronger risk-mitigation standards, even before a formal liability regime exists.
- Demonstrate more clearly that current AI systems already enable significant economic gains, reducing the perceived need to push the frontier further.
- Conduct worst-case planning for scenarios in which adversaries develop uncontrollable superintelligence, including consideration of extreme defensive measures such as disabling or destroying data centers.
- Increase high-level strategic geopolitical analysis (such as Aschenbrenner's *Situational Awareness*) to ensure that policy portfolios are coherent, aligned with a broader situational awareness, and avoid contradictory or counterproductive measures. Making a policy portfolio without a coherent perspective means that individual policies will not work together and could easily conflict (e.g. the 2025 France AI Action Summit).

In honor of the fastest animal on the planet