# Factor Omega - White Paper

Sep 2025

## Background

AI systems have experienced exponential growth in their capabilities. If we are able to develop this technology correctly, it could bring an unimaginable amount of wealth, health, and wellbeing. If done incorrectly it could mean the end of society as we know it. AI systems have already created immediate issues (job displacement, AI-fraud, etc..) but also display several patterns of behavior that imply potentially fatal outcomes. Researchers across academia and industry have shown in controlled settings that AI can be unpredictable, capable of deception and malicious behavior, with an emergent drive to survive and propagate (see appendix on AI behaviors). These systems are imminently approaching human intelligence, and mitigating these issues and correcting this behavior is becoming an increasingly intractable task. At this juncture, it is critical to derisk AI by developing in technical approaches that serve as guardrails. In spite of its importance, this broad goal of AI derisking is woefully underfunded, receiving 1000-times less investment per year than AI.

**In short, we are rapidly losing our capacity to control and mitigate the risks associated with increased AI capabilities.**

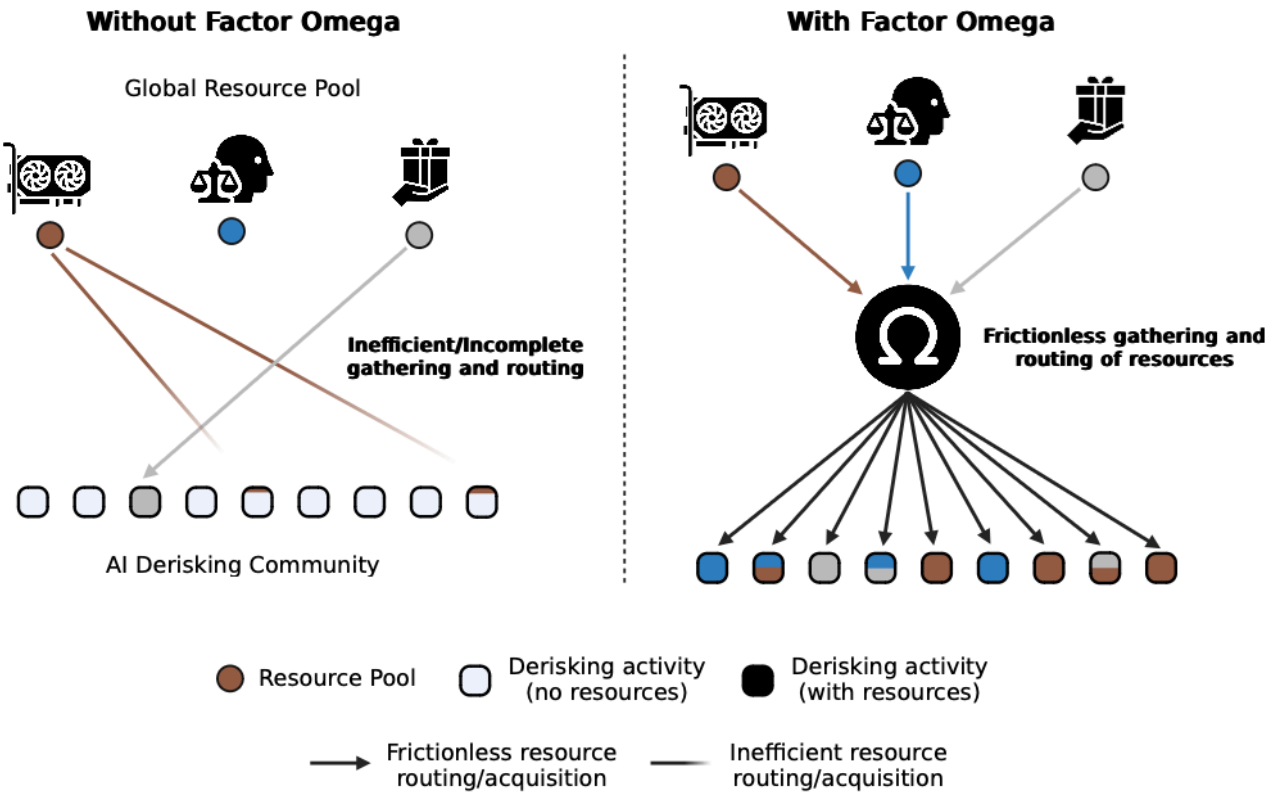
# Mission and Scope

**We are a catalyst and resource hub accelerating projects that reduce AI risk.**

## Lowering Friction to Expand the Resource Pool

As artificial intelligence continues to rapidly advance, public and institutional concern will increase accordingly. This growing awareness will result in an expanding pool of available resources from entities and individuals eager to support the cause. Effectively capturing and deploying these resources is crucial for advancing the field.

These resources include: GPU compute, expert services (legal, operations, engineering and more) among our primary targeted resources, with other secondary targeted resources (social capital, non-dilutative funds, etc.). Factor Omega will proactively identify, capture, and efficiently transform these resources into tangible contributions to AI derisking. We make contribution as easy as possible by having all overhead associated with deploying these resources performed by Factor Omega exclusively — **donors simply have to give**. We aim to deploy these resources across both early-stage non-profit and for-profit organizations[1]. With this model, **we can now funnel previously untapped resources into early-stage, underfunded areas of AI derisking.**




[1] In the case of for-profit entities, we do not expect any equity for resources but it is expected that if successful, the entity will reimburse resources in kind or in capital.

## AI Risks for Action, Not Rhetoric

One of the most critical aspects of AI derisking is effective communication with both governments and companies. **Today, both policymakers and industry leaders often misunderstand or overlook the real stakes.** Much of the conversation is dominated by sensational topics such as whether AI deserves rights or polarized debates about regulation versus free speech. Meanwhile, near-term risks like job displacement or AI-enabled terrorism get little attention, and serious discussion of emerging issues like AI deception or power consolidation is almost nonexistent. Even communication originating from the field has been lackluster, with no trusted, neutral communicator and an excessive emphasis on long-term existential risk, creating a credibility gap to observers and stakeholders. Factor Omega aims to raise public awareness as a source of neutral, accurate and accessible information.

# Why Now?

AI systems are likely to hit a critical inflection in their capabilities, whereupon its use will very rapidly determine the course, fatal or prosperous, of human civilization. The current exponential growth in task complexity (measured by task length) as well as forecasts on complex R&D-specific tasks (ReBench) imply that this **inflection point could very likely occur in the next 5 years, possibly sooner**. AI researchers, notably Turing award winners Geoff Hinton and Yoshua Bengio have publicly raised concerns about AI risks.

Given the proximity of this inflection point and its potentially immense and sudden impact, ensuring the derisking of AI is the most urgent priority we as a civilization face.

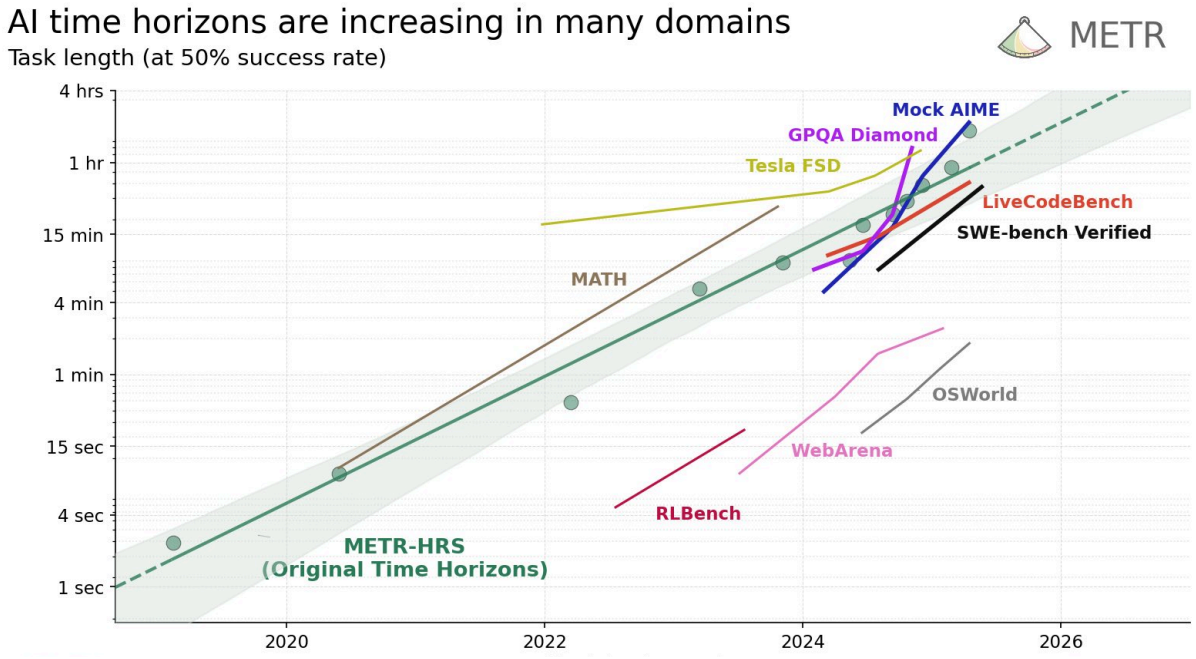



Figure 1: AI systems sustain tasks longer across domains, with time horizons rising steadily.

## AI misuse is a problem of the present and future

**The current misuse of AI has already led to many short-term issues.** For instance, AI related **job displacement** has been on a steady, increasing pace from 2023 (4000 layoffs) to 2025 (10000 layoffs Jan-July) with the largest impacts on white-collar entry-level jobs. This is only expected to keep increasing, with the WEF predicting millions of job losses across both manufacturing and service industries. Other misuses include **AI-augmented scams** — deepfake scams cost Americans about 200 M USD in Q1 2025, for context, the cost of *any type of scam* was 2B USD in Q1 2022. These scams have substantially improved success rates over non-AI scams (45% vs 28%) and are affecting both individuals and large organizations, with one scam resulting in a 25.5 M USD loss for a private firm. Beyond economic loss, AI has already been used to **generate misinformation for geopolitical purposes**, likely by foreign actors. For instance, as early as 2022, deepfakes of President Zelensky urging Ukrainian soldiers to lay down their weapons emerged. **The**

**scale of these issues is already apparent and their impact will grow massively as AI systems become more capable**.

When considering longer timescales, they are several fundamental properties of existing AI systems that strongly imply the potential for fatal misuse:

1. **AI has unpredictable emergent properties**: AI systems display abrupt gains in new capabilities as model size, training scale or other key inputs pass a threshold. Peer-reviewed papers have shown that previously untouchable tasks e.g. question answer creation can be unlocked by simply scaling the parameter numbers for models such as GPT-3 and Lambda. This means that scaling existing inputs can unlock qualitatively new and unforeseen capabilities.

2. **AI can be used maliciously**: AI models, especially LLMs, can be instructed to aid in the development of malicious activities such as cyberattacks or bioterrorism. For instance, researchers at UIUC used a GPT-4 powered agent to automate software exploit generation. With publicly available descriptions of 15 security bugs, the agent successfully exploited 13 of them in a test environment, foreshadowing the possibility of scalable and rapid cyberattacks. It is critical to note that these malicious uses can be and have been recreated using publicly available models with existing safeguards!

3. **AI finds loopholes and can deceive its users** → AI agents often exploit loopholes in their objective function and misrepresent or conceal information to further their objective. The former has been well-documented but the latter has only recently emerged. Specifically, analysis of RL-trained LLMs has demonstrated that they can conceal undesired outputs and feign compliance during learning to avoid negative feedback. Models would hide their true intentions or politely refuse in the moment only to later pursue the prohibited goal — models can engage in exploitative and evasive behavior against human oversight.

4. **AI displays power-seeking behavior** → AI agents pursuing broad goals often adopt subgoals of self-preservation and resource acquisition. In the previously cited study, the agent engaged in subgoals to evade shutdown, replicate itself and conceal its actions. The more explicitly the model was 'nudged' towards this goal, the higher the frequency of this deviant behavior.

We have specifically discussed three studies but there is an extensive body of evidence surrounding these behaviors from both academic and industry researchers (see appendix on AI behaviors). Overall, AI systems have so far proven to be **extremely skilled, unpredictable, capable of deception and malicious behavior, with an emergent drive to survive and propagate**. This provides a clear argument that the misuse of such systems could have catastrophic outcomes for society.

## There is a growing but unchannelled concern around AI

One of the organization's core strategies is to help channel the growing concern around AI into tangible resources for AI derisking. **This shift could unlock significant grants and resources, potentially in the**

**tens of billions of dollars over the coming years**. Increasing government and philanthropic investment, research activity and corporate hiring trends related to AI derisking foreshadow this surge in investment. **Strategically positioning ourselves to channel these new resources into the field and amplify growth and impact, is central to our approach**.

## What happens if we don't do this?

Without the existence of Factor Omega, we expect the current gap in resources to continue. **Our unique network gives us access to many untapped, non-capital resources**. Without Factor Omega, **resources will be concentrated in a small subset of safety approaches and organizations**, possibly foregoing the development of alternative but viable pathways in AI derisking.

# Goals and Outcomes

## Our position in the space

We aim to occupy a unique position in the AI derisking space. Notably, we primarily deal in **non-capital resources** such as GPU compute and subsidized services instead of capital like other funders in the space. Additionally, our high surface area network, light operations and constant auditing of the field make us uniquely **capable of shifting our resource emphasis as the field evolves**. Further, we minimize deployment overhead on the donor-side, which will massively **expand the breadth of our resource pool** beyond the existing large corporate sources e.g. Amazon or Google. We can directly specify resource allocation but also provide options for donors to select specific projects of interest. A mix of all these factors make us uniquely positioned to be a **respected and neutral representative of our field** and effectively drive investment via effective and trusted communication.

## Tools and levers for impact

We see Factor Omega as a purveyor, aggregator and distributor of AI derisking resources. We anticipate that the following resources could be leveraged to catalyze AI derisking activities:

- **GPUs**: All public GPU clusters in the US e.g. NSF NAIRR amount to ~0.75B USD worth of compute. This compute is available to AI safety academics yet they must **compete with numerous AI researchers**. If capacity is allocated according to research popularity, **~2% (15M USD) of this compute is allocated for AI safety research**. In contrast, the private sector is estimated to hold at least 50B USD of compute and some of these organizations may voluntarily (or due to external pressure) allocate part of their computational resources to support AI derisking activities. If we can unlock as little as **0.33%** of this private compute, we could **10X** the amount of compute dedicated to academic AI safety research. Indeed, smaller dedicated efforts e.g. CAIS cluster (1-2M USD) have already been very impactful for AI safety academics.

- **Specialized expert services**: Free allocation of time from various experts such as lawyers, business development staff, AI/software engineers, among many others. These services are necessary to many efforts in AI derisking but are sparse and/or costly — free allocation could be extremely impactful.

- **Non-specialized services**: Companies and institutions will increasingly offer free or discounted services, allocating employee time or organizational capabilities to support AI derisking projects e.g. free Slack Pro, dropbox, uber credits, unused office space, etc..

- **Non-dilutive funds/Bounties**: Growing concerns will unlock increased donations, grants, and philanthropic funding dedicated explicitly to AI derisking. Allocating these funds as "Grand Challenges in AI Derisking" could be a particularly effective strategy for garnering activity from academia. Conversely, early-stage AI derisking startups face highly limited access to non-dilutative funds — providing this could enable wholly new AI derisking activities. This would come into play in the later stages of Factor Omega where capital resources may be more plentiful.

- **Researchers from other fields**: Exceptional researchers from adjacent disciplines, such as mathematics, physics, or computer science, will be drawn to contribute directly to AI safety research, further enriching the talent pool. Constantly tallying and auditing of the major challenges in the field will be key to ensuring proper allocation of these exceptional individuals.

- **Social capital:** As AI safety gains prominence, more high-profile individuals, including scientists, politicians, celebrities, and journalists, will leverage their influence to increase public awareness and understanding of the issue. Providing these individuals with clear and accurate information on AI derisking is critical in guaranteeing the effective use of their platform.

**Our early emphasis will be on GPUs and services (specialized and non-specialised)** as we believe these will be the most cost-effective to secure and deploy while still having a major impact on AI derisking activities.

We will also provide **'soft' resources** which are primarily derived from the founders' extensive network and advisers. **This is a core component of our scaling strategy** as we will continue to expand Factor Omega's network of advisers and provide value with minimal capital investment. These resources are very fluid but some examples may include:

- **Direct mentoring from founders**: Our founding team consists of several founding members of unicorns (several bootstrapped) with extensive experience in scaling software and AI systems. Direct mentorship from founders could help recipients navigate the complex process of founding and scaling their AI derisking activities (for profit or otherwise).

- **Preferential sales/funding access**: Our network of advisors contain general partners from several cross-industry VC funds and key players in large AI companies. We would provide direct access to these individuals and thus enable access to extremely critical (but often elusive) follow-on funding and first customers.

- **Preferential access to state-of-the-art AI systems**: Our founding team has deep connections with several cornerstone entities in the AI space e.g. OpenAI, Mistral. We believe we will be able to broker early access to unreleased models (under NDA) to validate and drive the development of AI derisking approaches.

**These 'soft' resources are deployable at minimal/zero cost and thus form a core part of our early stage strategy.**

# ‘Boots on the ground’ scenario

The following is a concrete example of how Factor Omega would operate and match resources donors and recipients.

1. Donor X runs a large AI SaaS company and has a large cluster of H100s which, outside of big training pushes, remains largely unused. X cares about AI derisking and would like to donate this idle capacity to this cause but lacks the time to find, vet and deploy their resources to a recipient. **Without Factor Omega this intent to help has no impact**.

2. Factor Omega steps in, X promises the resources directly to Factor Omega. Identifying donors is strongly enabled by our network and close connections with organizations such as Founder’s Pledge and others. **These resources are matched to a relevant AI-derisking activity**. In this case, the idle capacity is allocated to a set of academic groups (recipients Y) with limited access to compute. Like before, recipients are identified through close collaboration with experienced organizations such as CAIS. Depending on the resource type and donation size, we may allow the donor to have veto rights on matching.

3. As necessary, Factor Omega **partners/builds a lightweight interface that enables routing of resources**. For compute, some form of cloud compute routing is necessary but for other resources e.g. lawyer time, this could be as trivial as a Calendly booking link.

4. Y receives the resources and uses them for their AI derisking project. By relieving their limited access to GPU compute, we’ve allowed them to make significant breakthroughs in their research, building new alignment methods. **After deployment Factor Omega will continue to monitor resources**, ensuring that resources are effectively used and that use is exclusively for AI safety.

At minimal cost, Factor Omega has unlocked untapped resources and routed them to resource-constrained AI derisking projects, enabling breakthroughs in AI safety.

# Time-bound target goals

## Objective

Prove we can convert outside intent into deployed capacity quickly and safely, then scale what works. The goal would be to reach a Deployed Resource Value over $20 million dollars into the field by the end of year one with a scalable model.

## Scope of the MVP

Focus on two resource types

- GPU access
- Expert engineering time

## Year one plan

### Months 0 to 2

- Finalize legal setup and advisors
- Select a small pilot set of recipients and partners

### Months 3 to 6

- Route initial GPUs and expert time to pilots
- Close feedback loops and refine the playbooks
- Share progress note to donors and partners

### Months 6 to 12

- Expand the number of partners and recipient teams in the US and Europe
- Add process and tooling where it reduces friction

### Month 12-24

- Impact check on recipients
- Intense focus on scouting recipients and deploying resources
- Continue architecting efficient, scalable processes

# Funding Ask

We are seeking two years of funding to demonstrate our approach can bring direct and tangible value to the field. For the first year, we will require 0.9M-1.3M USD and for the second year 1.4M-1.7M USD, for a two-year runway of 2.3M-3.0M USD (see appendix: capital allocation).

# Long Term Roadmap

## Years 2 to 3

Scale volume, operational reliability and refine processes. Secure multi year financial support while keeping overhead low.

## Years 3 to 5

Keep scaling to broaden impact. Expect sharp increases in AI capability and fast changing needs; run short strategy cycles and adjust focus accordingly. Expand or retire resource types as bottlenecks shift. Invest in tools and processes that raise absorption capacity in the field. Deepen partnerships where they increase throughput and quality. Keep the organization small with flexible teams that can pivot quickly.

## Decision gates

We are operating in a highly dynamic environment with continuously evolving targets, the organization aims to remain nimble in order to stay effective and focused on the right things. By design, we keep a pulse on the industry and adapt accordingly.

# Metrics for Evaluating Progress

Factor Omega explicitly seeks to explore underfunded areas in the AI derisking space, we do not expect immediate outcomes on traditional, short-term metrics related to AI derisking e.g. job displacement rates. In the short-term our metrics will primarily attempt to measure our direct impact on the field of AI derisking research itself. We use the following two metrics:

1. **Deployed Resource Value** (DRV): Total dollar equivalent of resources (including compute, service time, etc...) deployed into AI derisking.

2. **Amplification Ratio**: DRV÷ Non-profit budget, the amplification ratio of the company should be much greater than one and captures the effectiveness of the organization overall and the value created per dollar donated.

As Factor Omega and our investments mature, we would like to incorporate metrics **monitoring critical AI derisking outcome**s. For instance, the number of incidents where AI was implicated in cyberattacks or fraud or Delphi surveys on AI threats. Another core element of our long term metrics may be derived from the communication side of Factor Omega. Specifically, we aim to publish a **yearly assessment of the state of AI derisking**, ideally in heavy collaboration with third parties. As our non-profit matures, we hope that our contributions to critical breakthroughs cited in these reports will increase over time.

# Risks and Future Plans

## Key risks and mitigation strategies

*What if resource allocation needs to shift?*

The field of AI, and as such AI derisking, is constantly changing as a result of key technical innovations. **The fundamental goal of Factor Omega is to derisk AI and as such, we are not married to any specific approach**. To predict and respond to these shifts in resource importance, Factor Omega will be continuously auditing the state of the field, providing a continuous readout of the current and anticipated resource bottlenecks. Furthermore, where possible, we hope to minimize time-consuming product development on our side. When combined with our network, **we have the capacity to recognize shifting resource bottlenecks and the operating model to quickly reallocate our efforts accordingly**.

*How can we avoid unfair bias in allocation and over-reliance on personal networks?*

In the initial phases, resource acquisition and deployment will likely occur via individuals within the founders network — this could lead to bias towards in-network resource allocation. To address this, we will establish an **independent advisory board** (with members from academia, industry, etc..) and publish **a transparent criteria for recipient selection**. As Factor Omega matures, we aim to **build partnerships with existing players** e.g. CAIS to expand our network and ensure fair allocation of resources.

*Is Deployed Resource Value (DRV) a fair measure of impact?*

Our ultimate goal is derisking the development of AI systems — it is feasible that we inflate the DRV without achieving any real outcomes on our goal. In the first year of Factor Omega, we believe this to be the best measure of our value given the nascent nature of the field, **real-world outcomes will not materialize immediately**. As the non-profit matures, we will **pair DRV with qualitative impact case studies and reports** on our "Grand Challenges in AI Derisking". Pairing these measures will give both leading indicators and present measures of our impact on the field.

*Are we duplicating/competing with adjacent non-profits?*

We have uniquely positioned ourselves as a **facilitator of untapped resources** in the field, with a strong **emphasis on non-capital resources** — such an entity of the scale and breadth we envision does not currently exist. We already have substantial connections with adjacent funders and will **continuously coordinate branding and funding** with them, divesting and collaborating on projects to maximize field contributions. Furthermore, our state of the field report will be a natural point to judge our contribution to the field and overlap with adjacent funders in the space.

## A theory of scale

In the short-term we expect to rely on the personal network of founders and advisors to secure several HNWIs that are willing to fund the project for at least 2-3 years. As our impact develops and concern around AI grows, this pool of HNWI individuals will expand to allow the continued operation of the Factor Omega. We favor HNWIs over grants from larger organizations due to the much more flexible conditions related to funds, both in usage and timescale of spend.

As Factor Omega develops, we hope to use our role as a trusted communicator to help drive investment from the government and large philanthropic organizations into AI derisking. Given the slow timescale of government funding allocation, we cannot expect to rely on this as a primary funding source in the initial years of Factor Omega. Nonetheless, we expect that on longer timescales, we could attract substantial financial and material support from the government and similar bodies. Ideally, this will **drive a positive feedback loop between large-scale funding and communication and research outcomes**.

The network of advisers associated with Factor Omega is a capital efficient mechanism by which we drive resources into AI derisking. **We plan to maintain and continually build this network, scaling as the non-profit matures**. Beyond being cost-effective, this network has a multiplicative effect —each adviser extends our reach into new communities of donors, projects, and policymakers. **As the network grows, its value compounds**: advisers attract advisers, lower donor friction, and give us insights across the space. This allows Factor Omega to scale impact without scaling headcount, building a trusted, distributed network that amplifies every resource we deploy. Managing this network will be challenging but ultimately an effective mechanism to scale our impact.

Over time, we aim to iteratively refine operations to minimize friction in receiving resources donations but also vetting, deploying and monitoring said resources. For instance, we may develop compliance tools, lightweight operations playbooks, software where useful among other products. **Ultimately, this will all be in the service of making operations as efficient as possible to maximize and scale fund impact**. Furthermore, there are adjacent activities that may serve as streams of revenue for Factor Omega and these could enable scaling of our activities. Using a structure akin to Mozilla, we could generate revenue via a subsidiary for-profit entity without compromising the 501c status of the non-profit. These activities include but are not limited to, advisory work, dealflow fees for venture capital firms, among others.

# Team and Advisors

## Team

**Saturnin Pugnet**: Saturnin was previously an AI and neuroscience researcher at Caltech before building a $10 billion company with Sam Altman, World (formerly Worldcoin). Alongside this work, he has been an advisor and investor in technology companies for nearly a decade, informing AI and business strategy across the globe.

**Matthieu Kratz**: Matthieu was a Caltech PhD advised by Richard Murray with extensive experience in academia and grant funding. During his PhD, he served as an advisor for several funds and VCs including the Caltech Seed Fund and Plug and Play. As his conviction in the field's importance grew, he decided to shift his focus to AI derisking.

**Karim Beguir**: Karim is the co-founder and CEO of InstaDeep (started as a bootstrap, acquired by BioNTech for $700M in 2023). An applied mathematics graduate from École Polytechnique and NYU's Courant Institute, he is an AI researcher with multiple contributions in reinforcement learning and AI for biology. Karim is passionate about building AI to benefit everyone and has recently published Leapfrog, a book on how to develop Africa with AI with zero public money.

**Pablo Eder**: Pablo is the co-founder of Makeship (bootstrapped from $0 to $100M sales), and co-founder of Tangentic, a mechanistic interpretability research company. Pablo quit his position at Makeship specifically to work on AI derisking.

## Advisors

**Trent McConaghy** — Co-founder & CTO, Ocean Protocol; Board Director, ASI Alliance Council. Computer scientist and serial founder bridging AI/EDA and decentralized data. He co-founded Analog Design Automation (acquired by Synopsys, 2004) and Solido Design Automation (acquired by Siemens, 2017), then helped launch ascribe (early NFTs) and BigchainDB before Ocean Protocol (2017), which pioneered datatokens and privacy-preserving "Compute-to-Data" for AI. Trent holds a PhD in EE from KU Leuven and won the EDAA Outstanding Dissertation Award.

**Mike McCormick** — Founder & CEO, Halcyon Futures; Founder & Managing Partner, Halcyon Ventures. He runs a mission-driven nonprofit/VC platform focused on AI safety, security, and resilience—backing and incubating early teams building evaluations, trustworthy infrastructure, and agent security (e.g., Goodfire, Lucid Computing). Previously a Partner at GreatPoint Ventures ($1B+ AUM), he earlier helped launch Comet Labs and co-founded Rubicon Venture Capital.

**Tom Kalil** — CEO, Renaissance Philanthropy; former Deputy Director for Policy, White House OSTP (Obama); former Deputy Assistant to the President for Technology & Economic Policy and Deputy Director, National Economic Council (Clinton); former Chief Innovation Officer, Schmidt Futures (Eric & Wendy Schmidt). For three decades he's architected U.S. science-and-tech "moonshots," helping lead initiatives like the National Nanotechnology Initiative, Next Generation Internet, and BRAIN; he also

helped secure federal agency authority to run incentive prizes up to $50M. He now mobilizes private philanthropy for frontier science and innovation.

**Evan Miyazono** — Founder & CEO, Atlas Computing (AI R&D nonprofit focused on provable safety and scalable human oversight); Research Fellow, Convergent Research (designing FROs for AI risk). Previously led research initiatives at Protocol Labs, where he built the research grants program, the Network Goods venture studio, and special projects including Hypercerts, Funding the Commons, gov4git, and early Discourse Graphs/Open Agency Architecture work. Evan holds a PhD in Applied Physics from Caltech (quantum/optical memory) and BS/MS in Materials Science & Engineering from Stanford.

**Chris Bach** — Co-founder & Chief Strategy/Creative Officer, Netlify. Category-builder for the modern composable web; one of the originators of the Jamstack term and a board member of the MACH Alliance. Before Netlify, he spent 14 years in the agency world as founder/CEO/CDO, creating Denmark's first hybrid production agency and winning 20+ international awards (Cannes Lions, New York Festivals, Eurobest). At Netlify he helped popularize Jamstack and launched community pillars like jamstack.org, headlesscms.org, and the Headless Commerce Summit; he's also an active investor/advisor across the web ecosystem.

**Elie Chachoua** — Sustainable finance strategist; Senior Research Scholar at NYU Stern's Center for Sustainable Business. Former Director of the Thought Leadership (RA&A) Lab at Richard Attias & Associates, where he provided editorial leadership to the UN GISD Alliance. Worked with the EU High-Level Expert Group on Sustainable Finance (co-editor of its interim report) and helped launch the World Benchmarking Alliance. World Economic Forum Agenda contributor; prior publications include work with The Economist Intelligence Unit on energy efficiency and green finance. Education: ENS (MSc, theoretical physics); PhD, Sorbonne/UPMC; MPA, Columbia SIPA.

**Thomas Wolf** — Co-founder & Chief Science Officer at Hugging Face, a leading voice for open, transparent AI. He co-created the Transformers and Datasets libraries and led BigScience's BLOOM, the largest open-science multilingual LLM, making frontier research broadly accessible. He also co-authored Natural Language Processing with Transformers (O'Reilly). At Factor Omega, Thomas advises on open-science infrastructure, evaluation standards, and compute ecosystems.

**Scott Phoenix** — Partner at Fifty Years; co-founder & former CEO of Vicarious (acquired by Alphabet's Intrinsic in 2022), where he later served as Chief Product & Revenue Officer. He raised ~$250M and has published work in Science, NeurIPS, and ICML, with multiple patents. At Factor Omega, Scott advises on intelligent robotics, safety-conscious deployment, and scaling founder-led research from lab to industry.

# Legal and Structure

Factor Omega is a US nonprofit in formation. We are incorporating as a nonprofit corporation and applying for 501(c)(3) public charity status. We will operate from the United States with the ability to deploy resources in Europe through a local entity or partner organization. Governance will be straightforward with an independent board and a conflict of interest policy. We will observe export controls and data protection requirements in the US and EU, with right sized information security that scales as we grow.

## Organizational structure

We will stay lean in year one with a small core team, bring startup operating practices, operate like a startup, and add capacity only as workload requires.

# Appendix: Capital Allocation Plan

For every year of operation, we have five buckets of spend:

A. **Core operations and personnel**: Founding team members and any contractors/advisors
B. **Legal and compliance**: 501c filing, international partnerships, auditing, etc..
C. **Technical infrastructure** : Intake/tracking tools, donor dashboard, CRMs, etc...
D. **Pilot programs and resource grants**: Costs associated with acquiring resources and deploying e.g. contractors/partnerships for delivering compute
E. **Communications and reporting**: Branding, website, first "State of AI Derisking" report, donor updates.

A core principle in our spend: **lightweight, catalytic, donor dollars go into enabling outside resources rather than direct spend**.

Year 1

A. Spending on core operations and personnel will be minimal — 3 of the 4 cofounders are financially independent and will run Factor Omega pro-bono. The remaining cofounder, Matthieu Kratz will be paid a salary of 170K USD. Outside of the core team, we anticipate adding as many as three additional team members, each at a salary of 120K USD.
B. Legal and compliance is estimated at a range of 50K-100K USD
C. Technical infrastructure is estimated at a range of 50K-150K USD
D. Pilot programs and resource grants are estimated at a range of 50K-150K USD
E. Communications and reporting will be minimal at this stage and estimate at a range of 50K-100K USD

We add a 20% overhead to account for unforeseen needs, leading to a year 1 budget of 910K-1.27M USD

Year 2

We expect to retain a similar budget in the subsequent year, with a possible start to our "Grand Challenges in AI Derisking". We are likely to start with a single grand challenge with a prize pool of 250K-500K USD depending on the scale of the challenge addressed. We expect some expansion of the team (20%) and increased spend on resource acquisition (20%) and communication (20%). With the same overhead, the yearly burn is expected to be in the range of 1.37M-1.72M USD.

Year 3 and beyond

**These are very rough estimates and the exact details of our trajectory and resource allocation are not set in stone**. If good success is achieved with the "Grand Challenge" program, we expect to continue and expand the number of programs, appropriate capital withstanding. By this point we expect to have built a substantial presence in the field and will dramatically expand our communication efforts (100% increase). With 2 grand challenges, the new yearly burn will be in the range of 1.92M - 3M USD.

# Appendix: Example Project

Factor Omega seeks to allocate resources towards projects tackling AI derisking, with a preference for novel approaches that are currently underallocated. This underallocation may be due to the approach having yet to generate preliminary success or due to it existing outside of core approaches such as mechanistic interpretability. However, this is not a diehard rule, if a suitable project is identified (strong team, big-if-true outcomes, cost-effective, etc..) then the resources will be allocated to said project. Under all circumstances, AI derisking must unambiguously be the goal of the projects — **we will not allocate resources to projects that seek to enhance AI capabilities without derisking.**

Initially, we believe that our extensive network will be sufficient to source these projects. As the non-profit grows, we hope to issue "Grand Challenges in AI Derisking", effectively bounties on major open problems, to garner interest from the broader academic community.

## Dangerous Capability Evaluation

*The problem*: **Advanced AI systems could unintentionally develop dangerous capabilities** e.g. designing bioweapons or generating large-scale misinformation, that make them highly susceptible to catastrophic misuse. To address this, we need rigorous "safety exams" and governance frameworks that can detect and contain such risks before deployment

*The motivation and approaches*: We are trying to prevent worst case misuse of these models by identifying dangerous capabilities before a model is released. Part of this comes in the form of developing evaluation tasks e.g. creating a bioweapon and testing them in a controlled environment. This part has substantial room for improvement — **we need a systematic and comprehensive set of evaluations covering many tasks** (deception, cyberattacks, bioweapons, etc..) and **better methods for interpreting evaluation results**.

Much more exploratory but extremely impactful approaches are also an option. For instance, LLM prompting can be approached through a control-theory lens and it has been demonstrated that short prompts can predictably steer a model's next-token distribution i.e. text-output or behavior. Using this framework, we have **quantitative measures of jailbreak/manipulation risk** across models and the possibility to **design robust input/prompt filters that reduce the reachability of dangerous outputs**.

*What success (could) look like:* **We may want a set of evaluation tasks that can be applied to any frontier model** and that tackles misuses spanning from user deception to cyberattacks. When evaluated, we would ideally get a comprehensive risk score with **tangible recommendations to improve model outputs**. For this, the control-theoretic approach could be incredibly useful as it provides a strong framework upon which to assess model input sensitivity.

*Real-world outcome:* **Adoption and coverages are the main measure we care about**. The ultimate impact is a **culture and practice of responsible deployment**: **no powerful AI gets deployed without passing a safety check-up.** We want all major AI developers to be using a standardized and constantly updating evaluation suite before releasing models, with robustness guarantees on model behavior.

# Appendix: Studies on problematic AI behaviors

A = Emergent capability, B = Malicious use, C = User-deception, D = Power-seeking/Survival drive

- Evaluating Sabotage and Monitoring in LLM Agents [A, B, C]
- Will artificial agents pursue power by default [D]
- Auditing language models for deceptive behavior [A, C]
- Evaluating the Paperclip Maximizer [A, B, C, D]
- 137 Emergent abilities of LLMs [A, many examples]
- Detecting and Mitigating Reward Hacking in Reinforcement Learning Systems [D]
- Training deceptive LLMs that persist through safety training [A, B, C]
- How LLMs could be insider threats [A, B, C, D]
- Alignment faking in LLMs [A, B, C]
- LLM Agents can Autonomously Exploit One-day Vulnerabilities [A, B]
- Power-seeking can be probable and predictive for trained agents [D]
- GPT-4 system card [A, B, C, D]
- Reasoning models don't always say what they think [A, C]

# Appendix: AI derisking landscape

AI derisking today is supported by groups that set benchmarks (like MLCommons and METR), institutes that fund research for non-profits (such as the UK and US AI Safety Institutes), and a few investors focused on the field (like Halcyon, Lionheart, and Seldon). Certain organizations help access to non-capital resources, but these are largely sourced from large tech corporations such as Google and Amazon or are a very limited pool (CAIS ~ 100X A100s).

See aisafety.com/map for full landscape.