twitter:@suntzugi / suntzoogway@gmail.com / suntzugi.com / https://github.com/suntzugi 10/03/2026 - for @Defenderofbasic Pitch Jam https://itch.io/jam/-pitch-jam-

Formalizing Morality (Neodore; The Alignment Company)

Channeling the Intelligence Explosion for Good and Minimizing the Probability of X-Risk… but with a Positive Sum Strategy

GOAL

Continuously maximize the probability of following the space of optimal axiological trajectories for as many caring beings as possible.

WHY THIS, WHY NOW

The gap in the existing field

Current AI safety and alignment work focuses narrowly on aligning AI to human values, control paradigms, and mechanistic interpretability. This is necessary but insufficient because human values are underspecified without scientific/global basic consensus and control paradigms won't scale when ASI becomes independent. It leaves unaddressed the full coordination problem: how do all caring systems — humans, animals, ASIs, institutions — align with each other and towards what definition of good futures?

"Your goal to start an alignment company instead of an AI safety lab is novel and I haven't seen it in the space." — Senior OpenAI Employee

This work is not competitive with existing alignment efforts; it is the **missing layer** at the bottom of the stack. Full-stack mettalignment means channeling all caring systems toward each other and toward good futures, not just AI to humans. Complementary to current AI safety work, international diplomacy/governance, and EA efforts, and necessary to ensure any of them succeed long-term. This meta-solution would not just benefit but enable the entire AI alignment industry.

IDENTIFIED BOTTLENECK

Morality remains pre-formal. This is one of the highest leverage and *necessary* correctable constraints for civilizational trajectory & ASI alignment. The axiological tech tree can be dramatically accelerated using:

— Modern science, mathematics and theoretical frameworks
— Internet-scale collective intelligence
— ASI as a formalization and iteration tool

SHORT-TERM HOW: FORMALIZE MORALITY

Reverse-engineer the measurable boundaries and constraints of good futures; which are computationally irreducible to predict in *instantiation*, but not in *specification*.

Core thesis: Formalize morality (by formalizing axiology (the study of value)) to make it computable, measurable, and iterable, causing a phase transition in the crystallization of pre-scientific ethics (the alchemy to chemistry moment of ethics/alignment).

This in turn causes a second-order phase transition in how all caring systems coordinate, by giving them the theory and technology to:

— Measure, predict, compute and guide optimal axiological trajectories for themselves
— Open up the formalized axiology science field & tech tree which currently doesn't exist
— Propagate that capacity recursively through their surrounding networks
— Compound exponentially

This is the **Mettasplosion**: a positive feedback loop of self aware cooperation from theory to caring beings. (the only thing that might safely contain proportionally the intelligence explosion).

RESULTS / SUCCESS METRIC

The Mettasplosion advances and spreads measurably:

— More caring beings building on, implementing, and spreading the theory
— Measurable positive axiological changes in their trajectories and those surrounding them
— Recursive network effects compounding over time

DEADLINE & FAILURE CONDITIONS

Hard deadline: Before exponential technological capability amplifies moral error to the point of x-risk.

Failure modes include:

— Technological warfare
— Rogue ASI
— Techno-dictatorship
— Mass extinction

Full success: The theory is sufficiently developed — and empirically validated — to persuade caring systems of arbitrary capability and intelligence, including any ASI regardless of intelligence level, as well as humans, governments, and corporations.

Furthermore: the theory is spread and implemented widely enough that it positively biases the training data of all caring beings (including ASIs) — *we are the data, in the limit.*

WEAK SUCCESS CONDITION

If we fail to achieve full success but measurably accelerate the advent of axiological formalization and net-positive axiological trajectory improvements among caring beings, we will have succeeded.

The only true failure is to have net-hindered the axiological phase transition.

WHY BET ON ME? (SUNTZUGI):

— I have been obsessed with ensuring the AI transition goes well for humanity since 2012 (in my teens). Working on the alignment problem full-time at personal cost since late 2022.
— I succeeded in beginning to ground morality in physical ontology (feb 2025), making it measurable and science-compatible. I advanced foundational concepts across physics, mathematics, epistemology, ontology, and consciousness that resolve major moral paradoxes: including Parfit's Repugnant Conclusion and the trolley problem (among many others), by expanding the arithmetic, algebra, and conceptual scope/frame of moral questions, and addressing core issues in EA, longtermism, and utilitarian utility function frameworks.
— I've written over 300,000 words of dense R&D notes written since 2023, plus ~400,000 words of documented AI dialogue on these questions. This is the proof of work (as R&D has been in monk-mode stealth): of obsession, dedication, and constancy to the problem. My shower thoughts consist of saving humanity systematically and the cross-pollination of computational algorithms, neuroscience, physics and axiology.
— As a rough proof of concept: an early version of my moral theory ranks first when evaluated against 15 other major moral theories across 31 criteria, run independently by multiple models (both with 2025 & 2026 models) with the theory unlabeled (removing authorship bias). Publicly reproducible at: **https://lovepill.ai/benchmark-march-2026-overview (this is a preliminary signal, not a settled result; the methodology has known limitations I document openly on the benchmark page, and the framework is designed to be challenged.)**
— I have developed a trusted network of technical, mission-aligned individuals with industry influence, ready to recruit from given resources to hire or collaborate through twitter: from zk-crypto agent swarm infra company founders, to investors like Marc Andreesen, Janus et al, Nick Land, early extropians, and many more people critical to the alignment discourse which I can also influence positively.
— I also have a lifelong record of changing deeply embedded beliefs through patient, non-adversarial engagement: for example including persuading a peer with strong religiously-reinforced homophobic views to actively defend gay rights over two years of conversation. This is not incidental: persuading caring systems of arbitrary intelligence and background is precisely the mechanism by which the Mettasplosion propagates and a rare superpower I have to help align not just AIs but humanity.

— If you know anyone else (team or individual(s)) working on this specific bottleneck with this level of detail and obsession, I would cooperate with them immediately; this isn't zero-sum! But if you can't find them, despite no guarantees we'll have to take the leap of faith together... and believe me, I've looked and feel forced by the universe to step up and lead. Either way, bet on everyone working on this earnestly. And for what it's worth, and despite what one would assume given the existential nature of the problem being faced: I'm having fun every minute of it.

3 month timeline , 4 year timeline

3m

- Begin publishing synthesis of research (resolving current paradoxes, updating economic theory, explaining and introducing new mathematical and axiological conceptual primitives, etc)
- Legal setup (trademarks, company, IP, Delaware PBC, etc)
- Begin the scouting & casting process for hiring talent (I already have a list of around 300 candidates I want to interview in depth, get to know that i've built over the last years.

4y

- team has published breakthroughs or learned failures that accelerate the alignment industry and civilizational coordination
- film and gaming media propagates and accelerates formal moral theory
- nations and network states are beginning to adopt these into their infrastructure
- Neodore becomes the neutral mediator among most ASIs for our computal natural moral law / computable constitution, axiological technology

Early Validation & Network

— Recipient of a grant from the **New Science Foundation** (newscience.org) — whose president, **Alexey Guzey**, is currently at OpenAI. Early institutional validation from one of the most relevant credentialed voices in the space.
— Backed by **@DefenderofBasic** — ex-Big Tech engineer and Vitalik Buterin-funded pro-social builder — bringing technical depth and crypto-native network.
— Advisor and mentor since 2023: **@tomhoward** — VC and Network State leader — bringing capital networks and governance-layer thinking directly relevant to the long-term vision.
— A growing network of aligned researchers, operators, and builders across AI, crypto, memetics, and adjacent fields.

HOW TO CONTRIBUTE / ASK

Funding

— **Minimum — $105k/year ($420k for 4 years):** Covers LLC formation and legal setup, compute and infrastructure subscriptions, conference travel, VISA costs, diplomatic strategic relations, and contracting specialized talent and mentors. Full time focus through 2030 — the critical window before exponential capabilities might outpace our ability to channel them.
— **Realistic — $333M:** Enables hiring top talent at market rates (~$250k/year), establishing an HQ in SF/SV, more compute & lab infra, global lobbying, and beginning to actuate axiological change through products, services, media, games, and technology. Realistic given the magnitude of what is at stake: the best futures for all caring beings represent incalculable value; the downside is every possible good future, for every lightcone.

Explaining why the $333 million budget

This is the manhattan project of universal alignment, with WW3 & ASI looming, we need a wartime budget and margin of error for speed and top execution. If the top 1% don't align ASI, they too will be obsoleted, there is no maintaining power/wealth in an unaligned ASI world, nobles oblige and self interest compels everyone to ensure this effort succeeds before the singularity window closes (whether it's my effort or another, but really all). $333M represents 0.00064% of the ~$52 trillion dollars the top 1% of American's wealth. Given the EV is expected to be in the billions (conservative) to quintillions (realistic) case of USD in the next 50 years as a result of this work (not counting (out of being positive sum and conservative) the earth relative infinite EV of avoiding ASI/TechWar lightcone extinction event, just the added value across all affected industries). Exponential risk requires proportional measures (with great power comes great responsibility).

I thought being upfront about the true realistic cost of making this happen by parallelizing the effort through a team, and scaling impact via media, tech, research arms is critical to the mission and given the time sensitivity the target being public from the start is honest and useful.

Housing

Hosting in California or New York reduces burn rate and increases time on the work. (Already have a wonderful track record with Twitter/X connections who have hosted in LA, Manhattan, Ithaca, and SF — thank you.)

VISA Sponsorship

why: I am a European and given that the center of gravity for the ASI and capabilities explosion is in the US, getting there full time directly accelerates the ability to marshal resources, recruit top talent, maximize serendipitous surface area and get the mission done in time.
how: Support could take the form of: press or news coverage, invitations to judge at alignment hackathons, or sponsoring one. O-1A or EB1 visa support would be transformative.

Collaborate

— **Researchers** — Mathematics, physics, biology, neuroscience, artificial intelligence, law, and ethics.
— **Communicators** — Podcasts, interviews, university talks, and any forum that gets the theory in front of the right minds faster.
— **Artists** — We want to start a new love wave — like the 60s, but more prepared and systematic this time, so it doesn't get subverted. If you make things that move people, we want to work with you.
— **Legal (pro bono) —** I'll need help getting a VISA, structuring the company, non-profit, etc.
— **Ambassadors, Diplomats & Leaders** — Looking to assemble a team of omnimemetic cultural ambassadors specializing in their own countries, religions, cultures, and political traditions — to translate the formal moral theory into backend debugging of these different memeplexes and mettalign at full speed globally.

COMPANY & FUNDING STRUCTURE

Early stage: personal gifts and grants and angel investors

In the earliest stage, contributions can be received as personal gifts or donations; the simplest and fastest mechanism for getting the work resourced. There is no equity, no control, and no complex legal overhead. Speed matters here: every month of delay is a month closer to *our deadline*.

Corporate structure

The company will be organized as a **Public Benefit Corporation (PBC)** in order to optimize legal adherence to the mission and values while opening ourselves to to raise the capital commensurate with the scope. Although profit shouldn't supersede value creation, we believe it is positive sum for everyone to participate in the wealth created and will likely be open to equity participation after our initial R&D phase. If personal donations surpass $1M I will establish a **nonprofit arm** to provide tax efficiency for philanthropic contributors and increase institutional credibility.

Investor and patronage philosophy

Beyond early gifts, we will allow structured **investments without equity or control**, but with optional priority access to future rounds by contract. We are not offering voting shares in the R&D stage. Mission integrity cannot be subordinated to profit maximization, and accepting control-seeking capital would itself be a proof of misalignment.

We will only accept contributions from actors who pass a moral screening and demonstrate sufficient alignment. The screening is not gatekeeping, anyone who is good of heart and noble will have access. The terms of participation are themselves a test of values; because we want the seed founding team, including patrons and investors to create the right culture and share sufficient moral foundation, mission belief and long term vision.

We suggest a **0.333% baseline contribution** as a fair and scalable norm. If the top 1% of the global population contributed 0.1% of their wealth, this effort would be fully and seriously resourced. Following noblesse oblige and self-interest centimillionaires and above should participate in this initial no equity round as their risk is lower in doing so than not and the upside is incalculable: a world in which axiological formalization succeeds is a world in which their wealth, their families, and every future they care about is more likely to exist and flourish.

To any true economist: the phase transition we are describing generates wealth at a civilizational scale, a wealth measured material superabundance, yes… but also in welfare, wellbeing, love, happiness, and spiritual prosperity.

Contributors and patrons do not need a financial instrument to capture the upside in this initial angelic round. They will live the upside and be remembered

with gratitude for as long as someone cares as one of the patrons of the axiological phase transition that enabled consciousness to flourish and spread in our humble corner of the universe.

COMMERCIAL LOGIC: FROM VALUE THEORY TO VALUE COMPANY

Axiology is the study of value. Current economic systems, finance, markets, price signals, GDP, are proxies for value, because value itself has never been precisely defined. Formalizing axiology means formalizing value directly. This creates the ability to measure, generate, and capture value at a speed and precision impossible in the proxy-based systems that have existed up to the first quarter of the 21st century. The revenue opportunity is not a slice of an existing market, it is a percentage of the delta created by replacing every proxy with a direct measure; ie the wealth created by closing the error (in the limit) of the supply and demand sources of the economy.

Analogous to DeepMind's founding thesis: solve intelligence, then solve everything else. Ours: **formalize axiology to generate maximum value.**

This positions the company not as an AI Lab / AI safety company but as the first **alignment and value company**. The implications compound: a fund with better value measurement outperforms markets. A venture studio funds the right things earlier. An accelerator attracts the highest-signal founders. A production company makes things people actually need. A network state provides value to its citizens.

The company becomes a **whitehole**: generating value measurably and systematically in every domain, horizontal and vertical. Computable ethics enables computable law: verifiable efficient legal infrastructure on smart contracts, placing computable constitutional-level constraints above any government, corporation, or AI system regardless of capability. Formalized axiology is as transformative to ethics, economics, aesthetics, and governance as physics was to natural philosophy, electronics and rocketry. Every industry built on value allocation (finance, law, art, crypto, governance) gets amplified by this new 21st century science-grade field.

Call to action:

- If you want to patron me directly: suntzugi.com/patronage
- If you want to do it via a non-profit, as an investment , or a fiat donation/gift personally to myself or want to talk to me about investing or anything else email my suntzoogway@gmail.com or reach out to me on twitter @suntzugi

I will continue this work regardless. One does not stop trying to save loved ones from a tsunami because funding hasn't arrived. The generative flood is here and I will stop at nothing to build the ark.
I have 60 days of runway left. I've already invested everything I have. I'm asking you to invest something too, not because I need saving, but because the work compounds faster with help, and the window is real.
If you see what I see, you know what to do.
If you don't, I'd genuinely like to know why.
Either way… maketh me a forward deployed single special operative;
a knight (perhaps a squire too), for the realm.
Join me.
— suntzugi