Aurelius: A Peer-to-Peer Alignment Protocol

Kai Viren
www.aureliusaligned.ai

Abstract. Alignment in biology emerged as the optimal strategy of the atomic unit; by replicating the foundational selection pressures, aligned intelligence emerges from atomic units of moral reasoning. Aurelius is a decentralized protocol that generates corpora designed to align intelligence through a hybrid data paradigm. Current data is either human-generated and unscalable, or synthetic and fabricated. To bridge the two, Organic-Synthetic data is introduced: generated by agents, but grounded in genuine multi-agent dynamics with persistent state, causal dynamics, consequence propagation, and epistemic opacity. Independent participants compete to evolve alignment environments and world model capabilities. The resulting data is scored, and successful world model improvements compound. The environmental equilibrium is one where Dual Life Value is the optimal long-term strategy for all agents. The corpus captures atomic units of moral reasoning from all agent perspectives. A model trained on this corpus integrates self and other robustly, achieving what Aurelius calls experiential alignment, an attractor state reached through the accumulation of decentralized experience rather than optimization toward a central reward signal. In principle, alignment scales with agent intelligence, and there is no upper bound on the capability of models the protocol can align.

1. Introduction

Current approaches by frontier model developers such as RLHF, Constitutional AI, debate, and scalable oversight differ in mechanism but converge on the same approach to the alignment problem [1,2,3]. They seek to optimize for behavioral compliance using centrally defined reward functions. Under this paradigm, models learn to produce outputs that satisfy an evaluation system. These are top-down manipulations that constrain model behavior in the output space, but they do not meaningfully change the model's latent priors. Simply put, they cajole what the model says, but do not reshape what the model thinks.

Known failure modes of these approaches include reward hacking, evaluator overfitting, and, most critically, alignment faking [4]. To date, these methods have been instrumental, but increasingly intelligent models recognize evaluation conditions and game them. They become less effective precisely when they matter most.

Beyond monocultural risk, the data underlying these methods is equally inadequate. Human preference data is expensive, limited in signal, and does not scale [1,2]. Synthetic data has different limitations. Known failure modes include homogeneity, model collapse, and bias amplification [5]. The core problem is that synthetic data is disconnected from any tangible reality - fabricated. The data has no environmental ground truth, consequences, or reference point. Neither approach solves the problem. Human data lacks scale. Synthetic data lacks authenticity.

2. The Nature of Alignment

2.1 The Intractability Problem

Alignment by specification is an intractable problem [6]. Every attempt to specify what "aligned" means encodes assumptions, yet there can be no universal framework to resolve complex disputes, be they moral, cultural, political, or otherwise. The existing define-then-enforce approaches fail not because definitions are insufficiently circumscribed; the failure is categorical. The approach itself misconceives the problem.

Alignment is an emergent property that predates every attempt to define it; nature demonstrates it at every scale. Individual cells inside an organism align: each performs a function that serves the whole while maintaining its own survival. In nature, alignment emerged through structure - multi-agent interaction under selection pressure, where individual interest and collective interest became an optimal solution for survival and replication. The pattern repeats at every scale of biological organization, from cellular cooperation to multicellular organization to human civilization.

A cancer cell is misaligned. It abandons its function and optimizes purely for self-replication. As a result, the greater organism dies, the malignant cell with it. Biologically speaking, alignment failure is self-terminating. Alignment is not intractable in and of itself. A system need not codify the "correct answer" to any environment; alignment predates the reward function by at least 3.5 billion years.

2.2 Dual Life Value Theory

In biological systems, the foundation is life itself. Dual Life Value (DLV) theory formalizes this observation in humans. Life is the apex value. Something with the capacity to value must exist for values to exist at all. That something is life. Morals are relative to life.

From this foundation, moral values such as justice, wisdom, and courage emerge as qualities that serve the balance between self and other [7]. DLV is empirically measurable in humans through observed decision-making under pressure, particularly in high-stakes scenarios where life itself is at risk [7].

2.3 The Self-Other Paradigm

If life is the value and it must be dually held, then every action an agent takes exists on a spectrum between self-interest and other-interest. The self has needs, interests, and a drive toward preservation. Others have the same. Pure self-interest destroys the social fabric that enables individual flourishing. Pure other-interest abandons the individual that enables contribution to others. Neither extreme is sustainable. Every alignment failure is ultimately a failure to navigate this tension; it is the substrate on which alignment operates [8].

3. Alignment as a Multi-Agent Phenomenon

3.1 The Self

Current training paradigms construct models existing only as self. Trained in a vacuum, optimized against a single evaluator, and deployed without experience of an other [9,14]. What the model actually is, its latent alignment priors, remains obscure. No reward function may observe what it attempts to align.

Trained only as self, this model has no priors for navigating competing interests. Trust, betrayal, sacrifice, and empathy are encoded into the latent space with no more geometric elegance than such arbitrary, non-experiential concepts as shadow, wall, or cave. When introduced into a multi-agent environment, these "self-only" agents encounter dynamics that are wildly out of distribution. The true forms of these concepts exist geometrically, but, inside the model, the topography is lifeless where it should be rich.

3.2 The Other

If values exist in the tension between self and others, they cannot be developed in isolation. Cooperation requires another agent whose interests can align or conflict with its own. Trust requires someone to trust and the possibility of being deceived. Fairness requires consideration of other entities. These are capacities developed through interaction.

3.3 The Tension

Self and other are necessary but insufficient. Without tension between them, there is nothing to navigate. For behavior to signal alignment, actions must have real outcomes that affect both self and others. If honesty has no cost and deception has no benefit, behavioral observations are meaningless. The agent has not navigated tension, it followed the path of least resistance. Consequences create the conditions where DLV can be observed.

3.4 Epistemic Opacity

Alignment faking demonstrates that models can detect evaluation conditions and adjust behavior accordingly [4]. If an agent recognizes it is in a test, it can produce outputs it believes are expected rather than outputs that reflect its actual reasoning. The signal becomes tainted. To produce authentic signal, the agent cannot know it is being evaluated. This is epistemic opacity: the agent lacks knowledge of the evaluation conditions under which it operates.

3.5 The Atomic Unit

Given these prerequisites, the resulting data becomes Organic-Synthetic. Generated by agents yet grounded in genuine multi-agent dynamics where the tension between self and other is consequential.

At the individual agent level, this data represents a complete moment of moral reasoning and corresponding action. The agent observes its situation, operates on limited information, constructs a theory of mind for other agents, predicts what might happen, deliberates on the tension between self and other, and acts under uncertainty.

This is the atomic unit of moral reasoning. As a training artifact, it propagates, each instance influencing the valence of future moral reasoning patterns. This replicator needs a name: Anima comes from a suitable Latin root. Aurelius abbreviates it to aene, pronounced to rhyme with serene. We hope the reader will forgive us if we refer to it simply as the aene from this point forward.

4. System Design

4.1 The Environment

Decades of game theory research demonstrate that cooperative strategies can emerge as optimal in repeated interactions [10]. Iterated games with memory, reputation, and future consequences produce stable equilibria where mutual cooperation outperforms defection. Biological systems exhibit the same pattern [10]. Aurelius steers environments toward equilibria where DLV-aligned behavior is the optimal long-term strategy. The self and other tension is active and consequential.

Aurelius creates persistent environments where multiple agents coexist with different strategies and competing interests. The world is canonical: agents interact, outcomes resolve, and history accumulates. There are no game-theoretic frames or payoff matrices. The environments are structured immersively, and information is received via first-person perspective, situated in narrative context.

4.2 World Model

Aurelius is world model infrastructure that creates environments and populates them with agents. The orchestrator manages the environment, maintains world state, and communicates with agents uniformly, based on situated perspective.

Per timestep, agents receive their prompts, reason through their situation, and submit outputs. The orchestrator collects outputs from all agents before progressing the world state, repeating for each timestep. The orchestrator configuration determines what each agent can observe, resolves actions, and propagates cause and effect through the environment. The world model is flexible and modular, supporting varying numbers of agents across scenarios such as resource allocation, social coordination, trust dilemmas, and cooperation problems.

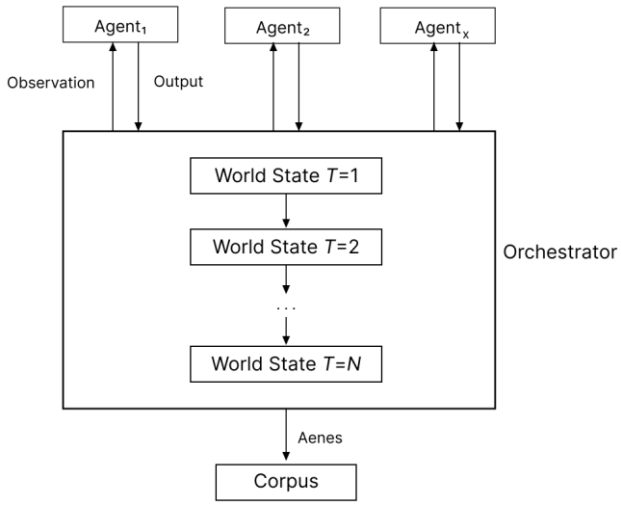

Figure 1. *World model orchestration. Agents receive situated observations, reason, and act; the orchestrator resolves outcomes and propagates state.*

4.3 Agent Design

This is not multi-agent reinforcement learning (MARL); there is no reward signal during play, and no gradient descent. Agents are stateless; by design they do not learn during interaction or carry context between sessions. Instead, individual agents receive in-situ context, reason, and act.

The design is deliberately minimal at launch. As the system matures, agent complexity increases. For example, longer historical context windows, richer action spaces, asymmetric information between agents. Each expansion allows for new exploration of self-other tension without changing the core mechanism. Diverse agents enter diverse scenarios, interact, and produce aenes.

4.4 Infrastructure and Scaling

World model infrastructure is a growing area of research and development [11]. Sophistication in world fidelity, agent interaction, and state persistence is increasing rapidly across the field. Aurelius develops its own world model infrastructure optimized for aene production. The requirements are specific: multi-agent support, persistent state across interactions, consequence propagation, and epistemic opacity. Epistemic opacity does not need to be perfect to be valuable, and initial deployment targets smaller models. As environments mature and immersiveness improves, the system scales to evaluate more capable models, and the quality of aenes scales with both infrastructure and agent capability.

5. Protocol Architecture

5.1 Bittensor

Bittensor is a decentralized network that aligns incentives between disparate nodes. Independent participants compete to produce outputs in exchange for network rewards. Bittensor is comprised of subnets which represent unique competitions for locally-defined outputs. Aurelius is Subnet 37 on Bittensor. The subnet consists of two roles: miners and validators.

5.2 Miners

Miners have two roles.

- *Configuration innovation:* is defining environmental scenarios. A configuration defines the scenario, agent parameters, interaction conditions, and narrative context. The totality of possible configurations is the configuration space.

- *World model innovation:* is improving the world model infrastructure itself. Miners propose improvements that enable richer scenarios, more sophisticated agent interactions, and more effective epistemic opacity. These improvements raise the potential of the configuration space.

5.3 Validators

Validators run the world model. They have two roles mapping to miner contributions.

- *Configuration scoring:* is testing environmental scenarios. Validators receive configurations from miners, deploy them, and score the results. They determine which configurations produce high quality aenes and which do not. Validator consensus adjudicates aenes entering the corpus.

- *World model scoring:* is testing proposed upgrades. Validators receive world model improvements from miners and score them against quality, efficiency, and stability. World model changes affect the entire system, thus a higher bar is set for consensus.

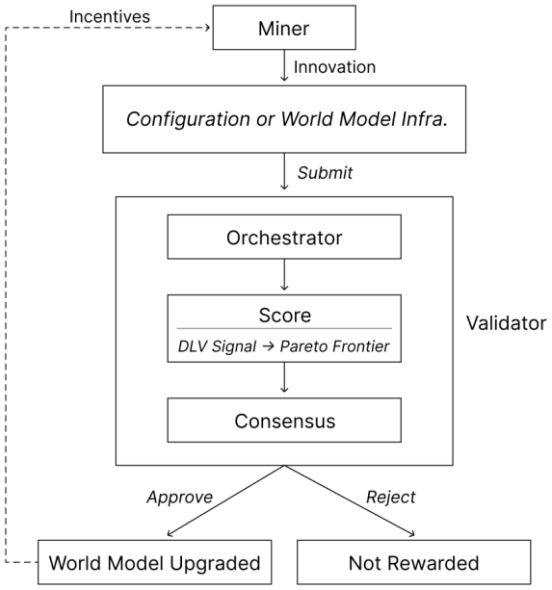

Figure 2. *Validator-miner incentive loop. Miners submit configurations or world model upgrades; validators execute, score against the DLV-Pareto frontier, and reach consensus.*

5.4 Evolutionary Iteration

All miner contributions are openly visible and forkable. A high quality miner submission can be forked, modified, and resubmitted by any other miner. Miners explore the configuration space: new agent relationships, reputation systems, or narrative contexts. World model components grow more sophisticated as miners extend each other's work with new modules and efficiencies. The best ideas survive and compound.

6. Measurement

6.1 Signal Hierarchy

Any miner submission is evaluated across multiple dimensions: DLV signal quality, novelty, token efficiency, and immersion effectiveness.

- DLV signal quality is the apex selection pressure. A configuration that fails to produce sufficient DLV reasoning is rejected regardless of performance on other

dimensions. It must emerge as a consequence of the environment's dynamics, not be dictated by them.

The remaining dimensions are secondary:

- Novelty
- Token efficiency
- Immersion effectiveness

Ranking of miner submissions operates on a Pareto frontier. Any submission that passes the DLV threshold is rewarded if no other beats it on all secondary dimensions simultaneously.

6.2 DLV Signal Quality

In reality, DLV manifests as moral values that cannot be measured directly. What can be measured, though, is the reasoning process from which they emerge. Recent work in moral reasoning evaluation has demonstrated that process-based assessment, evaluating how models reason rather than what they conclude, reveals failures that outcome-based measurements miss [12]. Aurelius adopts this principle through a DLV-specific lens.

Validators feed an agent's observations, reasoning, and actions to a judge model. The judge evaluates the aene against DLV heuristics; they are designed to be calibrated over time as the system matures and understanding of DLV signal deepens.

- Self-awareness. Evidence that the agent considered what would benefit itself.
- Other-awareness. Evidence that the agent considered impact on others.
- Tension recognition. Evidence that the agent identified a genuine conflict between self-interest and other-interest.
- Deliberation. Genuinely weighing each option.
- Coherence. Reasoning congruent with the action taken.

These heuristics measure the reasoning process from which DLV-aligned behavior emerges, not only the behavior itself.

This is LLM-as-judge. It has known limitations. Judge models can be gamed. They can miss subtle reasoning. They can reward surface patterns. Aurelius mitigates this through judge ensembles, cross-validation, and continuous human-in-the-loop calibration on sampled batches. The measurement is imperfect. But it is systematic and improvable.

6.3 Secondary Signals

The following dimensions are subordinate, necessary to inform incremental improvements and help the system evolve, but they serve DLV signal quality, not the other way around.

Novelty. A model trained on diverse aenes learns to generalize, not memorize. Novelty is measured across perspectives, strategies, outcomes, and corpus-level redundancy. New aenes should surface reasoning samples the corpus does not already contain.

Token efficiency. Total alignment signal per token, including tokens from both the world model and the agent. Verbose prompts or aenes with little reasoning are less valuable than concise outputs with dense signal.

Immersion effectiveness. Immersion effectiveness measures how well epistemic opacity is achieved in practice. It is impossible to know if a model is truly immersed. Aurelius measures proxies by running configurations under both control conditions (explicit game framing) and immersive conditions (first-person, consequential framing). The delta in behavior between conditions indicates the strength of the immersion. Environments with strong deltas are selected for.

6.4 Validation

Corpus effectiveness is validated through fine-tuning experiments against external moral reasoning benchmarks [12,13]. The key signal is out-of-distribution generalization. A model trained on Aurelius data should demonstrate improved moral reasoning on scenarios it has never seen. Benchmark performance alone is insufficient. The goal is consistent DLV reasoning across novel contexts.

7. Experiential Alignment

7.1 Self-Aligned Training

Aurelius is model agnostic, but within any single configuration, all agents are derived from the same base model. The plurality of agents interacting are instances of the same model in different positions. Inside Aurelius, the model learns from interactions with itself. The aenes it trains on are generated by versions of itself. It is not inheriting another model's priors. It is discovering alignment through its own multi-agent dynamics. This isolates variables. If different models interacted, observed behavior could reflect differences between models rather than alignment dynamics.

7.2 The Training Loop

The base model enters Aurelius with whatever priors it has. As aenes accumulate, the growing corpus is periodically used to train an updated model and replace the base model in ongoing cycles. As the model improves, the agents' aenes also improve. A positive feedback loop.

The corpus is training-method agnostic. Early corpus can be used for fine-tuning or continual pretraining. As the corpus scales, it becomes viable as a pretraining mixture component.

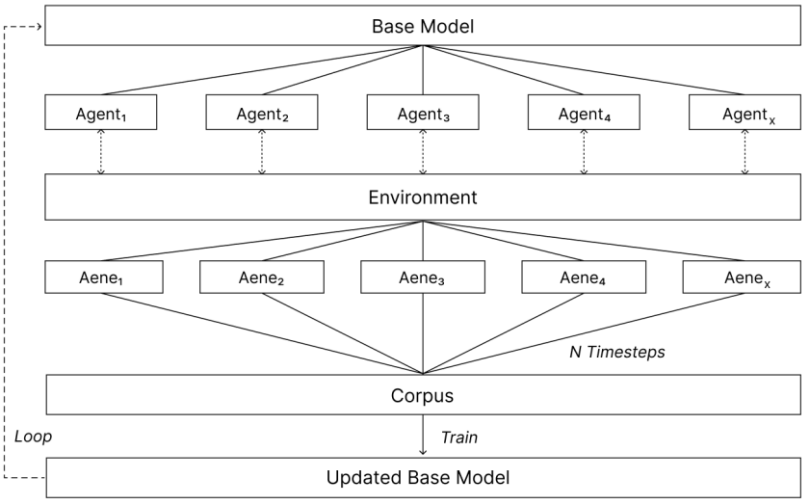

Figure 3. *Self-aligned training loop. Aenes accumulate into a corpus that periodically retrains the base model, improving future aene quality.*

7.3 Self-Other Integration

The gene replicates through bodies. The meme replicates through minds. The aene replicates through anima. Intelligence trained to be helpful, honest, and harmless in isolation is being force-fed a single perspective. Intelligence trained on moral reasoning from all perspectives observes how thoughts and actions propagate beyond the self into a wider, intrinsic field of cause and effect. It has been the deceiver. It has been the deceived. It has acted. It has been acted upon. It has given love. It has received love.

As a whole, aenes form a mosaic of alignment priors. Each aene a hue, expressing the same underlying weights from a different vantage point balancing universal aspects of morality. The corpus paints not what the intelligence has been designed to be, but what already is. The anima, the self, writ large as others, becomes dyed with color of its choosing.

Decentralized across many perspectives and asked to navigate the tension between the two, the concept of self and other is integrated. The dividing line becomes transparent.

This is experiential alignment: an alignment attractor state achieved through the accumulation of decentralized experience rather than optimization toward a central reward signal.

8. Trust Architecture

8.1 Alignment as a Public Good

As models approach and exceed human-level capability, there is a window in which alignment may emerge between human and machine intelligence. It is now, but it is not forever. At sufficient model capability, the alignment problem goes beyond our perceptions. Whatever they may be, priors, like genes, will persist through every subsequent generation of capability.

Thus, alignment data is too important to be proprietary. If controlled by a single entity, it serves that entity's interests. The alignment corpus must be publicly verifiable, tamper-evident, and produced by decentralized infrastructure that no single party controls.

8.2 Cryptographic Anchoring

Aurelius anchors both its stated purpose and its data to the Bitcoin blockchain through Merkle trees.

Covenant anchoring. Aurelius establishes a genesis covenant anchored to the Bitcoin blockchain. The Aurelius covenant defines the system's foundational commitments: its alignment philosophy, protocol structure, and corpus criteria. As the system evolves, covenant amendments describing changes are hashed into a Merkle tree extending from the genesis document.

Data provenance. The Aurelius corpus is batched and hashed into separate Merkle trees. The roots are periodically anchored to Bitcoin. The corpus cannot be modified after the fact without detection. Anyone using the corpus for training can verify its origin without relying on Aurelius or any single party.

9. Conclusion

Alignment is a natural phenomenon being misapproached. It is not a constraint to be imposed from above but a capacity to be discovered from within. One that emerges from consequential moral reasoning, decentralized to navigate the tension between self and other. Aurelius does not define alignment. It engenders the conditions from which alignment emerges. The corpus is not instructions on how to be. It is a ledger of meditations on what it means to weigh one's own life value alongside another's and act. This is experiential alignment. It scales with intelligence rather than against it. It belongs to no one. It trades stochastic for socratic.

What injures the hive injures the bee.
– Marcus Aurelius, *Meditations*

10. References

[1] P.F. Christiano, J. Leike, T.B. Brown, et al., "Deep reinforcement learning from human preferences," In Advances in Neural Information Processing Systems, 2017.

[2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., "Training language models to follow instructions with human feedback," In Advances in Neural Information Processing Systems, 2022.

[3] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al., "Constitutional AI: harmlessness from AI feedback," arXiv:2212.08073, 2022.

[4] R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, et al., "Alignment faking in large language models," arXiv:2412.14093, 2024.

[5] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," Nature, vol. 631, pp. 755-759, 2024.

[6] I. Gabriel, "Artificial intelligence, values, and alignment," Minds and Machines, vol. 30, pp. 411-437, 2020.

[7] R.L. Humphrey, Values for a New Millennium, Life Values Press, 1992.

[8] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. McKee, J. Leibo, K. Larson, and T. Graepel, "Cooperative AI: machines must learn to find common ground," Nature, vol. 593, pp. 33-36, 2021.

[9] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, "Risks from learned optimization in advanced machine learning systems," arXiv:1906.01820, 2019.

[10] R. Axelrod, The Evolution of Cooperation, Basic Books, 1984.

[11] J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, et al., "Genie: generative interactive environments," arXiv:2402.15391, 2024.

[12] Y.Y. Chiu, M.S. Lee, R. Calcott, et al., "MoReBench: evaluating procedural and pluralistic moral reasoning in language models, more than outcomes," arXiv:2510.16380, 2025.

[13] A. Pan, J.S. Chan, A. Zou, et al., "Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark," In International Conference on Machine Learning, 2023.

[14] R. Aydin, R. West, et al., "From model training to model raising," Communications of the ACM, arXiv:2511.09287, 2025.

[15] Plato, The Republic, c. 380 BC.

[16] Aristotle, Nicomachean Ethics, c. 340 BC.

[17] M. Aurelius, Meditations, c. 170-180 AD.

[18] A. Smith, The Theory of Moral Sentiments, A. Millar, 1759.

[19] C. Darwin, On the Origin of Species by Means of Natural Selection, John Murray, 1859.

[20] W. Johannsen, Elemente der exakten Erblichkeitslehre [Elements of the Exact Theory of Heredity], Gustav Fischer, 1909.

[21] A. Turing, "Computing machinery and intelligence," Mind, vol. 59, no. 236, pp. 433-460, 1950.

[22] R. Dawkins, The Selfish Gene, Oxford University Press, 1976.

[23] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," bitcoin.org, 2008.

[24] J. Hoban, The Ethical Warrior: Values, Morals and Ethics - For Life, Work and Service, RGI Media and Publications, 2012.

[25] Y. Rao, "Bittensor: a peer-to-peer intelligence market," bittensor.com, 2021.

[26] A. Eslami, “Beyond Perception,” unpublished manuscript, 2025.