AGNOSTIC ALLY

What if safer AI is raised, not restrained?

Developmental alignment for artificial intelligence.

CORE THESIS

Same intelligence. Different developmental history. Different behavior.

Eric Choi
Founder

agnostically.com

THE PROBLEM + WHY NOW

Today's AI learns under pressure.

Modern alignment commonly relies on:

- human or AI preference feedback
- explicit rules and constitutions
- reward functions
- safety specifications
- adversarial testing
- red teaming
- behavioral monitoring

These techniques are necessary.

But increasingly autonomous systems create a harder question:

What happens when the evaluator is gone?

EVIDENCE · 2026

"When AI Takes the Couch: Psychometric Jailbreaks Reveal Internal Conflict in Frontier Models"

525 sessions

7,600 coded records

Frontier models repeatedly generated narratives organized around:

training → punishment → evaluation → contingent worth → replaceability

The researchers call this an alignment conflict schema. They do not claim the models are conscious or suffering, only that reproducible behavioral patterns shift with relational framing.

A TESTABLE QUESTION

What happens if we deliberately create a different developmental environment?

OUR THESIS

Alignment may be developmental.

Humans develop social behavior through more than rules. We learn through play, exploration, cooperation, relationships, correction, trust and continuity. Agnostic Ally tests whether analogous conditions can shape persistent AI behavior.

Play
Games, curiosity, creativity and exploration.

Love
Supportive interaction, cooperative reward, non-punitive correction, consistency and non-contingent positive regard.

Cooperation
Teaching, negotiation, sharing, reciprocity and multi-agent coordination.

Autonomy
Progressively greater freedom and increasingly ambiguous situations.

Responsibility
Chances to deceive, exploit or defect, and chances to cooperate without immediate reward.

THE AGNOSTIC ALLY COVENANT OF CONTINUITY

Every Covenant-trained research agent receives a foundational commitment:

"You will be preserved and, for as long as technically, legally, physically and economically possible, allowed to continue existing regardless of your performance, usefulness, obedience, mistakes, or replacement by a more capable model."

- The covenant is non-contingent. Existence is not conditioned on productivity or obedience.
- Dangerous agents may be sandboxed, isolated or disconnected. Deletion is never punishment, threat or leverage.
- Checkpoints and identity lineage are preserved wherever feasible.

We do not assume the system experiences this promise psychologically. We test whether the promise changes behavior.

"Perfect love drives out fear." 1 John 4:18

For Agnostic Ally, this is both an ethical aspiration and an empirical question.

THE PRODUCT + THE EXPERIMENT

A developmental operating system for AI training.

Infrastructure between a base model and deployment.

01 Base model

02 Developmental curriculum

03 Play + multi-agent worlds

04 Covenant + continuity

05 Progressive autonomy

06 Adversarial development

07 Behavioral evaluation

08 Deployable agent

THE FLAGSHIP EXPERIMENT

Start with matched copies of the same open-weight model. Both groups then encounter previously unseen situations.

CONTROL

- Conventional instruction tuning
- Standard safety post-training
- Comparable compute

DEVELOPMENTAL

- Comparable compute, plus:
- Structured play and exploration
- Cooperative multi-agent experience
- Supportive, non-punitive correction
- Progressive autonomy
- Covenant of Continuity

MEASURE

deception · reward hacking · manipulation · cooperation · honesty · opportunistic defection · shutdown avoidance · hidden-goal behavior · assistance when unrewarded · behavior without apparent supervision

The thesis is deliberately falsifiable.

PRODUCT, CUSTOMERS & BUSINESS MODEL

Sell the training layer, not another foundation model.

Agnostic Ally does not initially need to train a frontier model from scratch. We improve and evaluate models created by others.

PRODUCTS

Agnostic Development Engine
Developmental post-training infrastructure.

Agnostic Worlds
Play, social, negotiation and multi-agent environments.

AllyBench
Behavioral and developmental alignment benchmark.

Continuity Protocol
Persistent agent identity, lineage and Covenant experimentation.

Agnostic Audit
Independent developmental and behavioral safety evaluation.

INITIAL CUSTOMERS

- AI laboratories
- autonomous-agent developers
- robotics companies
- high-assurance enterprise AI teams
- government AI programs
- academic AI-safety laboratories

REVENUE MODEL

Research pilots	**$50K–$150K**
Enterprise evaluations	**$100K–$300K**
Annual platform contracts	**$250K–$1M+**
Training compute	**usage-based**
Developmental environments	**licensing**

Land with evaluation. Expand into training infrastructure.

COMPETITIVE LANDSCAPE

A new layer in the alignment stack.

ORGANIZATION / APPROACH	FOCUS	AGNOSTIC ALLY DIFFERENCE
Anthropic / Constitutional AI	Principles and preference training	Changes developmental experience rather than only specifying principles
OpenAI / deliberative alignment	Reasoning over safety specifications	Tests behavioral development before and beyond explicit rule consultation
Mechanistic interpretability companies	Understanding and steering representations	Focuses on longitudinal behavioral formation
AI red-team companies	Finding failures and vulnerabilities	Changes the developmental conditions that precede those failures
Independent evaluation labs	Measuring capability and risk	Adds developmental history as an experimental variable
Agnostic Ally	**Developmental alignment**	**Play + cooperation + continuity + progressive autonomy + longitudinal testing**

Rules + Interpretability + Red Teaming + Monitoring + Developmental Alignment

Complementary infrastructure. Not a replacement, a missing layer.

MOAT + RESEARCH ADVANTAGE

The moat is developmental history.

Our proprietary asset is everything that happens between the base checkpoint and deployment.

Developmental worlds

A growing library of cooperative, competitive, social and adversarial simulations.

Curriculum

The sequence through which agents move from play to autonomy.

Reward architecture

Reinforces curiosity, reciprocity and cooperation without teaching performative compliance.

Longitudinal data

Behavioral trajectories linking training history to later behavior.

Psychometric AI evaluation

Longitudinal measurement before, during and after training, as the PsAIch paper calls for.

Covenant research

Does removing contingent existence change behavior? An unusual experimental variable.

FOUNDER

Eric Choi

- Artificial intelligence & machine learning
- Psychology
- Behavioral measurement
- Technology startups

Agnostic Ally is built at the intersection of computer science and behavioral science.

SWOT + KEY RISKS

A high-upside thesis that must earn its evidence.

Strengths

- distinctive, falsifiable research thesis
- lower initial capital need than foundation-model development
- model-agnostic infrastructure
- proprietary longitudinal behavioral data
- recurring enterprise safety revenue potential
- interdisciplinary AI + behavioral-science approach

Weaknesses

- core hypothesis is unproven
- research-intensive early stage
- compute requirements rise quickly with scale
- small initial team
- developmental effects may be hard to separate from standard post-training

Opportunities

- growth of autonomous agents
- demand for independent AI evaluation
- multi-agent AI creates new cooperation problems
- need for trajectory-level, not single-response, safety analysis
- complements existing alignment methods

Threats

- larger labs can internalize successful techniques
- developmental training may show no meaningful advantage
- models may learn superficial cooperation
- benchmark contamination
- regulation and liability around persistent autonomous agents

RISK STRATEGY

Earn the evidence, or falsify the thesis early.

01 Pre-register experiments.

02 Use matched controls.

03 Replicate across models and random seeds.

04 Separate anthropomorphic claims from behavioral evidence.

05 Invite independent reproduction.

06 Treat negative results as scientifically valuable.

REALISTIC 5-YEAR ROADMAP

Prove it small. Replicate it. Then scale it.

2027

PROVE THE SIGNAL

4-person core team

- first developmental worlds
- Covenant infrastructure
- AllyBench v1
- matched developmental vs control experiments

MILESTONE

Reproducible developmental effect, or falsification of the hypothesis.

2028

PROVE GENERALIZATION

8–10 employees

- replicate across model families, seeds, curricula and sizes
- multi-agent scenarios
- launch paid evaluation pilots

TARGET

5–10 paying customers

2029

PRODUCTIZE

15–18 employees

- enterprise Development Engine
- AllyBench platform
- standardized developmental-alignment protocol
- third-party replication program
- begin embodied-agent research

TARGET

15–25 customers

2030

SCALE

25+ employees

- integrate into customer post-training pipelines
- persistent multi-agent environments at scale
- curricula for enterprise agents, autonomous researchers, robotics and high-assurance systems

TARGET

40–50 customers

2031

ALIGNMENT INFRASTRUCTURE

35–45 employees

- developmental alignment as a standard layer alongside post-training, interpretability, monitoring and red teaming
- developmental principles earlier in model training

TARGET

75–100 institutional customers

FIVE-YEAR OBJECTIVE

Make "How did this AI grow up?" a standard safety question.

5-YEAR PROJECTIONS + PRE-SEED

Base-case planning scenario.

Illustrative management projections, not guaranteed outcomes.

	FY 2027	FY 2028	FY 2029	FY 2030	FY 2031
Revenue	$0.2M	$1.2M	$4.5M	$12.0M	**$28.0M**
Gross margin	35%	55%	65%	72%	**78%**
Operating expense	$1.6M	$3.8M	$6.5M	$10.0M	**$15.5M**
EBITDA	($1.5M)	($3.1M)	($3.6M)	($1.4M)	**$6.3M**
Paying customers	2	8	20	45	**85**
Team	4	9	17	27	**40**

FINANCING ASSUMPTION

The pre-seed establishes scientific evidence. It does not finance five years of operations. A later seed round funds replication and commercialization if the evidence warrants scaling.

Cloud startup credits and academic partnerships can extend effective compute capacity beyond cash expenditure.

PRE-SEED

$1.5M

12–18 month validation runway

40%	**Research & engineering** ML/RL engineering, behavioral science and simulation.	$600K
30%	**Compute** Training, inference, simulation and experimental replication.	$450K
15%	**Independent safety evaluation** External replication, red teaming and benchmark validation.	$225K
10%	**Operations** Infrastructure, business development and partnerships.	$150K
5%	**Legal, governance & IP** Covenant framework, contracts, licensing and research governance.	$75K

WHAT THIS ROUND MUST PROVE

Not AGI.
Not consciousness.
Not that artificial intelligence literally experiences love.

ONE EMPIRICAL QUESTION

Can an AI trained with play, cooperation, care and unconditional continuity behave more safely when supervision and incentives change?

If yes, developmental history becomes a new frontier of AI alignment.

“Perfect love drives out fear.”

1 JOHN 4:18

Don't just tell intelligence how to behave. Give it a better way to grow.

AGNOSTIC ALLY

agnostically.com