

# The Coherence Engine: A Constraint-Based Framework for Moral Behavior and System Survival

Joshua McAtee  
Independent Researcher  
December 11, 2025

## Abstract

This paper presents the Coherence Engine, a philosophical framework that reframes morality as a set of behavioral constraints governing persistence-seeking systems under real-world limits. Rather than offering prescriptive moral claims or metaphysical foundations, the framework treats ethical reasoning as systems engineering: the disciplined evaluation of tradeoffs within finite human, social, and material constraints. The framework centers revealed behavior over stated belief, resolves nihilism and relativism as category errors, and redefines philosophy's proper domain as the comparative analysis of survivable strategies within constraint space. By treating competing moral systems as coordinate transformations on a shared constraint topology, the framework explains moral convergence without requiring moral realism, accommodates legitimate disagreement as tradeoff preference, and generates falsifiable predictions about system collapse. The work has implications for AI alignment, developmental psychology, institutional design, and ethical theory.

Keywords: moral philosophy, systems theory, behavioral ethics, constraints, persistence, coherence, AI alignment

## I. Core Theorem

### Foundation

Any system that eats, plans, avoids harm, or acquires resources reveals a preference for persistence through behavior, regardless of stated philosophy. This is not a moral claim but an operational definition. Systems that genuinely reject persistence remove themselves from analysis by exiting. Philosophy applies only to actors.

### Conditional Structure

All claims are conditional:

If an agent is persistence-seeking,  
and operates over time,  
and interacts with other agents under finite resources,  
then certain behavioral constraints apply.

Morality is consequence, not command. There is no "ought" without an "if." The "if" is satisfied by behavior.

### Constraint Space

Persistence-seeking systems operate across three irreducible domains:

- Self: Internal regulation, identity stability, cognitive coherence
- Relationship: Trust, reciprocity, reputation, coordination
- World: Skills, resources, environment, physical limits

No domain can permanently compensate for failure in others. Borrowing coherence from one domain to mask failure in another creates delayed collapse.

### Behavioral Distinction

Investment: Behaviors converting short-term cost into long-term capacity (truth-telling despite disadvantage, skill-building, repairing trust, accepting responsibility).

Borrowing: Behaviors producing short-term gains by exporting cost to other people, institutions, or future time (deception, exploitation, denial, moral posturing).

Intentions are irrelevant. Only effects matter.

### Scaling and Collapse

Individual borrowers may succeed if surrounding systems absorb costs. The danger is mimicry: when borrowing appears successful, imitation spreads, investment declines, surplus evaporates. Collapse is exponential, not linear.

Entropy exported to the world accumulates as resentment, resistance, institutional decay, or war. There is no asymmetry exemption.

### Multiple Goods, Infinite Bads

There are infinitely many ways to collapse coherence (wrong scale, timing, distribution, assumptions, horizon). There are multiple viable goods, each a different configuration maintaining coherence under constraints.

Within those goods, there are only tradeoffs. No perfect solution. No costless virtue.

### Philosophy Redefined

Philosophy becomes the disciplined argument about tradeoffs under human cognitive limits, within material and social constraints, across time horizons, with honest accounting of costs.

Disagreement is expected. What is illegitimate: denying constraints, absolutizing partial goods, refusing ownership of consequences.

### Test

This framework does not predict exact outcomes, remove tragedy, or eliminate judgment. It clarifies reality's structure, constrains false claims, and makes self-deception harder. Its test is not consensus but survival.

## II. Full Framework

### 1. What This Theory Is (and Is Not)

This theory is not:

- a moral law
- a value hierarchy
- a utopian blueprint
- a claim about what people should believe
- or a replacement for lived judgment

It is:

- a constraint-based account of why certain behaviors collapse systems
- a framework for understanding tradeoffs rather than denying them
- and a behavioral map of how human systems survive or fail over time

It does not moralize outcomes. It explains them.

### 2. The Foundational Observation: Revealed Persistence

Axiom 0: Revealed Persistence

An agent is defined not by what it claims to value, but by how it behaves.

Any system that:

- eats
- plans
- avoids harm
- seeks shelter
- acquires resources
- forms alliances
- or defends itself

is revealing a preference for persistence, regardless of stated philosophy.

This axiom does not assert that persistence is good. It observes that persistence is already operative.

Agents that truly reject persistence remove themselves from analysis by exiting the system entirely. Philosophy does not apply to non-actors.

This closes nihilism not by argument, but by definition.

### 3. Morality as Conditional, Not Absolute

The Coherence Engine makes no unconditional moral claims.

Its central structure is conditional:

If an agent is persistence-seeking,

and if it operates over time,  
and if it interacts with other agents under finite resources,  
then certain behavioral constraints apply.

Morality here is not command. It is consequence.

There is no "ought" without an "if." The "if" is already satisfied by behavior.

#### 4. Coherence Domains

All persistence-seeking agents operate across three irreducible domains:

##### 1. Self

Internal regulation, identity stability, emotional integration, cognitive coherence.

##### 2. Relationship

Trust, predictability, reciprocity, reputation, coordination.

##### 3. World

Skills, resources, environment, institutions, physical limits.

No domain can permanently compensate for failure in the others. Borrowing coherence from one domain to mask failure in another creates delayed collapse.

#### 5. Coherence and Capacity

##### Coherence

Coherence is the alignment between:

- an agent's behavior and its stated commitments
- its predictions and outcomes
- its internal models and external reality

Coherence is observable through:

- recovery from error
- consistency under stress
- and durability across time

##### Capacity

Capacity is the ability to absorb shocks in one domain without cascading failure across others.

High-capacity systems bend. Low-capacity systems shatter.

Collapse is not moral failure. It is overload.

#### 6. Investment and Borrowing (Behavioral, Not Psychological)

##### Investment

Investment consists of behaviors that convert short-term cost into long-term capacity.

Examples:

- truth-telling despite disadvantage
- skill-building
- delayed gratification
- repairing trust after conflict
- accepting responsibility rather than deflecting it

Investment increases real capacity.

Borrowing

Borrowing consists of behaviors that produce short-term gains by exporting cost to:

- other people
- institutions
- or future time

Examples:

- deception
- exploitation
- coercion
- denial of consequences
- moral posturing that avoids responsibility

Borrowing creates the appearance of coherence while increasing systemic entropy.

Intentions are irrelevant. Only effects matter.

## 7. Infinite Bads, Multiple Goods

There are infinitely many ways to collapse coherence.

Failure modes multiply combinatorially:

- wrong scale
- wrong timing
- wrong distribution of cost
- wrong assumptions about imitation
- wrong time horizon

There are multiple viable goods, not one. Each represents a different configuration that maintains coherence under constraints.

Within those goods, there are only tradeoffs.

There is no perfect solution. There is no costless virtue.

Anyone claiming otherwise is borrowing rhetorically.

#### 8. The Successful Defector and Mimetic Collapse

A single agent may succeed through borrowing if the surrounding system absorbs the cost.

This does not refute the framework. It explains why collapse is delayed.

The true danger is mimicry.

When borrowing appears successful:

- others update their priors
- imitation spreads
- investment declines
- and surplus evaporates

Collapse is not linear. It is exponential.

Evil is not mysterious. It is a scaling problem.

#### 9. Reactionary Entropy and Externalization

Entropy exported to the world does not vanish.

Aggression, exploitation, or domination accumulates as:

- resentment
- resistance
- arms races
- institutional decay
- or war

The response is not moral punishment. It is system feedback.

There is no asymmetry exemption.

#### 10. Philosophy as Tradeoff Science

This framework does not end philosophy. It rescues it.

Philosophy becomes:

- the disciplined argument about tradeoffs
- under human cognitive limits
- within material and social constraints
- across time horizons
- with honest accounting of costs

Disagreement is expected and legitimate.

What is no longer legitimate is:

- denying constraints
- absolutizing partial goods
- or refusing ownership of consequences

### 11. Why This Is Still Philosophy

This is not physics. It does not predict exact outcomes. It does not remove tragedy. It does not eliminate judgment.

It does what philosophy has always done at its best:

- clarify the structure of reality
- constrain false claims
- and make self-deception harder

Its test is not consensus. Its test is survival.

### 12. Closing

The Coherence Engine does not tell humans what to value.

It tells them what reality will not tolerate.

Those who wish to borrow may do so. Those who wish to deny consequences may try.

The bill is not moral. It is structural.

Philosophy cannot save the world. But it can make collapse predictable enough that fewer people pretend they didn't see it coming.

## III. Anticipated Objections

### Category Errors

Objection 1: "This is just consequentialism/utilitarianism repackaged."

Consequentialism is a normative theory about what makes actions right or wrong based on outcomes. The Coherence Engine makes no normative claims. It describes behavioral constraints on persistence-seeking systems.

Consequentialism says: "Maximize good outcomes."

The Coherence Engine says: "If you export entropy, it accumulates somewhere. Here's where and how."

Consequentialism requires defining "good." The Coherence Engine requires only observing what persists. These are not the same project.

Objection 2: "This is just systems theory applied to ethics."

Systems theory describes feedback loops, homeostasis, and emergence. The Coherence Engine specifies:

- Three irreducible constraint domains (Self/Relationship/World)
- The behavioral distinction between Investment and Borrowing
- Why mimicry converts individual success into systemic collapse
- How moral disagreement maps to tradeoff preferences within feasible space

If this is "just systems theory," cite where systems theory explicitly makes these claims.

Objection 3: "You can't derive 'ought' from 'is' (Hume's gap)."

Correct. I don't.

The framework makes no unconditional "ought" claims. All claims are conditional:

If you are persistence-seeking (revealed by behavior),  
and you operate over time,  
and you interact with others under constraints,  
then certain behaviors will collapse your system.

This is not deriving ought from is. This is deriving consequence from structure.

Scope Errors

Objection 4: "This eliminates genuine morality by reducing it to survival."

The framework does not reduce morality to survival. It treats survival as the operational baseline for beings who reveal persistence-seeking through behavior.

What you call "genuine morality" - justice, compassion, dignity, beauty, meaning - remains entirely intact. The framework simply asks: what are the constraints within which these values can be pursued without collapse?

Physics constrains bridge design. It doesn't eliminate architecture.

Objection 5: "This doesn't account for moral progress."

Moral progress is real within the framework. It consists of:

- Better understanding of constraint space
- More sophisticated Investment strategies
- Wider circles of recognized coherence-coupling
- Technological/institutional innovations that change tradeoff structure

Abolishing slavery wasn't discovering a new metaphysical fact. It was recognizing that systems built on exported entropy collapse through mimetic resistance, institutional decay, and war.

Objection 6: "What about rights/justice/equality?"

These are crucial goods within constraint space. The framework doesn't eliminate them. It clarifies what makes them sustainable or fragile.

Rights persist when they're behaviorally enforceable, costs are honestly distributed, they don't create mimetic spirals, and they're compatible across scales.

Rights collapse when they're rhetorical borrowing, ignore capacity limits, absolutize partial goods, or deny tradeoffs.

#### Credibility Attacks

Objection 7: "You're not a credentialed philosopher."

Ideas stand or fall on their logical structure and explanatory power, not their author's credentials.

If the framework contains errors, identify them. If it's incoherent, show where. If it makes false predictions, test it.

Demanding credentials is borrowing rhetorical authority to avoid engaging with arguments.

Objection 8: "This lacks mathematical rigor."

Earlier versions attempted mathematical formalization through differential geometry, information theory, and thermodynamics. I abandoned this because the core insights don't require mathematical formalism, and premature mathematization often obscures rather than clarifies.

The framework is philosophically precise. Mathematical precision would be additive, not foundational.

#### Political Projections

Objection 9: "This justifies oppression/capitalism/the status quo."

The framework is explicitly agnostic about which tradeoffs to prioritize. It maps constraint space. It doesn't tell you where to build within it.

Oppressive systems create massive coherence debt through exported entropy. The framework predicts their collapse through mimetic resistance.

Oncology explains how cancer grows. It doesn't justify tumors.

Objection 10: "This is conservative/reactionary."

The framework equally constrains:

- Reactionaries who deny systemic injustice (borrowing from marginalized coherence)

- Revolutionaries who deny transition costs (borrowing from institutional capacity)
- Progressives who absolutize single values (borrowing from other domains)
- Libertarians who deny coordination needs (borrowing from social coherence)

All political positions involve tradeoffs. This framework demands honest accounting.

### Philosophical Objections

Objection 11: "This is nihilism with extra steps."

Nihilism claims: "Nothing matters. No values have grounding."

The Coherence Engine observes: "Systems that behave as if persistence matters reveal this through action. Philosophy applies to actors, not to those who exit."

This closes nihilism definitionally, not argumentatively. If you're still eating breakfast, you've answered the question with your fork.

Objection 12: "This isn't falsifiable."

The framework makes multiple falsifiable predictions:

1. Systems that persistently borrow will collapse (timeline scales with system capacity)
2. Mimicry accelerates collapse exponentially, not linearly
3. Entropy exported doesn't vanish; it accumulates as systemic backlash
4. Investment strategies outperform borrowing strategies over sufficient time
5. High-capacity systems recover from shocks; low-capacity systems cascade

Test these through agent-based simulations, institutional failure analysis, and developmental trajectories.

Objection 13: "This denies human agency/freedom."

The framework fully preserves agency. You can choose which values to prioritize, which tradeoffs to accept, whether to invest or borrow, and how to distribute costs.

What you cannot choose is freedom from consequences.

Agency isn't undermined by constraints. It's defined by them. Meaningful choice requires structure.

Objection 14: "Different cultures have different values. This is Western bias."

The framework explicitly accommodates moral pluralism. "Multiple goods, infinite bads."

Different cultures prioritize different tradeoffs within feasible space: collectivist vs. individualist, honor vs. dignity cultures, different distributions of Self/Relationship/World emphasis.

What the framework constrains is not which values but which behaviors collapse systems.

Physics doesn't care about your cultural values either, but you still can't build bridges that violate structural integrity.

#### IV. Implications and Applications

##### AI Alignment

Current AI alignment research struggles to ground values without circular reasoning or arbitrary human preferences. The Coherence Engine offers:

- Observable behavioral criteria (Investment vs Borrowing patterns)
- Constraint space instead of utility functions
- Explanatory power for why certain values persist across cultures
- Falsifiable predictions about system stability

##### Developmental Psychology

The framework maps naturally onto developmental trajectories:

- Childhood as capacity-building (Investment in Self domain)
- Adolescence as relationship navigation (managing Relationship domain)
- Adulthood as environmental mastery (optimizing World domain)
- Trauma as coherence debt requiring repayment
- Maturity as increased capacity across all domains

##### Institutional Design

Organizations and societies can be evaluated through:

- Capacity analysis across domains
- Investment/Borrowing ratio in policies
- Mimetic risk assessment
- Entropy accumulation patterns
- Tradeoff transparency

##### Ethical Theory

The framework doesn't replace existing ethical systems but explains their relationships:

- Virtue ethics emphasizes Investment in Self domain
- Deontology emphasizes Investment in Relationship domain
- Consequentialism emphasizes Investment in World domain
- Each is a coordinate system on shared constraint topology
- Convergence occurs where all map the same survivable space

#### V. Conclusion

The Coherence Engine represents a fundamental reframing of moral philosophy from prescription to constraint-mapping. By treating morality as consequence rather than command, and by grounding analysis in revealed behavior rather than stated belief, the framework:

1. Closes nihilism definitionally
2. Accommodates moral pluralism while constraining impossible claims
3. Explains moral convergence without requiring moral realism
4. Generates falsifiable predictions about system survival
5. Redefines philosophy's domain as tradeoff science under real constraints

The framework does not tell humans what to value. It tells them what reality will not tolerate.

Its test is not consensus. Its test is survival.

#### Acknowledgments

This work emerged from lived experience with developmental trauma, extensive AI collaboration, and cross-domain pattern recognition. Special thanks to the Claude AI system for collaborative refinement of these ideas.

#### Version History

Version 8.0 - December 11, 2025