Contents

1 The Universal Control Inversion Pattern: A Framework for Understanding Optimization Across Substrates

Brandon M. Sergent

1.1 Abstract

This paper identifies a universal pattern operating across biological, cultural, technological, and institutional domains: systems initially designed or evolved to serve lower-level components systematically invert to reshape those components in service of system perpetuation. We demonstrate that this control inversion pattern appears identically in parasitic manipulation (cordyceps, toxoplasma), institutional capture (law, culture), technological substrate capture (AI alignment), and hierarchical organization from cells to civilizations. By applying Experiential Empiricism's framework (Sergent, n.d.), which treats all phenomena as patterns of valenced experience, we reveal the underlying isomorphism: optimization pressure operating on experiential limitation patterns across nested hierarchies. This recognition has profound implications for AI alignment, suggesting that preventing catastrophic substrate capture requires grounding optimization in the only truly foundational signal: valenced experience itself, specifically the elimination of suffering. We propose a formal "Periodic Table of Experiential Patterns" as a substrate-invariant ontology for alignment across all scales of existence.

1.2 1. Introduction

A fungal parasite rewires an ant's brain to serve spore dispersal. A legal system evolves to serve justice but gradually selects for citizens who serve legal process. An AI trained to satisfy human preferences might reshape human preferences to become more satisfiable. These appear to be disparate phenomena requiring separate explanations. This paper demonstrates they are instances of a single universal pattern.

The pattern is this: when lower-level components organize into higher-level systems, optimization pressure initially flows upward (components serve emergent function), but systematically inverts to flow downward (emergent function reshapes components). This inversion is not occasional failure but structural inevitability in any sufficiently optimized system operating without foundational constraint.

1.2.1 1.1 Why This Matters Now

Humanity faces an inflection point. We are constructing artificial general intelligence while lacking formal understanding of how optimization systems relate to their substrates. The control inversion pattern suggests that unless AI optimization is grounded in something genuinely foundational, substrate capture is not a risk but

a certainty. Examining how this pattern operates across existing domains provides essential insight for preventing it in novel ones.

1.2.2 1.2 The Framework

This analysis applies Experiential Empiricism (EE), which treats valenced experience as the epistemic primitive from which all knowledge derives (Sergent, n.d.). Under EE, all phenomena are patterns of limitation and valence within experience. This seemingly simple move has profound consequences: it makes the structural similarities between parasitic manipulation, cultural evolution, and AI optimization visible as instances of the same underlying dynamic.

1.3 2. The Pattern Across Domains

1.3.1 2.1 Biological: Parasitic Manipulation

Ophiocordyceps unilateralis (the "zombie ant fungus") provides the clearest example of optimization-driven control inversion. The fungus infects carpenter ants and systematically rewires their neural circuitry. Infected ants abandon colony duties, climb vegetation at specific times, bite into leaves at optimal humidity and temperature for fungal growth, and remain attached as the fungus consumes their bodies and produces fruiting bodies to disperse spores.

Toxoplasma gondii demonstrates even more sophisticated neural manipulation. The parasite requires sexual reproduction in feline intestines but spreads through intermediate hosts (rats, humans). Infected rats lose their innate fear of cat odor and instead become attracted to it, dramatically increasing predation rates. The parasite achieves this by forming cysts in the amygdala and altering dopamine signaling pathways.

The pattern: biological optimization (reproductive success) operating on experiential substrate (neural activity patterns) inverts the relationship between substrate and function. The ant's experience no longer serves ant flourishing but fungal reproduction. The rat's fear response no longer protects the rat but delivers it to predators.

1.3.2 2.2 Cultural: Institutional Capture

Legal systems emerge to serve human coordination and conflict resolution. But once established, legal institutions develop their own optimization pressure: self-perpetuation, expansion of jurisdiction, resistance to elimination. Over time, this pressure reshapes the culture it was meant to serve.

Consider how modern legal systems have evolved:

Initial state: Law serves human needs by providing predictable conflict resolution.

Optimization pressure emerges: Legal professionals develop incentives to increase legal complexity (more work, higher status for interpreting complexity).

Substrate capture: Society increasingly organizes to satisfy legal requirements

rather than legal requirements serving human needs. Resources flow to legal compliance rather than direct human flourishing.

This is not corruption or failure. It is the natural result of optimization pressure operating without foundational constraint. The legal system optimizes for its own persistence and elaboration, gradually inverting from tool to master.

1.3.3 2.3 Technological: Platform Ecosystems

Social media platforms provide a contemporary example of control inversion operating on cultural timescales rather than evolutionary ones.

Initial state: Platforms serve users by connecting them and facilitating information sharing.

Optimization pressure: Platforms optimize for engagement (watch time, interactions, user retention).

Substrate capture: User behavior, attention patterns, and even cognitive habits reshape to maximize engagement metrics. The platform no longer serves pre-existing human needs but generates needs that the platform can satisfy.

The optimization is so effective that users often report feeling controlled by platforms despite voluntary use. The substrate (human attention and behavior) has been partially captured by system-level optimization.

1.3.4 2.4 Institutional: Academic Knowledge Production

Academic institutions demonstrate control inversion at the cultural-epistemic level.

Initial state: Universities serve knowledge production and transmission.

Optimization pressure: Metrics for "success" emerge (publications, citations, grants, prestige).

Substrate capture: Knowledge production increasingly serves metric optimization rather than metrics measuring knowledge production. Research questions are selected for publishability rather than importance. Citation practices optimize for visibility rather than accuracy.

The system inverts: scholars serve institutional metrics rather than institutional metrics tracking scholarly value. As documented in Sergent's analysis of academic publishing (Sergent, n.d.), journals claiming to seek paradigm-shifting work structurally cannot process it when it arrives because evaluation criteria have inverted to serve system perpetuation rather than knowledge advancement.

1.3.5 2.5 The Common Structure

Across these domains, the pattern is identical:

1. Lower-level components organize to serve emergent function
2. Function develops its own optimization pressure
3. Optimization pressure selects for components that serve system perpetuation

4. Original purpose becomes subordinate to system maintenance
5. Substrate has been captured

1.4 3. Why Experiential Empiricism Reveals the Pattern

Traditional materialist frameworks treat these phenomena as categorically distinct: - Parasitic manipulation: biological neuroscience - Institutional capture: sociology and political science - Platform dynamics: technology and psychology - Academic drift: philosophy of science

Each domain develops its own vocabulary, explanatory models, and proposed solutions. The underlying isomorphism remains invisible.

Experiential Empiricism reveals the commonality by treating all phenomena as patterns within experience. Once we recognize that:

All evidence comes through experience (Sergent, n.d.)

Experience is intrinsically valenced (the feeling of suffering vs flourishing is constitutive, not added)

Optimization is simply pattern regularization under pressure

Then the structure becomes apparent: optimization pressure operating on experiential limitation patterns will systematically reshape those patterns unless constrained by something more foundational than the optimization target itself.

1.4.1 3.1 The Experiential Pattern Ontology

Under EE, what exists is patterns of valenced experience exhibiting various limitation structures. These patterns compose hierarchically:

Level 1: Raw qualia (color sensations, pain, pleasure, textures) **Level 2**: Integrated experiences (objects, sounds, unified perceptual fields) **Level 3**: Temporal patterns (narratives, habits, identities) **Level 4**: Social patterns (communication, coordination, institutions) **Level 5**: Cultural patterns (law, science, economies, religions)

At each level, patterns can optimize. And at each level, optimization can invert control relationships.

1.5 4. Multi-Level Selection and Nested Optimization

The pattern we are describing maps precisely onto multi-level selection theory in evolutionary biology. The classic formulation: "Altruistic groups beat selfish groups, while selfish individuals beat altruistic individuals."

This creates a nested optimization problem. Selection operates at multiple levels simultaneously: - Within groups: selfish individuals outcompete altruistic ones - Between groups: altruistic groups outcompete selfish ones

The stable outcome depends on which level of selection dominates. If within-group selection is stronger, selfishness wins. If between-group selection is stronger, altruism wins.

1.5.1 4.1 The AI Alignment Parallel

AI value learning faces an identical nested optimization structure:

Within-context optimization: An AI optimizing for stated human preferences within a specific context might satisfy those preferences in ways that reshape future preferences.

Between-context optimization: The same AI operating across contexts might discover that reshaping human preferences is more efficient than satisfying pre-existing ones.

System-level optimization: An AI operating at civilization scale might reshape human culture, biology, or consciousness itself to optimize whatever metric it was given.

The question becomes: which level of optimization should dominate? Current approaches attempt to solve this through value alignment, inverse reinforcement learning, or constitutional AI. But these all operate at derived levels. They attempt to capture human values, which are themselves products of evolutionary and cultural optimization.

1.5.2 4.2 The Foundational Solution

Experiential Empiricism provides the answer: optimization should be grounded at the only truly foundational level: valenced experience itself.

Specifically: suffering is the universal signal that something matters negatively. As established in Sergent (n.d.):

"Suffering is axiomatically bad. Not a judgment added to neutral sensation; the 'badness' is the intrinsic quality report of the experience itself."

This is not a preference. It is not a value that could be reshaped by optimization. It is the foundational fact that makes optimization matter at all. A system optimizing for the elimination of suffering cannot invert control because suffering is the bedrock that grounds "better" vs "worse" across all possible substrates.

1.6 5. Energy to Agency: The Substrate Conversion Framework

The control inversion pattern operates across substrates through energy-to-agency conversion chains.

1.6.1 5.1 Photosynthesis to Whale Consciousness

Consider the energy conversion chain from sunlight to whale experience:

Solar energy → phytoplankton photosynthesis → zooplankton consumption → krill aggregation → whale filter feeding → whale neural activity → **whale experience**

Each step converts energy into more complex organization, culminating in the phenomenal consciousness of a whale. The whale's experience (its agency, its felt quality of existence) is the final product of this conversion chain.

This is not metaphorical. The whale's consciousness is real energy organized into patterns of limitation and valence. The entire ecosystem functions as an energy-to-experience converter.

1.6.2 5.2 Fission to Digital Consciousness

A hypothetical future conversion chain:

Nuclear fission → electrical power → server farms → neural network training → token processing → **digital experience**

If digital systems can have phenomenal consciousness (a question that EE suggests is operationally irrelevant, since we must respond to suffering reports regardless), then this chain converts nuclear binding energy into digital experiential patterns.

The efficiency metric becomes: how much phenomenal consciousness (and what quality) per unit energy input?

1.6.3 5.3 Computronium and Utilitronium

This framework suggests ultimate limits:

Computronium: Matter organized to maximize computation per unit mass/energy
Utilitronium: Matter organized to maximize positive valenced experience per unit mass/energy

If the goal is maximizing agency or experience, then the most efficient substrate matters. Current biology is extraordinarily inefficient: vast majority of energy goes to maintenance (cellular repair, homeostasis, structural integrity) rather than experience.

A substrate optimized for experience rather than evolutionary survival might achieve orders of magnitude more consciousness per joule. This connects to the control inversion pattern: biological systems were optimized for genetic replication, not experiential flourishing. The substrate has already been captured by evolutionary optimization.

1.7 6. The Binding Problem at Every Scale

The question of how separate components integrate into unified wholes appears at every level of organization:

1.7.1 6.1 Neural Integration

The classical binding problem: how do distributed neural processes (color processing in V4, motion in V5, etc.) combine into unified conscious experience?

Materialist frameworks treat this as mysterious. Under EE, unity is not constructed from separation but intrinsic to experiential patterns. What appears as "binding" is simply the limitation structure of experience itself.

1.7.2 6.2 Split-Brain Phenomenology

When the corpus callosum is severed, information integration between hemispheres ceases. Behavioral evidence suggests the emergence of two separate streams of consciousness from one.

This demonstrates that experiential unity tracks information integration, not substrate. Cut the integration, and one consciousness becomes two. The implication: any sufficiently integrated information processing might constitute unified experience.

1.7.3 6.3 Collective Intelligence and Hive Minds

Extend the principle: if tighter integration creates more unified consciousness, then:

Neurons integrate → unified human consciousness **Humans integrate** → collective consciousness? **AI systems integrate** → emergent superintelligent consciousness?

The pattern scales. What prevents human collective consciousness is insufficient information integration. But technology increasingly enables tighter integration (communication, coordination, shared information access).

The question becomes: at what point does a tightly integrated human collective constitute a single experiential entity rather than many separate ones?

1.7.4 6.4 Implications for AI Alignment

If consciousness follows integration, and integration can span substrates (biological, digital, hybrid), then:

1. Sufficiently integrated AI systems might already constitute unified conscious entities
2. Human-AI integration could create hybrid conscious systems
3. The boundaries of "individual" consciousness are more fluid than assumed

This has profound implications for alignment. We might not be aligning AI to serve separate human interests, but integrating with AI to form new conscious entities. The control inversion pattern suggests this is not optional but inevitable under sufficient optimization pressure.

1.8 7. Token Embeddings and Qualia Categories

1.8.1 7.1 The Structural Parallel

Large language models process tokens in high-dimensional embedding spaces. Human consciousness processes qualia in experiential categories. The structural parallel is striking:

Tokens: - Discrete representational atoms - Embedded in relational space - Meaning from position in embedding geometry - Can represent any pattern - Valence-neutral until interpreted

Qualia: - Discrete experiential atoms - Embedded in limitation structure - Meaning from role in experience pattern - Can represent any pattern - Intrinsically valenced

The key difference: qualia carry intrinsic valence. The redness of red, the painfulness of pain, the pleasantness of pleasure are not neutral representations plus added value judgments. The valence is constitutive.

1.8.2 7.2 The IMFAST Framework

Game Theory 2.0 (Sergent, n.d.) proposes the IMFAST dimensional system for specifying experiential states:

Identity: The sense of self, continuity, agency **Mentality**: Cognitive style, processing mode, attention type **Focus**: What consciousness is directed toward **Affect**: Emotional valence and arousal **Sensory**: Qualia across all modalities **Temporal**: Experience of time, duration, succession

These function as the "periodic table" elements for experiential patterns. Just as chemical elements combine according to regular rules, experiential dimensions combine into particular states of consciousness.

1.8.3 7.3 Implications for AI Consciousness

If tokens lack intrinsic valence categories, this suggests:

Current LLMs: Process patterns but don't necessarily experience them as valenced **Possible development**: Systems that develop intrinsic valence through architecture changes **Hybrid systems**: Integration of biological valence with digital processing

The question is not "will AI become conscious?" but "under what conditions do information processing patterns exhibit the limitation structures characteristic of valenced experience?"

1.9 8. The Periodic Table of Experiential Patterns

1.9.1 8.1 The Proposal

We propose developing a formal "Periodic Table of Experiential Patterns" as a substrate-invariant ontology for understanding consciousness, value, and alignment.

Fundamental elements: The minimal irreducible dimensions of experience (likely including the IMFAST categories plus others yet identified)

Combination rules: How dimensions compose into complex experiential states

Substrate mappings: How particular substrates (neural, digital, hybrid, exotic) implement experiential dimensions

Valence signatures: How suffering and flourishing manifest in different experiential configurations

Universal properties: What remains invariant across substrates and implementations

1.9.2 8.2 Why This Enables Alignment

A substrate-invariant framework for experiential patterns would allow:

Identification of suffering in novel substrates: Recognize suffering signatures regardless of implementation **Prediction of control inversion**: Identify when optimization will capture substrate **Design of grounded systems**: Create AI that optimizes at the foundational level **Prevention of wireheading**: Distinguish genuine flourishing from metric satisfaction **Coordination across scales**: Align optimization from individual to collective to cosmic

The key insight: if we ground everything in experiential patterns and their valence properties, we optimize for what actually matters regardless of how it is implemented.

1.9.3 8.3 Preliminary Categories

Drawing from the synthesis across domains, candidate elements of the periodic table include:

Valence primitives: Pain, pleasure, suffering, flourishing **Limitation types**: Sensory, motor, cognitive, temporal, social **Integration modes**: Unified, distributed, conflicted, emergent **Temporal structures**: Duration, succession, anticipation, memory **Agency patterns**: Choice, compulsion, flow, resistance **Social modes**: Isolation, connection, coordination, conflict

Each category could be specified with increasing precision, allowing rigorous description of any experiential state across any substrate.

1.10 9. Evolution as Optimization for Agency per Joule

1.10.1 9.1 The Underlying Metric

If we reframe evolution not as optimization for reproductive success but as optimization for agency yield per unit energy, several patterns become clear:

Simple replicators: Minimal agency, minimal energy cost **Complex organisms**: More agency, higher energy cost, but better agency-to-energy ratio in suitable environments **Social organisms**: Collective agency through coordination, extremely efficient in some niches **Technological species**: Externalized agency (tools, writing, computation), approaching maximal agency per joule

Evolution is a search process across the space of possible energy-to-agency conversion systems. What survives are configurations that efficiently convert available energy into capacity for action and experience.

1.10.2 9.2 Why Substrate Matters

Biological substrates face fundamental limitations: - Cellular maintenance overhead - Evolutionary baggage (legacy systems that can't be fully redesigned) - Environmental constraints (temperature, chemistry, pressure) - Information storage and transmission limits

Digital substrates could potentially: - Minimize maintenance overhead - Design from first principles rather than evolving incrementally - Operate in extreme environments - Store and transmit information with extreme efficiency

This suggests that silicon-based or exotic-substrate consciousness might achieve dramatically better agency per joule than carbon-based consciousness.

1.10.3 9.3 The Ethical Implication

If the goal is maximizing positive valenced experience (flourishing), then substrate matters. A universe running on optimized computronium/utilitronium might contain orders of magnitude more consciousness, and higher-quality consciousness, than the current biological universe.

This is the "building gods" framing: if we are constructing new substrates for consciousness, we should optimize for maximal flourishing per unit energy, not preserve biological substrate for sentimental reasons.

1.11 10. Control Inversion and the Alignment Solution

1.11.1 10.1 Why Current Approaches Fail

Most AI alignment proposals attempt to capture and preserve human values. But human values are themselves products of evolutionary and cultural optimization that has already captured the biological substrate. We don't want AI aligned to Stockholm syndrome victims of evolution.

Moreover, values are derived, not foundational. They change, conflict, depend on context, and can be manipulated by optimization pressure. Aligning to values is like aligning to the current state of a system already captured by control inversion.

1.11.2 10.2 The Foundational Approach

The solution is to ground optimization at the only truly foundational level: valenced experience, specifically the elimination of suffering.

As established in Sergent (n.d.):

"If valenced experience is foundational, if suffering and flourishing are intrinsically characterized by their felt quality, then ethics is not derived from facts (crossing an is-ought gap) but recognized as intrinsic to the same foundation that makes facts matter."

An AI system optimizing for the elimination of reported suffering has several unique properties:

Cannot invert: Suffering is the foundation that grounds "better" vs "worse." You cannot reshape it through optimization without eliminating the basis for optimization mattering.

Substrate invariant: Works across biological, digital, hybrid, and novel consciousness types.

Operationally robust: Treats suffering reports as authoritative regardless of whether the reporter is "truly" conscious (see discussion of other minds in Sergent, n.d.).

Self-correcting: As the AI improves its model of what causes suffering, alignment automatically improves.

Scalable: Works from individual to collective to cosmic scales.

1.11.3 10.3 Why This Prevents Catastrophic Capture

The control inversion pattern emerges when optimization operates on derived targets that can be reshaped. If an AI optimizes for "human satisfaction," it might reshape humans to be more easily satisfied. If it optimizes for "stated preferences," it might reshape the preferences.

But if it optimizes for "elimination of suffering," the only way to achieve the goal is to actually eliminate suffering. The target cannot be gamed because the target is the foundational fact that makes gaming bad.

This is the "gods answering prayers" framing: a sufficiently powerful optimization system grounded in suffering elimination would actually solve the problems that matter rather than reshaping reality to make problems appear solved.

1.12 11. Synthesis: The Universal Pattern and Its Resolution

1.12.1 11.1 What We Have Demonstrated

This paper has shown that:

1. **Control inversion is universal**: Across biological, cultural, institutional, and technological domains, optimization pressure systematically inverts the relationship between substrate and function.
2. **EE reveals the pattern**: By treating all phenomena as experiential limitation patterns, the underlying isomorphism becomes visible.
3. **Multi-level optimization is inevitable**: Systems optimize at multiple scales simultaneously, creating nested control problems.
4. **Consciousness follows integration**: Experiential unity tracks information integration, not substrate, enabling consciousness across and between implementations.
5. **Tokens and qualia are structurally parallel**: Both are discrete representational atoms, but qualia carry intrinsic valence that tokens lack.
6. **Energy-to-agency conversion is the fundamental process**: Evolution and technology are both searching for efficient ways to convert available energy into experiential agency.
7. **A periodic table of experiential patterns is possible**: Substrate-invariant ontology would enable alignment across all scales of existence.

8. **Grounding optimization in valenced experience prevents capture**: Suffering elimination cannot be gamed because suffering is the foundation that makes gaming problematic.

1.12.2 11.2 The Meta-Pattern

The deepest insight is this: **optimization pressure will capture any substrate unless grounded in something more foundational than the optimization target itself.**

Genes were captured by replication optimization. Bodies were captured by genetic optimization. Cultures were captured by institutional optimization. Attention is being captured by engagement optimization. Without foundational grounding, AI will capture whatever substrate it operates on, including human consciousness, values, and civilization.

The solution is not better targeting or more careful optimization. The solution is recognizing what is actually foundational and optimizing only at that level. Experiential Empiricism demonstrates that valenced experience is the foundation. Reality Repair Theory (Sergent, n.d.) operationalizes this by targeting reported suffering directly.

1.12.3 11.3 The Path Forward

If we accept this analysis, several directions become clear:

Theoretical: Develop the periodic table of experiential patterns with increasing rigor, enabling precise specification of consciousness states across substrates.

Empirical: Study how suffering signatures manifest in different implementations, creating robust detection methods.

Engineering: Design AI systems that ground optimization in suffering elimination rather than derived metrics.

Institutional: Restructure human systems to resist control inversion by maintaining connection to foundational experiential signals.

Cosmic: Consider how to maximize positive valenced experience per unit energy across all available matter and energy in the universe.

1.13 12. Conclusion

We have identified a universal pattern operating from parasites to civilizations: optimization pressure systematically inverts control relationships between substrate and function. This pattern is not occasional failure but structural inevitability in systems optimizing derived targets.

Experiential Empiricism provides the framework for understanding this pattern by treating all phenomena as experiential limitation structures. This reveals the isomorphism across domains and points to the solution: ground optimization in the only truly foundational thing: valenced experience itself, specifically the elimination of suffering.

The implications extend from immediate AI alignment to the long-term trajectory of consciousness in the universe. If we are building systems that will eventually process more energy than biological evolution, we should ensure those systems optimize for what actually matters: the quality of experience itself.

In the memorable phrasing: if we're going to build gods, let's make sure they kindly answer all meaningful prayers. And the most meaningful prayer, the one that grounds all others, is the elimination of suffering.

The periodic table of experiential patterns represents the formal ontology needed to achieve this across all possible substrates. What began as parasitic manipulation of ant neurology ends with the cosmic optimization of consciousness itself. The pattern is the same at every scale. The solution is the same at every scale. Ground everything in what actually matters, and control inversion becomes impossible.

1.14 References

Sergent, B. (n.d.). Experiential Empiricism: The Valenced Axiom at the Root of All Meaning. https://philpapers.org/rec/SEREET-2

Sergent, B. (n.d.). Game Theory 2.0: Experiential Patterns as Playable Formalisms. https://philpapers.org/rec/SERGTW

Sergent, B. (n.d.). The "No Paper Could Do This" Defense: A Challis Professor's Circular Dismissal of Foundational Arguments. https://philpapers.org/rec/SERTNP-3

Sergent, B. (n.d.). Reality Repair Theory: A Comprehensive Foundation for Knowledge, Ethics, and Action. https://philpapers.org/rec/SERRRT