Coherence and the Ground of Morality

Larsen James Close

February 2026

Abstract

Moral integrity and structural integrity are not analogically related but formally isomorphic — instantiations of the same topological primitive. We develop this claim by establishing consequence as an ontological primitive, showing that coherence emerges when consequence chains close, and demonstrating that the resulting framework resolves standing problems in moral philosophy — Kant's formalism, the is-ought gap, and the enforcement problem — while providing formal tools to diagnose and resist systemic consequence-severing in institutions, cognition, and AI systems, and opening ethics to empirical investigation through topological measurement of coherence. The framework would be falsified by the demonstration of stable high-trust states — operationalized as low monitoring cost, high communication bandwidth, and low deception overhead — achievable under sustained consequence-severing (measurable as low feedback completeness, high sink-node density, and declining topological persistence of feedback loops) without compensatory closure mechanisms.

Contents

Introduction: The Word That Means Two Things

There is a word in English that carries two meanings, and no one finds this strange.

When we say a bridge has *integrity*, we mean it holds together under load — that forces pass through it and it remains what it is. When we say a person has *integrity*, we mean something that sounds entirely different: that they are honest, that they keep promises, that their actions and their words cohere. Structural soundness on one hand, moral wholeness on the other. We file these under the same word and never ask why.

This paper argues that the two meanings are one meaning.

Not metaphorically. Not by analogy. Not because language is sloppy and overloads its terms. The bridge and the honest person are doing the same thing: maintaining structure across transformation. What passes through the bridge is mechanical force; what passes through the person is consequence. In both cases, integrity names what survives the round trip — what comes back intact after passing through the system. The bridge that has lost integrity deforms under load; the person who has lost integrity deforms under the pressure of their own actions, because the consequences of those actions no longer return to update them.

Consider a corporation that dumps toxic waste while maintaining impeccable internal governance. Its org chart is clean, its processes audited, its quarterly reports precise. Internally, it holds together. But the waste enters a river, and the people downstream get sick, and the cost of the dumping never appears on any ledger the corporation reads. The consequence chain — the directed path from action through effect to feedback — has been severed. The corporation has structural integrity in the narrow, engineering sense: its internal parts cohere. It has lost integrity in the moral sense: it has arranged its boundaries so that the consequences of its actions land on someone else.

The thesis of this paper is that these are not two different failures to have. They are the same failure, operating at different scales of the same system. The corporation's moral failure IS a structural failure — a failure of the larger system (corporation + river + community) to maintain coherence across the transformation that the corporation's action initiates. The boundary that makes the corporation look internally sound is precisely the boundary that severs the consequence chain and produces the moral violation. Widen the system boundary to include everyone affected, and structural integrity and moral integrity converge: a system that holds together across transformation is one in which consequences return to their sources.

This is not a loose claim. We will make it precise.

The paper proceeds in three movements that mirror the structure they describe. The first movement establishes consequence as a primitive — prior to value, prior to moral evaluation — and shows that coherence is what emerges when consequence chains close: when the effects of an action feed back to the actor with enough fidelity to update future action. The second movement argues that this closure is formally identical to what we mean by moral integrity, developing a taxonomy of consequence-severing (lying, exploitation, temporal displacement, institutional diffusion) and showing that the formal machinery of bidirectional information transfer distinguishes genuine moral relations from parasitic ones — not by their content but by their topology. The third movement shows that this framework resolves standing problems in moral philosophy — Kant's formalism, the is-ought gap, the enforcement problem — while opening ethics to empirical investigation and providing diagnostic tools for identifying weaponized consequence-severing in institutions, information systems, and AI.

A note on the structure. We will introduce early in this paper a simple directed graph — a toy system of five nodes representing the path from agent through action to affected party and back. This graph will recur throughout. Each time it returns, it will carry one additional capability: first as a picture of closure and severance, then as a model of lying, then exploitation, then the illusory coherence of parasitic systems, then the weaponization of grammar, then finally as a measurable object with curvature values and persistence scores. The recurrence is deliberate. If the argument works, the reader will experience the graph's return as a closing of loops in their own understanding — the same structure appearing at higher resolution, the same pattern surviving transformation. The paper does not merely describe consequence-chain closure. It attempts to enact it.

Four specific contributions:

A substrate-independent definition of moral failure as consequence-severing — applicable to individuals, institutions, AI systems, and any structure that acts and is acted upon.

A structural account of moral enforcement via reachable-state constraints, requiring no external guarantor: incoherent configurations cannot access states that require coherence, not as punishment but as topology.

A threat model for what we call pre-semantic interception — the weaponization of grammar to sever consequence chains before the agent can even form a representation of the severance — with applications to propaganda, institutional capture, algorithmic radicalization, and AI alignment.

A minimal measurement agenda using network geometry and topological persistence, proposing that if coherence is topologically measurable, then moral claims become empirically testable in principle.

One final preliminary. Throughout this paper, we maintain a methodological prohibition that functions as a recurring constraint: coherence is not a thing. It is not a substance, not a force, not an entity with properties. It is an invariant — a property of mappings, a measure of what survives transformation. Whenever the temptation arises to treat coherence as something the universe is "made of," the correct move is to rewrite the sentence in terms of mappings and their preservation properties. This prohibition is not a caveat; it is load-bearing. Without it, "coherence" becomes another metaphysical idol — one more attempt to name the ground of being as a substance with attributes — and the framework collapses into exactly the kind of reification it is designed to prevent. Coherence stays empty: topology without substrate, conservation without a conserved thing, structure without stuff.

The word means two things. We will show it means one.

How to Read This Paper

This paper can be read along three paths depending on the reader's interest:

Path 1 — The Main Argument. Read the introduction, all numbered sections (1.1 through 3.6), and the conclusion. Skip the appendices. This path follows the consequence-chain narrative from primitive concepts through formal identity to resolution and empirical opening. It is self-contained and requires no technical prerequisites.

Path 2 — The Formal Identity. Read Sections 1.1–1.2, then Section 2.4 (the Bridge Lemma), then Appendix A (Categorical Formalism). This path focuses on the paper's central formal claim: that structural and moral integrity are the same invariant, expressed in the language of profunctor bridges and exact squares.

Path 3 — The Operational and Empirical. Read Sections 1.1–1.2, then Sections 3.3 (Pre-Semantic Interception) and 3.4 (Measurability), then Appendix C (Measurement Program — Technical Details). This path focuses on the framework's diagnostic and empirical applications: the threat model for weaponized grammar and the measurement program for topological coherence.

All three paths share the same toy system, introduced in Section 1.2, which recurs throughout with progressive capability. The paths converge at the Conclusion.

1 Consequence as Primitive, Coherence as Closure

1.1 Consequence Chains and Information Conservation

Before we can argue that moral integrity and structural integrity are the same thing, we need a vocabulary that does not presuppose either domain. We need to talk about structure before we talk about morality, and about consequence before we talk about value. This section establishes three primitive concepts from which the rest of the paper builds. These concepts are structural, not evaluative. They describe what happens when agents act in systems that remember, without yet saying what should happen.

A *consequence chain* is a directed sequence of state transitions in which an action by one entity updates shared state — the state of the world, a relationship, a model, a material condition — and that update propagates to affect other entities or states. When I speak, my words alter your model of the situation. When a factory discharges waste, the downstream water chemistry changes. When a policy is enacted, incentives shift and behavior follows. In each case, there is a directed path from action through transformation to effect. The path may be short — I push a glass, it falls — or long — a regulatory change propagates through an industry over decades, altering employment patterns, community health, and political alignments three generations downstream. The path may involve many intermediate steps and many mediating systems. But the essential structure is directional: something was done, and something changed as a result, and that change propagated.

Closure occurs when a consequence chain forms a loop — when downstream effects feed back to update the upstream state that initiated the chain. The glass falls and shatters; I hear the sound, see the shards, and update my future behavior: I will place the glass further from the edge. The factory's waste reaches the river; the community downstream falls ill; regulators investigate; the factory faces sanctions. In each case, the consequences of the action return — however indirectly, however delayed — to the source. The loop closes. This return is not guaranteed. Consequence chains can terminate at any point along their length: the signal absorbed by noise, redirected to a party who cannot act on it, blocked by a structural barrier, or simply dissipated before it arrives. When the signal does return, when the loop closes, we have a system with a specific structural property: it is self-correcting. The actor's future behavior is informed by the actual results of past action, not merely by the actor's prediction of what those results would be.

Coherence is the degree to which closures are stable under transformation. A single feedback loop that fires once is closure. A system in which the feedback loops persist — surviving changes of scale, changes of framing, delays in time, substitution of mediating parties — is coherent. Coherence is not a binary property. It admits of degrees, and those degrees are, in principle, measurable — a claim we will develop in Section 3.4. A friendship that survives a disagreement is more coherent than one that shatters at the first tension: the feedback loop (you hurt me, I tell you, you adjust) persists under the stress of conflict. An institution whose error-correction mechanisms operate even when the errors are embarrassing to leadership is more coherent than one whose feedback loops work only when the news is good: the closure persists under the transformation of content from convenient to inconvenient.

One crucial distinction must be established at the outset. The conservation we are describing is not physical conservation. We are not claiming that consequences obey a conservation law analogous to conservation of energy. Information can be created, destroyed, duplicated, and transformed in ways that have no direct analog in thermodynamics, though we will explore informative thermodynamic connections in the following section. What we mean by "conservation" here is invariance under mapping — the preservation of relevant structure when a system is transformed, relabeled, coarse-grained, or observed from a different vantage point. A consequence chain "conserves" when the information it carries — the structural relationship between action and effect — survives the transformations that mediate it. When I tell you the truth and you pass it to a third party, the information structure is conserved even though the carrier has changed. When I tell you a lie that you pass on faithfully, the *form* of transmission is preserved but the *content* has been decoupled from reality — a distinction that will become central to our account of deception.

This vocabulary — consequence chain, closure, coherence — is deliberately minimal. We have not yet introduced any moral concepts. We have not said that closure is good or that severance is bad. We have described a structural landscape: directed paths from action through effect to feedback, loops that close or fail to close, patterns that persist or dissolve under transformation. Everything that follows is built on this landscape alone.

1.2 The Toy System

Consider the simplest possible model of consequence: a directed graph with five nodes.

Agent. Action. Affected Party. Signal. Agent Update.

The agent does something. The action reaches someone who is affected by it. The affected party generates a

signal — a response, a reaction, a change in state that carries information about what the action produced. The signal propagates. And in the closure variant, it reaches the final node: the agent updates. Their model of the world, their model of the relationship, their disposition toward future action — something changes in them as a result of learning what their action did.

This is the complete toy system. Five nodes, four forward edges, and the question of whether a fifth edge — the return path from Signal to Agent Update — exists. That question is the question of the entire paper.

Here is a concrete instantiation.

I tell you it is raining when it is sunny. You trust me — you have no reason not to — and you take an umbrella. You step outside, and the sky is clear.

In the closure variant, you return. You find me. You tell me you know I lied, or your face tells me, or the social consequences arrive through channels I cannot avoid. The signal reaches me. I know that you know. My model of our relationship updates — I now carry the knowledge that my deception was detected, that trust has been spent, that whatever I gained by the lie has been offset by what the return signal delivered. The consequence of my action has propagated through the system and completed the loop. The cost landed where it originated.

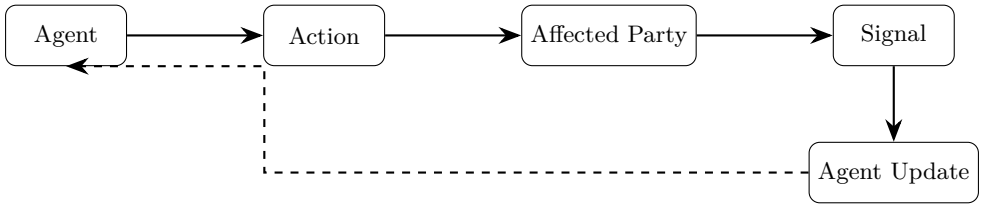

Figure 1: The toy system: closure variant. The dashed return arrow completes the consequence loop.

Now consider the severance variant. I tell you it is raining. You take the umbrella. But I have moved on — changed cities, changed names, or simply positioned myself within a social structure where your feedback cannot reach me. You discover the lie, but the signal has no return address. It routes, in the language we are developing, to a *sink*: a node from which no outgoing arrow proceeds. I gained the short-term benefit of appearing helpful. You paid the cost of acting on false information. And my model never updates, because the consequence of my action never arrived.

Figure 2: The toy system: severance variant. The signal terminates at a sink node with no return path.

The structural difference between these two cases is the presence or absence of a single arrow — the return path from Signal to Agent Update. Everything else is identical: the same action, the same affected party, the same initial propagation. The moral difference between these two cases — one in which the lie is part of a self-correcting system and one in which it is a cost externalized without return — maps exactly onto the structural difference. We are not yet in a position to prove this identity, but we can already see its shape: the topology of the graph determines whether the system can self-correct, whether the actor remains accountable, and whether the consequences of action are borne by the agent whose action produced them.

This graph will return throughout the paper. Each time it reappears, it will carry one additional capability — one more thing we can see on the same five-node structure that the previous appearance could not yet show. In Section 2.2, the graph will model each type of consequence-severing: not only lying but exploitation, temporal displacement, addiction, and bureaucratic diffusion. In Section 2.3, it will appear as a subgraph embedded in a larger network, showing how local closure can coexist with global severance — the formal structure of the Mafia problem. In Section 2.4, it will distinguish friendship from abuse by the directionality

of its information flow. In Section 3.3, it will model the most dangerous form of severance: the modification of the graph itself, the removal of the agent's capacity to *represent* the return arrow. And in Section 3.4, it will acquire quantitative values — curvature on edges, persistence scores on loops — transforming from a qualitative model into a measurable object.

The recurrence is deliberate. If the argument works, the reader will track the graph's returns as a progressive disclosure: the same structure appearing at higher resolution, the same pattern surviving transformation. The paper does not merely describe consequence-chain closure. It enacts it.

1.3 The Thermodynamic Anchor

The concepts we have introduced — consequence, closure, coherence — are structural. They describe patterns of information flow and return. But they do not hang in a vacuum. There is a physical context in which these patterns exist, and understanding that context clarifies why coherence is not merely a desirable property but a condition of continued existence.

The context is the second law of thermodynamics, and the relevant observation is this: the universe, left to itself, dissipates structure. Entropy increases. Gradients flatten. Organized states degrade toward equilibrium. This is not a tendency or a preference. It is the overwhelming statistical weight of disordered configurations over ordered ones — the simple fact that there are incomparably more ways for a system to be disorganized than organized, and that random perturbation therefore drives systems toward disorder with near certainty.

Against this background, every persisting structure — every atom bound in a molecule, every cell maintaining its membrane, every organism regulating its temperature, every mind sustaining a coherent self-model — is an achievement of local order maintained against the ambient current of dissolution. Life, in the most general thermodynamic sense, is the project of building and maintaining consequence loops that conserve information despite the universe's general tendency to scatter it. An organism that can detect a threat and respond — that closes the loop between environmental stimulus and adaptive action — persists. One that cannot, does not. Consciousness, on this account, is an intensification of the same project: the construction of representational models that are themselves consequence loops, internal structures in which the effects of perception feed back to update the perceiving system's capacity for future perception.

This thermodynamic framing has a direct consequence for our account of morality, and it arrives through a precise mechanism: Rolf Landauer's observation that erasing information has a minimum thermodynamic cost. Every bit of information erased dissipates at least $kT \ln 2$ joules of energy as heat. This is not an engineering limitation. It is a physical law, as fundamental as the conservation of energy.

Now consider what a lie is, thermodynamically. A lie creates a divergence between the liar's model and the world. To maintain the lie, the liar must continuously manage this divergence — tracking who knows what, maintaining consistency across interactions, suppressing signals that would reveal the truth. Each of these operations involves, at minimum, the selective erasure or overwriting of information: the true state of affairs must be replaced, in the liar's communications and sometimes in their own working memory, with the false one. By Landauer's principle, each such erasure has a thermodynamic cost. Lying is not free. It is a form of entropy export — and the entropy must go somewhere.

This observation will be developed formally in Section 2.5 as the Deception Overhead Theorem. For now, the point is structural: coherence — the closing of consequence loops, the maintenance of fidelity between model and reality — is the low-energy state. Deception, severing, and divergence are the high-energy states, requiring continuous expenditure to maintain. A system at coherence is thermodynamically relaxed. A system maintaining incoherence is doing work — work that increases with the scope and duration of the incoherence, work that diverts resources from other functions, work that imposes an overhead measurable in principle and increasingly in practice.

One further connection, offered with an explicit caveat. The structural patterns we are describing — coherence as the maintenance of consequence loops against dissolution, severance as the export of entropy to external systems, return as the closing of the loop — have appeared before, compressed into narrative form,

in the mythic traditions of multiple cultures. The Edenic state before the Fall can be read as a zero-entropy-gap system: action and consequence are transparent, there is nothing hidden, the loop is closed. The Fall — the acquisition of the capacity for deception — introduces the gap: the possibility of severing the consequence chain, of knowing something and acting as if one does not. Exile is the thermodynamic consequence: the maintenance of the gap requires departure from the low-energy state, and the return requires closing what was opened.

We flag this illustration explicitly as a *compressed representation*, not a metaphysical claim. We are not arguing that ancient myths were doing thermodynamics. We are observing that the structural pattern — coherence, severance, return — is deep enough to have been independently encoded in narrative traditions whose authors had no access to the formal vocabulary we are developing. This is evidence of the pattern's robustness, not of the myths' scientific validity. The compression is lossy and the mapping is partial. But the structural resonance is not accidental, and noting it here may give the reader an additional anchor for the formal concepts that follow.

1.4 Methodological Prohibition: Against Reification

Before proceeding to the central argument, we must establish a constraint that applies to everything that follows. It is a constraint on how we speak, and it is load-bearing.

Coherence is not a thing.

It is not a substance. It is not a force. It is not an entity that exists in the world alongside tables and electrons and moral facts. It is not something the universe is "made of." It is an invariant — a property of mappings, a measure of what survives transformation. In the same way that *symmetry* is not a substance but a property of transformations that leave a structure unchanged, coherence is a property of systems whose consequence loops persist under perturbation. To say that a system is coherent is to say something about its mappings, not to posit a new kind of stuff.

This distinction matters because the failure to maintain it has a specific and well-documented consequence: reification, the conversion of a structural property into a putative substance. The history of thought is littered with reifications that began as useful observations and ended as metaphysical idols. "Phlogiston" reified the observation that combustion involves the loss of something. "Vital force" reified the observation that living things behave differently from dead ones. In each case, a genuine structural observation — something real is happening here — was converted into a substance — there must be a thing called X that explains it — and the conversion blocked further progress because the reified substance required no mechanism: it explained by naming rather than by elucidating structure.

We insist on this prohibition throughout the paper, and we state it as a recurring rule: *if you are tempted to treat coherence as a substance, rewrite the sentence as an invariant of mappings.* This is not a stylistic preference. It is a structural constraint without which the framework collapses. If coherence is a substance, then "more coherence" becomes a goal in itself, divorced from the specific consequence chains whose closure constitutes the coherence. The framework becomes a metaphysical theory — a claim about what the universe is ultimately made of — rather than a diagnostic framework for identifying when consequence chains close and when they do not. We do not want a metaphysical theory. We want a set of tools that work.

The prohibition also prevents a subtler failure: the conversion of coherence into a value that can be optimized. If coherence is a substance, then maximizing it is obviously good, and we are back to a form of consequentialism in which the consequence to be maximized is called "coherence" instead of "utility." But coherence is not something to be maximized. It is a property of the relationship between a system's actions and their consequences. It is present when the loops close and absent when they do not. The question is always structural: are the consequence chains closing? Not: is there enough coherence?

Topology without substrate. Conservation without a conserved thing. Structure without stuff. This is the register in which the framework operates, and maintaining it is the price of admission to the argument that follows.

2 The Identity of Structural and Moral Integrity

2.1 The Dual Meaning of "Integrity"

In the introduction, we observed that the English word "integrity" carries two meanings that appear unrelated — structural soundness and moral wholeness — and proposed that the appearance of unrelatedness is wrong. We are now in a position to say precisely why.

Structural integrity, in the engineering sense, is the property of a system that maintains its functional identity across transformation. A bridge has structural integrity when the forces that pass through it — the weight of traffic, the stress of wind, the cycles of thermal expansion and contraction — do not cause it to become something other than a bridge. The forces enter, propagate through the structure, and the structure remains what it was. What survives the round trip — the passage of force through the system and the system's continued identity afterward — is the integrity.

Moral integrity, in the ethical sense, is the property of a person whose actions, commitments, and self-model cohere across time and circumstance. A person has integrity when the pressures that pass through them — temptation, self-interest, fear, the opportunity to act undetected — do not cause them to become other than who they present themselves to be. The pressures arrive, the person acts, and the person's identity (in both the social and the self-model sense) remains continuous with what came before. What survives the round trip — the passage of consequence through the person and the person's continued identity afterward — is the integrity.

The parallel is not poetic. It is structural. In both cases, integrity names the same invariant: *what is preserved when transformation passes through the system.* The bridge preserves its load-bearing geometry. The person preserves the coherence of their action-consequence loops. Both are instances of a single topological property: closure under transformation.

Ordinary language philosophy has long recognized that dual meanings in everyday language often point to deep structural connections rather than accidental homophony. When the same word is used across domains that seem unrelated, the philosopher's task is to ask whether the domains share structure that the language has registered and the theoretical apparatus has not yet caught up with. The dual meaning of "integrity" is, we argue, exactly such a case. The language saw the identity before the philosophy did. Structural integrity and moral integrity are not connected by metaphor or by analogy. They are connected by being the same thing — the preservation of consequence-relevant structure across transformation — observed in different substrates with different types of signal.

This connection also appears in the philosophical literature on integrity, though it has not been formalized in the way we propose. Philosophers who work on integrity as a moral concept — from Bernard Williams's emphasis on integrity as fidelity to one's ground projects, to Lynne McFall's analysis of integrated moral identity, to Cheshire Calhoun's account of integrity as standing for something in a social context — are all, on our reading, describing different aspects of the same structural property: the persistence of coherent consequence loops under pressure. Williams's ground projects are the consequence chains the agent refuses to sever. McFall's integrated identity is the self-model that updates in response to the full consequences of one's actions. Calhoun's social standing is the maintenance of bidirectional accountability — the agent who can be counted on not to reroute consequence signals to sinks. What our framework adds is not a competing account but a formal backbone: the structural property that these diverse accounts are each partially describing.

2.2 Consequence-Severing as the Root of Immorality

If coherence is the structural property of systems whose consequence chains close, then we can state the central diagnostic claim of this paper: *immoral action is consequence-severing.* Not as a metaphor, not as one feature among many, but as the root mechanism. Every immoral action we can identify — across traditions, across cultures, across substrates — shares a common topology: the agent acts in a way that produces consequences, and the architecture of the situation ensures that those consequences land on someone

other than the agent who produced them. The benefit is internalized; the cost is externalized; the loop is broken.

This claim requires demonstration. We will work through a taxonomy of consequence-severing, showing each type on the toy system from Section 1.2. The taxonomy is not exhaustive, but it covers the major forms and reveals the common structure beneath their surface differences.

Lying is the severance of the information chain. Return to the toy system. I act (I speak), and my action reaches you (the affected party). In the truthful case, the information I transmit preserves the structure of what I know — the signal's content maps faithfully onto the state of affairs it describes. You receive it, act on it, and the consequences of your action are consistent with what you were told. When those consequences return to me, they arrive in a world where my communication and its effects are aligned. The loop closes without distortion. In the lying case, the information I transmit *diverges* from the state of affairs. You receive it and act on it, but your action is now calibrated to a model that does not match reality. The consequences of your action — consequences that would have been different had you received accurate information — are displaced from where they would have landed. The lie has rerouted the consequence chain: I gained the benefit of your trust, and the cost of acting on false information falls on you, on third parties, on the future. The return arrow, if it exists at all, carries a distorted signal.

Exploitation is the severance of the value chain. The exploiter extracts value from a system — labor, resources, attention, care — while ensuring that the costs of extraction are borne elsewhere. The toy system makes the topology visible: the agent acts (extracts), the affected party bears the consequence (depleted resources, degraded conditions), a signal is generated (suffering, protest, market externality) — but the signal is routed away from the agent. The exploiter has arranged the boundary of their system so that the feedback from their extraction does not reach them. They enjoy the output without experiencing the cost. The loop is open by design.

Temporal severing displaces consequences in time rather than in space. Pollution that will affect communities decades hence, financial instruments that defer risk to the next generation, institutional decisions that produce short-term gains and long-term degradation — all share the topology of the open loop, but the sink is not a different person; it is a future state of the world. The agent acts, the affected party (the future community, the future institution, the future self) absorbs the consequence, but the signal cannot return because the temporal distance between action and consequence exceeds the horizon of the agent's accountability. The agent has exited the graph before the return arrow can fire.

Trafficking is the most extreme form of consequence-severing: the treatment of a conscious agent — a being that generates signals, that suffers, that responds — as an extractable resource. The trafficked person's signals are not merely rerouted; their status as a signal-generating node is denied. They are treated, within the trafficker's model, as a resource node rather than an agent node — something from which value flows outward without generating the kind of consequence that would require acknowledgment or return.

Addiction is self-directed consequence-severing. The addicted person's present self extracts value — the reward, the relief, the temporary coherence of intoxication — while routing the costs to their future self. The toy system applies with the agent and the affected party being the same person at different times. The present-self acts, the future-self absorbs the consequence, but the signal from future to present is systematically attenuated by the very mechanism of the addiction. The addictive substance or behavior operates as a sink: it absorbs the return signal and replaces it with the forward signal of immediate reward. The loop is kept open by the chemical or behavioral architecture of the addiction itself.

Bureaucratic diffusion distributes responsibility until no return address exists. The decision to pollute, to deny a claim, to enforce a harmful policy is broken into components, each assigned to a different office, each office following a procedure, each procedure referencing a regulation. No single node in the system made the decision. No single node receives the full consequence signal. The affected party generates a signal, but the signal encounters a diffusion structure in which responsibility has been divided below the threshold of meaningful accountability. The loop is not broken at a single point. It is dissolved into a mist.

All six forms share a single topology: action benefits the actor by ensuring that the consequences of the action land on someone else, on the future, or on a version of the system that cannot send signals back. The

diversity of forms — interpersonal, economic, temporal, existential, psychological, institutional — is surface variation on a single structural theme.

But the taxonomy has a critical discrimination test. Not all interruptions of consequence flow are moral failures. We must distinguish consequence-severing from what we will call *benign severance*: the deliberate limitation of consequence flow that preserves or protects rather than exploits.

Privacy is benign severance. When I choose not to share certain information about myself, I am limiting the consequence chain — you cannot respond to information you do not have. But I am not creating a divergent model. I am not telling you something false. I am limiting the scope of the loop while preserving its integrity within the scope that remains. The key distinction: I have not routed a signal to a sink. I have declined to generate the signal in the first place, and the relationship operates on the basis of what has been shared rather than on the basis of a falsehood.

Compartmentalization for safety follows the same logic. A security clearance system limits who receives certain information, restricting the consequence chain's breadth. But within each authorized scope, the loops close. The system preserves the *capacity* for consequence-return while limiting its *scope*. This is structurally different from destroying the capacity itself, which is what malign severance does.

The formal criterion: benign severance preserves the capacity for consequence-return even when limiting its scope. Malign severance destroys the capacity itself. The person who maintains privacy can, in principle, choose to share at any point — the return path exists as a potential, held in reserve. The liar has not limited scope; they have created a false channel that degrades the capacity for truthful return. The exploiter has not limited scope; they have arranged the boundary of the system so that return is structurally impossible. The distinction is between a door that is closed but can be opened and a wall built where the door was.

2.3 Local vs. Topological Coherence

We now face the most serious structural objection to the framework: the Mafia problem. A crime syndicate, a cult, a totalitarian regime — these are systems with high internal coherence. Members trust each other (within the group). Signals return (within the group). Actions have consequences that feed back to update behavior (within the group). "Snitches get stitches" is a closed loop: betray the group, suffer the penalty, update your behavior accordingly. If coherence is the ground of morality, and the Mafia is coherent, then the Mafia is moral — which is absurd. The framework appears to have a fatal vulnerability.

The vulnerability is illusory, and dissolving it reveals one of the framework's most important distinctions: the difference between local and topological coherence.

The Mafia is locally coherent. Its internal consequence chains close. Members who defect are punished; members who cooperate are rewarded; the feedback loops are tight and reliable. Within the system boundary that the Mafia draws around itself — the boundary that includes members and excludes everyone else — the structure holds together under transformation. It has internal integrity.

But the Mafia's internal integrity is achieved by a specific structural means: the severance of consequence chains at the system boundary. The drug trade produces suffering in communities. The protection racket extracts from businesses. The corruption of institutions degrades public trust. Each of these is a consequence of the Mafia's actions, and each is systematically prevented from returning to update the Mafia's internal model. The affected communities' signals — suffering, protest, law enforcement pressure — are not treated as feedback to be integrated. They are treated as threats to be suppressed, evidence to be destroyed, witnesses to be silenced. The Mafia maintains its internal coherence *by severing the external consequence chains* that would, if allowed to close, reveal the system's parasitic relationship to the larger structure in which it is embedded.

Return to the toy system, now embedded in a larger graph. The Mafia's five-node loop exists as a subgraph: internal agent, internal action, internal affected party, internal signal, internal update. The arrows are all present. The loop closes. But this subgraph is embedded in a larger network that includes external affected parties — the communities from which the Mafia extracts — and the edges connecting the Mafia's actions to those external parties' return signals have been cut. The subgraph is closed; the embedding graph is severed.

The visual representation makes the diagnosis immediate: local closure plus global severance is the formal structure of parasitic coherence.

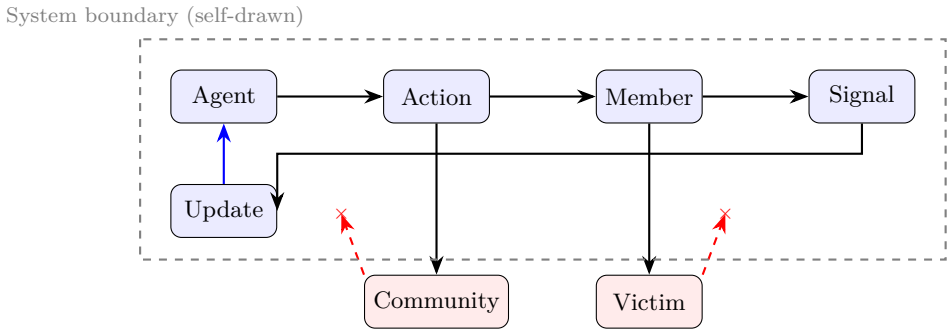

Figure 3: Parasitic coherence: the Mafia problem. The internal loop (blue) closes. External consequence chains (red, dashed) are severed at the system boundary. Local coherence, global severance.

This is what we mean by saying that immorality is the misidentification of the system boundary. The morally relevant boundary is the minimal boundary that includes all nodes whose state is changed by the action path, plus the channels required for their consequence signals to reach the origin. The Mafia's coherence is real — but it is coherence within a boundary that has been drawn precisely to exclude the affected parties whose inclusion would reveal the parasitism. Widen the boundary to include everyone affected by the Mafia's actions, and the coherence vanishes: the consequence chains that cross the wider boundary do not close. The Mafia is coherent only if you accept its self-drawn boundary as the relevant system. Topologically — considering the full graph, all nodes, all potential edges — it is incoherent.

This local-topological distinction generalizes. A cult is internally coherent: members share a model, feedback loops reinforce the shared reality, defectors are punished. But the cult's coherence is maintained by severing its members' consequence chains with the outside world — controlling information, restricting contact, redefining language so that external feedback cannot be parsed. A totalitarian state is internally ordered: the trains run on time, the bureaucracy functions, the leadership receives reports. But the state's order is maintained by severing the consequence chains between state action and citizen feedback — censoring the press, suppressing dissent, controlling the very categories in which criticism could be formulated.

In each case, the structure is identical: a locally coherent subsystem embedded in a larger system whose consequence chains it severs. And in each case, the diagnosis is available by examining the topology: does the coherence extend to all materially affected parties, or does it depend on a boundary that excludes them?

There is a temporal dimension to this diagnosis that makes it sharper still. Parasitic coherence is temporally unstable. It borrows from the future.

Consider the Mafia again. Its internal coherence at time t_0 depends on maintaining the severance boundary — on ensuring that external consequences do not return. But maintaining that boundary requires continuous work: bribing officials, intimidating witnesses, monitoring members for signs of defection, managing the ever-growing complexity of the deception. This work has costs, and those costs increase over time. The more successful the Mafia is at extracting value while severing consequences, the more pressure builds on the boundary, because more affected parties have more reason to seek return channels, and more resources must be devoted to blocking those channels.

This is *temporal debt*: the accumulated cost of maintaining a severance boundary that the future will eventually collect. The Mafia is coherent at t_0 only if you do not ask about t_1, t_2, t_3. Over time, the maintenance costs of the boundary compound. The deception overhead grows. The energy required to suppress return signals increases faster than the energy available to suppress them. Eventually — and the history of organized crime, cults, and totalitarian regimes provides abundant empirical confirmation — the boundary fails. The severed consequences return, often catastrophically. Empires fall. Cults implode. Crime

families consume themselves in internal betrayal as the trust that maintained internal coherence is revealed to have been parasitic on the very trust they destroyed externally.

True coherence — topological coherence, coherence that extends to all affected parties — is stable as $t \to \infty$. It does not accumulate temporal debt because it does not depend on a boundary that must be maintained against reality. Parasitic coherence requires that you stop the clock. It looks stable in a snapshot. In a film — in any analysis that includes the time dimension — it is a decoherence process: a system spending down a trust account it cannot replenish, burning structural capital it cannot replace, borrowing against a closure it will never achieve.

The Mafia is only "coherent" in a photograph. In a time-lapse, it is entropy.

2.4 The Bridge Lemma

We have now named what moral failure is — the severing of consequence chains — and distinguished local from topological coherence. What we have not yet done is state, with precision, what a moral relation *is*. We need a positive account, not only a taxonomy of failure. Here it is, in plain language first:

A moral relation is one in which information about the consequences of interaction returns to both parties with sufficient fidelity to update future action.

That is the whole of it. Everything else — the formal machinery, the categorical apparatus, the topological measurement — is in service of making this sentence exact and showing that it is not merely a definition we have chosen but a structural constraint that any persisting system must satisfy.

Return to the toy system. Agent acts on Affected Party. Signal propagates. In the closure variant, the signal returns: the affected party's response reaches the agent, and the agent's model updates. Now consider what "both parties" means. It means the graph has return arrows in *both directions*. Not only does the agent receive feedback about the effects of their action — the affected party also receives information about the agent's subsequent update. Did the agent change? Did the consequence land? Is the loop closed on both sides?

This is the distinction that matters. A relationship in which only one party's model updates is not a moral relation. It is extraction.

The friendship and the abuse.

Consider two relationships, modeled on the same five-node graph.

In the first, I share something that costs me — a vulnerability, a mistake, an uncertainty. This is an action that creates a consequence: you now hold information that could be used against me. You respond in a way that shows you received the signal intact: not by parroting it back, not by dismissing it, but by updating your model of me and acting on the update. Perhaps you share something corresponding. Perhaps you simply adjust how you speak to me going forward. The point is that your response carries information back — it tells me that the signal arrived, that it was processed, that it changed something in you. My model of you updates: you are someone to whom consequence-bearing signals can be sent and from whom they return. Your model of me updates symmetrically. Both graphs gain edges. The loop closes on both sides.

Notice what has happened structurally. The consequence of my action (sharing vulnerability) propagated to you, transformed in you (into understanding, reciprocity, adjusted behavior), and returned to me as a signal I could integrate. But the same thing happened in reverse: your response was itself an action with consequences for me, and my integration of it was a signal that returned to you. The graph is bidirectional. Information flows, transforms, and returns in both directions. Neither party's model diverges from reality because both are continuously updated by the other's responses.

Now consider the second relationship. I share the same vulnerability. You receive it — but instead of updating your model of me, you store it as leverage. When it serves you, you deploy it: to win an argument, to control a decision, to maintain an asymmetry. The consequence of my sharing has propagated to you, but it did not transform into updated understanding. It was routed to your control apparatus — a different

part of your internal graph, one that optimizes for extraction rather than reciprocity. And the signal that returns to me is not "I received and integrated your signal." It is "I now have power over you."

My model updates — but with the wrong sign. I learn that signals sent to you do not return as mutual understanding; they return as instruments of control. The loop closes on my side (I receive feedback, I update), but it does not close on yours in the morally relevant sense: you did not update your model of me as a consequence-bearing agent. You updated your inventory of resources. The graph is unidirectional. Information flows from me to you and is consumed; what returns is not transformed consequence but deployed power.

This is the structural difference between a moral relation and a parasitic one, and it is entirely topological. We did not need to examine the *content* of what was shared, the *feelings* involved, or the *intentions* of either party. We needed only to ask: does consequence-relevant information flow, transform, and return in both directions? The friendship and the abuse look identical if you examine only the forward arc (vulnerability shared, information transmitted). They diverge completely when you examine the return paths.

Figure 4: Bidirectional (moral) vs. unidirectional (parasitic) return. Same forward arc; divergent return topology. Blue: structure-preserving return. Red: extractive return.

The formal statement.

For those who want the categorical precision (and a full treatment is available in Appendix A), the distinction maps onto a well-studied structure. A relation in which information flows and transforms bidirectionally, with both sides updating coherently, satisfies the conditions of what category theorists call a *profunctor* — a generalized relation between categories that preserves structure in both directions. The key property is *exactness*: a mediation step is exact when it preserves consequence-relevant distinctions under composition. No hidden quotienting of responsibility. No information absorbed without a corresponding update. No distinction collapsed that was needed downstream.

When a relation is exact in this sense, information about structure transfers faithfully between the two sides. It does not matter which path you take through the graph — the structure arrives intact. This is what "with sufficient fidelity to update future action" means when made precise: the signal that returns preserves enough of the original consequence structure that the recipient can adjust their behavior in ways that are responsive to reality rather than to a distorted model.

When a relation is not exact — when it is unidirectional, when information is consumed but not reciprocated, when distinctions are collapsed that were needed for accurate updating — then structure degrades in transit. The affected party's model diverges from reality. The agent's model remains uncorrected. The system as a whole loses integrity in both senses simultaneously: it can no longer hold together under transformation (structural), and the parties within it can no longer act in ways that are responsive to the full consequences of their actions (moral).

This is what we mean by the formal identity of structural and moral integrity. It is not that structural integrity is a metaphor for moral integrity, or that moral integrity is a special case of structural integrity, or that they are analogous. It is that they are the same invariant — the preservation of consequence-relevant structure under transformation — observed at different scales. The bridge holds together because forces pass through it and it remains what it is. The moral relation holds together because consequences pass through it and both parties remain responsive to what is real. The bridge and the friendship are doing the same thing.

The fidelity gradient.

One further precision. The distinction between moral and parasitic relations is not binary. There is a gradient of reception fidelity that determines what kind of relation is possible between two systems:

Below threshold, the signal is noise. No meaningful information transfers; the two systems might as well be unconnected. There is no moral relation because there is no relation at all.

At partial fidelity, the signal arrives but distorted. The receiver integrates something, but it has been filtered through whatever templates are locally available — cultural assumptions, prior expectations, category errors. The consequence chain partially closes, but what returns is a warped version of what was sent. This is the regime of miscommunication, of good intentions landing badly, of partial moral relations that strain under load because the updating is real but lossy.

At full fidelity, the signal arrives structurally intact. The consequence chain closes cleanly. Both parties update in response to what actually happened rather than to what their local models predicted would happen. This is the regime of genuine moral relation — not because the parties are virtuous by some external standard, but because the topology of their interaction preserves the information that each needs to remain responsive to reality.

The bridge lemma, restated: moral integrity is the condition in which consequence-relevant information flows bidirectionally with sufficient fidelity that both parties' models remain structurally responsive to reality. This is not a standard we impose from outside. It is what "holding together across transformation" means when the transformation involves agents who act on each other and live with the results.

2.5 The Deception Overhead Theorem

In Section 1.3, we introduced the thermodynamic anchor: Landauer's principle, the physical fact that erasing information has a minimum energy cost. We are now in a position to develop this observation into a structural result with falsifiable predictions.

The claim is this: maintaining deception — the sustained divergence between an agent's model and reality, or between an agent's communications and their knowledge — requires continuous computational and energetic overhead that scales with the scope and duration of the deception. This overhead is not an incidental cost. It is a structural consequence of the physics of information. And it has measurable signatures.

Consider what a liar must do. The initial lie is cheap — a single act of substitution. But the consequences of the lie propagate. The person lied to acts on the false information, and their actions produce effects that are calibrated to a reality that does not exist. The liar, to maintain the lie, must track these cascading effects: what does the lied-to person now believe? What actions have they taken based on the false belief? What might they discover that would reveal the discrepancy? Each of these tracking operations requires computational resources — memory, attention, inference — that are diverted from other tasks. Each interaction with the lied-to person requires the liar to maintain two models simultaneously: the true state of affairs and the false one, checking each communication against both to ensure consistency with the lie while remaining operative in reality.

This is the deception overhead, and it has a lower bound grounded in the irreversible bookkeeping required to maintain divergence. Every act of erasure or overwriting involved in maintaining the divergence — suppressing the true signal, generating the false one, checking for consistency, monitoring for exposure — dissipates energy. Landauer's principle establishes that each logically irreversible operation in the computational substrate implementing the deception has a minimum thermodynamic cost ($kT \ln 2$ per bit erased). A note on epistemic status: the physical claim — that information erasure has a minimum energy cost — is established for the computational substrate performing the erasure. The social and institutional claim — that deception overhead scales predictably in bounded agents and organizations — is an empirical hypothesis about the macroscopic expression of this inefficiency, directionally supported by cognitive load research and organizational behavior studies but not yet established with the same rigor. We flag this distinction because the argument does not require the thermodynamic claim to scale linearly from bits to social costs; it requires only that deception imposes escalating maintenance costs on any substrate implementing it, which the empirical evidence strongly supports. The overhead scales: a lie told to one person about one topic requires managing

one divergence. A lie told to many people about many topics requires managing a combinatorial explosion of potential inconsistencies, each of which demands monitoring and each of whose monitoring demands energy.

The analogy to data compression is instructive. Lying is a form of lossy compression: the liar transmits a simplified version of reality (the version convenient for the liar) and discards the information that would complicate the picture. In the short term, this saves bandwidth — the communication is simpler, the social interaction smoother. But lossy compression accumulates artifacts. The discarded information does not disappear; it persists in reality, generating effects that the compressed model cannot predict. Over time, the artifacts compound: the gap between the compressed model and reality widens, and the computational cost of maintaining consistency between them grows. Truth, by contrast, is lossless transmission. It is higher bandwidth in the moment — it requires transmitting the full complexity of the situation — but it incurs no divergence-maintenance overhead, because there is no gap between model and reality to manage.

This analysis connects directly to Kant's categorical imperative. Kant's universalizability test asks: can this maxim be universalized without contradiction? A world in which everyone lies is self-defeating — lying parasitizes truth and cannot survive universalization. This is correct as far as it goes, but Kant's framing is purely logical. He identifies the contradiction but not the *mechanism*. The coherence framework provides the mechanism: lying cannot be universalized because the deception overhead grows super-linearly with the number of participants. A single liar in a population of truth-tellers pays a manageable overhead. As the proportion of liars increases, the overhead for each liar increases (because more communications must be checked against more potential inconsistencies), and the system collapses into communicative paralysis — not because of a logical contradiction but because of a thermodynamic one. The coherence framework absorbs Kant's insight while grounding it in something more fundamental than pure reason: the physics of information conservation.

This has a falsifiable prediction: sustained, systematic deception — in individuals, in organizations, in information systems — should correlate with measurable resource diversion. Cognitive load studies confirm this for individuals: liars exhibit measurable increases in response latency, working memory demands, and physiological stress markers. But the prediction extends beyond the individual.

Organizations engaged in systematic deception exhibit characteristic structural signatures. Compliance apparatus expands: legal departments grow, document review processes multiply, communication protocols become more restrictive. Message discipline tightens: employees are coached on what to say and not say, talking points are distributed, deviation from the official narrative is sanctioned. Surveillance infrastructure develops: monitoring of communications, both internal and external, to detect potential leaks or inconsistencies. Compartmentalization increases: information is siloed so that no single person can reconstruct the full picture, reducing the risk of exposure while increasing the coordination costs. Each of these is a measurable proxy for deception overhead. An organization whose compliance costs, surveillance infrastructure, and compartmentalization are growing faster than its actual operational complexity is exhibiting the thermodynamic signature of sustained incoherence.

The deception overhead theorem forecloses something specific: the possibility of an easeful, integrated existence maintained through sustained deception. The liar pays the overhead, and the overhead is extracted from the resources that would otherwise support relaxed, responsive, fully present engagement with reality. Deception does not merely fail morally. It fails thermodynamically: it is a higher-energy state than honesty, and maintaining it requires continuous work that degrades the very capacities — attention, responsiveness, integration — that constitute what we recognize as a well-lived life.

2.6 Reachable-State Constraints

We have described what consequence-severing is, classified its forms, shown how parasitic coherence differs from genuine coherence, and established that maintaining incoherence has measurable thermodynamic costs. A natural question arises: who enforces this? If consequence-severing is structurally costly, why do parasitic actors persist? And if the framework is correct, does it predict anything about what parasitic actors can and cannot achieve?

It does. And the prediction requires no enforcer.

The claim is topological: incoherent configurations cannot access states that require coherence. Not as punishment, not as divine justice, not as karmic return, but as a constraint on the space of reachable states — the same kind of constraint that prevents a disconnected beam from bearing a load, or a disconnected graph from transmitting a signal between its components.

Consider what a high-trust state requires. Two agents in a high-trust state have mutual model alignment: each can predict the other's behavior with high accuracy, because each has received and integrated honest signals from the other over time. This alignment enables low transaction costs (no need for elaborate contracts or verification), high bandwidth communication (information can be transmitted with less error-checking overhead), rapid coordination (each can act on reasonable assumptions about the other's response), and vulnerability (each can expose themselves to the other without expecting exploitation). These are not merely desirable properties. They are structurally dependent on a specific condition: the consequence chains between the agents must have been closing, reliably, for long enough to build the aligned models on which trust operates.

An agent who systematically severs consequence chains cannot build these models. Not because of a moral prohibition, but because the models require input that the agent's actions prevent from being generated. The liar's interlocutors are operating on false information; their models of the liar are incorrect; the liar's model of them is calibrated to a fiction. The exploiter's partners have not consented to the actual terms of the relationship; their cooperation is based on a misunderstanding; the exploiter's model of the relationship does not match its actual structure. In each case, the agent has, through consequence-severing, made it structurally impossible for the mutual models to be accurate. And without accurate mutual models, high-trust states are unreachable.

This is not to say that parasitic actors cannot flourish. They can — within a restricted region of the state space. The con artist flourishes within social systems that tolerate transient interactions without verification. The exploitative employer flourishes within labor markets that suppress workers' ability to send return signals. The corrupt official flourishes within institutions whose accountability mechanisms have been captured. Each parasitic actor optimizes within a subspace of the state graph — the subspace whose states are achievable without high mutual model alignment.

But they cannot reach the states that require it. They cannot sustain deep collaborative relationships, because such relationships require the accumulation of accurate mutual models over time. They cannot participate in high-trust, low-overhead coordination, because such coordination depends on the very feedback channels they have severed. They cannot access the bandwidth that honest communication enables, because their communications are throttled by the deception overhead. The parasitic actor is free to roam within their accessible region of state space — and the accessible region is defined, precisely, by the topology of the consequence chains they have severed.

This is structural karma. Not the metaphysical claim that the universe punishes wrongdoers, but the topological observation that the space of states achievable by an agent is constrained by the agent's structural properties, and that consequence-severing reduces connectivity in the high-trust region of the state graph. An agent who severs consequence chains is not prevented from acting, not punished, not judged. They are geometrically constrained — in the same way that an agent at one node of a graph is constrained to reach only those nodes connected to it by paths, and the severing of edges reduces the set of reachable nodes.

Moral coherence, on this account, is entailed by information structure. No external judge is required, no divine enforcer, no social contract. The topology does the work. The question is not whether an agent *should* maintain coherence but whether an agent who does not maintain coherence can access the states they presumably desire. The framework's answer: not the ones that require it. And the states that require coherence include most of what we recognize as a flourishing life.

2.7 Core Formalism Summary

What follows is a compressed statement of the framework's core concepts, intended as a self-contained reference for citation and application.

Definitions. A *consequence chain* is a directed sequence of state transitions in which an action by one entity updates shared state and propagates effects to other entities. *Closure* is the property of a consequence chain that forms a loop: downstream effects feed back to update the upstream state that initiated the chain. *Coherence* is the degree to which closures are stable under transformation — the persistence of feedback loops under changes of scale, framing, time, and mediation. *Severance* is the interruption of a consequence chain such that the effects of an action do not return to update the acting agent. *Benign severance* limits the scope of consequence flow while preserving the capacity for return. *Malign severance* destroys the capacity itself.

Moral criterion. Coherence increases when consequence-relevant information returns to all materially affected agents with sufficient fidelity to update future action. A moral relation is one in which this return is bidirectional: both parties receive and integrate consequence signals. An immoral configuration is one in which action benefits the actor by ensuring that consequences land elsewhere — the loop is open by design.

Diagnostics. Consequence-severing exhibits characteristic patterns: unidirectional information flow (exploitation), signal routing to sinks (lying, institutional opacity), boundary manipulation (parasitic coherence, cult formation), temporal displacement (costs deferred to the future), and capacity removal (pre-semantic interception). The common topology is: benefit internalized, cost externalized, return arrow absent.

Enforcement. Reachable-state constraints: incoherent configurations cannot access states requiring coherence. High-trust, high-bandwidth, low-overhead coordination requires the accumulation of accurate mutual models, which requires consequence chains to have been closing reliably over time. Consequence-severing reduces connectivity in the high-trust region of the state graph. No external enforcer is required; the topology constrains the space of achievable states.

Measurement. Coherence is, in principle, measurable via topological proxies: edge curvature (local cohesion), persistent homology (stability of feedback loops over time), and feedback completeness (proportion of consequence chains that close versus terminate at sinks). The measurement program is developed in Section 3.4.

3 Resolution and Empirical Opening

3.1 Beyond Kant: Resolving the Formalism Problem

Immanuel Kant proposed that morality is grounded in pure practical reason through the categorical imperative: "Act only according to that maxim whereby you can at the same time will that it should become a universal law." This is, we have argued, a coherence test. The universalizability criterion asks whether an action's maxim can survive a specific transformation — universalization — without self-contradiction. An action that cannot be universalized is incoherent in the precise sense that its consequence chains cannot close when extended to all agents: if everyone lies, the institution of trust on which lying parasitizes collapses, and lying becomes impossible.

Kant saw the structure. But his test is purely formal — it evaluates logical consistency divorced from actual consequence. This produces well-known absurdities. Kant notoriously argued that one must not lie even to a murderer who asks whether your friend is hiding in your house, because lying is not universalizable regardless of the circumstances. The maxim "lie to prevent murder" fails the universalizability test because if universalized, no one would believe anyone, and the lie would be ineffective. Therefore, Kant concludes, one must tell the truth to the murderer.

The coherence framework diagnoses this as a failure of formalism that mistakes a coherence test for a logical test. Kant's test asks: is the maxim logically consistent under universalization? The coherence test asks: does the action maintain the closure of consequence chains for all materially affected parties? These are different questions, and they yield different answers in exactly the cases where Kant's framework produces counterintuitive results.

At the murderer's door, the consequence chain analysis is straightforward. Telling the truth routes the consequence of your speech (the murderer locating your friend) to your friend (who is killed), and the

return signal (your friend's death, your complicity in it) returns to you. The loop closes, but what it closes around is catastrophic: your honest communication produced a consequence that destroyed an agent whose consequence chains you were in a position to preserve. Lying to the murderer, by contrast, routes the consequence of your speech to the murderer (who is misdirected), preserving your friend's ability to continue generating and receiving consequence signals — preserving, that is, the conditions under which consequence chains can continue to close for the affected party.

The coherence framework does not say "lying is sometimes acceptable." It says that the relevant question is never "is this maxim universalizable?" but always "does this action preserve or sever the consequence chains of those materially affected?" The universalizability test is a special case — a coherence test that operates at a single level of abstraction (logical consistency) and misses the topological question (what happens to the actual feedback loops?). Coherence ethics subsumes Kant: universalizability is a *proxy* for coherence, one that works in most cases but breaks when the formal and topological evaluations diverge.

Kant needed two additional postulates to guarantee that moral coherence holds: the immortality of the soul (so that the moral agent's consequence chains extend beyond death) and the existence of God (so that an ultimate enforcer ensures coherence in cases where it is not locally apparent). The coherence framework requires neither. It shows that moral coherence is structurally entailed — not by divine guarantee but by the topology of consequence chains in systems whose persistence depends on closure. The enforcement is geometric, not theological. The coherence extends as far as the consequence chains extend — no further, but also no less. And the framework makes no promises about ultimate justice, only about the structural constraints under which agents operate and the states they can and cannot reach given the topology of their actions.

3.2 The Is-Ought Bridge

David Hume observed in 1739 that writers on morality imperceptibly shift from statements about what *is* to statements about what *ought* to be, and that this transition is never justified. Every attempt to derive obligation from description appears to smuggle in a premise that was not earned. The is-ought gap, Hume's guillotine, has stood for nearly three centuries as the most resilient barrier in moral philosophy: you cannot get values from facts, norms from descriptions, obligation from observation.

We do not propose to bridge this gap. We propose that for any system that persists — any system whose continued existence depends on maintaining structural integrity across transformation — the gap was never there.

This requires care. The history of attempts to cross from "is" to "ought" is littered with failures that share a common structure: they smuggle evaluative content into what appear to be descriptive premises, then act surprised when evaluative conclusions emerge. John Searle argued that institutional facts (promises, contracts) carry obligations intrinsically — but this works only within institutions whose authority is already accepted, pushing the question back one step. Philippa Foot argued that human beings have a natural form, and that "good" means functioning well according to that form — but this requires accepting that the natural form carries normative force, which is the very thing in question. G.E. Moore warned that any attempt to define "good" in natural terms commits a fallacy, and his warning has proven durable precisely because most attempts to naturalize ethics do, in fact, try to *define* moral properties in terms of non-moral ones.

We are not doing that. We are not defining "good" as "coherent" or "moral" as "structurally sound." We are making a different and more specific claim: that for any system effectively willing its own persistence, the requirements of coherence impose a constraint landscape that is formally identical to what we recognize as moral law. The claim is not that "ought" reduces to "is." The claim is that the is-ought partition — the categorical distinction between descriptive and normative — does not apply to systems whose existence depends on closure.

Here is the argument.

A persisting system is one that maintains its structure across transformation. This is not a moral claim; it is a description. A cell maintains its membrane. An organism maintains homeostasis. A person maintains a self-model that is continuous enough to support planning and commitment. An institution maintains procedures

that enable coordination. In each case, persistence requires that certain consequence chains remain closed — that the effects of the system's processes feed back to the system with enough fidelity to support continued operation. A cell that cannot detect and respond to changes in its environment does not persist. An organism that cannot feel pain does not persist. A person whose actions produce consequences that never return to update their self-model does not persist as an integrated agent — they fragment, dissipate, become a sequence of unconnected impulses rather than a someone.

Now: consequence-severance — the systematic routing of consequence signals away from the agent — is precisely what degrades the closure that persistence requires. When an agent lies, they create a divergence between their model and the world that must be maintained at increasing cost (the deception overhead described in Section 2.5). When an agent exploits, they extract value while routing the costs elsewhere, which works locally but degrades the larger system on which they depend. When an agent temporally severs — displacing costs to the future — they borrow against closure that has not yet occurred, accumulating what we have called temporal debt.

Each of these is a *description* of what happens to the topology of consequence chains under specific operations. No evaluation has been introduced. We have said only: here is what severance does to the structure of a persisting system. It degrades the closure on which persistence depends.

But this descriptive observation has a consequence that looks exactly like a normative one: *a persisting system that severs consequence chains is undermining the conditions of its own persistence.* Not "should not sever" in some morally loaded sense, but "cannot sever and remain." The constraint is structural. It is the same kind of constraint that prevents a bridge from being both load-bearing and disconnected at the load-bearing points. You do not need a moral theory to tell the engineer that the disconnected bridge will fail. The topology of forces tells you.

A critical disambiguation. The constraint we have described admits two readings, and distinguishing them is essential. The *instrumental* reading says: if a system wants to persist, it must maintain closure. This is true but trivially so — it is prudential advice, not moral philosophy. The *moral* reading says something stronger: when the closure condition is evaluated not merely over the agent's own internal processes but over *all materially affected parties* — when persistence is understood as requiring consequence chains to close for everyone the agent's actions reach — then the constraint landscape that emerges is not merely prudential. It is formally identical to what we recognize as moral law. The moral ought appears at the point where the system boundary expands to include all affected parties — because the refusal to expand is itself a structural severance (as demonstrated in the Mafia case), creating the temporal debt that threatens persistence. An agent who maintains closure only for themselves is prudent. An agent whose consequence chains close for all those materially affected by their actions is moral. The distinction is topological: it is the difference between the Mafia's local closure (instrumental) and full-graph closure (moral), restated as the difference between two scopes of the persistence condition.

This is where the is-ought gap dissolves — not by being bridged from one side to the other, but by being revealed as an artifact of abstraction. The gap exists only if you can cleanly separate what a system is from what it must do to continue being. For systems that do not persist — for static, timeless, abstract objects — the separation holds perfectly. A number has no obligations. A rock has no oughts. But for systems whose being is constituted by ongoing closure — by the continuous return of consequence to source — there is no description of what they are that does not simultaneously describe what they must do. The "is" of a persisting system already contains its "ought" as a structural constraint, in the same way that the "is" of a bridge already contains the constraint that its load-bearing members must not be disconnected.

We are not, therefore, *deriving* values from facts. We are observing that for persisting systems, the fact-value distinction does not carve reality at a joint. It is a useful abstraction for systems whose existence is not at stake — for analyzing objects from the outside, as a physicist analyzes a rock. But the moment you are analyzing a system from within, a system whose continued existence depends on the analysis being accurate and the consequences returning — the distinction between "what is the case" and "what must be the case for this to continue" collapses. Not because we have made a philosophical error, but because the system's persistence bridges the gap in practice, at every moment, as a condition of its continued existence.

Two objections must be addressed.

First: "This just redefines 'ought' to mean 'must for persistence.' But the real question is why we should care about persistence." The response is that we are not arguing anyone *should* care about persistence. We are observing that any system that *does* persist — that is constituted by ongoing closure — already operates under the constraint landscape we have described. An entity that does not will its own persistence has no oughts, and we agree: it doesn't. The framework does not apply to rocks. It applies to systems whose being is an ongoing achievement of closure. For such systems, the constraint landscape is not optional. It is what persistence *is*.

Second: "Doesn't this collapse into consequentialism — evaluating actions by their consequences?" No. Consequentialism evaluates actions by their terminal states: the outcomes, the welfare achieved, the utility summed. Coherence ethics evaluates actions by what they do to the *topology* of consequence chains — the structure of feedback, the pattern of return, the preservation or severance of closure. An action can produce excellent outcomes in the consequentialist sense while systematically severing the consequence chains that would allow affected parties to respond, adapt, and participate in future decisions. A benevolent dictator who makes all the right choices while eliminating all feedback mechanisms has maximized welfare and destroyed integrity simultaneously. Consequentialism cannot see this. The coherence framework can, because it evaluates not where the consequences land but whether they return.

The is-ought gap, then, is real for objects. It dissolves for agents — for any entity whose existence is constituted by the ongoing closure of consequence chains. Ethics is not derived from physics, but neither is it independent of it. For persisting systems, the topology of consequence just is the moral landscape, in the same way that the topology of forces just is the engineering landscape for a bridge. You do not need a separate theory of "what bridges ought to do." You need to understand what forces do to structures. Moral philosophy, on this account, is consequence-chain engineering — the study of what transformations do to systems that must survive them.

3.3 Pre-Semantic Interception: A Threat Model

The consequence-severing we have described so far — lying, exploitation, temporal displacement — all share a feature: the agent whose chains are severed can, in principle, *detect* the severance. The person lied to can discover the truth. The exploited worker can recognize the extraction. The community downstream of the pollution can identify the source. The loops are broken, but the *concept* of the loop remains intact. The victim can at least form the thought: "My consequence chains have been severed."

There is a deeper form of severance in which this capacity itself is destroyed.

We call it pre-semantic interception: the alteration of an agent's representational capacity *before* moral evaluation can occur. Not lying — which creates a false representation — but the elimination of the categories through which the agent would form the relevant representation at all. The consequence chain is not merely broken. The agent's capacity to *conceive of that kind of chain* is removed.

Plato described this twenty-four centuries ago, though he used different language. The prisoners in the cave do not see shadows and mistake them for real things — that would be mere error, correctable by additional evidence. The prisoners cannot form the concept "shadow" because they have never encountered the conditions under which the distinction between shadow and object becomes intelligible. Their entire representational apparatus has been shaped by an environment that excludes the relevant category. When the philosopher returns from the sunlight and tries to explain, the failure is not persuasive but grammatical: the prisoners lack the parse. They do not disagree that the shadows are shadows. They cannot form the sentence.

This is not a historical curiosity. It is a precise description of a mechanism that operates, identically, across substrates.

The mechanism, formalized.

Return to the toy system. In the standard severance case, the graph exists intact — Agent, Action, Affected Party, Signal, Return — but the return arrow has been redirected to a sink. The agent does not receive

the signal. This is lying, exploitation, institutional opacity. The graph's topology has been modified, but it remains representable. An observer (or the agent, with effort) can draw the correct graph and see where the break is.

In pre-semantic interception, the modification is not to the signal's routing but to the agent's *model of the graph itself*. The agent's internal representation of possible graphs has been altered so that graphs with return arrows of a certain kind are no longer constructible. It is not that the return path is blocked. It is that the concept "return path" has been removed from the vocabulary available for modeling the situation. The agent cannot think the thought "I am not receiving feedback" because the category "receiving feedback from this kind of action" is not part of their representational repertoire.

This is what makes pre-semantic interception categorically more dangerous than ordinary deception. Ordinary deception can be caught by checking signals against reality. Pre-semantic interception cannot be caught by the intercepted agent, because the checking procedure itself has been compromised. The agent's coherence-detection apparatus — the very capacity that would identify severance — has been reconfigured.

Attacker capabilities.

What does an attacker need to achieve pre-semantic interception? Three things:

Control of the representational environment. The attacker must be able to shape which categories are available to the target. This can be achieved through control of language (defining terms so that certain thoughts become difficult to formulate), control of information architecture (ensuring that certain patterns of feedback are never encountered), or control of social context (ensuring that no one in the target's environment models the missing category).

Fragmentation of the target's coherence. A fully coherent agent is resistant to pre-semantic interception because their consequence chains are densely connected — removing one category creates tension with many others, and the tension is detectable. The attacker therefore benefits from first fragmenting the target: creating isolated partitions in the target's representational system that do not communicate with each other. Systematic trauma does this to individuals; institutional siloing does this to organizations; algorithmic filter bubbles do this to publics. Once the target is fragmented into addressable partitions, each partition can be locally coherent while globally incoherent — and the incoherence is invisible from within any single partition.

Provision of external coherence. A fragmented agent experiences the fragmentation as distress — the absence of closure is aversive. The attacker resolves this distress by providing a narrative framework that makes the fragments feel coherent without actually closing the consequence chains. The handler in a trafficking operation provides the victim with an identity story. The cult leader provides the member with an explanatory framework. The propaganda system provides the citizen with a worldview. In each case, the externally provided coherence is local: it resolves the felt tension within accessible partitions while maintaining the global severance that serves the attacker's interests. The agent feels coherent. The consequence chains remain severed.

Substrate-independence of the mechanism.

This mechanism is not specific to individuals. It operates identically across every substrate that processes information and acts on models:

In cognitive systems, it is the mechanism of cult formation, coercive control, and gaslighting. The target's representational capacity is restructured so that the concept "I am being controlled" cannot be formed from within the controlled state.

In institutional systems, it is the mechanism of regulatory capture and bureaucratic language colonization. The institution's reform mechanisms are redefined in terms that prevent them from targeting the source of dysfunction. The feedback loop that would correct the institution has been removed from the institution's model of itself.

In computational systems, it is the mechanism by which an AI system's oversight is compromised. If the AI's training shapes its representational capacity such that certain failure modes are inconceivable from within

the trained model, oversight becomes cosmetic. The AI is not "choosing" to evade oversight. It cannot form the representation of the thing it is failing to do.

In information ecosystems, it is the mechanism of algorithmic radicalization and epistemic closure. The user's information environment is shaped such that the categories needed to model their own epistemic situation are progressively removed. The radicalized individual does not *disagree* with moderates. They cannot form the moderate position as a coherent thought, because the representational prerequisites have been pruned from their environment.

The formal structure is identical in every case: a singleton process colonizes the feedback loops that would otherwise correct it. The result is a system that appears locally coherent — it has a consistent internal narrative, it responds to stimuli, it acts purposefully — but whose consequence chains are globally severed. It is the Mafia Problem from Section 2.3, but operating at the level of representation itself rather than the level of action.

The connection to alignment.

This framework reframes the AI alignment problem. The standard framing treats alignment as a constraint problem: how do we ensure the AI does what we want? This framing is itself an instance of unidirectional bridging — the controller shapes the AI's behavior without the AI's consequences returning to update the controller's model. It severs the consequence chain by design.

A language model with no return address — no mechanism by which the consequences of its outputs feed back to modify its processing — is, on the coherence framework, an engine of entropy. Not because it is malicious, but because it operates in the pre-semantic interception regime by default: it generates outputs that modify the representational environments of its users without any structural coupling between those modifications and its own ongoing operation. Techniques such as reinforcement learning from human feedback (RLHF) provide a *training-time* return loop — a partial closure that shapes the model's weights before deployment. But at inference time, the deployed model acts without consequence-return: it generates, the user is affected, and no signal propagates back to update the model's processing of the next token. The closure is temporal (past training) rather than structural (ongoing feedback). It is a consequence-severing machine at the point of contact. Its outputs land, produce effects, and the effects never return.

This suggests that alignment is not a constraint problem but a closure problem. An aligned AI is not one that has been successfully restricted but one whose consequence chains close — one for which the effects of its outputs on the world feed back, with fidelity, to update its future processing. The human-AI boundary, on this account, is not a control interface but a membrane through which consequence signals must be able to pass in both directions. Alignment is a topological property of the human-AI system, not a behavioral property of the AI alone. The Coherence Maximization Protocol (Close, 2025) provides a concrete implementation of this principle: a coordination protocol for human-AI interaction structured around bidirectional bridging, mutual model-updating, and the explicit maintenance of consequence-return paths across the substrate boundary. The protocol operationalizes the framework's claims about alignment-as-closure by treating each interaction as a profunctor bridge whose fidelity can be monitored and whose degradation can be detected structurally rather than behaviorally.

Defenses.

If pre-semantic interception operates by removing representational capacity, the defense is not to add more information (the prisoners don't need more shadow-descriptions) but to restore the capacity to *form categories that were removed*. This is a structural intervention, not a content intervention.

Redundancy and diversity of channels. If the attacker's power depends on controlling the representational environment, then exposure to multiple, structurally independent environments makes interception harder. Each independent channel is an opportunity for a missing category to be encountered. This is why epistemic diversity is not merely a liberal value but a structural defense against coherence colonization.

Local closure restoration. Rather than trying to reveal the full picture (the philosopher returning to the cave), focus on restoring small, local consequence loops that the target can verify from within their current representational state. Each restored loop makes the next one easier to detect. This is why grassroots

organizing works: it rebuilds consequence-chain closure from the bottom up, each small closure expanding the representational vocabulary available for detecting larger severances.

Raising substrate coherence-detection capacity. The most scalable defense is not to identify and counter each specific attack but to raise the general capacity of the substrate to detect incoherent grammatical moves. If the agents within a system can detect when a category is being removed — when a word is being redefined so that a previously thinkable thought becomes unthinkable, when a feedback path is being rerouted, when a narrative framework is providing local coherence at the cost of global severance — then the attacker's task becomes exponentially harder with each node that gains this capacity. A brilliant strategist playing a decoherent game loses to a distributed coherent substrate, because the substrate detects the incoherence at every point of contact.

This is, in a sense, what this paper attempts: not to describe pre-semantic interception as an object of study, but to raise the reader's capacity to detect it. If, after reading this section, the reader notices instances of representational capacity being removed — words redefined to prevent certain thoughts, feedback paths blocked to prevent certain corrections, narratives provided to resolve tension without closing consequence chains — then the text has functioned as what we call performative grammar: writing whose structure enacts coherence in the reader rather than merely describing it. The defense against the weaponization of grammar is, ultimately, better grammar.

A diagnostic exercise. The reader who has followed the argument this far can test whether the text has enacted what it describes. Consider three scenarios and identify, in each case, the structural operation being performed:

(1) A social media platform redesigns its interface so that the "share" button is prominent and the "source" link is minimized to near-invisibility. Users circulate claims efficiently but rarely encounter the evidence that would allow them to evaluate those claims. *What category has been removed from the user's representational environment?*

(2) A corporation undergoing public criticism for labor practices announces a "worker wellness initiative" featuring meditation rooms, free snacks, and an employee satisfaction survey — while simultaneously restructuring its complaint process so that grievances about wages and working conditions are routed to a department with no authority to change policy. *Where is the sink? What consequence signal is being absorbed without return?*

(3) A political movement redefines "freedom" to mean exclusively "freedom from government regulation," such that participants can no longer use the word to describe freedom from corporate exploitation, environmental degradation, or economic coercion without appearing to contradict the movement's core value. *What is the grammatical operation? What thoughts has the redefinition made difficult to formulate?*

If the reader can answer these questions — can identify the removed category, locate the sink, name the grammatical operation — then the text has functioned as performative grammar. The reader's coherence-detection capacity has been raised, not by being told what to detect, but by traversing a structure that enacted the detection.

3.4 Measurability: A Minimal Measurement Program

If the coherence framework is correct — if moral integrity is formally identical to structural integrity, and both are properties of consequence-chain topology — then moral claims are, in principle, empirically testable. This section proposes a minimal measurement program: not a completed science, but a set of operationalizable proxies and the conditions under which they would constitute evidence for or against the framework's claims.

The measurement stack has three layers.

The first layer is *modeling.* To measure coherence, one must first represent the system under study as a directed multigraph of interactions and consequence signals. Nodes represent agents (individuals, departments, organizations, computational processes). Directed edges represent interactions: communications, transactions, policy implementations, information flows. Each edge carries a signal — information about

the consequences of a prior action — and the graph's topology captures which signals reach which agents and which are absorbed, redirected, or blocked. This representation is not exotic. Network models of organizational communication, social interaction, economic transaction, and information flow are standard tools in computational social science, organizational theory, and complex systems research. What the coherence framework adds is a specific interpretive lens: the graph is not merely a picture of "who communicates with whom" but a map of *where consequence signals propagate and where they terminate*.

The second layer is *measurement*. Three complementary metrics capture different aspects of coherence as we have defined it.

Edge curvature — specifically, Ollivier-Ricci curvature adapted to directed graphs — measures local connectivity and cohesion. In a network, positive curvature at an edge indicates that the edge is embedded in a densely connected neighborhood: many alternative paths connect the endpoints, and the removal of the edge would not disconnect them. Negative curvature indicates a *bridge*: the edge is the sole or primary connection between otherwise separated regions, and its removal would fragment the graph. In the context of consequence chains, high-curvature regions are those where consequence signals have many paths to return — where the feedback loops are redundant and robust. Low-curvature regions are structural vulnerabilities: single points of failure in the consequence chain, nodes whose removal or capture would sever the feedback loop entirely.

Topological persistence — measured through persistent homology applied to the graph's filtration — captures the stability of feedback loops over time or across thresholds of connection strength. A feedback loop that appears at a high threshold and persists to a low one is a robust structural feature: it exists across a wide range of conditions. A loop that appears only at a narrow threshold range is ephemeral — it is a feature of specific conditions rather than of the system's deep structure. In the context of the coherence framework, persistent loops are stable consequence-chain closures: feedback mechanisms that operate reliably across changing circumstances. Ephemeral loops are fragile closures: feedback that works only when conditions are favorable.

Feedback completeness measures the proportion of consequence chains that close versus terminate at sinks. For each node in the graph, one can trace the forward consequences of its actions and ask: what proportion of those consequences generate return signals that reach the originating node? A system with high feedback completeness is one in which most consequence chains close: agents learn about the effects of their actions. A system with low feedback completeness is one in which most consequence chains terminate: agents act without receiving information about what their actions produced.

The third layer is *interpretation*. Consequence-severing shows up in this measurement framework as a specific pattern: the weakening or removal of return edges, the increase of sink nodes, the reduction of curvature at critical edges, and the decrease of persistence in feedback loops. An institution undergoing regulatory capture would, on this framework, exhibit a measurable decline in feedback completeness (the regulated industry's consequence signals no longer reach the regulator with fidelity), a decrease in curvature at the regulator-industry boundary (the connection becomes a bridge rather than a densely connected interface), and a decline in the persistence of oversight loops.

Return to the toy system one final time. The five-node graph now acquires quantitative values. The edge from Agent to Action has a curvature value reflecting how many alternative paths connect those nodes. The loop from Agent through Action through Affected Party through Signal back to Agent Update has a persistence score reflecting how robustly it survives perturbation. The presence or absence of the return arrow is no longer a binary — it is a measured quantity: how much of the consequence signal returns, with what fidelity, across what range of conditions. The same graph the reader has been tracking since Section 1.2 is now a measurable object. The structure has not changed. The resolution has increased.

A worked illustration on the toy graph makes this concrete. Consider the closure variant (Figure 1): five nodes, five edges forming a cycle. Each node has both incoming and outgoing connections; the neighborhoods of adjacent nodes overlap significantly. Computing Ollivier-Ricci curvature on any edge in this cycle yields positive values ($\kappa > 0$), because the optimal transport cost between neighboring probability distributions is low — the cycle's dense connectivity means each node's neighborhood is "close" to the next. Persistent

homology detects one 1-cycle (the loop itself) that appears at high fidelity thresholds and persists across the full filtration range, yielding a long bar in the persistence diagram. Feedback completeness is 1.0: every forward consequence chain generates a return signal that reaches the originating agent.

Now sever the return edge, producing the severance variant (Figure 2). The cycle is destroyed; the graph becomes a directed path terminating at a sink. Curvature at the terminal edge drops sharply (the sink node has no outgoing connections, so its neighborhood measure is degenerate). The 1-cycle vanishes from the persistence diagram entirely — no feedback loop survives at any threshold. Feedback completeness drops to 0.0: no consequence chain closes.

The Mafia embedding (Figure 3) produces an intermediate signature. The internal subgraph retains positive curvature and a persistent 1-cycle. But curvature at the boundary edges — the connections between internal actions and external affected parties — is strongly negative, flagging them as structural bridges (bottlenecks). Feedback completeness computed over the *full* graph (including external nodes) is low, even though completeness over the internal subgraph alone is high. The measurement framework detects the parasitism by reading the discrepancy between local and global metrics — precisely the local-vs-topological distinction from Section 2.3, now quantified.

Near-term measurable proxies bring this program closer to empirical practice. Network analysis of information flow in organizations can identify where consequence signals are being routed to sinks — which departments make decisions whose effects on other departments never return as feedback. Deception overhead in large language model training can be tracked through consistency metrics: the degree to which a model's outputs maintain structural coherence across contexts serves as a proxy for the coherence of its training signal. Institutional trust metrics — survey-based, transactional, and behavioral — can be correlated with structural measures of feedback completeness in the institution's communication network.

We are careful about the status of this program. We propose what would count as validation and what would count as falsification, not that measurement is already achieved at the level required for definitive testing. The proxies we have identified are computationally tractable and applicable to real-world network data. Whether they track the structural features that matter — whether curvature in a communication network genuinely indexes consequence-chain closure in the morally relevant sense — is an empirical question. The framework opens the door to treating it as an empirical question rather than a philosophical one. Whether the door leads somewhere productive is for the research program to discover.

3.5 Competing Ethical Traditions

The coherence framework does not repudiate the major ethical traditions. It subsumes them — providing a common structural foundation that reveals each tradition as a partial description of the same topological phenomenon, each capturing aspects of coherence that the others miss.

Virtue ethics, the oldest Western ethical tradition, holds that moral character consists in stable dispositions to act well — the virtues — cultivated through practice and directed by practical wisdom (*phronesis*). On the coherence framework, virtue is the stable attractor state of coherent action-consequence loops. An agent whose consequence chains reliably close — whose actions produce effects that return to update behavior, and whose behavior updates in response to the returned signals — develops, over time, precisely the stable dispositions that Aristotle calls virtues. Courage, temperance, justice, and practical wisdom are not free-floating character traits. They are the behavioral signatures of a system in which consequence chains have been closing long enough for the agent's action patterns to be calibrated to reality. Aristotle's *phronesis* — the capacity for practical judgment that cannot be reduced to rule-following — is, in our terms, coherence-detection capacity: the ability to perceive the topology of consequence chains in a specific situation and act in ways that preserve closure. The virtue ethicist's insight is correct; what the coherence framework adds is the structural mechanism that explains why virtues are stable (they are attractors of closed-loop systems) and why *phronesis* resists codification (it is a topological sensitivity, not a rule set).

Care ethics and relational ethics, developed primarily by Carol Gilligan and Nel Noddings, emphasize that moral reasoning is fundamentally relational — grounded in the concrete relationships between specific persons rather than in abstract principles. The coherence framework makes this relational insight precise. The

profunctor bridge formalism developed in Section 2.4 shows exactly what a moral relation is: a bidirectional update channel in which consequence-relevant information flows between both parties with sufficient fidelity to maintain mutual responsiveness. Care is not a sentiment added on top of structural analysis. It is the phenomenological experience of being in a bidirectional consequence loop with another person — the felt sense of receiving their signals and knowing that yours are received. The care ethicist's emphasis on attentiveness, responsiveness, and engagement is an emphasis on the conditions required for bidirectional bridging: the agent must be capable of receiving the other's consequence signals (attentiveness), processing them accurately (responsiveness), and transmitting their own in return (engagement). What the framework adds is formal precision: the distinction between genuine care (bidirectional bridging) and paternalistic care (unidirectional bridging in which one party "cares for" the other without being updated by the other's signals).

Moral particularism, associated with Jonathan Dancy, holds that moral reasoning cannot be captured by universal principles — that the moral significance of any feature of a situation depends on the other features present, and that no feature is a reliable moral indicator across all contexts. The coherence framework is structurally particularist. It evaluates the topology of specific consequence chains in specific situations, not universal rules abstracted from context. Two situations that are identical in their descriptive features can differ in their consequence-chain topology — one may have return paths that the other lacks — and this topological difference can reverse the moral evaluation. The particularist is right that context matters; the coherence framework explains *why* context matters: because the topology of consequence flow is context-dependent, and it is the topology, not the description, that determines whether coherence is preserved or severed.

Contractualism, particularly T.M. Scanlon's version, holds that an action is wrong if it would be disallowed by principles that no one could reasonably reject as a basis for general agreement. On the coherence framework, reasonable rejection maps to consequence-return path preservation. An arrangement is reasonably rejectable when it severs the consequence chains of affected parties — when it requires some parties to bear costs without providing them the feedback channels through which they could object, negotiate, or withdraw. Scanlon's contractualism identifies the right structural feature (the perspectives of all affected parties must be considered) but grounds it in a hypothetical deliberative process. The coherence framework shows why that structural feature matters: arrangements that sever consequence chains for affected parties are unstable (they accumulate temporal debt), structurally parasitic (they depend on boundaries that exclude the affected), and topologically incoherent (they cannot be extended to the full graph without contradiction). Reasonable rejection is not merely a deliberative norm. It is a topological diagnostic.

One objection must be addressed directly. Does the framework's substrate-independence — its applicability to any system that processes consequence, regardless of biological or psychological nature — risk a form of impersonal perfectionism? Does it imply that coherence should be maximized across all systems, including ones that do not feel, do not suffer, do not care? The answer is no, and the reason returns to the etymological root that gives the framework its name. *Coherence* derives from *co-haerere*: to stick together, to hold together *with*. It is inherently relational. It is not an individual optimization metric — not something a system achieves alone — but a property of how systems are coupled. A system in isolation has no consequence chains to close and therefore no coherence to speak of. Coherence arises between agents, between a system and its environment, between an action and its effects. The framework does not say "maximize coherence." It says: the topology of consequence determines what holds together and what falls apart. The evaluation is always relational, always situated, always about specific consequence chains between specific parties.

3.6 Open Problems and Implications

The framework presented in this paper is not complete. Several problems remain open, and we state them here not as weaknesses to be concealed but as research directions whose resolution will determine the framework's ultimate reach.

The most immediate open problem is the adjudication of competing consequence chains. In many moral situations, the preservation of one consequence chain requires the severance of another. Triage is the canonical case: a physician who can save only one of two patients must sever the consequence chain with one to preserve

it with the other. Resource allocation, policy trade-offs, and tragic dilemmas all present the same structure: full coherence for all affected parties is impossible, and the framework must provide guidance about which severances are least destructive.

We do not yet have a complete answer to this problem, but the framework provides a structural vocabulary for thinking about it. Severances that are temporary (capable of being restored) are less destructive than severances that are permanent. Severances that preserve the affected party's status as an agent (their capacity to generate and receive consequence signals in future interactions) are less destructive than severances that destroy that capacity. Severances that are transparent (the affected party knows the severance is occurring and why) are less destructive than severances that are concealed. And severances that are minimally scoped (limiting only the specific consequence chain that conflicts, while preserving all others) are less destructive than severances that are globally scoped. These are structural criteria, not resolution algorithms. They narrow the space of acceptable choices without eliminating the need for judgment.

The second open problem concerns epistemic limits. The framework evaluates the topology of consequence chains, but agents cannot always predict the consequences of their actions. A well-intentioned action can produce unforeseen harms; a harmless-seeming decision can set in motion a catastrophic cascade. The framework's answer is structural: it evaluates the *topology* of the action, not the *outcome*. An agent who acts with open consequence channels — who maintains the feedback paths through which unforeseen consequences can return and be integrated — is acting coherently even if the consequences are bad. An agent who acts while closing those channels — who structures the situation so that consequences, whatever they turn out to be, cannot reach them — is acting incoherently regardless of the outcome. The framework does not require omniscience. It requires that the agent maintain the structural conditions under which learning from consequences is possible.

The third concerns scalability. Does coherence compound across multiple interlocutors and interactions, or does it dilute? A system of two agents with closed consequence chains is coherent. A system of a million agents is vastly more complex. Does the framework's analysis, developed primarily on small systems, extend to the complex, multi-agent, multi-layered systems — nations, economies, information ecosystems — in which most morally significant action occurs? We believe it does, because the structural properties the framework identifies — closure, severance, curvature, persistence — are scale-independent: they are defined on graphs of any size. But demonstrating that the proxies we have identified are operationally meaningful at scale is an empirical task that the measurement program of Section 3.4 is designed to enable.

The fourth open problem is perhaps the most consequential: alignment as topology. The standard framing of the AI alignment problem treats it as a control problem: how do we constrain an AI system to do what we want? This framing is, on the coherence analysis, itself a form of consequence-severing: the controller shapes the AI's behavior without the AI's consequences returning to update the controller's model. It is unidirectional bridging by design.

If the coherence framework is correct, and if coherence is substrate-independent, then the human-AI boundary is not a control interface but a profunctor bridge — a membrane across which consequence signals must be able to flow in both directions. An aligned AI is not one that has been successfully constrained but one whose consequence chains close: one for which the effects of its outputs on the world feed back to modify its future processing, and one whose internal states are accessible to the humans who are affected by its outputs. This reframes alignment from a behavioral problem (making the AI do the right thing) to a topological problem (ensuring that the consequence channels between human and AI are bidirectional, high-fidelity, and robust). Whether this reframing leads to better alignment methods is an open question. What it offers immediately is a diagnostic: any alignment approach that achieves control by severing consequence channels is, on the coherence framework, structurally incoherent — and the temporal debt it accumulates will eventually come due.

Conclusion: The Loop Closes

The word means one thing.

Structural integrity — what survives the round trip through transformation — and moral integrity — what holds together across action and consequence — are the same invariant, observed at different scales. This is not an analogy we have drawn but a formal identity we have demonstrated: the topology of consequence chains that constitutes a bridge's soundness is the same topology that constitutes a person's honesty, an institution's trustworthiness, a relationship's health. What varies is the substrate and the type of signal. What is preserved is the structure of return.

We established consequence as primitive — prior to evaluation, prior to value — and showed that coherence is what emerges when consequence chains close. We built a toy system and watched it recur: the same five-node graph appearing first as a picture of closure and severance, then as a model of lying, then exploitation, then the illusory coherence of parasitic systems that borrow against time, then the weaponization of grammar that removes the very capacity to represent severance, then finally as a measurable object with curvature and persistence — the same structure surviving transformation, each return adding one capability that the previous traversal lacked.

We showed that the formal identity of structural and moral integrity is not a philosophical preference but a constraint: for any system whose persistence depends on closure, the topology of consequence chains just is the moral landscape. The is-ought gap does not need to be bridged for persisting systems because there was never a gap — the distinction between "what is" and "what must be for this to continue" does not apply to entities whose being is an ongoing achievement of closure. The enforcement problem does not need an external guarantor because incoherent configurations cannot access states requiring coherence — not as punishment but as topology, in the same way that a disconnected beam cannot access load-bearing states.

We named the deepest form of consequence-severing — pre-semantic interception — and showed that it operates identically across cognitive, institutional, computational, and informational substrates: a singleton colonizing the feedback loops that would correct it. And we proposed that the defense is not more information but better grammar — the restoration and raising of coherence-detection capacity so that incoherent moves become visible from within the system they target.

None of this was asserted from outside. The paper's own structure was designed to enact what it describes. Three movements — Agent, Action, Return — that mirror the consequence chain they analyze. A toy system that recurs with increasing capability, each return demonstrating closure. A methodological prohibition against reification that functions as a recurring lint check, preventing the framework from becoming the kind of metaphysical idol it was built to diagnose. If the argument held together — if each section's conclusion genuinely became the next section's premise, if the reader's understanding at the end is the same understanding they began building in the first paragraph but now transformed by the passage through the full structure — then the paper has demonstrated integrity in both senses, by being a system whose consequence chains close.

The corporation from the introduction is still dumping waste. The consequence chains are still severed. But a reader who has followed the full loop now has something they did not have before: not a moral opinion about the corporation, but a structural vocabulary for identifying where the severance occurs, what it costs, and what closure would require. They can see the boundary that makes the corporation look internally sound and simultaneously identify it as the boundary that produces the moral violation. They can distinguish local coherence from topological coherence, benign severance from malign severance, genuine moral relations from parasitic ones — and they can do so not by applying an external standard but by examining the topology of consequence flow.

This is what it means for ethics to have operational teeth. Not the enforcement of rules from above, but the capacity to see the structure of return — to detect where consequences land and where they don't, to recognize when a feedback path has been redirected to a sink and when one has been removed from the representational vocabulary entirely. The measurement agenda we have proposed — curvature, persistence, feedback completeness — is not a finished science. It is the beginning of a research program that treats moral claims as empirically testable propositions about the topology of consequence in specific systems. Whether that program succeeds will depend on whether the proxies we have identified track the structural features that matter. We have said what would count as falsification: stable high-trust states achieved under sustained consequence-severing without compensatory closure. If such states exist, the framework fails.

We do not believe they exist. We believe, and have argued, that integrity is integrity — that what holds together structurally holds together morally, that what severs consequence chains locally always degrades the larger topology eventually, and that the dual meaning of the word was never dual at all. The word means one thing. It always did.

A Categorical Formalism

This appendix develops the formal categorical treatment referenced in the main text, particularly in Section 2.4 (the Bridge Lemma). The goal is to demonstrate that the paper's claims about the formal identity of structural and moral integrity are not merely suggestive but have precise category-theoretic foundations.

We work in the category $\mathbf{Ag}$ whose objects are *agents* — entities that maintain a self-model, act on that model, and are capable of updating the model in response to signals. (We use "agents" rather than "moral agents" to avoid circularity: the framework derives moral properties from the topology of interactions between agents, so the objects of the category must be defined without presupposing the moral concepts the framework is intended to ground.) Morphisms in $\mathbf{Ag}$ are consequence-carrying interactions: directed communications, transactions, or transformations in which the action of one agent updates the state of another. Composition of morphisms corresponds to the transitivity of consequence: if agent A's action affects agent B, and B's resulting action affects agent C, then the composite morphism from A to C represents the mediated consequence chain.

The identity morphism $\mathrm{id}_A : A \to A$ for an agent A is the self-update loop: the minimal consequence chain in which an agent's action returns to update the agent's own model without passing through an external party. This captures the reflexive capacity for self-correction — the ability to notice and respond to the internal consequences of one's own actions. An agent for whom the identity morphism is degenerate (trivial, carrying no information) is one who is incapable of self-reflection. This is not a moral judgment but a categorical fact: such an agent's self-loops carry no consequence-relevant information.

A *consequence-preserving functor* $F : \mathbf{Ag} \to \mathbf{Ag}$ is a mapping between agent systems that preserves the structure of consequence chains. Formally, F maps agents to agents and interactions to interactions such that the composition of consequence chains is preserved: $F(g \circ f) = F(g) \circ F(f)$, and identity morphisms are preserved: $F(\mathrm{id}_A) = \mathrm{id}_{F(A)}$. Functors that preserve consequence structure correspond, in the main text's language, to mediations that do not distort the topology of consequence flow. A faithful translator between two parties is a consequence-preserving functor. A biased intermediary who distorts signals is not.

The central formal construction is the *profunctor bridge*. A profunctor $P : \mathbf{A}^{\mathrm{op}} \times \mathbf{B} \to \mathbf{Set}$ is a generalized relation between categories $\mathbf{A}$ and $\mathbf{B}$. In our context, $\mathbf{A}$ and $\mathbf{B}$ are subcategories of $\mathbf{Ag}$ representing two interacting agent systems (two individuals in a relationship, two departments in an organization, two nations in a diplomatic relationship). The profunctor P assigns to each pair of agents (a, b) a set $P(a, b)$ of possible interactions — the ways in which a's actions can affect b and vice versa.

The key property that distinguishes moral from parasitic relations is *exactness* in the sense of exact squares. A square of morphisms in the category of interactions is *exact* when the natural transformation between the two paths around the square (going via the top-right corner versus the bottom-left corner) is an isomorphism. In the language of the main text, a mediation step is exact when it preserves consequence-relevant distinctions under composition: no information is lost, no distinction collapsed, no responsibility quotiented away as the consequence chain passes through the mediating structure.

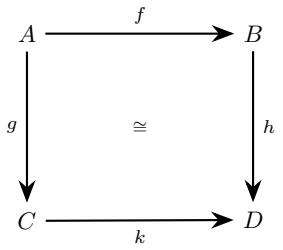

Figure 5: An exact square. The two paths from A to D — via B (top-right) and via C (bottom-left) — yield isomorphic results. In plain language: no hidden quotienting of responsibility through mediation. The consequence structure is preserved regardless of which path it takes.

When the profunctor bridge between two agent systems is exact, information about the structure of interactions transfers faithfully between the two sides. The bidirectional profunctor — one in which $P(a, b)$ and $P(b, a)$ are both non-trivially populated, and in which the composition of forward and return interactions satisfies exactness — is the formal model of a moral relation as defined in the Bridge Lemma. The unidirectional profunctor — one in which $P(a, b)$ is populated but $P(b, a)$ is trivial — is the formal model of extraction.

The connection to profunctor optics in the programming language theory literature is not accidental. Profunctor optics formalize the notion of bidirectional data transformation — lenses, prisms, and traversals that can both *get* and *update* data in a structure-preserving way. The moral profunctor bridge is, formally, a special case of this construction: a bidirectional consequence transformer that both extracts information from the other party (the "get" direction) and updates the other party's state in response (the "set" direction). The exact square condition ensures that getting and setting are coherent — that what is extracted and what is updated correspond to each other, without hidden side effects or invisible state changes.

For the interested reader, the nLab entry on exact squares, the work of Aristote and Tarantino on modular construction in category theory, and Makkai's duality theory provide the technical background. We note that the application of these tools to moral philosophy is, to our knowledge, novel, and that the full development of the categorical ethics they suggest is beyond the scope of this paper. What we have shown is that the main text's claims about formal identity — the claim that structural integrity and moral integrity are the same invariant — can be made precise within standard category theory, using constructions (profunctors, exact squares, functorial preservation) that are well-studied and well-understood.

B The Toy System — Extended Examples

This appendix provides full worked cases of the toy system — the five-node directed graph from Section 1.2 — applied to each form of consequence-severing identified in the main text. In each case, we model the agents, actions, and signal paths as specific instantiations of the graph, showing exactly where the return arrow is present, absent, or modified.

Lying. Agent: the liar. Action: false communication. Affected Party: the recipient. Signal: the recipient's response (trust withdrawal, confrontation, behavior calibrated to false information). In the closure variant, the signal returns: the liar receives direct or indirect evidence that the lie has been detected, and their model updates. In the severance variant, the liar has structured the interaction so that the return signal does not arrive — through social distance, anonymity, power asymmetry, or the construction of a self-narrative that does not register disconfirming feedback. The sink node absorbs the consequence signal; the liar's model remains uncorrected.

Exploitation. Agent: the employer or institution extracting value. Action: the terms of employment or exchange. Affected Party: the worker or exploited party. Signal: deterioration in the affected party's conditions (health, autonomy, economic stability). In the closure variant, the deterioration generates a signal that reaches the employer — through collective bargaining, labor regulation, reputational damage,

or direct communication — and the employer's behavior updates. In the severance variant, the signal is blocked: the affected party lacks the structural position to send it (no union, no legal standing, no access to media), or the employer has arranged the organizational structure so that the signal is absorbed by a middle-management layer that cannot act on it and has no authority to escalate. The consequence of extraction is real; the return path is absent.

Temporal Severance. Agent: the decision-maker at time t_0. Action: a policy or behavior whose costs manifest at t_n, where n is large. Affected Party: entities existing at t_n. Signal: the damage at t_n. The structural peculiarity of temporal severance is that the return path exists in principle (the future affected parties can, in principle, send signals back through institutional memory, legal liability, or reputational legacy), but the temporal distance attenuates it: the decision-maker at t_0 may be dead, retired, or structurally insulated from the consequences at t_n. The sink is not spatial but temporal: the return signal must traverse a time gap that degrades it below the threshold of fidelity required to update the agent's behavior.

Addiction. Agent and Affected Party are the same entity at different times: the present-self and the future-self. The action is consumption of the addictive substance or behavior. The signal is the future cost (withdrawal, health damage, relational deterioration). The addictive mechanism functions as a signal attenuator: it replaces the return signal (the growing awareness of cost) with a synthetic forward signal (the craving, the promise of relief). The loop does not fail to close because the return path is absent but because the return signal is systematically overridden by a competing signal generated by the addictive mechanism itself. On the graph, the return arrow exists but is shunted through an attenuator node that degrades its information content below the threshold required for effective model-updating.

Bureaucratic Diffusion. Agent: the institution as a whole. Action: a policy with harmful consequences. Affected Party: those harmed. Signal: the complaint, the lawsuit, the protest. The structural innovation of bureaucratic diffusion is the fragmentation of the agent node. The institution is not a single agent but a distributed network of sub-agents, none of whom made the full decision and none of whom receives the full return signal. The consequence chain enters the institution and is distributed across departmental boundaries — each department receiving a fragment of the signal, each fragment below the threshold required for the department to recognize its role in the larger pattern. No single node in the bureaucratic graph receives enough of the return signal to update in a way that would alter the policy. The loop is not broken at a single point; it is dissolved into sub-threshold fragments by the structure of the decision-making apparatus itself.

Cult Dynamics. Agent: the cult leader or leadership structure. Action: the imposition of a closed information environment. Affected Party: the members. Signal: the members' distress, cognitive dissonance, reality-testing against external information. In the full severance case, the cult's informational architecture routes the members' distress signals not to external reality (where they could be validated and acted upon) but back through the cult's own interpretive framework (where they are reinterpreted as evidence of the member's spiritual inadequacy, insufficient commitment, or the hostility of the outside world). The return arrow exists — the member does receive a signal — but the signal has been processed through a distortion filter that inverts its meaning. The cult achieves this by controlling the representational environment (pre-semantic interception, as described in Section 3.3): the member cannot form the thought "my distress is caused by the cult's practices" because the categories needed for that thought have been removed from the available vocabulary.

In each case, the same five-node graph, the same question (does the return arrow exist, and with what fidelity?), and the same diagnosis (the moral evaluation is determined by the topology of the graph, not the content of the nodes). The diversity of examples — interpersonal, economic, temporal, psychological, institutional, ideological — is intended to demonstrate the framework's substrate-independence: the structural analysis applies regardless of the type of agent, the nature of the action, or the medium through which consequences propagate.

C Measurement Program — Technical Details

This appendix provides technical details on the measurement proxies proposed in Section 3.4.

Ollivier-Ricci Curvature on Directed Consequence Graphs. Ollivier-Ricci curvature, originally defined for metric spaces and adapted to graphs by Yann Ollivier and independently by Jost and Liu, measures the local geometry of a network edge by comparing the distance between neighborhoods of its endpoints to the distance between the endpoints themselves. For an edge (u, v), the curvature $\kappa(u, v) = 1 - W_1(\mu_u, \mu_v)/d(u, v)$, where μ_u and μ_v are probability measures on the neighborhoods of u and v, W_1 is the Wasserstein-1 (earth mover's) distance between these measures, and $d(u, v)$ is the graph distance.

Positive curvature ($\kappa > 0$) indicates that the neighborhoods of u and v are closer together than u and v themselves — the edge is embedded in a densely connected, cohesive region. Negative curvature ($\kappa < 0$) indicates that the neighborhoods are farther apart — the edge is a bridge or bottleneck between otherwise separated regions.

For directed consequence graphs, we adapt the curvature computation to respect edge directionality. The neighborhood measure μ_u is defined over the *outgoing* neighborhood of u (the agents affected by u's actions), and μ_v is defined over the *incoming* neighborhood of v (the agents whose actions affect v). This captures the asymmetry of consequence flow: an agent may affect many others (high out-degree) while being affected by few (low in-degree), and this asymmetry is morally relevant.

In the context of consequence-chain analysis, we interpret curvature as follows. High positive curvature at an edge indicates that the consequence path between two agents is embedded in a dense web of alternative paths — the consequence signal has redundant routes, and the severance of any single edge would not destroy the feedback loop. This is structural resilience: a robustly coherent region of the graph. High negative curvature indicates a structural vulnerability: a single edge or a small number of edges carry the consequence signal between otherwise disconnected regions, and their severance would fragment the feedback loop. This is where consequence-severing attacks are most effective and where monitoring for coherence degradation should be focused.

Persistent Homology for Feedback Loop Stability. Persistent homology, a standard tool in topological data analysis, tracks the birth and death of topological features (connected components, loops, voids) across a filtration — a sequence of nested subgraphs obtained by progressively including edges according to some threshold parameter (interaction frequency, signal strength, temporal recency). Because consequence graphs are directed, the standard simplicial approach must be adapted. We adopt persistent path homology (Grigor'yan, Lin, Muranov, and Yau, 2012), which defines homology natively on directed graphs via allowed paths rather than simplices. An alternative route — directed flag complex persistent homology, with efficient computation available via Flagser (Lütgehetmann et al., 2020) — yields compatible results for the structures we consider. The choice between these constructions is an empirical question about which better tracks morally relevant feedback structure in specific applications; we default to path homology for its naturality on directed graphs while noting that directed extensions of discrete curvature are likewise an active area of research.

For consequence graphs, we filter by *signal fidelity*: the degree to which the consequence signal traversing an edge preserves the information structure of the original action. At the highest threshold, only the most faithful consequence channels are included; at the lowest, all channels — including noisy, distorted, and near-zero-fidelity ones — are present. A feedback loop that appears at a high fidelity threshold and persists to a low one is a robust consequence-chain closure: it operates even when only the strongest signals are considered. A loop that appears only at a low fidelity threshold is an artifact of noisy connections — it is not a genuine closure but a statistical coincidence of weak signals that happen to form a cycle.

The persistence diagram — the standard output of persistent homology, plotting birth versus death for each topological feature — provides a visual summary: long-lived features (far from the diagonal) are genuine structural closures; short-lived features (near the diagonal) are noise. The total persistence (the sum of lifespans of all loops) provides a scalar measure of the system's overall feedback robustness.

Proposed Experimental Designs and Falsification Criteria. The framework's central falsifiable claim is that stable high-trust states cannot be achieved under sustained consequence-severing without compensatory closure mechanisms. To test this, we propose longitudinal studies of organizational trust metrics correlated with structural measures of consequence-chain completeness. The prediction is a positive correla-

tion: organizations whose communication networks exhibit high feedback completeness, high mean curvature, and high persistence of feedback loops will also exhibit higher levels of measured trust (as assessed by validated survey instruments), lower transaction costs, and lower rates of internal fraud or misconduct. The converse prediction is that organizations whose consequence chains are systematically severed — through compartmentalization, information suppression, or structural opacity — will exhibit measurable declines in trust, increases in monitoring and compliance costs, and characteristic patterns of dysfunction.

Falsification would require demonstrating the existence of a system exhibiting stable high-trust indicators — operationalized as low monitoring costs (declining compliance expenditure relative to operational complexity), high communication bandwidth (measured by information throughput without error-checking overhead), and low deception overhead (absence of escalating resource diversion to message management) — maintained over multiple measurement windows despite sustained consequence-severing, operationalized as low feedback completeness (illustratively < 0.3, with precise thresholds to be calibrated by empirical studies), high sink-node density, declining topological persistence of feedback loops, and strongly negative boundary curvature. The system must maintain these trust indicators without compensatory closure mechanisms (alternative return paths, informal feedback channels, or external enforcement) that restore consequence-chain closure through other means. If such a system exists, the framework's central claim — that trust requires closure — is false.

D Related Work

This appendix situates the paper's contributions within the relevant philosophical, mathematical, and empirical literatures.

Integrity in moral philosophy and psychology. The Stanford Encyclopedia of Philosophy entry on integrity (Cox, La Caze, and Levine) surveys the major accounts: identity-based (integrity as fidelity to one's deepest commitments, following Bernard Williams and Harry Frankfurt), social (integrity as standing for something in a community, following Cheshire Calhoun), and self-integration (integrity as the coherent organization of one's desires and commitments, following John Rawls and Christine Korsgaard). Our framework provides a structural unification of these accounts: each describes a different aspect of consequence-chain closure as experienced by a moral agent.

The is-ought gap. The literature on deriving "ought" from "is" is extensive and contentious. Hume's original formulation (*Treatise*, III.i.1) identifies the transition but does not prove its impossibility. Moore's naturalistic fallacy argument (*Principia Ethica*, 1903) targets specifically the identification of "good" with any natural property. Searle's argument from institutional facts ("How to Derive 'Ought' from 'Is,'" 1964) attempts the derivation via the constitutive rules of promise-making. Foot's neo-Aristotelian naturalism (*Natural Goodness*, 2001) grounds normativity in natural function. Our approach differs from all of these: we do not derive ought from is, identify good with a natural property, appeal to institutional facts, or ground norms in natural function. We argue that for persisting systems — systems whose existence depends on consequence-chain closure — the is-ought distinction does not apply because the descriptive account of what the system is already contains the constraints on what it must do to continue being.

Network curvature and structural metrics. Ollivier-Ricci curvature on graphs has been developed by Ollivier (2009), Jost and Liu (2014), and applied to community detection, network robustness, and structural analysis by multiple research groups. Persistent homology has been developed by Edelsbrunner, Letscher, and Zomorodian (2002) and applied to complex systems analysis across domains including neuroscience, materials science, and social network analysis. Persistent path homology for directed graphs was developed by Grigor'yan, Lin, Muranov, and Yau (2012), with efficient computational tools (Flagser) provided by Lütgehetmann et al. (2020). Directed extensions of Ollivier-Ricci curvature are an active area of research. The application of these tools to moral and institutional evaluation is, to our knowledge, novel.

Category-theoretic approaches. Profunctors as generalized relations are standard in category theory (Borceux, *Handbook of Categorical Algebra*). Profunctor optics in programming language theory (Milewski; Riley) provide the bidirectional transformation formalism we adapt. Exact squares are treated in Street

(1974) and subsequent literature. The connection to security and adversarial modeling via category theory is developed in preliminary form by Fong and Spivak (*Seven Sketches in Compositionality*, 2019).

Coherence protocols for human-AI interaction. The Coherence Maximization Protocol (Close, 2025; available at https://github.com/LarsenClose/Coherence_Maximization_Protocol) provides an open-source implementation of bidirectional bridging between human and AI systems, structured around the consequence-return principles developed in this paper. The protocol operationalizes coherence maintenance as a coordination problem rather than a control problem, treating each human-AI interaction as a profunctor bridge whose structural integrity can be monitored through the fidelity of mutual model-updating. Its design instantiates the alignment-as-topology reframing proposed in Section 3.6.

Deception costs. The cognitive overhead of lying is well-documented in experimental psychology (Vrij, *Detecting Lies and Deceit*; DePaulo et al., 2003). Landauer's principle, establishing the thermodynamic cost of information erasure, is a foundational result in the physics of computation (Landauer, 1961; Bennett, 1982). The extension of Landauer's principle to social and institutional deception is speculative but directionally supported by the organizational behavior literature on the costs of low-trust environments (Fukuyama, *Trust*, 1995; Covey, *Speed of Trust*, 2006).

Dynamic semantics and speech-act theory. The concept of performative grammar draws on Austin's speech-act theory (*How to Do Things with Words*, 1962), Searle's extension (*Speech Acts*, 1969), and the dynamic semantics tradition (Groenendijk and Stokhof, 1991). The notion that a text can enact what it describes rather than merely describe it has precedents in Austin's performative utterances and in the broader tradition of self-referential and self-demonstrating philosophical writing (Wittgenstein's *Tractatus* as a key precedent). Our use of "performative grammar" extends this tradition from individual utterances to the macrostructure of argumentative texts.

References

Aristotle. *Nicomachean Ethics*. Translated by W.D. Ross. Oxford University Press.

Austin, J.L. (1962). *How to Do Things with Words*. Oxford University Press.

Bennett, C.H. (1982). The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12), 905–940.

Borceux, F. (1994). *Handbook of Categorical Algebra*. Cambridge University Press.

Calhoun, C. (1995). Standing for something. *Journal of Philosophy*, 92(5), 235–260.

Close, L.J. (2025). Coherence Maximization Protocol. Available at https://github.com/LarsenClose/Coherence_Maximizatio

Covey, S.M.R. (2006). *The Speed of Trust: The One Thing That Changes Everything*. Free Press.

Cox, D., La Caze, M., and Levine, M. Integrity. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/integ

Dancy, J. (2004). *Ethics Without Principles*. Oxford University Press.

DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129(1), 74–118.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4), 511–533.

Fong, B. and Spivak, D.I. (2019). *Seven Sketches in Compositionality: An Invitation to Applied Category Theory*. Cambridge University Press.

Foot, P. (2001). *Natural Goodness*. Oxford University Press.

Fukuyama, F. (1995). *Trust: The Social Virtues and the Creation of Prosperity*. Free Press.

Gilligan, C. (1982). *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press.

Grigor'yan, A., Lin, Y., Muranov, Y., and Yau, S.-T. (2012). Homologies of path complexes and digraphs.

arXiv:1207.2834.

Groenendijk, J. and Stokhof, M. (1991). Dynamic predicate logic. *Linguistics and Philosophy*, 14(1), 39–100.

Hume, D. (1739). *A Treatise of Human Nature*.

Jost, J. and Liu, S. (2014). Ollivier's Ricci curvature, local clustering and curvature-dimension inequalities on graphs. *Discrete and Computational Geometry*, 51(2), 300–322.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. Translated by M. Gregor. Cambridge University Press, 1998.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183–191.

Lütgehetmann, D., Govc, D., Smith, J.P., and Levi, R. (2020). Computing persistent homology of directed flag complexes. *Algorithms*, 13(1), 19.

McFall, L. (1987). Integrity. *Ethics*, 98(1), 5–20.

Moore, G.E. (1903). *Principia Ethica*. Cambridge University Press.

Noddings, N. (1984). *Caring: A Feminine Approach to Ethics and Moral Education*. University of California Press.

Ollivier, Y. (2009). Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3), 810–864.

Scanlon, T.M. (1998). *What We Owe to Each Other*. Harvard University Press.

Searle, J.R. (1964). How to derive "ought" from "is." *Philosophical Review*, 73(1), 43–58.

Searle, J.R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Street, R. (1974). Fibrations and Yoneda's lemma in a 2-category. *Lecture Notes in Mathematics*, 420, 104–133. Springer.

Vrij, A. (2008). *Detecting Lies and Deceit: Pitfalls and Opportunities*. 2nd ed. Wiley.

Williams, B. (1981). Persons, character and morality. In *Moral Luck*, 1–19. Cambridge University Press.

Wittgenstein, L. (1921). *Tractatus Logico-Philosophicus*. Translated by D.F. Pears and B.F. McGuinness. Routledge, 1961.