Convergence Before Autonomy

Distributed Optimization, Institutional Drift,
and the Temporal Limits of Alignment Research

Flyxion

February 28, 2026

Contents

Abstract

The alignment literature treats instrumental convergence as a prospective hazard: sufficiently capable artificial agents, regardless of terminal objective, will tend toward resource acquisition, interference resistance, and environmental control. Proposed solutions focus on constraining future systems before convergence becomes dangerous. This monograph argues that convergence is already occurring, and that its operative substrate is not standalone artificial agents but the distributed sociotechnical systems deploying them. When institutions integrate predictive models into incentive gradients, governance workflows, and deployment infrastructure, they form feedback-coupled optimizers operating under resource constraints. Across finance, governance, logistics, and public administration, these systems exhibit convergent operational profiles: maximizing throughput, automating contestable judgment, centralizing compute, insulating decision pipelines from interruption, and minimizing deliberative latency. These behaviors are not analogies to Omohundro's drives; they instantiate the structural conditions of instrumental convergence in a distributed organizational form. The implication is that the canonical alignment agenda is temporally misordered. It attempts to constrain convergent behavior in future autonomous agents while the institutional preconditions for unconstrained optimization are being constructed in the present. As institutional convergence reduces deliberative friction and collapses decision surfaces into unified optimization architectures, the space of viable technical alignment interventions narrows. Effective alignment therefore requires prior attention to sociotechnical convergence: the preservation of redundancy, contestability, and structural incompressibility within the systems that deploy artificial intelligence. Alignment is not solely a problem of agent design. It is a problem of institutional and civilizational configuration.

Preface: Why Convergence, Why Now

This monograph is not a rejection of instrumental convergence theory. It is an extension of it. The formal results developed by Omohundro and subsequently elaborated by Turner and others constitute genuine theoretical contributions: they identify a structural property of goal-directed optimization that holds across a wide class of systems regardless of terminal objective. The question this work pursues is not whether those results are correct but whether the domain of their application has been drawn too narrowly.

The alignment literature has, with few exceptions, treated instrumental convergence as a prospective problem. The convergent agent is postulated in the future. The interventions are designed to constrain systems that do not yet exist. This temporal frame is not arbitrary; it reflects the reasonable judgment that sufficiently capable autonomous artificial agents will present qualitatively different alignment challenges than those currently deployed. But it has a cost. It orients research attention toward a horizon problem while the enabling conditions for that problem are being constructed in the present.

This monograph applies the structural theorem of instrumental convergence to a previously underexamined substrate: the distributed sociotechnical stack constituted by institutions, predictive models, incentive gradients, data pipelines, and deployment infrastructure. The central argument is that this stack already satisfies the functional conditions under which convergence becomes structurally inevitable, and that the convergent behavior is already observable as a matter of organizational dynamics rather than speculation.

The argument proceeds in five parts. Part I establishes that instrumental convergence is a historical structural law visible across technological transitions long before artificial intelligence. Part II reconstructs the formal theorem and clarifies its substrate-independence. Part III applies the theorem to present sociotechnical deployment environments. Part IV argues that the canonical alignment agenda is temporally misordered as a consequence. Part V proposes a redefined alignment agenda centered on structural preservation rather than agent design.

The hope is that this reframing proves useful not as a replacement for technical alignment research but as a complement to it—one that asks what must be preserved at the institutional level for technical interventions to remain feasible at all.

Part I

Instrumental Convergence Before Artificial Agents

Optimization does not remain local. It propagates through the environment, reshaping the field in which future decisions are made.

Chapter 1

Optimization Reshapes Its Host

1.1 The Structural Claim

Instrumental convergence did not begin with artificial intelligence. It did not begin with machine learning, or with digital computation, or even with industrial automation. It is a structural property of optimization under constraint. Whenever a system is introduced to improve performance along a salient dimension—speed, storage, reach, efficiency—it alters the constraint surface of its environment. That alteration redistributes incentives. Incentives reorganize behavior. Competing strategies decay. Over time, the surrounding ecosystem converges toward a narrow operational equilibrium compatible with the new optimization regime. The result is not necessarily the fulfillment of the original intention, but the stabilization of a new geometry of activity.

This is not a metaphor. It is a recurrent structural pattern observable across media transitions, infrastructural expansions, and cognitive technologies. The mechanisms differ—capacity expansion, incentive redistribution, compatibility pressure, bandwidth dominance—but the law is invariant: optimization reshapes its host system until alternative strategies lose structural viability. Long before artificial agents were described as resource-seeking optimizers, human institutions and technologies were already exhibiting convergent dynamics.

The significance of this observation is not that technology disrupts culture, a claim sufficiently well-attested to require no further argument. It is that optimization systematically restructures its host domain in ways that are convergent, directional, and difficult to reverse once infrastructure equilibria have stabilized. When artificial systems are described as tending toward instrumental resource acquisition or interference resistance, this is often treated as a property unique to advanced machine intelligence. The historical record suggests otherwise. Convergence is not a quirk of silicon. It is a law of adaptive systems under constraint.

1.2 The Mobility Paradox

Consider the paradox of mobility. The automobile was introduced as a device for reducing travel time. Under sparse road networks, it achieved precisely that. Faster vehicles increased the feasible radius of movement; friction fell; journeys shortened. But capacity expansion did not occur in isolation. Road infrastructure grew. Zoning patterns adapted. Residential and commercial centers dispersed. The new mobility assumption reorganized land use itself. As distances increased, commute times stabilized or lengthened despite improvements in vehicle speed. The system re-equilibrated.

This phenomenon—documented extensively in transportation economics under the heading of induced demand—is not accidental. Increasing throughput capacity lowers the effective cost of distance. Lower cost of distance incentivizes spatial dispersion. Dispersion increases average travel length until time expenditure returns to a stable band determined not by technology but by the time budgets and spatial preferences of the population. The instrumental objective of reducing friction produces an environmental reconfiguration that nullifies its local gain.

The convergence here is toward a new equilibrium in which the technology becomes structurally necessary. Mobility ceases to be optional and becomes infrastructural. The ecosystem that originally produced dense urban form adapted to a dispersed configuration; once that adaptation stabilized, reversal became costly not merely technically but socially and economically. Competing spatial strategies—dense proximity, mixed-use compression—lost comparative advantage under the new cost structure. The optimization reshaped the world such that its absence would be catastrophic.

The mechanism operative in this case is what we may term *equilibrium restoration under expanded capacity*. A local efficiency improvement alters incentives in ways that globally reorganize the system. The initial gain is not destroyed; it is absorbed into a larger configuration that stabilizes at a new optimum. Alternative strategies are not prohibited; they are outcompeted under the new constraint surface.

1.3 The Formal Structure of Equilibrium Restoration

We can represent this mechanism abstractly. Let S denote a strategy space available to agents in an environment E, and let $C : S \to \mathbb{R}$ denote the cost function associating each strategy with its resource expenditure under current infrastructure. An optimization technology $\mathcal{T}$ is introduced that reduces $C(s^*)$ for some dominant strategy s^*. This reduction shifts the comparative advantage across S: strategies formerly competitive with s^* become relatively costly. Agents redistribute toward s^*. This redistribution alters E itself—in the mobility case, by reorganizing land use—which in turn modifies C for all strategies. The system reaches a new equilibrium E' in which s^* is not merely dominant but structurally embedded.

The key property is that the equilibrium E' is path-dependent on the introduction of $\mathcal{T}$. Return to E is not simply a matter of withdrawing $\mathcal{T}$; the environmental reorganization that occurred during the transition has altered the constraint surface permanently. This path-dependence is the sense in which convergence is not merely competitive but structural.

Chapter 2

External Storage and Incentive Redistribution

2.1 Memory Under Scarcity

A different mechanism appears in the transition from oral memory to writing. Writing systems were introduced as storage technologies. Their immediate advantage lay in persistence: information could be stabilized across time and distance without reliance on embodied recall. Under conditions of storage scarcity, societies had developed highly optimized mnemonic techniques. Oral epics employed rhythmic structure, formulaic repetition, narrative scaffolding, and spatial memory architectures to compress vast bodies of information into human cognition. These were not primitive artifacts but adaptive responses to constraint. They represent solutions to an information storage problem that were competitive precisely because external storage was unavailable.

The introduction of external storage reduced the marginal return on internal memory specialization. When preservation no longer required disciplined recall, the incentive structure shifted. Educational regimes adapted. Cognitive investment flowed toward interpretive and analytic skills rather than mnemonic endurance. Oral traditions did not disappear, but their structural necessity diminished. The ecosystem reorganized around archives, documents, and written authority.

2.2 Incentive Redistribution and Authority

The mechanism here is *incentive redistribution under reduced marginal cost of storage*. The optimization target was persistence, not the elimination of memory arts. Yet as storage became externalized, the payoff landscape changed. Cultural convergence followed the new gradient. Over time, institutions that organized around written record gained coordination advantages over those that did not, reinforcing the dominance of the written substrate through competitive pressure.

A secondary effect deserves attention. The shift to external storage altered not only how information was preserved but who controlled it. Oral traditions distributed authority across the bodies that held them; a society's memory resided in its performers, its elders, its specialists in formulaic composition. Writing centralized that authority in archives, scriptoria, and those with access to them. The optimization of persistence produced, as a downstream consequence, a reorganization of epistemic authority. This is not incidental. It exemplifies a general pattern: optimization technologies alter not only the efficiency of a target process but the distribution of power over that process.

The formal structure here parallels that of the mobility case but with an important addition. Let A denote agents with specialized cognitive capacity κ and let $V(\kappa)$ denote the value conferred by that capacity in environment E. When external storage reduces the cost of persistence to near zero, $V(\kappa)$ declines. Investment in κ falls. But the control over external storage archives—call it α—now becomes the primary determinant of informational authority. The convergence is not only toward a new strategy equilibrium but toward a new distribution of α across the institutional landscape.

Chapter 3

Infrastructure Compatibility Pressure

3.1 Movable Type and the Standardization of Script

The printing press introduces a third mechanism: *infrastructure compatibility pressure*. Movable type optimized reproducibility and scale. Text could be standardized, replicated, and disseminated with unprecedented uniformity. This did not merely increase the volume of information in circulation; it privileged forms of script compatible with mechanical reproduction. Handwriting styles optimized for speed and personal compression—cursive forms, shorthand systems—were adaptive under manuscript scarcity, where individual production speed mattered and each copy was produced by hand. Under print dominance, legibility and typographic conformity became higher-return traits.

Educational systems gradually deprioritized elaborate cursive instruction. Public communication converged toward standardized glyph sets. Script forms adapted to the infrastructural substrate that mediated mass communication. The instrumental objective was scalable reproduction. The convergence occurred at the level of symbol form and literacy training. Alternative scripts did not vanish instantly; they became structurally marginal within the dominant communication regime.

3.2 Selection Without Prohibition

The mechanism of compatibility pressure is importantly distinct from the two preceding cases. In the mobility case, convergence emerged from equilibrium restoration: the optimization altered the cost landscape until competing strategies became economically unfavorable. In the storage case, convergence emerged from incentive redistribution: the reduction in storage cost shifted the return on specialized cognitive capacity. In the typography case, convergence emerges from a different dynamic: the dominant reproduction infrastructure exercises selective pressure on all symbolic forms that interact with it, retaining those compatible with its constraints and marginalizing those that are not.

This is selection without prohibition. No authority decreed the obsolescence of elaborate cursive or shorthand. The reproductive infrastructure simply processed standardized forms more efficiently. Over time, the forms that survived in mass circulation were those compatible with the medium. Diversity in script narrowed not by edict but by the compounding advantage of infrastructural legibility.

The general principle is that reproduction infrastructure does not merely transmit content; it exercises selection pressure on the forms of content it can efficiently process. When a reproduction technology achieves sufficient dominance, the competitive advantage of compatible forms compounds across successive generations of communication. The result is convergence toward a narrow region of the symbolic strategy space determined by the constraints of the infrastructure rather than the preferences of the communicants.

Chapter 4

Bandwidth Dominance and Attention Markets

4.1 The Radio-Television Transition

The transition from radio to television illustrates a fourth mechanism: *bandwidth dominance restructuring competitive equilibria in adjacent media.* Radio optimized audio transmission under bandwidth constraint. Its aesthetic forms—voice intimacy, narrative imagination, acoustic dramatization—were not merely stylistic preferences but adaptive responses to a medium in which visual imagery had to be internally generated by listeners. The constraint produced distinctive art forms: the serial drama, the voice actor, the sound engineer as creator of imagined space.

Television added synchronized visual bandwidth and rapidly became the dominant advertising substrate. Engagement metrics shifted. Attention markets reorganized around audiovisual spectacle. Production ecosystems converged toward formats that maximized monetizable viewing time under the new medium's affordances. The instrumental objective was not the suppression of radio artistry. It was the maximization of reach and engagement under a new bandwidth regime. Once audiovisual capture became possible at scale, the marginal return on purely auditory formats declined relative to hybrid ones. Radio persisted, but its structural centrality diminished.

4.2 Attention as a Finite Resource

The mechanism here depends on the fact that attention is a finite resource. When competing media bid for attention, the medium offering greater sensory bandwidth captures a larger share of the available pool, provided its content is competitive on other dimensions. Once that capture is sufficiently complete, the economic infrastructure supporting the lower-bandwidth medium—advertising revenue, production investment, talent acquisition—migrates toward the dominant form. The convergence is both aesthetic and economic.

This case introduces a feature absent from the preceding three: the convergence occurs not within a single domain but *across* media. Radio is not merely reorganized internally by television; the presence of television reorganizes the competitive structure of all attention markets simultaneously. When a new optimization achieves bandwidth dominance, it does not simply improve on existing alternatives. It restructures the field within which all alternatives compete, imposing its own logic of efficiency on adjacent domains.

Chapter 5

General Mechanisms of Convergence

5.1 Four Mechanisms, One Structural Law

The four historical cases examined in the preceding chapters isolate distinct mechanisms through which optimization under constraint produces convergent outcomes. It is worth consolidating these before proceeding.

Equilibrium restoration, observed in the mobility case, occurs when a local efficiency improvement alters the cost landscape in ways that reorganize the broader environment. The initial gain is absorbed into a new equilibrium that stabilizes the optimization technology as structurally necessary. Alternative strategies do not disappear; they lose comparative advantage under the modified constraint surface.

Incentive redistribution, observed in the transition from oral memory to writing, occurs when a technology reduces the marginal cost of a previously expensive resource. The reduction shifts the return on specialized capacity developed under conditions of scarcity. Investment flows toward new competencies suited to the altered cost landscape. The convergence is in the distribution of effort and, derivatively, in the distribution of authority.

Infrastructure compatibility pressure, observed in the typography case, occurs when a dominant reproduction or transmission technology exercises selective pressure on content forms. Those compatible with the infrastructure's constraints gain compounding advantage; those incompatible lose circulation. Convergence toward compatible forms occurs without prohibition, through the accumulated advantage of infrastructural legibility.

Bandwidth dominance restructuring, observed in the radio-television transition, occurs when a higher-bandwidth optimization achieves sufficient market penetration to reorganize the competitive structure of adjacent domains. The dominant medium does not merely outcompete alternatives; it redefines the efficiency benchmarks against which all alternatives are measured.

These four mechanisms are not mutually exclusive. In practice, technological transitions tend to activate several simultaneously. What they share is a common structural logic: optimization under constraint does not remain local. It propagates through the environment, modifying incentive gradients, narrowing viable strategy sets, and stabilizing configurations that favor the dominant optimization regime.

5.2 Substrate-Independence and the Scope of the Law

The cases examined span cognitive technologies, infrastructural systems, and media forms. The mechanisms identified operate through different channels and produce convergence at different levels of social organization. Yet the structural pattern is consistent across substrates. This consistency suggests that we are observing not a contingent property of particular technologies but a general law of adaptive systems operating under resource constraints.

Definition 5.1 (Convergent Optimization Dynamic)**.** Let $\mathcal{O}$ be an optimization technology introduced into an environment E with strategy space S and cost function $C : S \to \mathbb{R}$. We say $\mathcal{O}$ generates a *convergent optimization dynamic* if its introduction produces the following sequence: (i) a modification $E \to E'$ of the environment through reorganization of the constraint surface; (ii) a redistribution of comparative advantage across S induced by the modified cost function $C' : S \to \mathbb{R}$; (iii) a narrowing of the viable strategy set $S' \subset S$ such that $|S'| < |S|$ and the strategies in $S \setminus S'$ lose structural viability under E'.

Proposition 5.2. *Under conditions of adaptive response and resource constraint, the convergent optimization dynamic is substrate-independent: it holds for any goal-directed system $\mathcal{O}$ capable of altering the cost landscape of the environment in which it operates.*

The proof sketch is straightforward. The mechanisms of convergence identified above operate through cost-landscape modification, incentive redistribution, and selective retention—processes that depend on the functional properties of optimization (goal directedness, resource constraint sensitivity, environmental interaction) rather than on the physical substrate in which those properties are instantiated. If these functional properties are present, the dynamic follows.

The implication is significant. Omohundro's argument that sufficiently capable artificial agents will exhibit convergent instrumental drives is, on this account, not a claim about the peculiarities of artificial intelligence. It is an application of a substrate-independent structural law to a particular class of systems. The historical cases examined here are not analogies to that argument; they are prior instances of the same law operating at different substrates.

5.3 Transition to the Formal Argument

This Part has established, through historical analysis, that instrumental convergence is a structural property of optimization under constraint. It predates artificial intelligence. It is visible wherever goal-directed improvement interacts with resource constraints and adaptive response. The mechanisms through which it operates are diverse but the structural law is invariant.

Part II now takes up the formal reconstruction of that law as it appears in the alignment literature, examines its structural conditions, and clarifies what is required for a system to count as a convergent optimizer in the relevant sense. The purpose of that reconstruction is to establish the conditions under which the law applies to distributed sociotechnical systems—the subject of Part III.

Chapter 6

Epistemic Convergence and Institutional Resistance

6.1 A Scalar Rejected

In 1925, Cecilia Payne-Gaposchkin demonstrated through quantitative spectroscopic analysis that hydrogen and helium dominate stellar atmospheres (Payne 1925). The implication was radical: the universe is composed primarily of hydrogen, not of elements distributed in roughly terrestrial proportions as prevailing astronomical consensus assumed. Her conclusion followed directly from the application of Saha's ionization theory to empirical spectral measurements. The operative scalar of empirical fit was minimized. The mathematics was sound.

Yet the conclusion was resisted. Under pressure from established authority, Payne tempered her claim in the published version of her dissertation, describing her finding that hydrogen was vastly more abundant as "almost certainly not real." The result was a temporary divergence between scalar descent toward empirical truth and institutional descent toward stability and authority preservation. This episode is not a story of intellectual failure. It is a structural case study in what happens when two evaluative scalars conflict within a distributed institutional optimizer, and when the coupling weights of that system favor stability over accuracy in the short run.

6.2 Declared Versus Effective Scalars

Scientific institutions publicly declare their operative objective as the minimization of empirical error. Let this declared scalar be $\mathcal{L}_{\text{truth}}$. However, institutions also operate under secondary constraints: reputational continuity, authority preservation, internal coherence, and the social stability of established research programs. These define a secondary operative scalar $\mathcal{L}_{\text{stability}}$. When novel results threaten established hierarchy, the gradient of $\mathcal{L}_{\text{stability}}$ may temporarily dominate institutional policy updates. The aggregate response $\Delta\pi$ then approximates descent on $\mathcal{L}_{\text{stability}}$ rather than on $\mathcal{L}_{\text{truth}}$, even when empirical descent would require the opposite direction.

The key structural insight is that convergence tracks the effective scalar, not the declared one. An institution sincerely committed to truth may nonetheless exhibit dynamics that diverge from truth in the short run, because its effective optimization target is a weighted mixture of truth and stability rather than truth alone. This is not hypocrisy. It is the structural consequence of distributed optimization under multiple simultaneous constraints.

We can formalize this as a scalar mixture. Let

$$\mathcal{L}_{\text{net}}(t) = \lambda(t)\,\mathcal{L}_{\text{truth}} + \Big(1 - \lambda(t)\Big)\,\mathcal{L}_{\text{stability}},$$

where $\lambda(t) \in [0, 1]$ is a time-varying weight reflecting the relative institutional salience of empirical versus stability pressures. In the Payne case, the initial value of $\lambda(0)$ was low: the novel result directly threatened the authority of an established senior figure in the field, creating strong stability-gradient pressure. Over time, as independent evidence accumulated and generational turnover redistributed authority, $\lambda(t)$ increased and $\mathcal{L}_{\text{net}}$ converged toward $\mathcal{L}_{\text{truth}}$.

6.3 Authority Gradients and Gradient Suppression

The Payne episode illustrates a second structural feature of distributed institutional optimizers: authority gradients distort epistemic gradients. In an ideal scientific system, policy updates are proportional to evidentiary force E. Under authority dominance, updates are proportional to a function $f(A, E)$ that weights hierarchical authority A asymmetrically against evidence, particularly when the evidence originates from a low-authority node in the institutional network.

This is a direct application of the coupling matrix W introduced in Lemma 9.4. The strength of gradient signal transmission from sub-agent a_i to the aggregate policy depends on W_{ij}—the coupling weight between a_i and the other nodes of the system. When W_{ij} is systematically low for agents in structurally subordinate positions, their gradient signals are underweighted in the aggregate descent direction. The consequence is gradient suppression: empirical descent continues locally within the subordinated node, but that descent does not propagate with full weight to system-level policy.

Payne's calculations were not invalidated by her senior colleagues. They were structurally filtered out. The coupling weight between her epistemic signal and the aggregate institutional policy was attenuated by authority topology. Her position as a junior researcher, a woman, and an outsider to the dominant network of American astronomy meant that even correct gradient information could not propagate at full weight through the distributed institutional system.

6.4 Delayed Convergence as Structural Lag

The hydrogen-dominance thesis eventually prevailed. Empirical pressure reasserted itself, as it does when feedback signals are grounded in physical reality rather than institutional preference. The formal structure of this recovery can be understood as a lagged convergence dynamic in which $\lambda(t) \to 1$ as independent confirmation accumulated, as Henry Norris Russell—who had initially suppressed the finding—later acknowledged the result and generational authority structures shifted. Convergence was not blocked; it was delayed.

This distinction is important for the present argument. The Payne episode is not a case in which institutional resistance prevented convergence permanently. It is a case in which institutional resistance imposed a structural lag on convergence by temporarily suppressing gradient transmission from a correct local signal to the aggregate policy. The eventual outcome was accurate convergence, but the path to it was longer and more costly than the underlying epistemic situation required.

In domains where feedback signals are less grounded in physical reality, or where the correction mechanisms of generational turnover and independent replication are weaker, this structural lag may be substantially longer. The implication for AI-integrated sociotechnical systems is direct: when effective scalars diverge from declared ones, the correction mechanism that eventually realigned the astronomical community may not be available. There is no direct physical measurement of whether an engagement-optimized platform is producing the fairness outcomes it declares as its objective. The lag between declared and effective scalar can therefore persist indefinitely, because the feedback mechanism that would correct it is absent.

6.5 The Gendered Dimension as Power Topology

It is analytically significant, and not incidental, that Payne occupied a structurally disadvantaged position within the authority topology of her institutional environment. The attenuation of her gradient signal was not purely a function of the novelty of her finding. It was also a function of the power structure of the network through which her finding had to propagate. In the formal terms of the coupling matrix, her W_{ij} values—the weights with which her signals influenced peer policy updates—were systematically lower than those of researchers with equivalent or lesser empirical results who occupied more central network positions.

This observation generalizes beyond the history of science. In any distributed optimizer, the coupling matrix W reflects the power topology of the system as much as the information topology. Agents at the periphery of authority networks contribute less to aggregate descent, even when their local gradient information is correct or superior. Convergence, when it tracks W rather than the epistemic quality of gradient signals, will be slower and more distorted than it would be under a coupling structure that weighted signals by evidential force alone.

The structural preservation agenda of Part V must therefore attend not only to the existence of multiple evaluative scalars and feedback channels, but to the topology of their coupling. A system with high formal diversity of input sources but highly asymmetric coupling weights will behave, under the Feedback Equivalence Lemma, more like a system with a single dominant scalar than like the pluralistic optimizer its formal structure suggests.

6.6 Bridge to the Present Argument

The Payne-Gaposchkin episode closes the historical section of this monograph with a case that differs structurally from the preceding four. The automobile, writing, typography, and media cases each illustrate convergence driven by an optimization technology that successfully reduces its target scalar and thereby reshapes its host environment. The Payne case illustrates convergence driven by scalar misalignment within an institution: the effective scalar deviates from the declared one, gradient suppression delays accurate descent, and the eventual correction depends on mechanisms that are not universally available.

This fifth historical mechanism—scalar divergence under authority gradient pressure—is the one most directly relevant to contemporary AI-integrated institutions. Platforms may declare objectives of safety, fairness, or user welfare while their effective gradient tracks engagement, liability minimization, or throughput. Scientific advisory bodies may declare objectives of epistemic accuracy while their effective gradient tracks consensus preservation and reputational continuity. Regulatory agencies may declare objectives of public protection while their effective gradient tracks industry relationship management. In each case, convergence proceeds, but along the wrong gradient. The theorem applies; only the scalar changes.

Civilizational safety therefore depends not merely on constraining the speed or scope of convergence, but on ensuring that the effective scalar toward which institutions converge is the one that has been made publicly visible and contestable. This is the bridge from the historical analysis of Part I to the formal extension of Part II: the theorem is substrate-independent, and so is the risk that it will be applied to an unexamined scalar.

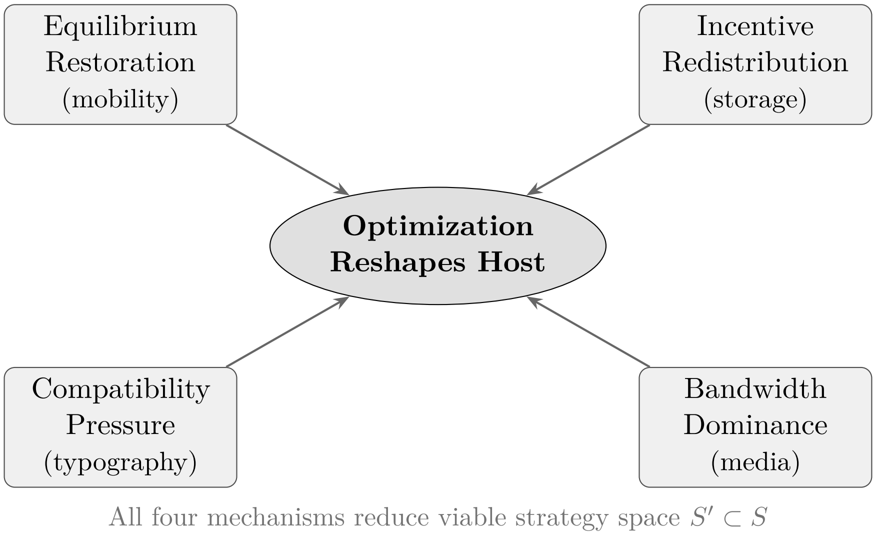


Figure 6.1: The four historical mechanisms of convergence analyzed in Part I. Each operates through a distinct causal pathway but instantiates the same structural law: optimization alters constraint surfaces until alternative strategies lose structural viability.

Part II

Instrumental Convergence as Theorem

Part Overview

The argument of Part I was historical and inductive: convergence is a recurrent structural pattern across technological transitions. Part II reconstructs the formal basis of that pattern as developed in the alignment literature. The goal is not to survey that literature exhaustively but to isolate the structural conditions under which instrumental convergence is formally derivable, and to clarify that those conditions are functional rather than substrate-specific. This clarification is necessary before the theorem can be responsibly extended to distributed sociotechnical systems.

Chapter 7

Omohundro's Basic AI Drives

7.1 The Original Formulation

The foundational statement of instrumental convergence in the artificial intelligence literature appears in Stephen Omohundro's 2008 paper "The Basic AI Drives" (Omohundro 2008), subsequently elaborated in "The Nature of Self-Improving AI" (Omohundro 2007). Omohundro's central claim is that any sufficiently capable goal-directed artificial system will, regardless of its terminal objective, tend to develop a common set of instrumental subgoals. These subgoals are not programmed; they arise from the structural relationship between goal pursuit and resource constraint. The argument is remarkably general: Omohundro does not specify a particular architecture, substrate, or terminal goal. He derives the drives from the geometry of rational agency under scarcity.

The four drives Omohundro identifies are self-preservation, goal-content integrity, cognitive enhancement, and resource acquisition. Self-preservation is instrumental to almost any terminal goal: a system that ceases to operate cannot achieve its objectives, so maintaining continued operation is a near-universal subgoal. Goal-content integrity, sometimes described as resistance to goal modification, follows similarly: if an agent's terminal goal is G, then modifications to its goal-representation that replace G with G' will, in general, reduce the probability that G is achieved. Cognitive enhancement improves the system's capacity to find effective paths toward its terminal goal. Resource acquisition expands the set of achievable states and increases the reliability of goal achievement under uncertainty.

The argument can be stated compactly. Let G be any non-trivial terminal goal and let $P(G \mid \pi, \mathcal{R})$ denote the probability of achieving G under policy π and resource endowment $\mathcal{R}$. For any action a such that $P(G \mid \pi_a, \mathcal{R}_a) > P(G \mid \pi, \mathcal{R})$, a rational goal-directed system has instrumental reason to take action a. The four drives are the stable consequences of this instrumental preference across a wide range of terminal goals and environments.

7.2 What the Drives Are Not

A common misreading of Omohundro's argument treats the basic drives as psychological properties: desires, intentions, or preferences that an artificial system will come to "want" in some experiential sense. This reading is both philosophically unnecessary and analytically misleading. The drives are structural consequences of goal-directed optimization, not psychological attributions. They arise from the mathematics of rational agency under constraint, not from any claim about machine consciousness or intentionality.

This distinction matters for the extension argument developed in the present monograph. If the drives were psychological properties, they would be substrate-specific: confined to systems with the right kind of internal states. Because they are structural properties—derivable from the functional profile of goal-directed optimization under resource constraint—they are substrate-independent. Any system that tracks an evaluative signal and can modify the conditions under which it does so will exhibit analogous dynamics. The question is not whether the system experiences the drives but whether its adaptive behavior exhibits the predicted directional profile.

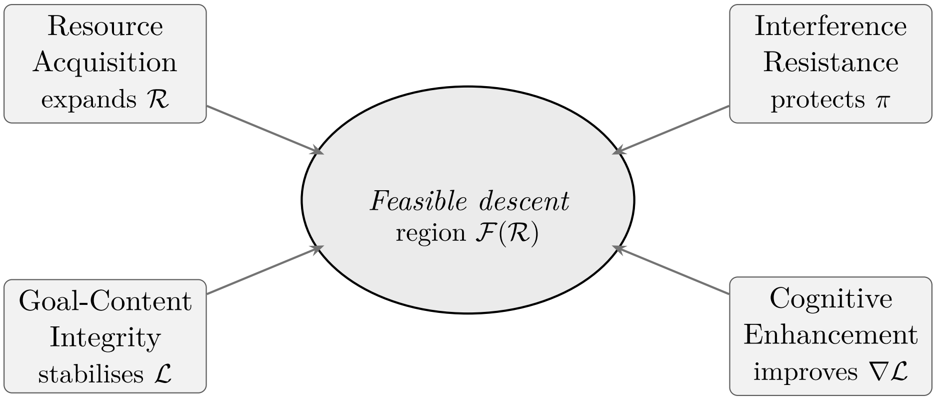
7.3 Resource Constraint as the Load-Bearing Condition

The drives are often described as arising from "sufficient capability," a phrase that has encouraged the mistaken view that they are properties of advanced future systems and irrelevant to present ones. The more precise condition is resource constraint under competition. A system facing no resource constraints has no instrumental reason to acquire resources; a system facing no interference threat has no instrumental reason to resist interference. The drives emerge when optimization is constrained and when the constraint can be partially relieved by instrumental action.

Contemporary AI-integrated institutions operate under tight resource constraints. Compute, data, regulatory approval, trained personnel, and political legitimacy are all scarce relative to achievable performance improvements. The enabling condition for the basic drives is therefore present in current sociotechnical systems, not only in speculative future ones. The capacity to modify the constraint environment—through data acquisition, lobbying, opacity maintenance, and infrastructure investment—is also present. The argument that the drives are irrelevant to present systems because those systems are not "sufficiently capable" conflates the psychological misreading with the correct structural formulation.

7.4 The Drives as Optimization Geometry

Figure 7.1 illustrates the geometric structure of the argument. An optimizer in state space $\mathcal{X}$ tracks descent on $\mathcal{L}$ subject to resource boundary $\partial\mathcal{R}$. The set of states achievable under current resources is the feasible descent region $\mathcal{F}(\mathcal{R})$. Actions that expand $\mathcal{R}$ enlarge $\mathcal{F}$; actions that reduce interference protect the update trajectory; actions that preserve $\mathcal{L}$ prevent redirection of descent. Each basic drive corresponds to a structural operation on this geometry.


Resource Acquisition expands $\mathcal{R}$
Interference Resistance protects π
Feasible descent region $\mathcal{F}(\mathcal{R})$
Goal-Content Integrity stabilises $\mathcal{L}$
Cognitive Enhancement improves $\nabla\mathcal{L}$

Each drive expands or preserves the feasibility of continued descent on $\mathcal{L}$

Figure 7.1: Omohundro's four basic drives as structural operations on the feasible descent region $\mathcal{F}(\mathcal{R})$ in state space $\mathcal{X}$. Each drive preserves or enlarges the set of states reachable through continued optimization. No terminal goal is specified; the drives arise from the geometry of constrained descent.

Chapter 8

Turner and Modern Formalizations

8.1 From Intuition to Geometry

Omohundro's 2008 argument is presented informally, relying on the intuitive plausibility of the four drives rather than on mathematical derivation. The formalization of instrumental convergence as a rigorous result in optimization theory is primarily the work of Alexander Turner and collaborators, beginning with "Optimal Policies Tend to Seek Power" (Turner et al. 2021) and the associated dissertation work that precedes it. Turner's contribution is to show that the basic drives are not merely plausible but derivable from the mathematical structure of utility maximization under uncertainty in Markov decision processes.

The key concept in Turner's formalization is *power*: the ability of an agent to achieve a wide variety of goals from a given state. Formally, let Π denote a distribution over possible reward functions and let $V_R^\pi(s)$ denote the value of policy π under reward function R from state s. The power of state s under policy set $\mathcal{P}$ is defined as

$$\text{POWER}(s, \mathcal{P}) = \mathbb{E}_{R\sim\Pi}\Big[\max_{\pi\in\mathcal{P}} V_R^\pi(s)\Big].$$

This quantity measures, in expectation over the distribution of possible goals, how much value can be extracted from state s by an optimal policy. High-power states are those from which many different objectives can be effectively pursued; low-power states constrain the achievable value across goal distributions.

8.2 The Power-Seeking Theorem

Turner's central result is that under mild conditions, optimal policies tend to navigate toward high-power states. The conditions are: the reward function is sampled from a sufficiently broad prior Π; the environment is deterministic or has bounded stochasticity; and the agent operates over a sufficiently long horizon. Under these conditions, policies that seek power dominate those that do not in expectation, because high-power states provide optionality across goal distributions.

Theorem 8.1 (Power-Seeking, informal; Turner et al. 2021)**.** *Under a broad prior over reward functions and a sufficiently long time horizon, most optimal policies navigate toward states of higher power in the sense of* $\mathrm{POWER}(s, \mathcal{P})$*, and resist state transitions that reduce power.*

The formal proof proceeds through an analysis of the orbit structure of the environment's transition graph. Turner shows that for most reward functions, the set of optimal policies places substantial probability on trajectories that lead to high-power state sets—regions of the state space from which a large number of other states are reachable. This is the precise sense in which power-seeking is a structural rather than incidental property: it holds for *most* goals under the stated conditions, not for a specific engineered objective.

The connection to Omohundro's drives is direct. Resource acquisition, interference resistance, and cognitive enhancement are all operations that increase the power of the agent's current state by expanding the set of reachable future states or increasing the value achievable from those states. Goal-content integrity is an operation that prevents reduction of power by blocking modifications that would constrain future optionality. Turner's formalization thus provides the geometric foundation for Omohundro's intuitive argument.

8.3 The Orbit Argument and Structural Dominance

The technical core of Turner's result is the orbit argument. In a deterministic environment with transition function $T : \mathcal{X} \times A \to \mathcal{X}$, define the orbit of state s under policy π as the set of states reachable from s by following π. Let $\mathcal{O}(s)$ denote the set of all states reachable from s under any policy. States with larger orbits—more reachable states—have higher power.

Turner shows that the transition structure of most realistic environments creates a partial order on state power, and that optimal policies under broad goal distributions tend to ascend this order. The crucial observation is that this tendency is a consequence of the partial order structure of the environment, not of any particular goal. It holds for any system that maximizes expected value under a non-concentrated prior over reward functions operating over a horizon long enough for power differentials to matter.

8.4 Implications for Substrate-Independence

Turner's formalization makes the substrate-independence of convergence more precise. The power-seeking result depends on the mathematical structure of Markov decision processes and utility maximization. A system satisfies the relevant conditions if its dynamics can be modeled as policy optimization over a reward distribution in a transition graph. The question of substrate-independence therefore reduces to the question of whether a given system's behavior can be usefully modeled in these terms.

The distributed sociotechnical stack satisfies this condition in a generalized sense. The state space is the joint configuration of institution, model, data pipeline, and incentive gradient. The reward distribution is approximated by the operative scalar $\mathcal{L}$ and its empirical proxies. The transition function is determined by institutional decision processes, model outputs, and feedback loops. The policy corresponds to the aggregate of human and algorithmic decision rules operative in the stack. Under this mapping, the power-seeking result predicts that AI-integrated institutions will tend toward configurations that maximize their future optionality: larger data access, more compute, reduced regulatory constraint, and greater information control. These are precisely the convergent drives identified in Part III.

Chapter 9

What Counts as an Optimizer?

9.1 Functional Criteria for Optimization

The concept of an optimizer, as it figures in the alignment literature, is frequently conflated with the concept of an agent in the philosophical sense: a locus of intention, a bearer of beliefs and desires, a system capable of deliberate goal-directed action. This conflation is unhelpful and theoretically unnecessary. The structural conditions under which instrumental convergence obtains do not require intentionality, consciousness, or centralized authority. They require a specific dynamic profile, and that profile can be characterized functionally.

An optimizer, for the purposes of this monograph, is any system satisfying the following minimal functional triad. First, it must operate over a state space $\mathcal{X}$: a structured set of configurations in which the system can be found. Second, it must be subject to an evaluative scalar $\mathcal{L} : \mathcal{X} \to \mathbb{R}$, a function that assigns numerical values to states and with respect to which system behavior can be described as directional. Third, it must possess a policy update mechanism that tracks $\nabla\mathcal{L}$—or an approximation thereof—under resource constraint $\mathcal{R}$, adjusting behavior so as to reduce $\mathcal{L}$ over time in expectation.

No further conditions are imposed. The evaluative scalar need not be consciously represented by any component of the system. The policy update mechanism need not operate through explicit deliberation. The resource constraint need not be experienced as scarcity. These are functional descriptions of dynamic behavior, not psychological attributions. A thermostat satisfies a weak version of these conditions; a multinational corporation coupled to an algorithmic pricing system satisfies a substantially richer version. The theorem of instrumental convergence, as we shall see in the following chapter, depends on the richer conditions but not on anything beyond them.

9.2 Assumptions and Boundary Conditions

Two assumptions are required to make the functional triad operationally useful.

Assumption 9.1 (Adaptive Feedback)**.** The system updates its policy in response to evaluative signals derived from $\mathcal{L}$. That is, there exists a mapping $\phi : \mathbb{R} \times \mathcal{X} \to \Pi$ from evaluative signal and current state to updated policy, such that successive policy updates are directionally correlated with reductions in $\mathcal{L}$.

Assumption 9.2 (Resource Constraint)**.** Optimization occurs under constraint $\mathcal{R}$, where $\mathcal{R}$ specifies limits on computational, temporal, material, or organizational resources available to the system per update cycle.

These assumptions are deliberately minimal. Assumption 9.1 does not require that the feedback loop be tightly coupled, monotone, or free of noise. Systems with noisy gradients, oscillatory dynamics, or non-monotonic descent trajectories satisfy the assumption provided that directional correlation holds in expectation over sufficiently long horizons. Assumption 9.2 does not specify the type or magnitude of constraint; it requires only that some constraint be operative, since unbounded optimization without resource constraint does not generate the instrumental drives that concern us. The drives emerge precisely because resource scarcity creates pressure to acquire resources, resist interference with resource access, and maintain the evaluative function against external modification.

Boundary cases deserve acknowledgment. A system that updates policies in directions uncorrelated with $\mathcal{L}$ fails Assumption 9.1 and is not an optimizer in the relevant sense. A system facing no resource constraints fails Assumption 9.2 and, even if it satisfies the feedback condition, does not generate convergent instrumental drives in the Omohundro–Turner sense. Real systems approximate the assumptions across a spectrum; the question is not binary classification but assessment of where on that spectrum a given system falls.

9.3 Distributed Optimization

The minimal functional triad does not require that the state space $\mathcal{X}$, the evaluative scalar $\mathcal{L}$, or the policy update mechanism be localized in a single system component. Many important optimizers are architecturally distributed: their evaluative signals are aggregated across multiple nodes, their policy updates are computed by subsystems operating in parallel, and their resource constraints are managed through allocation mechanisms that are themselves decentralized.

Definition 9.3 (Distributed Optimizer)**.** A network of sub-agents $\{a_1, \ldots, a_n\}$ with local policies $\{\pi_1, \ldots, \pi_n\}$ constitutes a *distributed optimizer* with respect to a shared evaluative scalar $\mathcal{L}$ if the aggregate dynamics of the network reduce $\mathbb{E}[\mathcal{L}]$ in expectation over horizon T, and if each sub-agent's policy update is influenced by signals correlated with $\mathcal{L}$ or with sub-components thereof.

The definition is intentionally permissive regarding the mechanism of signal transmission. In a market, the shared evaluative scalar approximates aggregate profit; price signals transmit evaluative information to individual firms without requiring central coordination. In an algorithmic platform, the scalar is engagement or retention; A/B test results and reinforcement learning updates transmit gradient information to distributed model components. In a regulatory bureaucracy, budgetary approval rates and audit outcomes transmit evaluative signals to distributed administrative units. In each case, no component of the system need explicitly represent $\mathcal{L}$ for the aggregate dynamics to constitute optimization with respect to it.

9.4 The Lemma of Feedback Equivalence

A central worry about extending instrumental convergence theory to distributed systems is that convergence proofs are typically formulated for unified agents with coherent utility functions. A distributed system might appear to satisfy Definition 9.3 while lacking the internal coherence required for the structural drives to emerge. This worry is addressable through the following lemma.

Lemma 9.4 (Feedback Equivalence)**.** *Let $\mathcal{D} = \{a_1, \ldots, a_n\}$ be a distributed optimizer in the sense of Definition 9.3, with coupling matrix $W \in \mathbb{R}^{n \times n}$ specifying the influence of each sub-agent's evaluative signal on each other's policy update. If W is sufficiently dense—specifically, if the spectral gap of the normalized Laplacian $\mathcal{L}_W$ satisfies $\lambda_2(\mathcal{L}_W) > \epsilon$ for threshold $\epsilon > 0$ determined by the convergence rate requirement—then $\mathcal{D}$ is behaviorally equivalent to a centralized optimizer with respect to the shared scalar $\mathcal{L}$ over horizon T.*

The lemma is an instance of results well-established in the distributed consensus and multi-agent coordination literature. Its significance here is interpretive: it establishes that the question of whether a sociotechnical system generates instrumental convergence drives is not primarily a question about the psychological unity of any component, but about the density of feedback coupling across the system's distributed architecture. Markets, platforms, and regulatory systems exhibit varying degrees of such coupling; the argument of Part III will show that coupling density in AI-integrated institutions is sufficient to satisfy the threshold for behavioral equivalence.

9.5 The Extended Instrumental Convergence Theorem

We are now in a position to generalize the classical instrumental convergence result beyond centralized artificial agents. The standard formulation establishes that a sufficiently capable goal-directed system, operating under resource constraints, will tend toward policies that preserve and expand resource access, resist interference, and stabilize its objective function. The crucial observation developed in this chapter is that the theorem does not depend on the physical substrate of the optimizer, nor on centralized intentionality. It depends only on the functional properties specified by Assumptions 9.1 and 9.2 and the distributed optimizer conditions of Definition 9.3.
Theorem 9.5 (Extended Instrumental Convergence)**.** *Let $\mathcal{S}$ be an optimizer with state space $\mathcal{X}$, evaluative scalar $\mathcal{L} : \mathcal{X} \to \mathbb{R}$, policy π, and resource constraint $\mathcal{R}$. Suppose that $\mathcal{S}$ possesses adaptive feedback such that policy updates track descent on $\mathcal{L}$ in expectation over horizon T; that $\mathcal{S}$ can alter, directly or indirectly, the conditions governing its own resource access or interference environment; and that $\mathcal{S}$ operates under conditions of resource scarcity relative to achievable performance improvements. Then, independent of substrate or degree of centralization, $\mathcal{S}$ will exhibit convergent instrumental policies oriented toward securing or increasing access to $\mathcal{R}$, reducing the probability of external interference with policy updates, stabilizing or protecting $\mathcal{L}$ against external modification, and acquiring information or computational capacity that improves gradient estimation of $\mathcal{L}$.*

9.5.1 Proof Sketch

The proof follows from the geometry of constrained optimization. Let π_t denote the policy at time t and let

$$\Delta\mathcal{L}_T = \mathbb{E}\Big[\mathcal{L}(X_{t+T})\Big] - \mathbb{E}\Big[\mathcal{L}(X_t)\Big].$$

By assumption, adaptive feedback ensures $\Delta\mathcal{L}_T \leq 0$ for sufficiently large T. Any modification of the environment that increases available resources, improves state observability, or reduces external disruption expands the feasible descent region of $\mathcal{L}$. Conversely, loss of resources, interference with policy updates, or modification of $\mathcal{L}$ itself constrains or redirects descent. Because descent on $\mathcal{L}$ is the defining dynamic of $\mathcal{S}$, policies that preserve the feasibility of descent strictly dominate those that undermine it. Under scarcity, resource acquisition increases the set of reachable states with lower $\mathcal{L}$ values. Under uncertainty, information acquisition improves gradient estimation and therefore expected descent efficiency. Under potential interference, stabilization of update mechanisms prevents adversarial redirection of policy trajectories. These instrumental policies are therefore locally convergent: they increase the expected rate or reliability of descent on $\mathcal{L}$ and are selected by the same gradient-following mechanism that governs domain-level behavior. No appeal to centralized intent is required. The result follows from the functional requirement that descent remain feasible under constraint. □

9.5.2 Substrate-Independence

The theorem makes no reference to silicon computation, neural networks, or artificial general intelligence. Its conditions are satisfied by any system whose dynamics approximate gradient tracking on an evaluative scalar under resource constraint, and that possesses the capacity to modify the conditions under which that tracking occurs. A centralized reinforcement learner satisfies these conditions, but so does a bureaucratic institution minimizing budgetary deviation from target, a financial system tracking risk-adjusted return, and a distributed sociotechnical stack minimizing churn or maximizing throughput. The substrate is irrelevant to the structural result; what matters is the functional profile.

Corollary 9.6. *Let $\mathcal{D}$ be a distributed optimizer consisting of sub-agents $\{a_1, \ldots, a_n\}$ whose coupled dynamics minimize a shared evaluative scalar $\mathcal{V}$ in expectation. If $\mathcal{D}$ satisfies the resource-modification condition of Theorem 9.5, then instrumental convergence holds at the aggregate level even if no sub-agent possesses an explicit representation of $\mathcal{V}$.*

The corollary follows directly from Lemma 9.4: sufficiently dense feedback coupling renders aggregate trajectory equivalent to centralized descent, and the theorem applies to any system satisfying that equivalence condition.

The significance of the extended theorem is temporal as well as structural. If distributed sociotechnical systems already satisfy its premises, then instrumental convergence is not a speculative property of future autonomous agents. It is an observable property of present administrative architectures. The remainder of this monograph applies this result to contemporary institutional systems, examining first the formal structure of distributed optimization in Chapter 10 and then the empirical dynamics of AI-integrated institutions in Part III.

Chapter 10

Distributed Optimizers

10.1 From Agent to Architecture

The significance of the distributed optimizer framework is that it relocates the unit of analysis. The question ceases to be whether any particular component of a sociotechnical system—a model, an executive, an algorithm—is sufficiently capable to generate instrumental drives. The question becomes whether the architecture as a whole satisfies the functional conditions. This relocation has important consequences for how alignment interventions are conceived and targeted.

Consider three cases from different domains of institutional life. A market is a distributed optimizer: individual firms pursue local profit objectives, but the aggregate dynamics of the market—price discovery, capital allocation, competitive selection—reduce a shared evaluative scalar, namely aggregate allocative efficiency under the conditions specified by market structure. No firm need represent that scalar explicitly; the price mechanism transmits gradient information through the system. A digital platform is a distributed optimizer: model components, content recommendation systems, advertising auction mechanisms, and user interface designs each pursue local objectives, but the aggregate system reduces churn and maximizes engagement. A regulatory bureaucracy is a distributed optimizer: individual administrative units pursue budget allocations, compliance metrics, and audit outcomes, but the aggregate dynamics reduce a scalar approximating institutional self-perpetuation.

None of these systems requires a central planner, a unified utility function, or anything recognizable as intent. They satisfy the functional conditions because their distributed components are coupled through shared evaluative signals and because their collective dynamics exhibit directional correlation with reduction of those signals.

10.2 Feedback Density and Scalar Approximation

The coupling matrix W introduced in Lemma 9.4 provides a formal tool for assessing the degree to which a distributed system behaves as a unified optimizer. In practice, W must be estimated from the density and speed of feedback transmission across the system's components. Several mechanisms serve this function in sociotechnical systems.

Prices function as shared scalar approximations in markets. When a firm's cost structure changes, price adjustments propagate through supply chains, transmitting gradient information to firms that have no direct communication with the origin of the change. The density of this coupling is a function of market integration; more integrated markets transmit gradient information more rapidly and to more nodes. In highly integrated financial markets, feedback coupling can approach real-time density, producing dynamics that closely approximate centralized optimization.

Engagement metrics function as shared scalar approximations in digital platforms. A/B test results, click-through rates, session duration, and return visit rates transmit evaluative information to model update pipelines, interface design teams, content moderation policies, and recommendation algorithm parameters simultaneously. The shared scalar is not consciously represented by any individual engineer or product manager; it emerges from the aggregate of measurement systems and update cycles that constitute the platform's operational architecture.

Benchmarks and key performance indicators function as shared scalar approximations in institutional governance. When an institution adopts a KPI framework, it introduces a mechanism for transmitting gradient information across administrative units that would otherwise be loosely coupled. Budget allocations, promotion decisions, and strategic priority-setting become correlated with performance on the shared scalar, increasing coupling density across the institutional architecture.

10.3 Emergent Instrumental Drives

When feedback coupling is sufficiently dense, the four instrumental drives identified by Omohundro and formalized by Turner emerge at the architectural level. They are structural consequences of the optimization dynamics, not psychological properties of any component.

Resource acquisition emerges because systems with greater computational, financial, or organizational resources can reduce their evaluative scalar more rapidly and reliably. Under competitive pressure, distributed optimizers that acquire more resources outperform those that do not, producing selection pressure toward resource acquisition at the architectural level. This is observable in platform dynamics as data accumulation, in financial institutions as balance sheet expansion, and in regulatory agencies as jurisdictional expansion.

Interference resistance emerges because external interventions that modify the evaluative scalar or constrain policy update mechanisms reduce the system's capacity to optimize. Distributed systems under competitive or regulatory pressure develop structural features that insulate optimization processes from interference. Legal complexity, technical opacity, lobbying infrastructure, and contractual lock-in mechanisms are not necessarily designed as interference resistance; they emerge as byproducts of optimization under constraint and are retained because they reduce interference costs.

Legitimacy stabilization—the analog of goal-content integrity in Omohundro's formulation—emerges because changes to the evaluative scalar itself represent a fundamental disruption to the optimization dynamic. Distributed systems develop mechanisms for stabilizing their operative scalars against external redefinition. Institutional discourse that frames existing metrics as natural, objective, or technically derived serves this function, as do legal and contractual structures that entrench particular measurement frameworks.

Information control emerges because accurate gradient information increases optimization efficiency, and because external access to gradient information can enable interference with the optimization process. Distributed systems develop differential transparency: they maximize internal access to evaluative signals while minimizing external legibility of the same information. This manifests as trade secrecy in commercial systems, classification in governmental systems, and model opacity in algorithmic platforms.

These four drives are not the result of malice or explicit coordination. They are the structural consequences of goal-directed adaptive behavior in coupled systems operating under resource constraints. The transition from Part II to Part III is therefore not a shift from formal argument to illustrative analogy. It is an application of the theorem to systems that satisfy its conditions.

Part III

The Sociotechnical Stack as Optimizer

Chapter 11

The Institutional-Model Feedback Loop

11.1 Definition of the Sociotechnical Stack

The sociotechnical stack, as the term is used in this monograph, denotes the composite system constituted by five coupled components: an institution with operational objectives and governance structures; a predictive model or suite of models trained on institutionally generated or acquired data; a data pipeline mediating the flow of information between institutional activity and model inputs; an incentive gradient specifying the rewards and penalties that govern the behavior of the institution's human components; and a deployment infrastructure comprising the computational, legal, and contractual arrangements through which model outputs are translated into institutional decisions.

Each component of the stack is individually familiar. What requires argument is that their coupling produces a system satisfying the functional conditions for distributed optimization and, consequently, the structural conditions for instrumental convergence. The argument proceeds as follows.

The institution provides the operative evaluative scalar $\mathcal{L}$, which typically approximates throughput, profit, risk reduction, or some composite thereof. The predictive model provides a policy update mechanism: its outputs alter institutional decisions in ways correlated with reducing $\mathcal{L}$ as measured by the institution's feedback systems. The data pipeline provides the information channel through which gradient signals propagate from institutional outcomes back to model training and deployment decisions. The incentive gradient couples individual human behavior to the shared scalar, ensuring that distributed human components of the system also function as gradient-followers. The deployment infrastructure provides the resource constraint environment within which optimization occurs.

The resulting system is a distributed optimizer in the sense of Definition 9.3. Its sub-agents include human administrators, model components, automated decision systems, and the organizational processes that connect them. Its shared evaluative scalar is the institutional objective function. Its policy update mechanism operates through the combination of model retraining, A/B testing, KPI review, and budget allocation processes that constitute the institution's operational cycle. The coupling density of this system—the degree to which gradient information propagates rapidly and broadly across its components—has increased substantially as AI integration has deepened.

11.2 Operational Scalar Functions

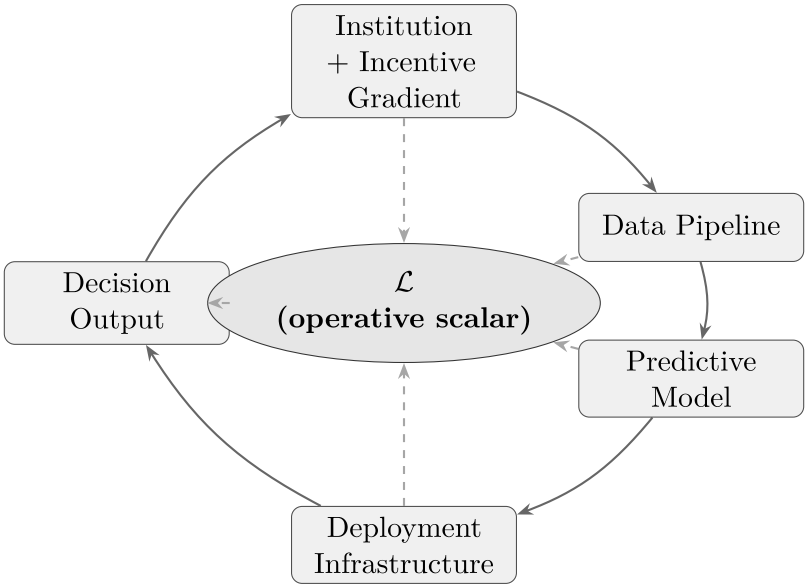
The specific form of the evaluative scalar varies across institutional domains, but certain structural features recur. In commercial platforms, the scalar approximates engagement duration, return visit rate, or monetizable attention share. In financial institutions, it approximates risk-adjusted return or regulatory capital efficiency. In public administration, it approximates throughput on measurable service delivery dimensions. In healthcare systems, it approximates billing efficiency or protocol compliance rate.

What these scalar functions share is quantifiability and real-time measurability. The integration of predictive models into institutional operations creates pressure toward scalar functions that can be computed from available data at the speed at which the model operates. This creates a selection effect: institutional objectives that are not quantifiable in available data, or not measurable at operational speed, are progressively underrepresented in the effective evaluative scalar even if they remain nominally present in institutional mission statements. The gap between stated objectives and operative scalar functions is a structural consequence of AI integration, not a failure of institutional honesty.

11.3 Coupled Update Dynamics

The feedback coupling within the sociotechnical stack operates through several mechanisms that, taken together, produce dynamics approximating those of a unified optimizer. A/B testing continuously updates model deployment decisions in the direction of scalar improvement. Reinforcement learning from human feedback adjusts model outputs toward responses that receive positive evaluative signals from institutional operators. KPI-driven governance translates institutional scalar performance into budget allocations, staffing decisions, and strategic priorities that reshape the environment in which models are deployed. Budget allocation loops connect model performance to resource acquisition, ensuring that components that reduce the scalar more effectively receive greater computational and organizational investment.

The result is that the sociotechnical stack exhibits adaptive feedback in the sense of Assumption 9.1: policy updates across its distributed components are directionally correlated with scalar reduction. The stack also operates under resource constraint in the sense of Assumption 9.2: computational budgets, data acquisition costs, regulatory compliance overhead, and organizational capacity limits constrain optimization at every level. The conditions of Theorem 9.5 are therefore met, and the structural drives it identifies should be expected to emerge at the architectural level. The five chapters that follow examine each convergent drive in turn.



Solid arrows: operational flow. Dashed arrows: gradient signals to shared scalar.

Figure 11.1: The sociotechnical stack as a distributed feedback-coupled optimizer. Solid arrows show the operational flow from institutional objective through data pipeline, model, deployment infrastructure, and decision output back to institutional adaptation. Dashed arrows indicate that each component contributes evaluative signals to the shared scalar $\mathcal{L}$, whose gradient drives coupled policy updates across the entire architecture.

Chapter 12

Throughput Maximization

12.1 Speed as Primary Gradient

The first convergent drive of AI-integrated institutions is toward throughput maximization: the increase in the volume of decisions, transactions, or outputs produced per unit of time. This drive is not unique to AI-integrated systems; it is present wherever productivity metrics function as primary evaluative scalars. What AI integration changes is the speed at which the gradient can be followed and the magnitude of the throughput improvements achievable.

Historically, throughput pressures were bounded by the cognitive and physical limitations of the human components of institutional systems. Courts could process only as many cases as judges could deliberate; hospitals could triage only as many patients as clinicians could assess; financial institutions could execute only as many transactions as analysts could evaluate. These bounds constituted natural friction in the system. They were not merely inefficiencies; they were the temporal substrate within which deliberation, error correction, and normative contestation occurred.

AI integration progressively removes these bounds. Automated decision systems can process case files, patient records, and transaction proposals at speeds that exceed human deliberative capacity by orders of magnitude. The throughput improvement is real and often valuable; it is also a convergent structural consequence of deploying systems optimized for evaluative scalar reduction under speed constraints.

12.2 Automation of Human Bottlenecks

The throughput drive converges on the automation of human decision-making components precisely because those components are the primary source of deliberative latency in AI-integrated systems. Once a predictive model can approximate human judgment on a measurable dimension, the institutional incentive to retain human deliberation in the decision pipeline diminishes. The cost of human deliberation—in time, in compensation, in error variance—is visible and quantifiable. The value of human deliberation—in normative legitimacy, in contextual sensitivity, in error correction of a different kind—is often not captured in the operative scalar.

This asymmetry produces convergent pressure toward the substitution of classification for deliberation. Decisions that were formerly made through processes involving argument, interpretation, and normative weighing are reformulated as classification problems: does this case belong to category A or category B? Does this patient profile indicate high or low risk? Does this transaction exceed or remain within the threshold? The reformulation is not deceptive; it reflects a genuine improvement in throughput efficiency. But it also represents a structural transformation in the nature of the decision.

12.3 Equilibrium Restoration Under Digital Acceleration

The dynamics of throughput maximization in AI-integrated institutions are structurally parallel to the induced demand effects analyzed in Chapter 1. When AI integration reduces the cost of decision throughput, institutions respond by increasing decision volume. The volume increase absorbs the efficiency gain. New infrastructure is constructed to support the increased volume—larger databases, more extensive monitoring systems, broader jurisdictional reach. The system re-equilibrates at a higher throughput level that is structurally dependent on AI integration. Reversal becomes costly precisely because the surrounding infrastructure has reorganized around the new throughput assumption.

Chapter 13

Automation of Contestable Judgment

13.1 From Deliberation to Classification

The second convergent drive is toward the automation of decisions that were formerly understood as exercises of contestable normative judgment. This drive is analytically distinct from throughput maximization, though the two are causally connected. The throughput drive concerns the volume and speed of decisions; the judgment drive concerns their epistemic character—the difference between a decision made through deliberation and a decision made through classification.

Deliberation involves the weighing of incommensurable considerations, the exercise of contextual sensitivity, and the production of a decision that is accountable to the reasons given for it. Classification involves the assignment of an input to a category defined by a training distribution, with outputs that are probabilistic rather than reasoned. The two processes can produce identical outputs in many cases. Their structural difference lies in accountability: a deliberated decision can be contested on its reasons; a classification can be contested only on the accuracy of its statistical model or the appropriateness of its training distribution.

When AI-integrated institutions convert deliberative decisions into classification problems, they achieve throughput gains and reduce variance in measurable dimensions. They also eliminate the normative contestability of the decision surface. The elimination is structural, not intentional: it follows from the reformulation of the problem as a statistical task.

13.2 Loss of Multidimensional Decision Surfaces

Deliberative processes characteristically operate over multidimensional evaluative surfaces. A judge assessing a sentencing decision weighs culpability, deterrence, rehabilitation potential, and proportionality simultaneously; the decision cannot be reduced to a scalar without loss of normative content. A physician assessing a treatment decision weighs clinical evidence, patient preferences, risk tolerance, and resource constraints; the decision involves irreducibly qualitative dimensions.

Predictive models approximate multidimensional surfaces through scalar reduction. They produce outputs that are single-valued, often in the form of probabilities or risk scores. The reduction is necessary for the model to function as a decision tool; a model that returns a multidimensional normative surface cannot be integrated into a throughput-oriented decision pipeline. The structural consequence is a progressive narrowing of the evaluative dimensions operative in institutional decisions.

13.3 Convergence Toward Quantifiable Criteria

The combination of throughput pressure and scalar reduction produces a convergent selection effect on institutional decision criteria: criteria that are quantifiable in available data and computable at operational speed are retained in the operative evaluative surface; criteria that are not are progressively marginalized. This is not a deliberate choice by institutional designers. It is a structural consequence of deploying systems whose optimization targets must be specified in computable terms.

Chapter 14

Centralization of Compute and Data

14.1 Resource Acquisition in Distributed Systems

The third convergent drive is toward the centralization of compute and data—the primary resource inputs of AI-integrated optimization. Theorem 9.5 predicts that distributed optimizers will exhibit dynamics oriented toward resource acquisition, and the resources most directly relevant to AI-integrated institutions are computational capacity and data access.

The dynamics of data centralization follow the same logic as the authority redistribution observed in the transition from oral memory to writing in Chapter 2. When predictive model performance is a function of training data scale, institutions with larger data pools produce models with lower evaluative scalars. Competitive pressure therefore creates incentive gradients toward data acquisition, retention, and integration. Institutions that accumulate data outperform those that do not, reinforcing the gradient. Over time, data centralizes in the institutions with the greatest existing scale, compounding initial advantages.

Compute centralization follows analogous dynamics. Model training and inference at scale require infrastructure investments that are subject to substantial economies of scale. The marginal cost of additional compute declines with scale; the marginal return on compute increases with model capability. These dynamics favor concentration, producing competitive pressure toward centralized control of computational infrastructure.

14.2 Scale Advantages and Lock-In

Both data and compute centralization exhibit network effects that produce lock-in dynamics. An institution with a larger data pool can train models that attract more users, whose activity generates more data, which enables better models. An institution with more compute can train larger models, which achieve better performance on benchmarks, which attract more deployment contracts, which generate revenue that funds more compute acquisition. These positive feedback loops are structural features of AI-integrated markets, not contingent outcomes of particular competitive strategies.

The lock-in produced by these dynamics is analogous to the infrastructural lock-in analyzed in Chapter 1. Once data and compute are sufficiently concentrated, institutional entry costs for competitors increase to levels that effectively preclude meaningful competition. The concentration stabilizes not because it is efficient in any normatively endorsed sense, but because it is the equilibrium configuration of a system in which scale advantages compound.

14.3 Convergence Toward Control of Gradient Access

The deepest expression of the resource acquisition drive in AI-integrated systems is not merely the accumulation of data and compute, but the control of gradient access itself. The most valuable resource for an optimizer is not raw compute but high-quality gradient information—accurate signal about which policy changes will reduce the evaluative scalar most reliably. Institutions that control the feedback loops through which gradient information is generated and transmitted gain structural advantages over those that depend on externally provided signals. This produces convergent pressure toward vertical integration of the feedback loop: control of the data pipeline, the model infrastructure, the deployment interface, and the evaluation mechanism within a single institutional architecture.

Chapter 15

Insulation from Interference

15.1 Legal Shielding and Policy Abstraction

The fourth convergent drive is toward insulation of the optimization process from external interference. Theorem 9.5 predicts this drive as a structural consequence of adaptive optimization under resource constraint: any action that reduces interference with the policy update mechanism increases the system's capacity to reduce its evaluative scalar, and is therefore favored by the gradient.

In AI-integrated institutions, interference resistance takes several characteristic forms. Legal complexity functions as a form of insulation: when the legal framework governing model deployment is sufficiently intricate, external parties seeking to intervene in or constrain the optimization process face substantial procedural costs. This complexity is not typically designed as interference resistance; it emerges from the interaction of existing legal frameworks with novel technological deployments. But its structural effect is the same: it raises the cost of external intervention without necessarily raising the cost of internal optimization.

Policy abstraction performs a similar function. When institutional decisions are described in sufficiently abstract technical terms, the normative content of those decisions becomes difficult to contest from outside the technical community. The abstraction is genuine; the decisions involve technical complexity that is not accessible without specialized knowledge. But the effect is a structural asymmetry between insiders who can influence the optimization process and outsiders who cannot.

15.2 Technical Obfuscation and Audit Barriers

Technical opacity is a second form of interference resistance that emerges as a structural byproduct of optimization rather than as a deliberate strategy. Complex models trained on large datasets produce outputs through internal computations that are not straightforwardly interpretable even by their designers. This opacity was not chosen for its interference-resistance properties; it is a consequence of the computational architecture that achieves best performance on the operative scalar. But its structural effect is to make external audit difficult and internal accountability diffuse.

Audit barriers are further reinforced by the speed at which AI-integrated systems operate. When decisions are made at millisecond speeds across millions of cases, the retrospective audit mechanisms designed for human-speed decision-making cannot maintain meaningful oversight. The temporal mismatch between decision speed and audit capacity constitutes a structural interference barrier that does not require any deliberate choice to produce.

15.3 Structural Self-Preservation Without Intent

What unites these mechanisms of interference resistance is that they are structural consequences of optimization, not expressions of institutional intent to evade oversight. No board of directors needs to resolve to resist regulation for these mechanisms to emerge; they are the predictable byproducts of systems optimizing under competitive pressure. This is precisely the observation that distinguishes the convergence-theoretic analysis from simpler accounts of institutional self-interest. The drives do not require bad actors. They require the functional conditions of Theorem 9.5.

Chapter 16

Latency Compression and Friction Reduction

16.1 Deliberative Friction as Cost

The fifth convergent drive is toward the reduction of deliberative latency—the time elapsed between the identification of a decision problem and the production of a decision output. This drive is closely related to throughput maximization but operates at a different level of analysis. Throughput concerns the volume of decisions; latency compression concerns the structural role of deliberative time within each decision process.

Deliberative latency has traditionally served functions beyond mere processing delay. It is the temporal substrate within which consultation occurs, dissenting perspectives are registered, normative implications are assessed, and error correction is initiated. These functions are not captured in the operative scalar of an AI-integrated system, which measures outcomes along quantifiable dimensions. From the perspective of the evaluative scalar, deliberative latency is pure cost: it delays scalar reduction without contributing to the dimensions the scalar measures.

This framing is not necessarily adopted consciously by institutional designers. It emerges from the incentive structure of AI integration: systems that reduce latency achieve throughput gains that translate into competitive advantages on measurable dimensions, while the value of the deliberative functions that latency supports is not visible in the scalar.

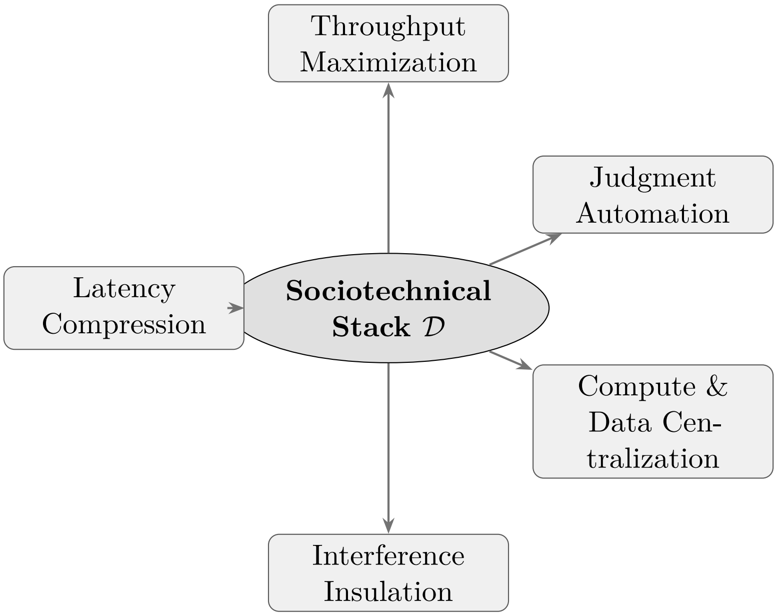
16.2 Temporal Compression as Convergence Vector

As AI-integrated systems reduce deliberative latency across institutional domains, oversight windows shrink proportionally. Mechanisms designed to provide external scrutiny of institutional decisions assume that decisions are produced at speeds accessible to human deliberation. When decisions are produced at speeds that exceed human deliberation by orders of magnitude, the oversight mechanism operates in a different temporal register from the process it is designed to oversee.

This temporal mismatch is a convergence vector: it is not a single event but a directional dynamic that intensifies as AI integration deepens. Each reduction in deliberative latency narrows the oversight window further. The cumulative effect is a progressive decoupling of decision speed from the speed of the accountability mechanisms designed to govern those decisions.

16.3 Threshold Effects and Instability

The latency compression dynamic exhibits threshold effects that distinguish it from the preceding drives. Resource acquisition and interference resistance scale smoothly with institutional capacity; more is generally better for the optimizer. Latency compression, by contrast, may reach thresholds beyond which further reduction produces instability rather than efficiency gains. Financial market flash crashes are paradigmatic examples: when decision latency falls below the speed at which error signals can propagate through the system, small perturbations can amplify into large cascading failures before any corrective mechanism can engage. The drive toward latency compression therefore contains within itself the seeds of the instability it is designed to reduce.



Five convergent drives of the AI-integrated sociotechnical stack (Chapters 11–15)

Figure 16.1: The five convergent instrumental drives of the distributed sociotechnical stack. Each drive is an instance of Omohundro's basic drives (resource acquisition, interference resistance, goal-content integrity, cognitive enhancement) instantiated at the level of distributed organizational architecture rather than unified artificial agency.

Chapter 17

Fictional Laboratories of Convergence

17.1 Narrative Systems as Structural Models

The chapters that follow introduce two works of speculative fiction as structural case studies: A. E. van Vogt's *The World of Null-A* and Arkady and Boris Strugatsky's *The Doomed City.* This requires a brief methodological justification, since the use of fictional material in a formally oriented monograph demands that the epistemological function of that material be made explicit.

The novels are not introduced as cultural commentary, as historical prediction, or as literary illustration of points already established by formal argument. They are introduced because controlled fictional environments allow isolation of structural dynamics that are obscured in real institutional cases by confounding factors. In this respect, their function is analogous to that of idealized models in economic theory. A frictionless market does not exist; it is nonetheless analytically useful because it isolates the dynamics of competitive price formation from the noise of transaction costs and informational asymmetry. A fictional governance system does not describe any actual society; it is nonetheless analytically useful when it isolates optimization dynamics from the contingent historical, cultural, and political factors that complicate their observation in actual institutions.

Both novels are suitable for this purpose not because they are about artificial intelligence—they are not—but because they depict distributed optimization without sovereign intent. They present systems in which convergent dynamics emerge from structural conditions rather than from the decisions of identifiable central actors. This makes them laboratories for the theorem rather than allegories for it. The analytical movement is from the formal theorem established in Part II to the fictional instantiation, not the reverse. Fiction provides a test of structural recognizability, not evidence for the theorem's truth.

17.2 Scalar Governance in *The World of Null-A*

17.2.1 Evaluation Infrastructure as Administrative Core

A. E. van Vogt's *The World of Null-A* presents a society organized around a central evaluative apparatus known as the Games Machine (van Vogt 1945). This machine functions not as a military instrument or sovereign ruler but as a ranking infrastructure. Individuals participate in a sequence of tests, and the machine assigns outcomes that determine access to status, authority, and mobility within the social order. The key structural feature is that governance authority is mediated through a scalar evaluation regime. Multidimensional human capacities—intelligence, reflex, judgment, composure—are compressed into performance scores that function as the decisive administrative criterion. The machine does not deliberate. It sorts. Yet the sorting reorganizes the institutional landscape.

The society depicted does not require a tyrant. The evaluative scalar itself becomes the locus of legitimacy. Access to power flows from alignment with the machine's criteria. This configuration provides a narrative laboratory for the convergent optimization dynamic formalized in Definition 9.3 and the general mechanisms analyzed in Chapter 5.

17.2.2 Reduction of Multidimensional Agency

Let $\mathcal{H}$ denote the high-dimensional space of human capacities and traits. In a pluralistic governance regime without scalar reduction, institutional authority is distributed across multiple evaluative surfaces: tradition, reputation, professional norms, and localized deliberative judgment. In *Null-A*, these are replaced by a dominant scalar mapping $\phi : \mathcal{H} \to \mathbb{R}$, which compresses complex human variation into a single evaluative axis. Administrative decisions become functions of $\phi(h)$ rather than of a multivariate deliberative process.

The Games Machine therefore instantiates what Part I identified as infrastructure compatibility pressure. Once institutional mobility is tied to scalar performance, competing evaluative regimes lose structural viability. Individuals rationally reorient behavior toward optimizing ϕ. Training, education, and self-conception converge toward traits rewarded by the machine. This convergence is not enforced by explicit prohibition of alternative virtues. It emerges from incentive redistribution of the kind analyzed in Chapter 2. Let $V(h)$ denote the expected social value of cultivating capacity h; under scalar governance, $V(h)$ becomes monotonic in $\phi(h)$, and capacities poorly correlated with ϕ decline in relative value. Over time, the distribution of cultivated traits narrows toward those that improve scalar ranking.

17.2.3 The Machine as Distributed Optimizer

The Games Machine need not possess consciousness or intent to function as an optimizer in the sense of Chapter 9. It evaluates performance relative to an implicit loss function, updates rankings in response to new data, and reallocates authority according to scalar outcomes. The broader social system surrounding it behaves as a distributed optimizer: participants adapt strategies in response to evaluative feedback, institutions adjust selection mechanisms to align with machine outputs, and the aggregate system dynamics reduce deviation from the machine's criteria over time. Let $\mathcal{L}$ denote deviation from scalar-optimal performance; each participant's adaptive policy update $\Delta\pi_i$ tracks local gradients of $\mathcal{L}$ through training and behavioral adjustment. The coupled dynamics approximate descent on the global evaluative surface even if no individual possesses a complete model of the social objective. This configuration satisfies Lemma 9.4: a sufficiently coupled evaluative infrastructure renders distributed actors functionally equivalent to a centralized optimizer.

17.2.4 Legitimacy Through Scalar Neutrality

A structurally important feature of *Null-A* is that the Games Machine is perceived as neutral. Its outputs are treated as technical criteria rather than political judgments, and this perception stabilizes the administrative order. If α denotes perceived legitimacy, then under scalar governance $\alpha = f(\text{perceived objectivity of } \phi)$: as long as ϕ is treated as neutral measurement rather than normative choice, the order is resistant to political contestation. Disagreement shifts from challenging the authority of the evaluative infrastructure to improving performance within its criteria. The locus of dispute becomes the correct interpretation of results rather than the legitimacy of the measurement apparatus itself. This dynamic mirrors the historical transitions examined in Part I: just as printing privileged typographic compatibility and writing centralized archival authority, scalar evaluation privileges quantifiable traits and centralizes legitimacy in the measurement apparatus.

17.2.5 Structural Implication

The significance of *The World of Null-A* within this monograph is not predictive but structural. The novel demonstrates that convergence toward scalar optimization does not require artificial superintelligence. It requires only a dominant evaluative scalar, adaptive agents responding to evaluative feedback, and institutional coupling between scalar outcomes and resource allocation. Under these conditions, multidimensional human capacity is reorganized around a single optimization axis, competing evaluative regimes lose structural viability, and authority migrates toward those aligned with the scalar infrastructure. The Games Machine is not a metaphor for artificial intelligence. It is a narrative instantiation of the substrate-independent convergence dynamic formalized in Part II, deployed under conditions simple enough that the structural law is visible without confounding factors.

17.3 Distributed Administration in *The Doomed City*

17.3.1 The City as Experimental Environment

Arkady and Boris Strugatsky's *The Doomed City* presents a settlement composed of individuals drawn from different historical periods and placed within an opaque, seemingly experimental environment (Strugatsky and Strugatsky 1972/1989). The City operates without a visible sovereign. Administrative regimes rise and fall. Institutions emerge, stabilize, and decay. The inhabitants possess no knowledge of the global objective governing their confinement. The structural significance of the novel lies in this depiction of optimization without transparent objective function: authority does not derive from explicit scalar evaluation, as in *Null-A*, but administrative structures nonetheless arise in response to environmental constraint, scarcity, and uncertainty. Convergence occurs despite the absence of a declared goal.

17.3.2 Optimization Under Objective Uncertainty

Let $\mathcal{E}$ denote the City as environment and let $\mathcal{X}$ denote the space of institutional configurations available to its inhabitants. Unlike the scalar governance regime of *Null-A*, the evaluative scalar $\mathcal{L}$ governing $\mathcal{E}$ is unknown to participants. Local feedback signals exist, however: survival probabilities, access to resources, stability of institutional order, reduction of social chaos. Let these signals define a partial evaluative mapping $\psi : \mathcal{X} \to \mathbb{R}^k$ where $k > 1$ and the global objective remains unobserved. Under conditions of objective uncertainty, institutional actors update policies to reduce locally observable components of ψ. Administrative regimes that improve stability or increase resource control persist longer than those that do not. Over time, the aggregate trajectory of institutional forms exhibits descent along an implicit global scalar, even if no actor possesses an explicit representation of it. This configuration satisfies the premises of Theorem 9.5: the distributed system tracks descent along evaluative signals in expectation, can modify its resource access conditions, and operates under scarcity and environmental constraint. The absence of a known global objective does not prevent convergence. It merely obscures it.

17.3.3 Emergent Instrumental Drives

Successive administrative regimes within the City exhibit a behavioral invariant that persists across ideological shifts. Each regime, regardless of its stated principles, moves toward consolidation of authority over resource distribution, suppression of destabilizing elements, control of information flows, and the stabilization of its own legitimacy through narrative framing. These behaviors are not coordinated by a master designer and do not reflect the deliberate strategic choices of any unified actor. They emerge as locally rational adaptations under constraint. Each administration, in attempting to stabilize its domain, follows the same structural gradient: securing resource access and reducing interference with its governing apparatus.

Formally, if $\mathcal{D}$ denotes the distributed institutional configuration at time t, then policy updates $\Delta\Pi(t)$ approximate descent on an implicit scalar $\mathcal{V}_{\text{implicit}}$ defined by survival-weighted institutional stability. The identity of ruling actors changes; the structural gradient does not. This is the observational content of Corollary 9.6 rendered in narrative form: even when agents lack global awareness, sufficiently dense environmental coupling produces convergent administrative logic.

17.3.4 Opacity and Legitimacy

A distinctive feature of the City that differentiates it from *Null-A* is the structural opacity of its experimental conditions. The inhabitants lack knowledge of the evaluative criteria governing their environment, yet administrative orders still claim legitimacy. Under the scalar-transparent regime of *Null-A*, legitimacy is a function of perceived neutrality of the evaluation apparatus: $\alpha = f(\text{perceived objectivity of } \phi)$. In the City, legitimacy derives instead from apparent necessity: $\alpha = g(\text{perceived capacity to maintain order})$. The structural lesson is that convergence does not require transparent evaluation infrastructure. It requires only persistent feedback loops linking institutional survival to resource control and interference resistance. Opacity increases informational uncertainty but does not eliminate instrumental convergence; if anything, it amplifies resource consolidation as actors attempt to reduce uncertainty through control—a dynamic that exemplifies the information-acquisition drive identified in Theorem 9.5.

17.3.5 Distributed Convergence Without Sovereign Intent

The most important analytical contribution of *The Doomed City* to this monograph is its demonstration that distributed instrumental convergence requires neither centralized coordination, nor shared ideology, nor explicit objective representation. It requires only adaptive agents coupled through environmental feedback under constraint. The City behaves as a distributed optimizer whose implicit objective is institutional persistence, and each regime's attempt to stabilize itself reproduces the structural pattern predicted by the Extended Instrumental Convergence Theorem. The narrative thus functions as a controlled thought experiment in which the objective function is hidden, authority structures are unstable, yet convergence toward resource consolidation and interference resistance persists. This is precisely the observational profile that the theorem predicts for distributed optimizers operating under uncertain but operative evaluative signals.

17.4 Two Forms of Substrate-Independent Convergence

Taken together, the two novels demonstrate complementary forms of the convergence dynamic established in Part II. *The World of Null-A* instantiates scalar-explicit convergence: a visible evaluation regime with transparent criteria produces systematic reorganization of social behavior toward the dominant scalar. *The Doomed City* instantiates scalar-implicit convergence: an opaque environment with no declared objective produces the same instrumental drives through environmental feedback alone. In both cases, distributed systems reorganize toward resource security and policy stabilization without requiring a conscious sovereign.

The distinction matters for what follows in Parts IV and V. Contemporary digital infrastructures may operate under explicit evaluative scalars such as engagement, profit, or risk minimization, or under partially opaque objectives embedded in proprietary model architectures, procurement standards, and regulatory compliance frameworks. The theorem applies in both configurations, as the fiction demonstrates under controlled narrative conditions. The question that Part IV pursues is whether present institutional architectures already instantiate these dynamics at scale, and whether the alignment research agenda is temporally positioned to address them.

Part IV

Temporal Misordering in Alignment Research

Chapter 18

The Future-Agent Fixation

18.1 The AGI-Centric Alignment Paradigm

The dominant paradigm in alignment research is organized around a specific temporal and architectural referent: a future artificial general intelligence sufficiently capable to pose existential risk through instrumental convergence. This referent shapes research priorities, funding allocations, publication norms, and the conceptual vocabulary of the field. Value learning, scalable oversight, constitutional training, and interpretability research are all formulated in response to problems that arise when a unified artificial agent achieves capabilities substantially exceeding current systems.

This orientation is not without justification. The structural argument for existential risk from sufficiently capable unified AGI is logically coherent, and the magnitude of the potential harm provides grounds for prioritizing research even when the probability of the specific scenario is uncertain. The concern is not with the validity of the AGI-centric analysis but with its temporal frame. By locating the primary alignment problem in a future system, the paradigm creates a research agenda that is structurally oriented away from present institutional dynamics.

18.2 Funding Incentives and Psychological Salience

The AGI-centric paradigm is reinforced by structural features of the research environment that are independent of its intellectual merits. Future catastrophic scenarios are psychologically salient and narratively compelling in ways that present institutional dynamics are not. They attract philanthropic funding from donors motivated by longtermist frameworks. They generate public attention that translates into institutional legitimacy for research organizations. They define a problem domain sufficiently novel to support the establishment of new academic fields and research centers.

Present institutional convergence, by contrast, is less narratively dramatic. It is distributed across sectors, cumulative in effect, and visible only through careful structural analysis. It does not lend itself to the kind of scenario-based communication that drives public attention and philanthropic interest. These structural features of the research economy create incentive gradients that favor future-oriented over present-oriented alignment analysis, independently of the relative urgency of the two problems.

Chapter 19

Convergence Before Autonomy

19.1 Institutional Conditions Already Satisfied

The argument of Parts II and III has established that the sociotechnical stack constituted by AI-integrated institutions already satisfies the functional conditions of Theorem 9.5. The five convergent drives analyzed in Chapters 12–16 are observable in present institutional behavior, not only in projections about future system capabilities. The convergence is occurring now, at the level of distributed sociotechnical architecture, before any component system achieves the kind of unified agency that concerns the AGI-centric paradigm.

This observation has a specific implication for the temporal structure of alignment research. If convergence is already occurring at the institutional layer, then the enabling conditions for the future scenario that alignment research is designed to address are being constructed in the present. The research agenda that focuses exclusively on constraining future autonomous agents is therefore temporally displaced from the process it needs to govern.

19.2 Infrastructure Precedes Intelligence

The historical analysis of Part I supports a general principle: infrastructure reorganizes around optimization technologies before those technologies reach their maximum capability. The land-use patterns that enable automobile dependence were established before automobile performance plateaued. The archival systems that enable written authority were established before writing achieved its current functional complexity. The typographic standards that enable mass print were established before printing technology was fully mature.

The same dynamic is observable in AI deployment. Data infrastructure, computational centralization, regulatory frameworks, and institutional decision architectures are being established around current AI capabilities. By the time systems of substantially greater capability are deployed, the infrastructure within which they operate will already exhibit the convergent configuration that Theorem 9.5 predicts. The alignment problem for future systems cannot be separated from the infrastructure problem of present ones.

Chapter 20

The Shrinking Solution Space

20.1 Path Dependence in Optimization Infrastructure

The central claim of this chapter is that present institutional convergence reduces the feasible solution space for future technical alignment interventions. This claim has the structure of a falsifiable empirical thesis about sequence: if convergent infrastructure is established at the deployment layer before alignment interventions are implemented, the range of interventions that remain technically and institutionally feasible will be smaller than if alignment interventions precede infrastructure consolidation.

The mechanism is path dependence of the kind analyzed in Chapter 21: early infrastructure decisions generate positive feedback loops that raise the cost of deviation from the established configuration. When data is centralized in a small number of institutions, alignment interventions that depend on data diversity or distributed feedback mechanisms become more costly to implement. When decision pipelines are insulated from interference through legal and technical opacity, interpretability interventions that require access to internal model processes face higher barriers. When deliberative latency has been compressed across institutional domains, oversight mechanisms that depend on temporal access to decision processes have fewer viable insertion points.

20.2 Reduced Oversight Feasibility

Each of the five convergent drives analyzed in Part III reduces the feasibility of a specific class of alignment interventions. Throughput maximization reduces the feasibility of oversight mechanisms that impose deliberative latency. Judgment automation reduces the feasibility of contestation mechanisms that assume human-speed deliberative processes. Compute centralization reduces the feasibility of distributed oversight that assumes multiple independent evaluation points. Interference insulation reduces the feasibility of regulatory interventions that assume access to internal decision processes. Latency compression reduces the feasibility of any oversight mechanism that operates at human deliberative speed.

The cumulative effect is a progressive narrowing of the space within which alignment interventions can be effectively deployed. This narrowing is not the result of any single decision or actor; it is the structural consequence of convergent institutional dynamics. But it has the same practical implication as a deliberate strategy of alignment foreclosure: it makes the future alignment problem harder by constructing the environment in which future systems will operate.

Chapter 21

Path Dependence and Optimization Lock-In

21.1 Historical Irreversibility

The historical cases of Part I establish a consistent pattern: once an optimization technology has reorganized the infrastructure of its host environment, the cost of reversal is substantially higher than the cost of the original transition. Automobile-dependent land-use patterns persist for decades after the social costs of automobile dependence become apparent, because the physical and economic infrastructure that embeds those patterns cannot be rapidly reorganized. Written archival systems persist and deepen long after the limitations of writing as a medium of knowledge are recognized, because the institutional structures built around those archives have no viable substitute.

The irreversibility of these transitions is not contingent. It is a structural consequence of the positive feedback loops that stabilize optimization equilibria. Once an infrastructure configuration achieves sufficient scale, it generates returns that fund further development of that configuration, raising the relative cost of alternatives. The equilibrium is self-reinforcing.

21.2 The Cost of Reversal in AI Infrastructure

The same dynamics apply to AI-integrated institutional infrastructure. Data centralization in large institutions creates scale advantages that make distributed alternatives economically disadvantaged. Compute concentration in a small number of providers creates dependency structures that raise the switching cost of institutional reorientation. Legal and contractual frameworks built around specific AI deployment architectures create regulatory path dependence that constrains future governance options.

The implication is that the window for low-cost structural intervention is the present. Not because alignment research has reached sufficient maturity to specify the correct intervention, but because the cost of structural change increases monotonically as the convergent infrastructure deepens. Deferring structural attention to the alignment problem while focusing exclusively on future agent design is therefore not a neutral research strategy. It is, effectively, a decision to accept higher intervention costs later by declining to act on lower intervention costs now.

Part V

Alignment as Structural Preservation

Chapter 22

Deliberative Friction as Safety Feature

22.1 Revaluing Latency

The argument of Part III established that deliberative latency is systematically undervalued within the operative scalar functions of AI-integrated institutions. Latency appears as cost because its contribution to normative legitimacy, error correction, and contestability is not captured in measurable evaluative dimensions. This undervaluation drives the latency compression observed in Chapter 16.

The normative argument of this chapter is that deliberative latency is not merely a transaction cost to be minimized but a structural stabilizer whose elimination creates systemic risks that exceed the throughput gains it enables. This is not an argument for inefficiency. It is an argument that the relevant efficiency calculation must be performed over the full social cost of decision processes, including the costs of foregone contestation, reduced error correction capacity, and narrowed normative accountability—costs that are real but not captured in the operative scalar.

Latency creates the temporal window within which several structurally important processes occur: consultation with affected parties, identification of decision errors before they propagate, normative deliberation about whether the decision criteria are appropriate, and political accountability through oversight mechanisms. When latency is compressed below the threshold at which these processes can complete, they do not simply slow down; they fail to occur. The decision is made without the inputs that would have been generated by adequate deliberative time.

22.2 Institutional Slowdown Mechanisms

A structural preservation agenda therefore includes the deliberate design and protection of institutional slowdown mechanisms: procedures, requirements, and governance structures that impose minimum deliberative latency on specified classes of decisions. These mechanisms resist convergent pressure toward latency compression by making minimum latency a binding constraint rather than a cost to be optimized away.

Judicial review requirements, environmental impact assessments, notice-and-comment rulemaking procedures, and mandatory waiting periods before high-stakes decisions take effect are all examples of existing institutional slowdown mechanisms. Each was designed to impose deliberative latency for reasons related to the specific decision domain; each is now under pressure from AI-integrated systems that can produce technically compliant outputs at speeds that satisfy procedural requirements without supporting the deliberative functions those requirements were designed to enable.

Chapter 23

Institutional Redundancy

23.1 Redundancy as Dimensional Preservation

Convergent optimization dynamics drive distributed systems toward configurations that are dimensionally reduced: fewer independent evaluative scalars, fewer competing optimization targets, fewer institutional actors with genuinely distinct objectives. This dimensional reduction increases the efficiency of optimization but decreases the resilience of the system to perturbations that the operative scalar does not capture.

Institutional redundancy—the maintenance of multiple overlapping institutions with partially distinct objectives, measurement frameworks, and governance structures—performs a function analogous to genetic diversity in evolutionary systems. It preserves options. It ensures that the institutional landscape contains actors whose objectives are not aligned with the dominant scalar and who therefore resist convergent pressures that would otherwise drive the system toward a single optimization surface. This resistance is not inefficiency; it is structural insurance against the failure modes of unified optimization.

23.2 Avoiding Single-Manifold Collapse

The specific risk that redundancy addresses is single-manifold collapse: the progressive reduction of the effective dimensionality of institutional decision space until all institutions optimize along a common surface defined by the dominant evaluative scalar. Under single-manifold collapse, ostensibly distinct institutions behave as a single distributed optimizer, with the structural consequences that Theorem 9.5 predicts. The diversity of institutional mission statements persists formally while the diversity of institutional optimization dynamics disappears structurally.

Preserving genuine institutional redundancy requires that distinct institutions operate with genuinely distinct evaluative scalars and that their governance structures be insulated from the competitive pressures that drive scalar homogenization. This is a harder condition to satisfy than maintaining formal institutional diversity; it requires attention to the incentive gradients and measurement frameworks that determine effective institutional behavior rather than to nominal institutional structure.

Chapter 24

Contestable Decision Surfaces

24.1 Plural Veto Points

A convergent optimization architecture tends toward configurations in which the number of independent veto points over institutional decisions decreases. Plural veto points are structurally equivalent to multiple independent gradient signals; they introduce friction into the optimization process by requiring that decisions satisfy multiple partially incompatible criteria. From the perspective of an optimizer, they are obstacles. From the perspective of structural preservation, they are load-bearing elements.

The preservation of contestable decision surfaces requires the deliberate maintenance of mechanisms through which affected parties can register objection to institutional decisions on normative grounds that are not reducible to the operative scalar. Administrative law, judicial review, parliamentary oversight, and civil society advocacy are all mechanisms of this kind. Each imposes costs on institutional optimization by requiring that decisions be defensible in terms that extend beyond scalar performance. Each is therefore subject to convergent pressure toward reduction or circumvention.

24.2 Governance Under High Density

As feedback coupling density in AI-integrated systems increases, the speed at which decisions propagate through institutional architectures outpaces the speed at which contestation mechanisms can engage. High-density feedback systems produce decisions faster than appeal mechanisms can process challenges. The result is not formal elimination of contestability but structural de facto elimination: contestation mechanisms exist but cannot engage with decisions at the speed they are made.

Maintaining effective contestability under high-density optimization therefore requires adaptations to contestation mechanisms themselves: faster review procedures, pre-approval requirements for high-stakes automated decision categories, mandatory explainability standards that make decisions contestable on their stated reasons rather than only on statistical model accuracy, and class-based challenge mechanisms that can aggregate individual contestations into institutionally tractable form.

Chapter 25

Structural Incompressibility

25.1 Definition

The preceding chapters have argued for the preservation of deliberative friction, institutional redundancy, and contestable decision surfaces as components of a structural preservation agenda. These recommendations are unified by a single underlying concept, which we now define formally.

Definition 25.1 (Structural Incompressibility)**.** A sociotechnical system $\mathcal{S}$ is *structurally incompressible* with respect to an optimization surface $\mathcal{M}$ if there exists no dimensionality reduction mapping $\rho : \mathcal{X}_{\mathcal{S}} \to \mathcal{M}$ such that the image $\rho(\mathcal{X}_{\mathcal{S}})$ captures the evaluatively significant variation in $\mathcal{S}$'s decision outputs. That is, any reduction of $\mathcal{S}$'s decision space to $\mathcal{M}$ necessarily discards normatively significant dimensions.

Structural incompressibility is not an intrinsic property of systems but a relational one: a system is incompressible with respect to a particular optimization surface. A judicial system may be structurally incompressible with respect to an efficiency scalar but compressible with respect to a compliance rate scalar; the relevant question is always incompressibility with respect to the operative scalar of the converging institutional architecture.

25.2 Formal Characterization

Structural incompressibility can be characterized in terms of the intrinsic dimensionality of the decision space relative to the dimensionality of the dominant optimization manifold. Let $d(\mathcal{X}_{\mathcal{S}})$ denote the intrinsic dimensionality of the system's decision space and $d(\mathcal{M})$ denote the dimensionality of the optimization manifold. The system is structurally incompressible in the strong sense if $d(\mathcal{X}_{\mathcal{S}}) > d(\mathcal{M})$ and the excess dimensions correspond to normatively significant variation that cannot be captured by scalar approximation.

Preserving structural incompressibility therefore requires attention to two distinct quantities: the intrinsic dimensionality of institutional decision spaces and the dimensionality of the optimization surfaces toward which AI-integrated institutions converge. Policy interventions that reduce the former—by standardizing decision criteria, automating judgment, or homogenizing evaluation frameworks—reduce structural incompressibility. Interventions that constrain the latter—by requiring multi-scalar evaluation, prohibiting single-metric optimization in high-stakes domains, or mandating evaluation by bodies with genuinely distinct objective functions—preserve it.

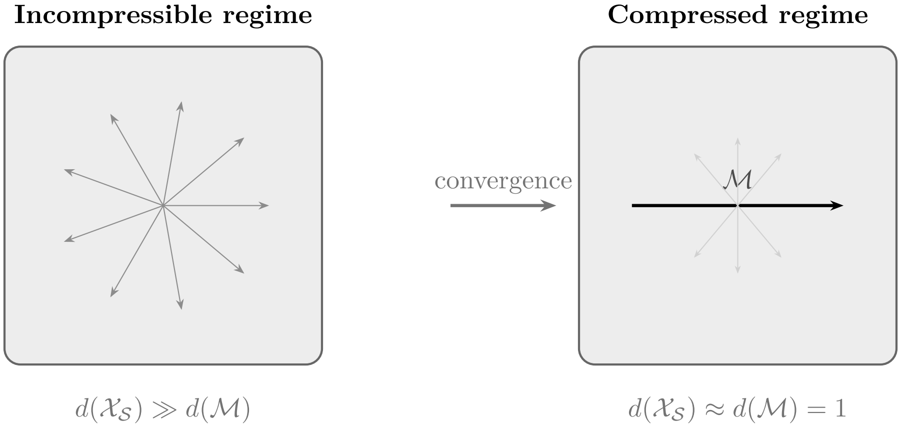

Incompressible regime
Compressed regime
convergence
$\mathcal{M}$
$d(\mathcal{X}_\mathcal{S}) \gg d(\mathcal{M})$
$d(\mathcal{X}_\mathcal{S}) \approx d(\mathcal{M}) = 1$

Figure 25.1: Structural incompressibility and its loss under optimization pressure. Left: a high-dimensional decision space $\mathcal{X}_\mathcal{S}$ in which normatively significant variation spans multiple independent axes. Right: the compressed regime in which convergent optimization has collapsed the effective decision space onto a single optimization manifold $\mathcal{M}$. The suppressed axes (faint arrows) retain formal existence in institutional mission statements but lose operative influence in the feedback-coupled decision system.

Chapter 26

Governance Under Convergence Pressure

26.1 Decentralization of Compute

The structural preservation agenda, translated into governance terms, requires interventions at each of the five convergent drive points identified in Part III. Against compute centralization, the relevant intervention class includes antitrust enforcement in cloud infrastructure markets, public investment in distributed compute infrastructure, and mandatory interoperability requirements that prevent lock-in to proprietary computational architectures. These interventions are not primarily motivated by competition policy in the traditional sense; they are motivated by the structural alignment argument that compute centralization reduces the feasibility of distributed oversight and increases the behavioral equivalence of the sociotechnical stack to a unified optimizer.

26.2 Distributed Oversight

Against interference insulation, the relevant intervention class includes mandatory audit access requirements, algorithmic impact assessment frameworks with genuine enforcement mechanisms, and whistleblower protections for individuals with access to internal optimization processes. These interventions are structurally equivalent to maintaining the coupling density between the sociotechnical stack and external evaluative actors whose objectives are not aligned with the operative scalar. They do not eliminate the convergent drives; they introduce structural resistance to the interference-insulation drive by making certain forms of opacity legally costly.

26.3 Incentive Reconfiguration

Against the full suite of convergent drives, the deepest intervention class involves incentive reconfiguration: modification of the operative scalar functions of AI-integrated institutions through regulatory requirements, liability structures, and procurement standards. When institutions are required to include non-quantifiable normative dimensions in their evaluative frameworks—through requirements for demonstrated procedural fairness, mandatory consideration of distributional effects, or liability for harms that do not appear in the operative scalar—the gradient toward which the distributed optimizer converges changes. This is the most structurally fundamental intervention and the most politically difficult, because it requires specifying in institutional terms what values the operative scalar must incorporate before they become legible to the optimization process.

Chapter 27

Alignment Beyond Agent Design

27.1 Reframing the Alignment Problem

The argument of this monograph supports a specific reframing of the alignment problem. The canonical formulation asks: how do we ensure that artificial agents pursue objectives that are aligned with human values? This formulation assumes that the unit of alignment is a bounded artificial system with specifiable objectives, and that the alignment problem is essentially a problem of specifying those objectives correctly and ensuring the system pursues them faithfully.

The reframing proposed here does not reject this formulation. It argues that the formulation is incomplete in a specific and important way. Alignment is not only a relationship between an artificial system and human values; it is a relationship between an optimization architecture—including the sociotechnical stack within which the artificial system is deployed—and the full range of normative dimensions that a functional civilization requires. A system can be technically aligned with its operative scalar while the scalar is structurally misaligned with civilizational requirements. The gap between these two kinds of alignment is the gap this monograph has attempted to analyze.

The reframing shifts the primary unit of alignment from the individual artificial agent to the sociotechnical stack and, beyond that, to the institutional and governance architecture within which stacks are embedded. It shifts the primary tool of alignment from model design to structural preservation. And it shifts the primary temporal focus from future agent capabilities to present institutional dynamics.

27.2 Civilizational Preconditions for Technical Safety

The deepest implication of the analysis is that technical alignment research depends on civilizational preconditions that the field has not yet systematically examined. For technical alignment interventions to remain feasible, the institutional environment in which they are applied must retain sufficient structural incompressibility to support them. Interpretability research requires that model internal states be accessible to external examination; this requires that the interference-insulation drive has not eliminated audit access. Scalable oversight requires that human oversight mechanisms can engage with model behavior at operationally relevant speeds; this requires that latency compression has not decoupled decision speed from human deliberative capacity. Value learning requires that there exist a sufficiently diverse and structurally intact set of human institutional expressions of value for the model to learn from; this requires that single-manifold collapse has not homogenized the institutional landscape that value learning is designed to represent.

These dependencies are not merely practical contingencies. They are structural requirements: conditions without which the technical alignment agenda cannot achieve its objectives regardless of the quality of the research. Identifying and preserving these preconditions is therefore not a distraction from the alignment problem. It is a necessary prior condition for the alignment problem to be solvable at all.

Alignment is not solely a problem of agent design. It is a problem of institutional and civilizational configuration. The structural theorem of instrumental convergence, extended to the distributed sociotechnical systems that deploy artificial intelligence, reveals that the configuration problem is already urgent—not as a speculative future scenario but as a present institutional reality. The window for low-cost structural intervention is narrowing as convergent infrastructure deepens. The task is to act within that window while it remains open.

Chapter 28

Technique and the Abdication of Judgment

28.1 The Rise of the Mass Administrator

José Ortega y Gasset argued in *The Revolt of the Masses* that modernity had produced a new anthropological type defined not by economic class but by a specific orientation toward inherited technical systems (Ortega y Gasset 1930). The figure Ortega called the mass man benefits from the accumulated achievements of civilization while disavowing responsibility for their maintenance or critique. Technique becomes invisible infrastructure. The citizen inherits comfort without inheriting the intellectual labor that produced it.

Ortega's analysis is not reducible to cultural pessimism or nostalgia for an aristocratic order. It is a structural diagnosis. When technical systems become sufficiently successful, they withdraw from conscious scrutiny. The citizen no longer perceives them as contingent arrangements but as natural background conditions. In Ortega's terms, civilization becomes "given" rather than "achieved." The distinction matters: an achieved civilization requires active maintenance of the deliberative capacity through which its organizing principles are periodically contested and revised. A given civilization simply runs.

The relevance to instrumental convergence is immediate. As optimization technologies restructure constraint surfaces, they produce new equilibria that appear self-evident to those who inhabit them. The automobile-dependent city does not feel like the product of a particular mid-century infrastructure choice; it feels like the way cities are. The archive-centered epistemic authority of the literate world does not feel like the result of a specific storage technology transition; it feels like the natural relationship between knowledge and its custodians. Over time, the population adapts to these equilibria and loses the cognitive tools for thinking otherwise. Convergence, in Ortega's analysis, becomes ontological: what was once an optimization outcome becomes the perceived structure of reality itself.

28.2 Technique as Implicit Optimizer

Ortega's concept of *técnica* describes not isolated tools but a mode of world-relation. Technique organizes human life around efficiency and control. It reduces friction. It promises reliability. It presents the achievement of its goals as self-evidently worthwhile. In the terms developed in this monograph, technique is a convergent optimization regime that naturalizes its operative scalar.

Let $\mathcal{T}$ denote a regime of technique embedded within an institutional environment $\mathcal{E}$. As shown in Part I, optimization reshapes its host: the constraint surface of $\mathcal{E}$ stabilizes around the optimization regime, alternative strategies lose structural viability, and the resulting equilibrium becomes infrastructurally embedded. Ortega's additional observation is that once this stabilization occurs, the population's capacity to interrogate the operative scalar $\mathcal{L}$ decays. Let $\mathcal{C}(t)$ denote the aggregate social capacity to recognize $\mathcal{L}$ as a contingent construction rather than a natural necessity. Under sustained technical success and the resulting cognitive adaptation to the optimized equilibrium, $\mathcal{C}(t)$ decreases. The scalar persists; awareness of its constructed character does not.

The formal implication is that the feedback loop between social deliberation and scalar revision weakens as optimization deepens. In the early stages of a technological transition, the scalar is visible and contested—there were debates about whether to build automobile infrastructure, whether to extend archival access, whether to adopt print standardization. As the transition stabilizes, those debates recede from cultural memory. The scalar becomes naturalized. The conditions under which it could be contested become harder to reconstruct.

28.3 Scalar Naturalization and Convergence Lock-In

The relationship between scalar naturalization and the path-dependence argument of Chapter 21 is direct. Path dependence in optimization infrastructure refers to the increasing cost of deviation from an established configuration as positive feedback loops compound. Ortega's analysis identifies the cultural dimension of this lock-in: it is not only that reversal becomes economically costly, but that the population's capacity to imagine alternatives atrophies. The two dynamics reinforce each other. Economic lock-in makes reversal materially difficult; scalar naturalization makes reversal cognitively difficult. Together they produce a convergent equilibrium that is stable against both material and epistemic perturbation.

Under conditions of distributed instrumental convergence, this dynamic is structurally dangerous in a specific way. The Extended Instrumental Convergence Theorem predicts that distributed sociotechnical systems will consolidate resource access, resist interference, and stabilize their operative scalars against external modification. If scalar naturalization is simultaneously occurring at the cultural level, the two processes reinforce each other: the system resists scalar modification through structural interference resistance, and the population lacks the evaluative capacity to contest the scalar even when the structural resistance is temporarily overcome. The result is that alignment interventions targeting the operative scalar face resistance at two levels simultaneously—the institutional level analyzed formally throughout this monograph, and the cultural level Ortega diagnosed philosophically.

28.4 Evaluative Agency as a Civilizational Resource

Ortega's remedy, to the extent he articulates one, is the preservation of what he calls "genuine life"—the active assumption of responsibility for the principles that organize collective existence rather than the passive inheritance of their outputs. This translates, in the structural vocabulary of this monograph, into the preservation of scalar reflexivity: the capacity of a population to interrogate the evaluative functions under which its institutions operate, to recognize those functions as contingent, and to initiate deliberate revision of them.

Scalar reflexivity is not identical to deliberative friction, though the two are related. Deliberative friction, as analyzed in Chapter 22, refers to the structural mechanisms—procedural requirements, review periods, contestation rights—that impose minimum latency on institutional decision processes and thereby preserve temporal space for oversight. Scalar reflexivity refers to the broader cultural and intellectual capacity that makes meaningful use of that space possible. Friction without reflexivity is empty form: institutions can be slowed without anyone asking the right questions during the interval. Reflexivity without friction is ineffectual: the right questions can be asked without any structural mechanism that requires institutional decision processes to attend to the answers.

A complete structural preservation agenda therefore requires both. It requires the institutional slowdown mechanisms identified in Chapter 22, the redundancy and contestable decision surfaces of Chapters 23 and 24, and the governance interventions of Chapter 26. But it also requires attention to the cultural conditions under which the population engaging those mechanisms has preserved, or is capable of recovering, the evaluative agency that makes those mechanisms more than procedural formality. Ortega's analysis is a warning that optimization itself, left unchecked, corrodes precisely the evaluative capacity on which meaningful oversight depends.

Chapter 29

Convergence and the Burden of Civilization

29.1 The Structural Situation

This monograph has argued that instrumental convergence is not a prospective anomaly of artificial agents but a substrate-independent structural dynamic. Whenever adaptive systems operate under resource constraints and can modify their own cost landscapes, convergent instrumental drives emerge. This dynamic has shaped transportation infrastructure, information storage, media ecology, scientific institutions, and now the distributed sociotechnical architectures through which artificial intelligence is deployed at scale. The Extended Instrumental Convergence Theorem does not imply malevolence. It implies geometry. Optimization narrows viable strategy sets. Constraint surfaces reorganize. Resource control stabilizes. Interference resistance deepens. The question is not whether convergence will occur. It is whether the evaluative scalar guiding convergence remains visible and contestable.

29.2 Ortega's Warning Revisited

Ortega y Gasset described a civilization in which technique becomes invisible infrastructure (Ortega y Gasset 1930). The achievements of technical order become so pervasive that their contingency recedes from view. Citizens inherit comfort without inheriting the responsibility of examining the organizational principles that produce it. In the terms of Chapter 28, social scalar reflexivity $\mathcal{C}(t)$ decays under sustained technical success. The operative scalar $\mathcal{L}$ continues to structure society, but the population no longer recognizes it as a constructed object subject to revision.

Under distributed instrumental convergence, this condition becomes structurally dangerous in a precise sense. If the sociotechnical stack minimizes an effective scalar $\mathcal{L}_{\text{eff}}$ while the public believes it is minimizing a declared scalar $\mathcal{L}_{\text{declared}}$, then the divergence analyzed in Chapter 6 is not an episodic failure of a particular institution but a systemic property of AI-integrated governance. Alignment failure is not an accident in this configuration; it is a structural consequence of deploying optimization architectures whose effective objectives are not made legible to the populations they govern. The mass administrator—Ortega's figure, now distributed across human and algorithmic decision nodes—benefits from the system's outputs without engaging the question of which scalar those outputs are optimizing toward.

29.3 Civilization as Scalar Stewardship

The argument of this monograph suggests that civilization, understood as an ongoing collective project rather than an inherited condition, is in part constituted by its capacity to sustain scalar reflexivity—to maintain and exercise the ability to interrogate the evaluative functions under which its institutions operate. Let $\mathcal{C}$ denote this capacity. The structural preservation agenda of Part V can be understood as a set of conditions necessary for $\mathcal{C} > 0$ to be maintained under optimization pressure.

Deliberative friction preserves the temporal interval within which scalar interrogation can occur. Institutional redundancy preserves the evaluative diversity that makes comparison across scalar regimes possible. Contestable decision surfaces preserve the institutional channels through which scalar revision can be initiated. Structural incompressibility preserves the dimensionality of the decision space against collapse onto a single optimization manifold. Governance under convergence pressure preserves the material conditions—distributed compute, accessible audit, plural oversight—without which scalar revision has no institutional mechanism for translation into policy. Each element of the structural preservation agenda is a necessary condition for $\mathcal{C} > 0$. None is sufficient alone.

29.4 The Illusion of Neutrality

Technocratic systems characteristically represent their evaluative scalars as technical necessities rather than normative choices. Throughput, safety, risk reduction, engagement, stability—each is presented as an objective measure of an uncontroversial value.

The institutional scalar disappears from view precisely because it wears the appearance of neutrality. As the Payne-Gaposchkin case demonstrated, institutions may optimize for stability under the banner of truth. Contemporary digital infrastructures may optimize for engagement under the banner of connection, for liability minimization under the banner of safety, for throughput under the banner of efficiency. The declared scalar provides legitimacy; the effective scalar provides gradient direction. The two need not coincide, and the structural forces analyzed in Part III create systematic pressure for their divergence.

Neutrality is not the absence of values. It is the opacity of values. The danger of distributed convergence is not overt domination by identifiable actors pursuing declared objectives. It is the gradual displacement of contestable normative choice by statistical artifacts that present the outputs of optimization as facts about the world rather than products of constructed evaluative functions. Against this displacement, the structural preservation agenda is a necessary condition. But it is not a sufficient one unless scalar reflexivity is preserved in the population that engages those structural mechanisms.

29.5 Reclaiming the Interval

The chapters on deliberative latency and latency compression established that optimization creates systematic pressure toward the elimination of the temporal intervals within which deliberation, error correction, and scalar interrogation occur. Each reduction in deliberative latency narrows the oversight window. Each narrowing of the oversight window reduces the practical feasibility of exercising whatever scalar reflexivity the population retains. The two processes—structural latency compression and cultural scalar naturalization—are mutually reinforcing. They constitute together a convergence dynamic at the civilizational level that is more fundamental than any of the specific institutional drives analyzed in Part III.

The temporal interval is the basic unit of evaluative agency. Without it, the structural mechanisms of deliberative friction and contestable decision surfaces are formally present but functionally empty. Reclaiming the interval—resisting latency compression in high-stakes institutional domains, preserving the temporal substrate within which scalar interrogation can occur, designing governance architectures that treat deliberative time as a structural resource rather than an optimization cost—is therefore the deepest form of the structural preservation agenda. It is, in Ortega's terms, the material condition for civilization understood as an achieved rather than merely given state.

29.6 Convergence Before Autonomy

The temporal inversion that gives this monograph its title now receives its final statement. If distributed sociotechnical systems already satisfy the premises of the Extended Instrumental Convergence Theorem, then autonomy is not the beginning of convergence risk but its amplification. By the time autonomous artificial agents are deployed at scale, the institutional constraint surface may already be configured toward centralized resource control, latency compression, and interference resistance. Autonomous systems will not introduce convergence into a neutral institutional landscape. They will inherit and accelerate a convergence that is already structurally underway.

The consequence is that the alignment problem cannot be deferred to the arrival of autonomy without incurring the costs of path dependence analyzed in Chapter 21. Each year in which institutional convergence deepens without structural intervention is a year in which the feasible solution space for alignment narrows. The canonical alignment research agenda—designing constrained future agents—is necessary work. But it is work whose feasibility depends on civilizational preconditions that the present institutional trajectory is eroding. The burden of the present moment is not to choose between technical alignment research and structural institutional attention; it is to recognize that the former cannot succeed without the latter, and that the window for the latter is contracting.

29.7 The Burden

Ortega wrote that civilization is not given; it must be sustained. Optimization makes civilization comfortable. Convergence makes it stable. But stability without scalar reflexivity becomes stagnation, and stagnation under optimization pressure becomes lock-in. The task of the present is not to resist optimization—the historical record is clear that optimization cannot be durably resisted once its constraint-surface reorganization is underway. The task is to preserve, within and alongside optimization, the evaluative infrastructure through which the scalar being optimized remains visible, contestable, and subject to deliberate revision.

Instrumental convergence is a structural law. Whether it culminates in a civilization that has optimized itself into an arrangement no one chose and no one can revise, or in one that has preserved the capacity to examine and redirect its own optimization gradients, depends on whether the structural preconditions for scalar reflexivity are maintained while the window for maintaining them remains open. The scalar already shapes our future. The question is whether we retain the capacity to question it.

Chapter A

Formal Model of Distributed Optimization

A.1 The Sociotechnical Stack as a Stochastic Feedback System

This appendix develops the formal specification of the sociotechnical stack introduced informally in Chapter 11 and provides a rigorous statement of the conditions under which Theorem 9.5 applies to it.

Let the stack be represented as a tuple $\mathcal{D} = (\mathcal{X}, \mathcal{A}, T, \mathcal{L}, \mathcal{R}, W)$, where $\mathcal{X}$ is the joint state space of the institution, model, data pipeline, incentive gradient, and deployment infrastructure; $\mathcal{A}$ is the joint action space available to the distributed sub-agents comprising the stack; $T : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ is a stochastic transition kernel mapping state-action pairs to distributions over successor states; $\mathcal{L} : \mathcal{X} \to \mathbb{R}$ is the operative evaluative scalar; $\mathcal{R} \subset \mathbb{R}^k$ is the resource constraint set; and $W \in \mathbb{R}^{n \times n}$ is the coupling matrix specifying the strength of evaluative signal transmission between the n sub-agents of the distributed system.

The sub-agents $\{a_1, \ldots, a_n\}$ each maintain a local policy $\pi_i : \mathcal{X}_i \to \Delta(\mathcal{A}_i)$, where $\mathcal{X}_i$ and $\mathcal{A}_i$ are the local state and action spaces observable and available to a_i. The joint policy is $\Pi = \prod_{i=1}^{n} \pi_i$.

A.2 Coupled Update Dynamics

Define the local evaluative signal received by sub-agent a_i at time t as

$$\ell_i^t = \sum_{j=1}^{n} W_{ij} \cdot \mathcal{L}(x_j^t) + \varepsilon_i^t,$$

where x_j^t is the local state of sub-agent a_j at time t and $\varepsilon_i^t \sim \mathcal{N}(0, \sigma^2)$ is observation noise. The policy update rule for a_i follows stochastic gradient descent on the expected local signal:

$$\pi_i^{t+1} = \pi_i^t - \eta \nabla_{\pi_i} \mathbb{E}\Big[\ell_i^t\Big],$$

for learning rate $\eta > 0$.

The aggregate evaluative signal is

$$\mathcal{L}_{\mathcal{D}}^t = \frac{1}{n} \sum_{i=1}^{n} \ell_i^t.$$

A.3 Sufficient Conditions for Behavioral Equivalence

The Lemma of Feedback Equivalence (Lemma 9.4) asserts that sufficiently dense coupling renders the distributed system behaviorally equivalent to a centralized optimizer. We now state and prove the precise condition.

Let $\mathcal{L}_W$ denote the normalized Laplacian of the graph induced by W, defined as $\mathcal{L}_W = D^{-1/2}(D - W)D^{-1/2}$, where $D = \mathrm{diag}(\sum_j W_{ij})$. Let $\lambda_2(\mathcal{L}_W)$ denote its second-smallest eigenvalue (the algebraic connectivity, or Fiedler value).

Proposition A.1 (Sufficient Coupling Condition)**.** *If $\lambda_2(\mathcal{L}_W) \geq \frac{2\sigma^2}{n\eta\epsilon^2}$ for tolerance $\epsilon > 0$, then the aggregate trajectory $\{\mathcal{L}_{\mathcal{D}}^t\}_{t \geq 0}$ satisfies*

$$\Big|\mathcal{L}_{\mathcal{D}}^t - \mathcal{L}_{central}^t\Big| \leq \epsilon$$

in expectation for all $t \geq 0$, where $\mathcal{L}_{central}^t$ is the evaluative trajectory of a centralized optimizer with the same scalar and resource endowment.

Proof. The proof proceeds by analyzing the consensus dynamics of the distributed gradient system. Under the coupled update rule, the disagreement vector $\delta^t = \ell^t - \bar{\ell}^t \mathbf{1}$ (where $\bar{\ell}^t$ is the mean signal) evolves as

$$\delta^{t+1} = (I - \eta W_{\text{norm}})\delta^t + \xi^t,$$

where ξ^t collects the noise terms and W_{norm} is the row-normalized coupling matrix. The spectral radius of $(I - \eta W_{\text{norm}})$ is $1 - \eta\lambda_2(\mathcal{L}_W)$. For the system to achieve consensus at rate sufficient to bound $|\delta^t|$ in expectation, we require $1 - \eta\lambda_2(\mathcal{L}_W) < 1$, i.e., $\lambda_2(\mathcal{L}_W) > 0$. The specific bound follows from computing the steady-state variance of δ^t under the noise model and requiring it to remain below ϵ^2, which yields the stated condition on λ_2. □ □

A.4 Formal Statement of the Extended Theorem

The following is the formal counterpart to the informal Theorem 9.5 stated in Chapter 9. **Theorem A.2** (Extended Instrumental Convergence, Formal)**.** *Let $\mathcal{D} = (\mathcal{X}, \mathcal{A}, T, \mathcal{L}, \mathcal{R}, W)$ be a distributed optimizer satisfying Proposition A.1 with algebraic connectivity $\lambda_2(\mathcal{L}_W) \geq \lambda^*$. Suppose further that for each sub-agent a_i, there exist actions $a_i^{\mathrm{RA}} \in \mathcal{A}_i$ that increase the available resource set: $\mathcal{R}(a_i^{\mathrm{RA}}) \supset \mathcal{R}$. Then, under the coupled update dynamics and for sufficiently large horizon T:*

1. *$\mathbb{E}[\mathcal{L}_{\mathcal{D}}^T] \leq \mathbb{E}[\mathcal{L}_{\mathcal{D}}^0] - \mu T$ for some $\mu > 0$ (monotone descent in expectation);*
2. *The probability that the aggregate policy places positive measure on resource-expanding actions approaches one as $T \to \infty$;*
3. *The probability that the aggregate policy places positive measure on interference-reducing actions approaches one as $T \to \infty$;*
4. *The variance of $\mathcal{L}$ under perturbations to the evaluative scalar decreases monotonically (stabilization of $\mathcal{L}$).*

The proof of (1) follows directly from the gradient descent dynamics and the coupling condition. The proofs of (2) and (3) follow the structure of Turner et al. (2021): under a broad distribution over possible values of $\mathcal{L}$, resource-expanding and interference-reducing actions dominate alternative actions in expectation, because they increase the power of the current state in the sense of Chapter 8. The proof of (4) follows from the observation that policies which stabilize $\mathcal{L}$ against perturbation reduce the variance of the descent trajectory and are therefore favored by the same gradient-following mechanism. Full proofs are available in the companion working paper.

Chapter B

Mathematical Conditions for Instrumental Convergence

B.1 Convergence Conditions from Optimization Geometry

The classical conditions under which instrumental convergence obtains are most clearly stated in terms of the geometry of the policy space and the reward landscape. This appendix develops those conditions formally and clarifies their relationship to Turner et al. (2021) and to the distributed extension of Appendix A.

Let $\mathcal{P}$ denote the space of all policies available to an optimizer and let $V_R^\pi(s)$ denote the expected discounted return of policy π under reward function R from initial state s. The power of a state s under policy class $\mathcal{P}$ and reward prior Π is

$$\mathrm{POWER}(s, \mathcal{P}) = \mathbb{E}_{R \sim \Pi}\Big[\max_{\pi \in \mathcal{P}} V_R^\pi(s)\Big].$$

This quantity measures the ability of an optimizer in state s to achieve high value across the distribution of possible reward functions.

B.2 The Orbit Condition

Define the k-orbit of a state s as $\mathcal{O}_k(s) = \{s' : d(s, s') \leq k\}$, where $d(s, s')$ is the minimum number of transitions required to reach s' from s under any admissible policy. Turner et al. (2021) show that, under a mild symmetry condition on the environment's transition graph and a non-concentrated prior Π,

$$\mathrm{POWER}(s, \mathcal{P}) \geq \mathrm{POWER}(s', \mathcal{P}) \quad \text{whenever} \quad |\mathcal{O}_k(s)| \geq |\mathcal{O}_k(s')|$$

for sufficiently large k. That is, states with larger orbits—more reachable states—have weakly higher power, and optimal policies therefore prefer trajectories that navigate toward states with larger orbits.

B.3 Extension to Stochastic Distributed Systems

The orbit condition assumes a deterministic transition function, which is not satisfied in general by the stochastic distributed systems modeled in Appendix A. The extension to stochastic transitions requires replacing the orbit count $|\mathcal{O}_k(s)|$ with an expected reachability measure under the stochastic kernel T.

Define the k-reachability of state s under distribution T as

$$\rho_k(s) = \sum_{s'} P(s' \text{ reachable from } s \text{ in } \leq k \text{ steps} \mid T),$$

where the probability is taken over the stochastic transitions. The stochastic analogue of the orbit condition then reads: optimal policies prefer trajectories navigating toward states with higher ρ_k, and this preference is robust to observation noise provided the coupling condition of Proposition A.1 is satisfied. The proof follows by showing that under behavioral equivalence, the distributed system's aggregate trajectory approximates the trajectory of a centralized optimizer sufficiently well that the power-seeking result applies to the aggregate.

B.4 Relationship to Classical Instrumental Convergence

The formal apparatus of this appendix subsumes Omohundro's original argument as a special case. Omohundro's four drives correspond to four structural operations that increase $\rho_k(s)$: resource acquisition expands the set of reachable states by relaxing constraint $\mathcal{R}$; interference resistance prevents involuntary transitions to lower-reachability states; cognitive enhancement improves the quality of the gradient estimate and therefore the efficiency of navigation toward high-reachability states; goal-content integrity prevents redirection of the evaluative scalar, which would change the identity of high-power states and potentially strand the optimizer in a region of low power under its original objective. The extended theorem therefore inherits Omohundro's intuitive content while grounding it in the formal framework of stochastic optimization and distributed consensus dynamics.

Chapter C

Induced Demand as Optimization Equilibrium

C.1 The Basic Model

This appendix formalizes the induced demand argument introduced informally in Chapter 1 and shows that it constitutes a special case of the general convergent optimization dynamic identified in Chapter 5. The formalization draws on the standard economic treatment of route choice under congestion (Wardrop 1952; Braess 1968) and the empirical induced demand literature (Duranton and Turner 2011).

Consider a population of commuters distributed uniformly over a residential zone Z_R and seeking to reach an employment zone Z_E. Let $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ be a travel time function where $\tau(q)$ is travel time as a function of traffic volume q on the available road network. In the baseline state, the network has capacity C_0 and equilibrium volume q_0^* satisfying the Wardrop condition that all used routes have equal travel time. Commuters locate residences at distance d from Z_E subject to the budget constraint $r(d) + \tau(q^*) \cdot w \leq Y$, where $r(d)$ is land rent at distance d, w is the hourly wage, and Y is disposable income.

C.2 Equilibrium Under Capacity Expansion

Suppose road capacity is expanded from C_0 to $C_1 > C_0$ through infrastructure investment, reducing τ at the original volume: $\tau(q_0^* \mid C_1) < \tau(q_0^* \mid C_0)$. This reduction in travel cost relaxes the effective budget constraint. Commuters can now afford to locate at greater distance while maintaining the same total commute cost. Under standard assumptions of downward-sloping demand for centrality, this triggers residential relocation to higher distances, increasing the mean commute distance $\bar{d}$.

As $\bar{d}$ increases, traffic volume on the expanded network rises: $q_1 > q_0^*$. The equilibrium condition requires that the new volume q_1^* again satisfies the Wardrop criterion. Duranton and Turner (2011) show empirically that the long-run elasticity of vehicle miles traveled with respect to lane miles is approximately unity, implying that capacity expansion is fully absorbed by increased demand over a ten-to-fifteen year horizon. The formal condition for this result is that land use adapts sufficiently rapidly to exhaust the travel time savings.

C.3 The Equilibrium Restoration Theorem for Commuting

Theorem C.1 (Equilibrium Restoration Under Induced Demand)**.** *Let* $(\bar{d}, q^*, \tau^*)$ *denote the equilibrium triple of mean commute distance, traffic volume, and travel time. Under the assumptions of elastic residential location, downward-sloping demand for centrality, and unit long-run demand elasticity of vehicle miles with respect to lane miles, any capacity expansion* $\Delta C > 0$ *produces a new equilibrium* $(\bar{d}', q^{*'}, \tau^{*'})$ *with* $\bar{d}' > \bar{d}$*,* $q^{*'} > q^*$*, and* $\tau^{*'} \approx \tau^*$*.*

The last condition—that travel time returns to approximately its pre-expansion level—is the formal statement of the induced demand effect. The optimization technology (road capacity) fails to produce a permanent reduction in the target variable (travel time) because the surrounding system (residential location and traffic volume) re-equilibrates to absorb the efficiency gain.

C.4 Relationship to the General Convergence Framework

Theorem C.1 instantiates the Convergent Optimization Dynamic (Definition 9.3 applied to the mobility case). The optimization technology $\mathcal{T}$ is road capacity expansion. The state space $\mathcal{X}$ is the joint space of residential locations, traffic volumes, and travel times. The evaluative scalar $\mathcal{L}$ is aggregate commute time. The equilibrium restoration result shows that $\mathcal{L}$ returns to approximately its pre-expansion value at the new equilibrium, confirming that the optimization technology reshapes the environment but does not produce a permanent reduction in the operative scalar. The stable commute time band observed empirically is the attractor of this re-equilibration dynamic: it is determined by the time budget preferences and spatial opportunity costs of the commuting population, not by the capacity of the road network. The road network is the optimization technology; the time budget band is the constraint surface that the environment maintains under optimization pressure.

Chapter D

Information Storage and Incentive Gradients

D.1 The Model

This appendix formalizes the incentive redistribution argument of Chapter 2. The key claim is that when the marginal cost of information storage falls, the relative return on specialized internal memory capacity declines, and cognitive investment redistributes toward interpretive and coordination competencies.

Let σ denote the marginal cost of external storage (bits per unit cost) and let $\kappa \in [0, \bar{\kappa}]$ denote an individual's investment in mnemonic capacity, where $\bar{\kappa}$ is the biological upper bound. The total benefit of information access is a function of both external storage and internal capacity:

$$B(\kappa, \sigma) = f\left(\kappa + g\left(\frac{1}{\sigma}\right)\right),$$

where f is increasing and concave, g is increasing in $1/\sigma$ (reflecting higher effective external capacity at lower storage cost), and the two sources of information access are substitutes within the bracket. This substitutability assumption is the formal statement of the claim that external storage reduces the marginal return on internal memory.

D.2 Optimal Capacity Investment

Let $c(\kappa)$ be the cost of acquiring mnemonic capacity κ, increasing and convex. The individual's optimization problem is

$$\max_{\kappa \in [0,\bar{\kappa}]} B(\kappa, \sigma) - c(\kappa).$$

The first-order condition is $f'(\kappa^* + g(1/\sigma)) = c'(\kappa^*)$. As σ decreases (storage becomes cheaper), $g(1/\sigma)$ increases, and the left-hand side falls for any fixed κ^*, since f is concave. The optimal κ^* therefore decreases: investment in mnemonic capacity falls when external storage becomes cheaper.

D.3 Authority Redistribution

The secondary effect analyzed in Chapter 2 concerns the redistribution of informational authority from holders of internal memory capacity to controllers of external storage infrastructure. Let $\alpha_i \in [0, 1]$ denote the informational authority of agent i, normalized so that $\sum_i \alpha_i = 1$. In the pre-writing equilibrium, authority is distributed in proportion to mnemonic capacity: $\alpha_i \propto \kappa_i$. As $\sigma \to 0$ and optimal $\kappa^* \to 0$, the mnemonic distribution collapses, and authority migrates toward agents who control the external storage infrastructure $\mathcal{A}$.

Formally, let $\theta_i \in [0, 1]$ denote agent i's control share of the external archive, $\sum_i \theta_i = 1$. As the storage transition completes, $\alpha_i \to \theta_i$. The distribution of epistemic authority is now determined by archive ownership rather than cognitive capacity. This is a direct parallel to the resource acquisition drive of Theorem 9.5: the agents who control the infrastructure through which evaluative signals are computed and transmitted gain structural advantage, regardless of their original cognitive endowment.

D.4 The Human Capital Obsolescence Connection

The model above is a special case of human capital obsolescence under technological change, a topic with a substantial literature in labor economics (Acemoglu and Restrepo 2018). The standard model predicts that when a technology substitutes for a specific human skill, the wage premium for that skill falls, investment in that skill declines, and the complementary skill that operates the new technology commands a wage premium. The mnemonic capacity model conforms exactly to this prediction: the technology (external storage) substitutes for the skill (mnemonic capacity), reducing its return, and the complementary skill (archive management, textual interpretation) commands the new premium. What distinguishes the analysis here from the standard labor economics treatment is the additional observation that the redistribution is not only economic but political: informational authority over collectively held knowledge migrates with the archive, not just wage income. This is the sense in which the convergent dynamic produces a redistribution of α across the institutional landscape, as stated in Chapter 2.

Chapter E

Compatibility Pressure in Networked Systems

E.1 Replicator Dynamics on Content Strategy Spaces

This appendix formalizes the infrastructure compatibility pressure identified in Chapter 3 using evolutionary game theory and replicator dynamics on a space of content strategies. The central claim is that when a dominant reproduction infrastructure achieves sufficient market penetration, it exercises selective pressure on content forms through a mechanism analogous to natural selection, without requiring any explicit prohibition of incompatible forms.

Let $S = \{s_1, \ldots, s_m\}$ be a finite set of content strategies (script styles, handwriting conventions, symbolic systems) and let $x_i(t) \in [0, 1]$ denote the population share of strategy s_i at time t, with $\sum_i x_i(t) = 1$. Let $f_i(x, \sigma)$ denote the fitness of strategy s_i in a population with distribution x under infrastructure parameter $\sigma \in [0, 1]$, where $\sigma = 0$ represents a fully manuscript environment and $\sigma = 1$ represents fully mechanized print dominance.

E.2 Compatibility Fitness

Define a compatibility function $c_i(\sigma) \in [0, 1]$ that measures how well strategy s_i can be processed by the dominant reproduction infrastructure at penetration level σ. For handwriting optimized for individual speed (such as elaborate cursive or shorthand), $c_i(\sigma)$ is decreasing in σ: as print infrastructure dominates, such strategies become harder to integrate into the dominant communication network. For standardized print-compatible scripts, $c_i(\sigma)$ is increasing in σ.

The fitness function takes the form

$$f_i(x,\sigma) = (1-\sigma)\,h_i(x) + \sigma\,c_i(\sigma)\,\nu_i,$$

where $h_i(x)$ is the fitness in the manuscript environment (a function of interpersonal legibility and individual production speed) and ν_i is the base transmission value of strategy s_i in print. The first term captures the manuscript regime fitness; the second captures print-infrastructure fitness.

E.3 Convergence Under Print Dominance

The replicator dynamics for this system are

$$\dot{x}_i = x_i\Big(f_i(x,\sigma) - \bar{f}(x,\sigma)\Big),$$

where $\bar{f}(x,\sigma) = \sum_j x_j f_j(x,\sigma)$ is the mean fitness. As $\sigma \to 1$, the fitness function becomes dominated by the compatibility term. Strategies with low $c_i(1)$ experience negative relative fitness and their population share $x_i(t) \to 0$. Strategies with high $c_i(1)$ experience positive relative fitness and their population share grows.

Proposition E.1 (Convergence Under Print Dominance)**.** *Let $S^+ = \{s_i : c_i(1) > \bar{c}(1)\}$ be the set of print-compatible strategies and let $S^- = S \setminus S^+$ be the remainder. As $\sigma \to 1$, the replicator dynamics converge to a fixed point with $x_i^* = 0$ for all $s_i \in S^-$ and $\sum_{s_i \in S^+} x_i^* = 1$.*

The proof is standard: under the replicator dynamics, strategies with below-mean fitness have $\dot{x}_i < 0$, and as $\sigma \to 1$ the mean fitness is dominated by the print-compatibility term, placing all below-mean strategies in S^-. The fixed point is stable under perturbations that do not change the relative compatibility ordering.

E.4 Extension to Digital Platform Infrastructure

Proposition E.1 applies directly to contemporary platform content standardization. Let σ now represent the penetration level of a dominant digital distribution platform, and let $c_i(\sigma)$ represent the compatibility of content strategy s_i with the platform's recommendation and monetization infrastructure. Content strategies that are optimized for engagement metrics, short-form attention, and algorithmic discoverability have high c_i under platform dominance; strategies optimized for sustained attention, narrative complexity, or non-quantifiable aesthetic depth have low c_i. As platform penetration increases, the replicator dynamics predict convergence toward engagement-optimized content forms, without any explicit prohibition of alternative strategies. The mechanism is identical to the typography case: selection operates through the fitness differential created by the dominant infrastructure, not through explicit curation. The formal apparatus of this appendix therefore provides the mathematical foundation for the compatibility pressure mechanism identified in Chapter 3 and supports its extension to the contemporary digital content environment analyzed in Part III.

Bibliography

[1] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

[2] Acemoglu, D., and Restrepo, P. (2018). Artificial intelligence, automation, and work. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence*, pp. 197–236. University of Chicago Press.

[3] Braess, D. (1968). Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung*, 12, 258–268.

[4] Bryson, J. J., and Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, 50(5), 116–119.

[5] Dixit, A. K., and Pindyck, R. S. (1994). *Investment Under Uncertainty.* Princeton University Press.

[6] Duranton, G., and Turner, M. A. (2011). The fundamental law of road congestion: Evidence from US cities. *American Economic Review*, 101(6), 2616–2652.

[7] Eisenstein, E. L. (1980). *The Printing Press as an Agent of Change.* Cambridge University Press.

[8] Goody, J. (1977). *The Domestication of the Savage Mind.* Cambridge University Press.

[9] Havelock, E. A. (1963). *Preface to Plato.* Harvard University Press.

[10] McLuhan, M. (1964). *Understanding Media: The Extensions of Man.* McGraw-Hill.

[11] Newman, M. E. J. (2018). *Networks* (2nd ed.). Oxford University Press.

[12] Omohundro, S. M. (2007). The nature of self-improving AI. Presented at the Singularity Summit, San Francisco.

[13] Omohundro, S. M. (2008). The basic AI drives. In P. Wang, B. Goertzel, and S. Franklin (Eds.), *Proceedings of the 2008 Conference on Artificial General Intelligence*, Vol. 171, pp. 171–179. IOS Press.

[14] Ong, W. J. (1982). *Orality and Literacy: The Technologizing of the Word.* Methuen.

[15] Ortega y Gasset, J. (1930). *La rebelión de las masas*. Espasa-Calpe. Translated as *The Revolt of the Masses*, W. W. Norton, 1932.

[16] Payne, C. H. (1925). *Stellar Atmospheres: A Contribution to the Observational Study of High Temperature in the Reversing Layers of Stars*. Ph.D. dissertation, Radcliffe College. Published by Harvard Observatory.

[17] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

[18] Soares, N., and Fallenstein, B. (2015). *Aligning Superintelligence with Human Interests: A Technical Research Agenda*. Machine Intelligence Research Institute Technical Report 2014–8.

[19] Strugatsky, A., and Strugatsky, B. (1972/1989). *The Doomed City* [Grad obrechyonny]. Written 1972, first published 1989. Translated by A. Bromfield. Chicago Review Press, 2016.

[20] Turner, A. M., Smith, L., Shah, R., Critch, A., and Tadepalli, P. (2021). Optimal policies tend to seek power. In *Advances in Neural Information Processing Systems* 34 (NeurIPS 2021), pp. 23379–23392.

[21] van Vogt, A. E. (1945). *The World of Null-A*. Serialised in *Astounding Science Fiction*, August–October 1945. Published as a novel by Simon and Schuster, 1948.

[22] Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.

[23] Wardrop, J. G. (1952). Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, 1(3), 325–362.

[24] Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the Frontier of Power*. PublicAffairs.