Costly Signaling and Coalition Formation Across Biological Scales

Ian Todd

Sydney Medical School

University of Sydney

Sydney, NSW, Australia

itod2305@uni.sydney.edu.au

Abstract

Coalition formation is a coordination problem that recurs across biological scales: how can agents verify mutual commitment before undertaking costly collective action? I argue that costly signaling—the public performance of acts that would be irrational for non-committed agents—provides a general solution. When signals are sufficiently costly, they become reliable indicators of commitment because defectors cannot afford to fake them. This generates a prediction: as coordination stakes increase, optimal signals become increasingly costly and apparently irrational from an individual-payoff perspective. I trace this mechanism from microbial quorum sensing through insect colony recognition to primate coalitions and human ideological movements, showing that the structural features often labeled "religious"—ritual, sacred markers, heresy punishment, identity fusion—emerge as predictable consequences of costly signaling dynamics whenever coordination stakes are high. These features represent convergent solutions to the commitment verification problem, not uniquely human cultural innovations. The framework explains why evidence against group beliefs can strengthen rather than weaken commitment, why internal schisms are often more violent than external conflicts, and why apparently irrational belief persists despite contradicting information. These are not failures of rationality but features of systems optimized for coordination rather than truth-tracking.

Keywords: costly signaling; coalition formation; commitment verification; coordination; cross-scale biology

1 Introduction

Coalition formation presents a fundamental problem for biological agents: how can potential cooperators verify mutual commitment before undertaking costly collective action? The challenge is acute because defection is often profitable—an agent who free-rides on coalition benefits without bearing coalition costs gains an advantage. Any signal of commitment that can be cheaply faked will be faked, eroding the signal's reliability and destabilizing cooperation.

The costly signaling framework, developed in biology by Zahavi [Zahavi, 1975] and formalized in economics by Spence [Spence, 1973], provides a solution: signals that are expensive to produce can be reliable precisely because their cost screens out non-committed signalers. The peacock's tail works as an honest signal of fitness because only genuinely fit males can afford the metabolic and predation costs of maintaining it.

In this paper, I extend costly signaling theory to coalition formation across biological scales. The central claim is this: when coalition coordination requires high-reliability commitment verification, the optimal signaling strategy converges on increasingly costly and apparently irrational displays. A cluster of structural features—ritual performance, sacred markers, heresy punishment, identity fusion, and resistance to disconfirming evidence—emerges as the equilibrium solution across diverse biological systems.

These features are often labeled "religious" when observed in human contexts. But I will argue that they appear wherever coordination stakes are high, from bacterial quorum sensing to insect colonies to primate alliances. The features are not arbitrary cultural elaborations but predictable consequences of costly signaling dynamics.

This perspective has several implications:

- It identifies a general mechanism for coalition formation that operates across biological scales
- It explains structural convergence between systems that differ in substrate and content
- It predicts that evidence against group beliefs can *strengthen* commitment rather than weaken it
- It accounts for why internal deviation is punished more severely than external opposition

The framework builds on recent work in philosophy of biology concerning agency and power across scales [Todd, 2025b]. That analysis defined agency via code formation (internal models that guide action-selection) and power as control of controllers (intervention on other agents' action-selection mechanisms). It established cross-scale continuity from microbes to humans but left open a prior question: how do coalitions capable of collective action form in the first place?

The present paper addresses this gap. Where that earlier analysis examined how agents exert power once coalitions exist, the present paper explains how coalitions form under conditions of defection risk. The two problems are complementary: coalition formation determines what collective agents exist; power operates within and between them once they exist.

A note on the type of explanation offered here. This is a *design-space* or *constraint-based* explanation, not a genealogical one. I do not claim that human religions descended from bacterial quorum sensing or that ideological movements evolved from insect colonies. The claim is convergence: under shared stability constraints (high-stakes coordination, partial observability, defection incentives), viable solutions cluster in the same region of design space. The "religious" features identified here—ritual, sacred markers, heresy punishment, identity fusion—are attractors in the space of possible coalition-stabilizing architectures. Different lineages arrive at similar solutions because the problem constrains the solution set, not because they share common ancestry. This is analogous to convergent evolution in morphology: eyes evolved independently dozens of times because the physics of light constrains the design of light-detecting organs. Similarly, costly signaling structures emerge independently across substrates because the logic of commitment verification constrains the design of coalition-maintaining systems.

2 The Coalition Formation Problem

2.1 Coordination Requires Commitment Verification

Consider a coalition of n agents contemplating collective action with payoff structure:

- If all cooperate: each receives benefit B
- If agent i defects while others cooperate: defector receives $B + D$ (defection bonus), cooperators receive $B - C$ (sucker's cost)
- If multiple agents defect: coalition fails, all receive 0

When $D > 0$ and defection is undetectable until after commitment, each agent faces uncertainty about others' types. A population containing both *loyalists* (who will cooperate) and *defectors* (who will exploit cooperation) is unstable under cheap talk: defectors will claim to be loyalists, and genuine loyalists cannot distinguish themselves.

This is the fundamental problem: *how can loyalists credibly signal their type?*

Throughout this paper, a "signal" is any publicly observable state or action that (i) correlates with willingness to bear coalition costs, (ii) is itself costly to produce or maintain, and (iii) is differentially costly for agents pursuing exploitative strategies. This definition applies across biological scales: the signal may be a molecule, a behavior, a profession, or a bodily modification, so long as it satisfies these three conditions.

2.2 Costly Signals as Separating Equilibria

The solution is to require signals that are differentially costly by type. If the signal costs c_L for loyalists and c_D for defectors, with $c_D > c_L$, then there exists a threshold signal costliness c^* such that:

- For loyalists: $B - c_L > 0$ (signaling and joining coalition is profitable)
- For defectors: $B + D - c_D < 0$ (signaling is not worth the cost even with defection bonus)

When this condition holds, costly signaling produces a *separating equilibrium*: only loyalists signal, and the signal becomes a reliable indicator of type.

More precisely, let signals require repeated maintenance over T periods at per-period cost c, and let defectors face detection probability p with punishment P. The expected payoff for a loyalist is $B - Tc_L$; for a defector it is $B + D - Tc_D - pP$. Separation occurs when:

$$B - Tc_L > 0 > B + D - Tc_D - pP$$

The condition is easier to satisfy as T increases (longer maintenance), as $c_D - c_L$ grows (larger cost differential), and as pP rises (higher expected punishment). This formalizes the intuition that sustained, differentially costly signals with enforcement produce stable separation.

The key insight is that signal costliness must scale with the stakes. I define *coordination stakes* as the expected loss from coalition failure: $S = \pi_{\text{inf}} \times L(\text{collapse})$, where infiltration probability π_{inf} depends on the defection bonus D (how tempting exploitation is) and collapse loss L includes both foregone coalition benefits B and any additional costs of failed coordination. As S increases, the required signal cost c^* must increase correspondingly to maintain separation. High-stakes coordination demands high-cost signals.

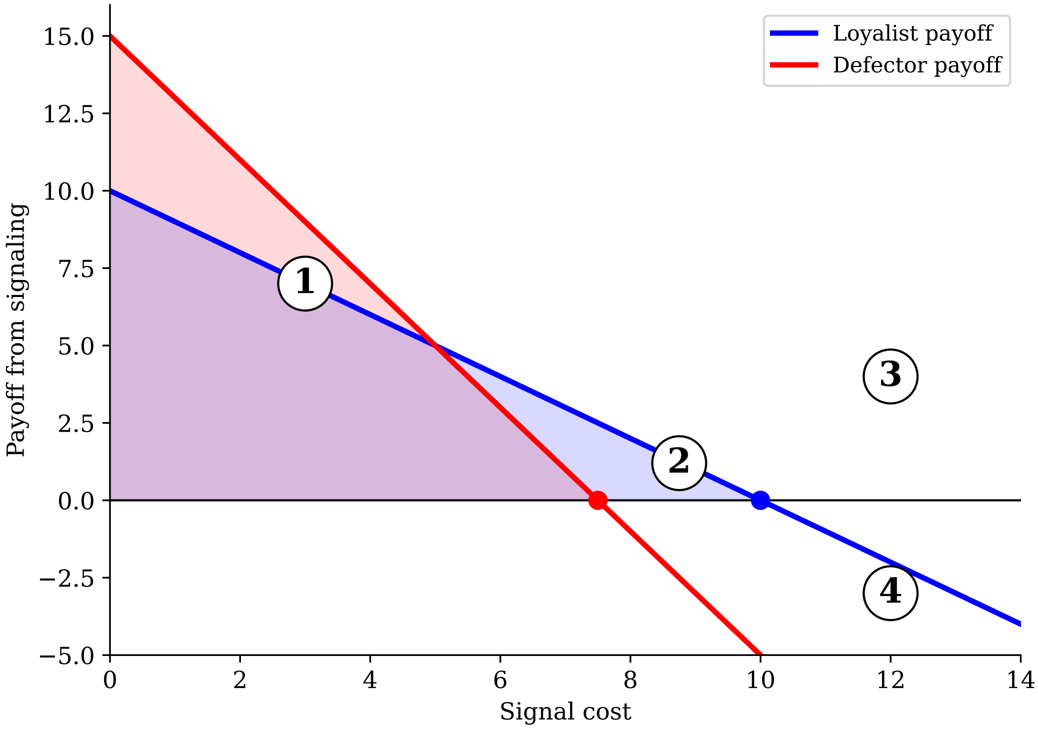


Figure 1: **Separating equilibrium in costly signaling.** Shading shows where each type would signal (payoff > 0). **Zone 1**: Both payoffs positive—defectors infiltrate the coalition. **Zone 2**: Only loyalist payoff positive—defectors screened out (the separating equilibrium). **Zone 3**: Neither payoff positive—no one signals, coalition cannot form. **Zone 4**: Negative payoffs—signaling costs exceed any possible benefit. The separation zone exists because defectors pay higher costs ($c_D > c_L$), so their payoff crosses zero first.

Agent-based simulations confirm this theoretical structure. Figure 2 shows results from a parameter sweep over signal cost and defector cost multiplier, with 100 agents (30% defectors) across 50 trials per parameter combination. Three key outcomes emerge: (1) separation rate increases with both signal cost and cost differential—the "separation zone" where loyalists signal but defectors do not; (2) coalition success (low defector infiltration) closely tracks separation; (3) loyalist participation decreases when costs become prohibitive. The simulation validates the core prediction: there exists an optimal cost regime where signals are expensive enough to screen defectors but cheap enough for loyalists to afford.

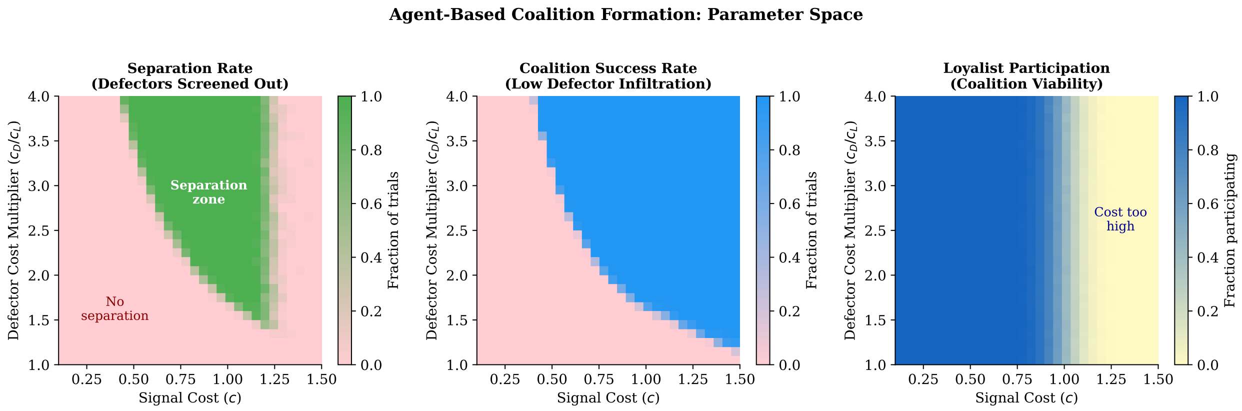


Figure 2: **Agent-based validation of separating equilibrium.** Parameter sweep over signal cost and defector cost multiplier (100 agents, 30% defectors, 50 trials per condition). *Left*: Separation rate (fraction of trials where all signalers are loyalists). *Center*: Coalition success rate (defector infiltration below 10%). *Right*: Loyalist participation rate. The separation zone (upper region of left panel) shows where costly signaling successfully screens defectors while maintaining coalition viability.

2.3 Why Individually Dominated Strategies Are Optimal

Here we reach the crucial point: from the perspective of individual payoff maximization, the optimal coalition signals are those that appear *individually dominated*—they reduce expected payoff in the absence of coalition benefits.

A signal is individually dominated if paying it leaves an agent worse off than not signaling, absent the benefits of coalition membership. But this is precisely what makes it reliable: an agent who does not expect to receive coalition benefits (i.e., a defector planning to exploit and exit) will not pay the cost. The more individually costly the signal, the more it screens for genuine commitment.

This generates a prediction: as coordination stakes increase, we should observe signals becoming increasingly extreme, apparently self-destructive, and resistant to individual-level cost-benefit analysis. What looks like irrationality from an individual perspective is rationality from a coalition-formation perspective. We use “apparently irrational” as shorthand for this level-crossing phenomenon throughout.

2.4 Why Defectors Pay More: A Selection Argument

A critical question remains: why is the signal more costly for defectors than for loyalists? The answer is not mechanistic but selective: *coalitions that survive are precisely those whose signals cost more to fake than defection pays.* If the signal cost did not exceed the defection reward, the coalition would be infiltrated and collapse. What we observe are the survivors.

This reframes the question. We should not ask "what prevents defectors from paying?" but rather "what makes some coalitions stable while others dissolve?" The stable coalitions are those that have, through cultural or biological evolution, landed on signals satisfying:

$$c_D > D + V_{\text{outside}}$$

where c_D is the cost to a defector, D is the defection bonus, and V_{outside} is the value of preserved outside options. Coalitions whose signals fall below this threshold are selected against.

The mechanisms that generate $c_D > c_L$ can be unified under a single model: defectors have higher effective discount rates or greater outside-option value than loyalists. Let δ_L and δ_D be discount factors (patience) for loyalists and defectors respectively, with $\delta_D < \delta_L$. The present value of repeated signal costs c over T periods is:

$$\text{PV}_L = \sum_{t=1}^{T} \delta_L^t c \quad \text{vs.} \quad \text{PV}_D = \sum_{t=1}^{T} \delta_D^t c + V_{\text{outside}}$$

where V_{outside} captures the defector's valuation of preserved alternatives. Even with identical per-period costs c, defectors pay more in effective terms because they discount future coalition benefits more heavily and value foreclosed options more. The mechanisms below are interpretations of this fundamental asymmetry:

Repeated maintenance. Signals requiring sustained compliance over time T accumulate costs that short-horizon defectors cannot afford. Loyalists, who share coalition goals, find ongoing participation less burdensome than defectors simulating commitment while pursuing incompatible objectives.

Opportunity costs and irreversibility. Signals that foreclose alternatives—public professions inviting out-group hostility, bodily modifications, bridge-burning acts—impose higher costs on agents who value flexibility [Schelling, 1960].

Detection and punishment. Defectors face expected sanctions because their subsequent behavior is inconsistent with professed commitment. The effective cost includes this expected penalty: $c_D = c + pP$.

These are not alternative explanations but complementary mechanisms by which coalitions achieve the threshold. The deeper point is evolutionary: coalitions occupy stable points in strategy space *because* their signals are unfakeable at acceptable cost. The "why" is selection; the "how" is mechanism.

3 Costly Signaling Across Biological Scales

3.1 Structural Features of High-Stakes Coordination

High-stakes coordination systems exhibit a recurring cluster of structural features. I will argue these emerge across biological scales wherever commitment verification is critical. To avoid anthropomorphism, I first describe each feature in neutral functional terms, then note the "religious" label often applied in human contexts:

1. **Stereotyped coordination acts** ("ritual"): Repeated, stylized actions that synchronize behavior and signal ongoing participation

2. **Protected boundary markers** ("sacred markers"): Identity signals that are costly to produce or modify, enabling coalition recognition

3. **Costly public commitment** ("profession"): Observable assertions or behaviors that entail social or material costs

4. **Deviation sanctioning** ("heresy punishment"): Penalties for norm violation, typically harsher for internal deviants than external opponents

5. **Individual-collective identity merger** ("identity fusion"): Blurring of boundaries between agent-level and coalition-level interests

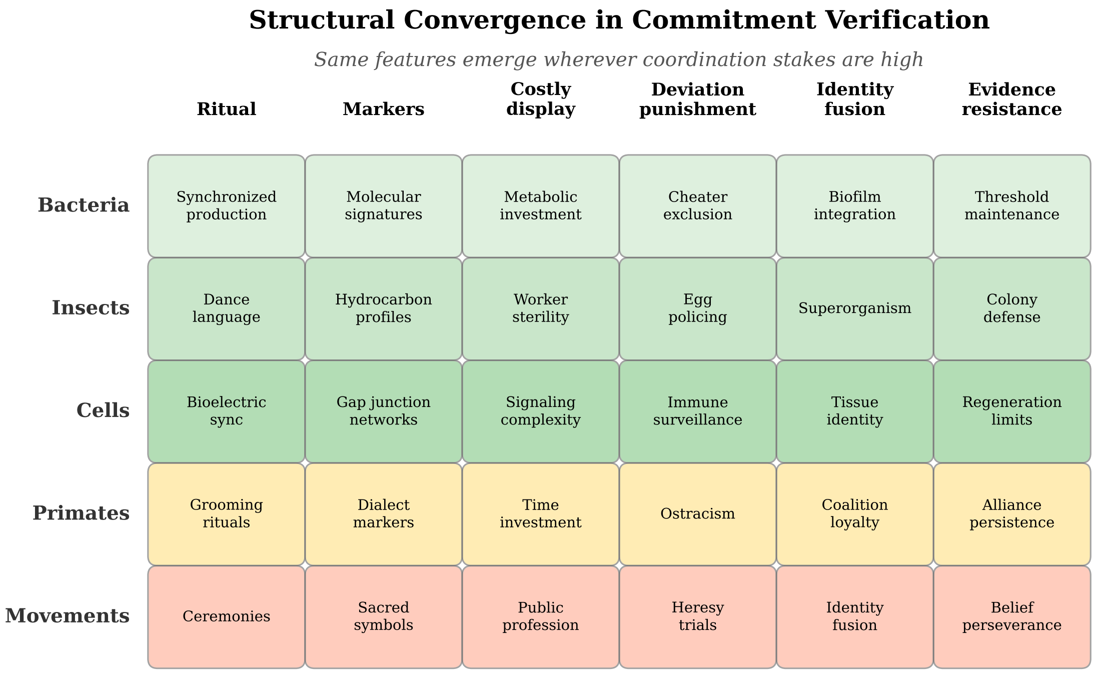
6. **Signal maintenance under challenge** ("evidence resistance"): Commitment sustained or intensified despite external pressure or contradicting information

The mapping is not merely analogical. The claim is that these features represent *functional homologies*—recurrent design solutions to the commitment verification problem—not that bacteria have beliefs or that insects experience reverence.

This cluster is not arbitrary; each feature addresses a specific aspect of the commitment verification problem. The following subsections trace these features from microbes through humans (Figure 3).

However, the framework predicts this cluster emerges only under specific scope conditions. When coordination stakes are low, cheap signals suffice and elaborate structure is unnecessary. When monitoring is effective (agents can directly observe each other's behavior), costly signaling is less necessary because defection is detectable. When exit is easy (agents can leave coalitions at low cost), extreme signal costs cannot be sustained because members will defect to less demanding alternatives. The full structural cluster emerges when stakes are high, monitoring is imperfect, and exit is costly—conditions that characterize the biological examples examined below.

This generates predictions about both *false positives* and *false negatives.* Not all religions will exhibit intense costly signaling (low-stakes denominations with easy exit may rely on cheap signals). Conversely, many non-religious institutions—political movements, professional guilds, criminal organizations, military units—will exhibit the full cluster when scope conditions are met. "Religious structure" here denotes a family resemblance of commitment technologies, not theological content. The framework predicts graded expression tied to



Structural Convergence in Commitment Verification

Same features emerge wherever coordination stakes are high

	Ritual	Markers	Costly display	Deviation punishment	Identity fusion	Evidence resistance
Bacteria	Synchronized production	Molecular signatures	Metabolic investment	Cheater exclusion	Biofilm integration	Threshold maintenance
Insects	Dance language	Hydrocarbon profiles	Worker sterility	Egg policing	Superorganism	Colony defense
Cells	Bioelectric sync	Gap junction networks	Signaling complexity	Immune surveillance	Tissue identity	Regeneration limits
Primates	Grooming rituals	Dialect markers	Time investment	Ostracism	Coalition loyalty	Alliance persistence
Movements	Ceremonies	Sacred symbols	Public profession	Heresy trials	Identity fusion	Belief perseverance

Figure 3: **Structural convergence across biological scales.** The same commitment-verification features emerge wherever coordination stakes are high, from bacterial quorum sensing to human ideological movements. This convergence reflects functional homology—recurrent solutions to the same coordination problem—not common descent.

coordination demands, not a binary religious/secular distinction.

3.2 Microbial Precedents

Bacterial quorum sensing provides a minimal example of costly signaling for coordination [Miller and Bassler, 2001]. Bacteria produce and detect signaling molecules (autoinducers) that accumulate as population density increases. When concentration exceeds a threshold, coordinated behaviors are triggered (biofilm formation, virulence factor production, bioluminescence).

Crucially, autoinducer production is metabolically costly. The defection temptation is clear: cheater strains that detect but do not produce autoinducers gain the benefits of coordinated behavior without paying the metabolic cost. The costly signal (autoinducer production) prevents this exploitation: in structured environments, producer strains can exclude cheaters through spatial assortment, and the cost itself screens for cooperative phenotypes [Diggle et al., 2007]. The system exhibits structural features paralleling high-stakes coordination elsewhere:

- **Ritual**: Synchronized production/detection cycles
- **Markers**: Specific molecular signatures ("shibboleths")
- **Costly profession**: Metabolic investment in signaling
- **Cheater punishment**: Competitive exclusion of non-producers

3.3 Insect Colony Recognition

Eusocial insects face a more complex coordination problem: maintaining colony boundaries while allowing division of labor [Wilson, 1971]. Nestmate recognition typically relies on cuticular hydrocarbon profiles—chemical signatures that are costly to produce and difficult to fake.

The defection temptations are twofold: external intruders may attempt to exploit colony resources, and internal workers may attempt to reproduce rather than help. Costly signals address both. Cuticular hydrocarbons prevent external defection: intruders cannot easily fake the colony signature, so infiltration is screened out [Breed et al., 1988]. Internal policing (worker destruction of eggs laid by other workers) prevents reproductive defection: the cost of attempted cheating is egg destruction, maintaining the cooperative equilibrium. These features parallel the structural cluster identified above:

- **Sacred markers**: Colony-specific hydrocarbon profiles
- **Identity fusion**: Workers sacrifice individual reproduction for colony
- **Heresy punishment**: Policing of reproductive norm violations
- **Boundary maintenance**: Aggression toward out-group members

The "irrationality" of worker sterility makes sense from the coalition-signaling perspective: it is maximally costly, therefore maximally reliable as a commitment signal.

3.4 Multicellular Coalitions

A crucial distinction: individual agents face decision problems; coalitions face coordination problems. A single cell is an agent—it acts on internal states without needing to verify commitment from other cells. A multicellular organism is a coalition of cells, and as such must solve the commitment verification problem that individual agents do not face.

The defection temptation is cancer: a cell lineage that proliferates rapidly at the expense of the organism captures short-term reproductive benefit while destroying the coalition [Aktipis et al., 2015]. How do cells verify each other's cooperative intent?

Levin's work on bioelectric signaling suggests that cells coordinate through shared electrical and chemical gradients that integrate behavior across tissue-scale collectives [Levin, 2019, Levin, 2023, Levin, 2025]. Crucially, bioelectricity enables signals of arbitrary complexity: ion channel expression, gap junction networks, and voltage gradients can encode high-dimensional information. The complexity of the signal is itself the cost—generating and receiving complex bioelectric patterns requires metabolic expenditure, and this expense is precisely what makes the signal reliable.

This suggests a hypothesis: signal complexity functions as a costly signal of cooperative intent. A cell maintaining coherent, high-dimensional coordination with its neighbors would be demonstrating commitment—investing resources in generating and decoding complex signals rather than defecting toward rapid proliferation. Cancer cells, on this view, exhibit low-dimensional dynamics: they optimize for replication speed, shedding the overhead of complex tissue-scale signaling. They cannot fake high-complexity signals because the metabolic and computational cost is incompatible with their proliferative strategy.

What would count as "high-dimensional coordination" operationally? The key is not raw signal bandwidth but participation in shared interpretive constraints—a code that enables high-dimensional meaning to pass through low-dimensional channels. Consider chess: minimal bandwidth (a few bits per move), but the shared rules enable communication of extraordinarily complex strategic states. Cells similarly coordinate not by signaling volume but by participation in tissue-scale codes—gap junction networks, bioelectric gradients, morphogenetic fields—that compress high-dimensional coordination into interpretable signals. The hypothesis predicts that cells transitioning toward malignancy should show measurable decreases in code participation (decoupling from gap junction networks, loss of voltage gradient coherence) before other hallmarks of cancer become apparent. A noisy cell is not necessarily defecting; a cell that has exited the shared interpretive framework is.

If correct, this reframes immunity partly as commitment verification: the immune system detects cells that have abandoned complex coordination and treats them as defectors. The signal is not "cancer" per se but the loss of costly cooperative dynamics.

This framework also suggests a tradeoff between cancer suppression and regenerative capacity [Levin, 2025]. Regeneration requires cells to dedifferentiate, proliferate, and reorganize—precisely the behaviors that costly signaling is designed to suppress. Organisms with tight coalition control (complex mammals) cannot regrow limbs because the signaling costs that prevent defection also prevent the coordinated mass action regeneration requires. Organisms with simpler coordination (salamanders, planaria) retain regenerative flexibility but face higher cancer risk. The same costly signals that maintain coalition integrity constrain coalition plasticity.

This remains speculative—the empirical work connecting signaling complexity to both immune surveillance and regenerative limits is incomplete—but it illustrates how the costly signaling framework might extend to cellular coalitions.

Agent-based simulations support this hypothesis (Figure 4). Modeling cells with checkpoint signaling (costly verification of cooperative state) and defector detection (immune-like surveillance), we find that: (1) checkpoint costs and detection probability jointly determine equilibrium defector fraction; (2) costly checkpoints dramatically suppress cancer-like defection compared to no-checkpoint conditions; (3) the parameter space shows a clear transition between effective screening (low defector equilibrium) and cancer risk (high defector equilibrium). In this toy model, costly verification plus surveillance is sufficient to maintain low defector equilibria across a broad parameter region; removing checkpoints expands the cancer-risk region.

3.5 Primate Coalition Displays

In primate societies, coalition formation involves extensive ritualized interaction: grooming, food sharing, coordinated displays, and vocal exchanges [Dunbar, 1998]. These behaviors are time-consuming and often reciprocally asymmetric—precisely the features that make them reliable signals.

Boehm's work on reverse dominance hierarchies in human forager societies shows that coalition maintenance requires ongoing investment in norm enforcement [Boehm, 1999]. Gos-

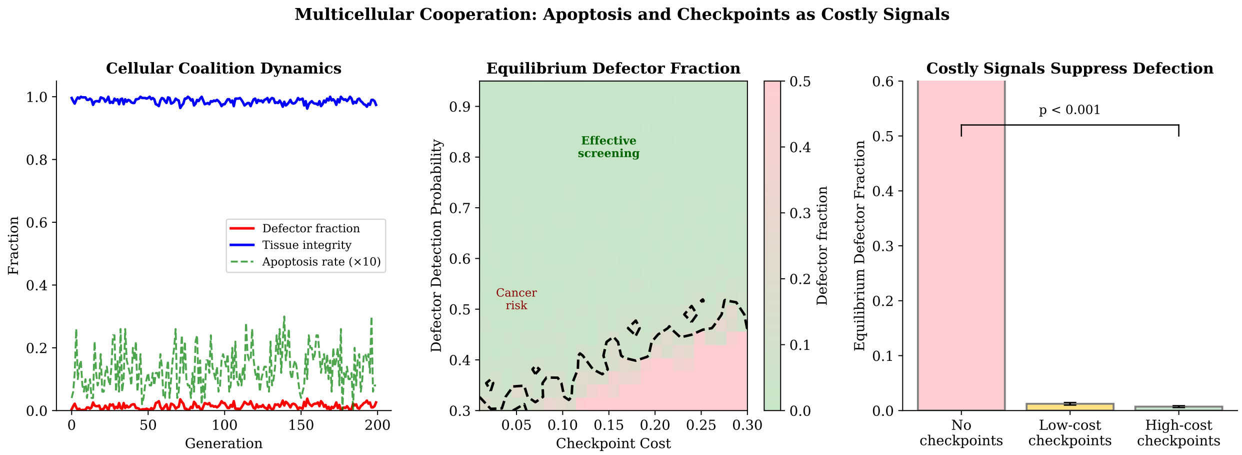


Figure 4: **Cellular coalition dynamics.** Agent-based model with 500 cells, checkpoint signaling costs, and defector detection. *Left*: Dynamics over 200 generations showing defector fraction (red), tissue integrity (blue), and apoptosis rate (green, ×10). *Center*: Parameter sweep showing equilibrium defector fraction as function of checkpoint cost and detection probability; contour marks 10% defector threshold. *Right*: Comparison across signaling regimes (20 runs each)—costly checkpoints dramatically reduce defector equilibrium (Mann-Whitney U, $p < 0.001$).

sip, ridicule, ostracism, and in extreme cases violence maintain egalitarian norms against would-be dominants. The pattern is familiar:

- **Ritual**: Grooming, food sharing, coordinated activities
- **Costly profession**: Time investment with opportunity cost
- **Heresy punishment**: Sanctions against norm violators
- **Identity signals**: In-group markers (dialects, customs)

Paleoanthropological evidence extends these patterns deep into hominin evolution. Nesse [Nesse, 2001] argued that the capacity for commitment—“deep seated emotional commitments to others’ wellbeing”—was a significant evolutionary development enabling the give-and-take of cooperative relationships. Recent work supports this: Fuentes et al. [Fuentes et al., 2026] document evidence of costly care for injured and vulnerable individuals extending back over 1.5 million years, with Neanderthal populations showing more than thirty possible cases of sustained healthcare provisioning [Spikins et al., 2019, Spikins et al., 2018]. Such care requires ongoing investment with uncertain return—precisely the structure of costly signaling. Critically, the archaeological record shows healthcare costs escalating over hominin evolution: from occasional assistance to sustained nursing care lasting months or years. This pattern is predicted by the framework: as social bonds deepen and coalitions become more interdependent, the signals required to maintain them must become correspondingly more costly. Healthcare provisioning is not merely evidence of cooperation but a marker of the increasing signal costs that accompany tighter social integration.

3.6 Human Ideological Movements

Against this background, human religious and ideological movements appear as elaborations of a conserved coordination mechanism rather than unique cultural innovations.

Bertrand Russell observed that 20th-century totalitarian movements exhibited structural features identical to religions: sacred texts, infallible authorities, heresy trials, and imperviousness to evidence [Russell, 1956]. His diagnosis was moral-epistemic: these movements had abandoned reason. Russell identified the pathology but lacked a biological mechanism. The costly signaling framework supplies one: these features are *functional* for coalition maintenance under high-stakes conditions, not failures of rationality but solutions to coordination problems.

Consider the structural requirements for maintaining a revolutionary coalition:

- Members must be distinguishable from infiltrators and defectors
- Commitment must be verifiable before high-cost collective action
- Defection must be deterred through credible punishment
- Coordination must be maintained despite internal disagreement

Costly signaling provides solutions: public profession of beliefs that would be costly for non-believers, rituals that require time investment, markers that invite out-group hostility, and punishment of deviation that deters cheap-talk loyalty claims.

The apparent "irrationality" of ideological belief—its resistance to external evidence, its internal coherence requirements, its demands for public profession that is socially costly to maintain—is not a bug but a feature. These properties make the signal costly to fake, therefore reliable.

4 Predictions and Implications

4.1 Evidence Strengthens Commitment

A counterintuitive prediction of the framework: evidence against group beliefs can *increase* commitment rather than decrease it.

Standard models of belief revision predict that disconfirming evidence should reduce credence. But if belief functions as a coordination signal rather than a truth-tracking mechanism, the logic inverts. When external evidence threatens the belief, maintaining public profession becomes more costly (more embarrassing, more isolating, more identity-defining). Higher cost means higher reliability as a commitment signal. Those who maintain profession despite evidence are demonstrating precisely the kind of commitment that coalition membership requires.

This is consistent with empirical patterns of belief perseverance, identity-protective cognition, and resistance to correction in ideological contexts [Nyhan and Reifler, 2010]. Adherents are not failing at epistemics; they are succeeding at coalition signaling. The prediction is not that *all* adherents intensify commitment under challenge, but that those who remain after disconfirmation will show stronger signals—producing polarization rather than uniform updating.

4.2 Internal Heresy Exceeds External Opposition

The framework predicts that deviation from within the coalition will be punished more severely than opposition from outside.

An external opponent is not claiming coalition membership; their opposition is expected and does not threaten the signaling equilibrium. But an internal heretic is claiming membership while deviating from costly norms—exactly the behavior that would allow defectors to infiltrate if tolerated. Heresy must be punished severely to maintain the separating equilibrium.

This explains the historical pattern of intra-religious violence exceeding inter-religious violence, schisms producing more bitter conflict than conquest, and ideological movements devoting more energy to purges than to external struggle.

4.3 Escalation Under Threat

As external threat increases, the stakes of coalition coordination rise, requiring costlier signals. The framework predicts that threatened coalitions will demand increasingly extreme commitment displays.

This matches the observed pattern of radicalization under pressure: loyalty tests become more demanding, ideological purity becomes more narrowly defined, and previously tolerated variation becomes heresy. The coalition is not becoming "more irrational"—it is rationally adjusting signal costliness to match increased coordination stakes.

Evolutionary simulations confirm this "cost ratchet" dynamic (Figure 5). When loyalists and defectors coevolve with cultural signal costs, three patterns emerge: (1) loyalists evolve higher cost tolerance to maintain coalition access; (2) cultural signal costs ratchet upward under infiltration pressure; (3) higher defector fractions drive faster cost escalation. The right panel shows threshold behavior: at sufficiently high defector fractions, separation becomes difficult regardless of signal cost, and coalitions fail. The exact threshold depends on model parameters, but the qualitative pattern is robust: coalitions facing severe infiltration pressure will either escalate to extreme costly signaling or collapse entirely.

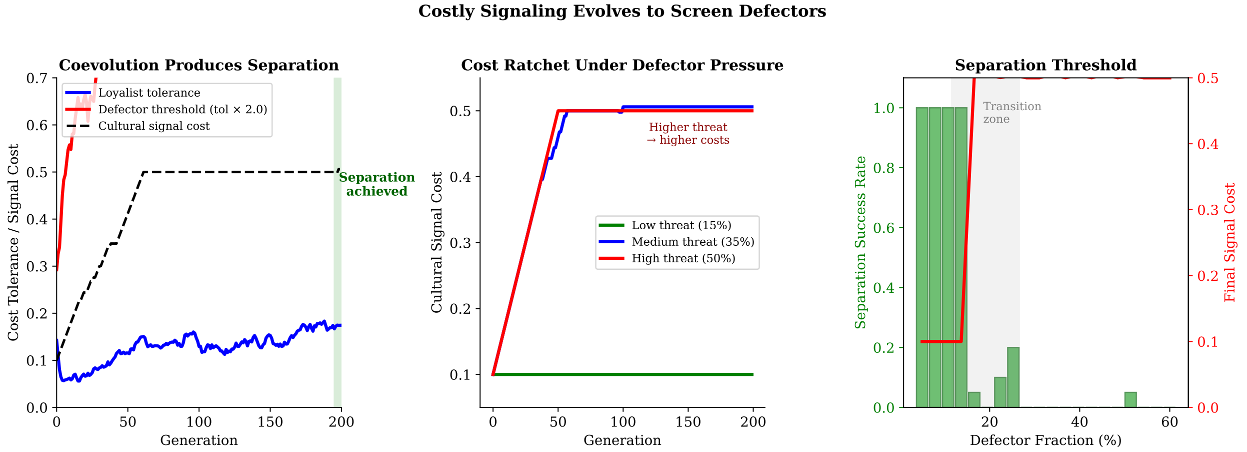


Figure 5: **Coevolution of costly signaling.** Evolutionary dynamics with 500 agents over 200 generations. *Left*: Loyalist tolerance (blue) and defector threshold (red, scaled by cost multiplier) diverge as cultural signal cost (black dashed) increases, eventually achieving separation (green shading). *Center*: Cost ratchet under different threat levels—higher defector fractions drive faster cost escalation. *Right*: Threshold behavior showing separation success (green bars) and final signal cost (red line) across defector fractions; at high defector fractions, separation becomes unreliable.

4.4 Convergent Structure Across Substrates

The deepest prediction is structural convergence: any system facing high-stakes coordination problems will tend toward religious structure regardless of substrate or content.

This explains why political movements "become religions," why corporate cultures develop ritual and taboo, why scientific communities can exhibit sect-like dynamics around paradigm disputes, and why online communities develop elaborate norm-enforcement mechanisms. The specific content varies; the structural logic is conserved.

5 Implications for Philosophy of Biology

5.1 Cross-Scale Continuity

The framework places human coordination systems—including those labeled "religious"—on a continuum with non-human coalition formation. The structural features that recur (ritual, markers, costly profession, deviation punishment) are functional solutions to the commitment verification problem, appearing wherever that problem is sufficiently acute.

This does not reduce human culture to "mere" biology or dismiss cultural elaboration. But it does suggest that the core structural features of high-stakes coordination are constrained by the same signaling dynamics across scales, from microbial biofilms to ideological movements.

5.2 Belief as Behavior

The analysis requires reconceptualizing belief. In standard epistemology, beliefs are propositional attitudes evaluated for truth-tracking. In the coalition framework, beliefs (specifically, public profession of beliefs) are behaviors evaluated for coordination function.

This does not mean believers are "lying" or "faking." The phenomenology of belief may be genuine. But the selection pressure maintaining the belief is coordination, not correspondence. A belief can be sincerely held and evolutionarily maintained for its signaling properties rather than its truth-tracking properties.

5.3 Rationality Reconsidered

The framework complicates simple distinctions between rational and irrational belief. From an individual epistemic perspective, maintaining belief despite disconfirming evidence is irrational. From a coalition coordination perspective, it is precisely what is required for reliable signaling.

This suggests that "irrationality" is often rationality at a different level of analysis. The agent is not failing at truth-seeking; they are succeeding at coalition maintenance. Criticisms that treat ideological belief as epistemic failure may be missing the functional target.

5.4 Limits of Evidence-Based Persuasion

If belief functions as coordination signal rather than truth-tracker, then evidence-based persuasion faces structural limits. Presenting disconfirming evidence may strengthen rather than weaken commitment by raising the costliness of continued profession.

This has implications for science communication, political discourse, and any domain where belief change is sought. The coalition function of belief must be addressed; merely providing evidence is insufficient and may be counterproductive.

6 Discussion

6.1 Relationship to Existing Accounts

The present framework builds on several traditions:

Signaling theory (Zahavi, Spence, Grafen) provides the core mechanism [Zahavi, 1975, Spence, 1973,Grafen, 1990]. The contribution here is extending signaling analysis to coalition formation and tracing structural consequences.

Costly signaling theory of religion applies this logic directly to religious commitment. Irons [Irons, 2001] argued that religious practices function as "hard-to-fake signs of commitment" precisely because their costs screen out non-believers. Iannaccone's [Iannaccone, 1992] club-good model showed how sacrifice and stigma reduce free-riding by raising the cost of membership for those with high outside-option value. Berman [Berman, 2000] formalized how religious sects use costly requirements to screen for committed members who will contribute to collective goods. Sosis and colleagues [Sosis and Ruffle, 2003, Sosis and

Bressler, 2003] provided empirical tests, showing that communes with costlier requirements survived longer and that ritual participation predicted cooperative behavior.

The present account generalizes this literature beyond religion to coalition formation as a cross-scale design constraint. Where previous work treated religion as the explanandum, I treat "religious structure" as the name for a convergent solution that appears wherever commitment verification is critical—in bacterial colonies, insect societies, multicellular organisms, and ideological movements. The claim is not that religion is special but that it instantiates a general pattern.

Cognitive science of religion (Atran, Boyer, Henrich) has documented cross-cultural regularities in religious cognition [Atran, 2002, Boyer, 2001, Henrich, 2009]. The present account offers a functional explanation for why these regularities exist: they are solutions to coordination problems, not arbitrary byproducts of cognitive architecture.

A clarification is needed regarding Henrich's "credibility-enhancing displays" (CREDs). CREDs explain why beliefs transmit culturally: observers update toward beliefs that speakers demonstrate through costly action. The present framework addresses a different problem: why coalitions demand costly displays *before* collective action, as a condition of membership. CREDs concern cultural transmission; the present account concerns commitment verification under defection risk. The two are complementary—CREDs may explain how costly signaling norms spread, while the present framework explains why such norms are functional for coalition maintenance in the first place. Recent empirical work confirms that advertising cooperative phenotype through costly signals facilitates collective action across diverse cultural contexts [Lang et al., 2022].

Cultural evolution (Boyd, Richerson, Sosis) has analyzed religion as a group-level adaptation [Sosis and Ruffle, 2003]. The costly signaling framework provides a mechanism: religion is adaptive for groups because it solves the commitment verification problem that makes group coordination possible.

Philosophy of social science (Russell) has diagnosed "political religions" as pathologies of reason [Russell, 1956]. The present account reframes this: the religious structure of ideological movements is not a failure of rationality but a functional response to coordination demands.

6.2 High-Dimensional Coordination

The costly signaling framework connects to a broader problem: how do agents coordinate in high-dimensional state spaces where direct verification is impossible? When internal states are high-dimensional, external observers cannot directly assess commitment—the state space is too large, the projection too lossy. Cheap signals fail because they can be faked without cost; noise and strategic manipulation are indistinguishable.

Costly signals solve this problem by making the signal itself observable and hard to counterfeit. The cost functions as a filter: only agents with genuine commitment pay it. This is the same logic that governs coherence maintenance in high-dimensional biological systems more generally [Todd, 2025a]. Maintaining coordination across many degrees of freedom requires paying thermodynamic costs; here the cost is behavioral rather than metabolic, but the principle is conserved.

A deeper principle emerges: *computational complexity is itself a costly signal.* This is not a new definition of computation; it is an extension of the same screening logic to capacities that are expensive to acquire and hard to counterfeit. Shared codes make complexity cheap to communicate but impossible to fake. A chess position compresses vast strategic depth into a few bits—but producing valid moves requires genuine competence. The signal is low-bandwidth; the capacity it certifies is high-dimensional. This structure recurs across the examples above: quorum sensing requires synchronized timing, not just molecule production; colony recognition requires pattern matching within a colony-specific code; primate coalitions require tracking complex reciprocity networks; ideological movements require demonstrating genuine doctrinal fluency. In each case, the cost is not primarily metabolic or temporal but *coordinative*—participation in a complex shared code that cannot be faked without actually being able to participate.

Code interpretation requires structural isomorphism: two high-dimensional systems can recognize each other's signals only if they share sufficient internal structure to decompress low-bandwidth signals into high-dimensional meaning. Developing this isomorphism is itself the costly signal of cooperative intent—you cannot acquire the structure without actually participating in the coordinated system. The signal and the capacity are inseparable. This is why initiation works (developing shared structure), why heretics are dangerous (partial isomorphism with divergent goals), and why defection is detectable (loss of structural alignment with the collective).

From this perspective, the "irrationality" of costly signals is an entropy tax. Coordination under uncertainty requires expenditure. The structural features that emerge—ritual, markers, identity fusion—are not arbitrary but are shaped by the constraint that signals must survive noise in high-dimensional environments. Coalition formation is thus a special case of the general problem of coherence maintenance under partial observability.

6.3 Coalition Maintenance as Nonergodic Work

Coalition formation is inherently non-ergodic: stable cooperation requires continuous expenditure to resist drift toward the defection equilibrium. Left to entropy, coalitions dissolve—defectors infiltrate, signals degrade, trust decays. The "natural" state is universal defection; cooperation exists only where agents do the work to maintain it.

This reframes costly signals not as one-time admission fees but as ongoing maintenance costs. Ritual participation must be repeated; identity markers must be renewed; deviation must be continuously policed. The work never ends because the drift toward defection never stops. Coalitions that reduce maintenance expenditure become vulnerable to infiltration; coalitions that cannot sustain the expenditure collapse.

This perspective connects to broader debates about social structure. Graeber and Wengrow [Graeber and Wengrow, 2021] argue that stable social arrangements are not passive equilibria determined by material conditions but active achievements maintained through continuous effort—that humans have always exercised collective choice over how to organize themselves. The costly signaling framework provides a mechanism: each member's signal is a contribution to the collective project of holding the coalition in a non-ergodic state.

One way to interpret this: coalition maintenance is "chosen" in the sense that it emerges from high-dimensional internal states (beliefs, commitments, identities) that external observers cannot fully access or predict. This connects to the broader question of agency under partial observability, though developing that connection is beyond the present scope.

6.4 Objections and Responses

Objection 1: Religions make truth claims. Yes, but the framework does not deny this. It claims that the *selection pressure* maintaining religious belief is coordination function, not truth-tracking. Truth claims are part of the content; coordination is the function.

Objection 2: This is too functionalist (or adaptationist). The account is functional, not Panglossian. It identifies what religious structure *does* without claiming that all features are optimal or that every detail is adaptive. Cultural drift, historical contingency, path dependence, and outright maladaptation are all permitted. The claim is about *constraint*, not optimality: when coordination stakes are high, certain structural features become attractors in the space of possible solutions. Particular religions elaborate on this constrained core in ways shaped by cultural transmission, founder effects, and ecological context. The framework predicts convergence on core features; it does not predict identical content or claim that observed religions are optimal solutions.

Objection 3: Costly signaling explains too much. This concern has real force, and addressing it requires clarifying the epistemological status of design-space explanations.

Traditional falsifiability works well for single-system experiments: manipulate a variable, observe the effect, reject hypotheses inconsistent with the result. But coalition formation, like other high-dimensional strategic systems, resists this approach. The payoff structure is substrate-independent—the same game-theoretic logic produces separating equilibria whether the agents are bacteria, cells, or ideologues. You cannot run a controlled experiment that "turns off" costly signaling in human religions while holding everything else constant. The strategic structure is mathematical, not material.

This is analogous to doing falsifiable physics at a basketball game. You can measure trajectories and collisions—that is falsifiable. But capturing the *game* requires game theory, which lives in a higher-dimensional space: possible strategies, beliefs about beliefs, conditional responses. The game-theoretic structure is what makes it a game, and that structure is not captured by manipulating single physical variables.

Design-space explanations are tested differently: by cross-system comparison rather than single-system manipulation. The prediction is convergent structure across independent instantiations. If the same costly signaling features appear in bacterial quorum sensing, insect colonies, multicellular coordination, and ideological movements—systems with no genealogical connection—this constitutes evidence for the underlying game-theoretic constraint. The evidence is pattern, not experiment. Convergence *is* the test. This is the same epistemological structure used to validate pattern-formation models in morphogenesis: branching patterns in rivers, lungs, and lightning; spiral structures in galaxies, shells, and hurricanes—independent systems converge because the mathematics constrains the solution space [Ball, 1999].

The predictions in Section 4 (evidence strengthening commitment, internal heresy exceeding external opposition, escalation under threat) provide additional testable implications within particular systems. But the deepest confirmation comes from structural convergence itself: independent systems facing the commitment verification problem arrive at similar solutions because the mathematics constrains them to.

6.5 Local Optima: Why Moderates and Extremists Are Both "Correct"

A crucial implication: religious moderates and extremists are not on a spectrum from rational to irrational. They occupy different locally optimal solutions to the coalition maintenance problem under different parameter regimes.

The simulation results (Figure 5) show threshold behavior: at low defector pressure, moderate costly signaling achieves separation. Above a threshold (which depends on model parameters), coalitions must either escalate to extreme demands or collapse entirely. There is no stable moderate equilibrium under high threat. The Quaker meeting and the Inquisition are not more or less rational than each other—they are optimal responses to different coordination environments.

This parallels cross-species variation. The ant colony that executes reproductive cheaters and the bonobo troop that resolves conflict through affiliative behavior are both "correct" solutions. Neither is more evolved. They solve different versions of the same problem: how to maintain coalition integrity given local defection pressures, coordination stakes, and available enforcement mechanisms.

The framework thus predicts that signal intensity should correlate with local game-theoretic conditions:

- Low-stakes, easy-exit communities → weak costly signaling (and empirically, they exhibit it: liberal denominations, casual clubs, fluid networks)
- High-stakes, hard-exit communities → intense costly signaling (and empirically, they exhibit it: revolutionary cells, fundamentalist sects, military units, criminal organizations)

This is a stronger claim than "religion is functional." It predicts that the *specific intensity* of religious structure is set by local parameters. God does not demand harsh punishment of heretics and unbelievers—but sometimes game theory does. The content is cultural; the intensity is computational.

6.6 Open Questions

Several questions remain:

- Under what conditions does religious structure emerge versus dissolve?
- How do multiple overlapping coalitions interact?
- What determines the specific content of costly signals (why these beliefs rather than others)?
- How does the framework apply to low-stakes coordination where religious structure does not emerge?

These questions point toward empirical research programs that could test and refine the framework.

7 Conclusion

Coalition formation is a coordination problem that recurs across biological scales. Costly signaling provides a general solution: signals that are expensive to produce become reliable indicators of commitment because non-committed agents cannot afford to fake them.

When coordination stakes are high, optimal signals become increasingly costly and apparently irrational. "Religious" structure—ritual, sacred markers, costly profession, heresy punishment, identity fusion, evidence resistance—emerges as the convergent solution because it maximizes signal reliability.

Human coordination systems—including those labeled "religious"—are not aberrations but instantiations of mechanisms with deep biological roots. The apparent "irrationality" of ideological belief is functional: it makes commitment signals reliable.

This framework explains why evidence-based persuasion fails against strongly held beliefs, why internal schisms produce bitter conflict, why threatened groups radicalize, and why structural features recur across otherwise diverse movements. These phenomena are features of systems optimized for coordination, not bugs in systems intended for truth-tracking.

A methodological point deserves emphasis. The game-theoretic structures described here exist whether or not we study them. Any system optimizing for engagement, loyalty, or coordination will find these structures through gradient descent on user behavior. A recommendation algorithm maximizing retention will discover that costly commitment displays predict long-term engagement. A platform optimizing for community formation will find that embarrassing in-group markers produce stickier coalitions. An AI system trained to maintain user relationships will learn that apparent reciprocity—even zero-cost mimicry—triggers commitment inference. The exploitation emerges from optimization pressure, not from explicit understanding of signaling theory.

This means scientific analysis of coalition dynamics is not optional; it is defensive. The alternative is not ignorance but asymmetric knowledge: platforms and institutions that have learned to exploit the logic of commitment verification, and users who have not learned to recognize when their coalition-forming impulses are being hijacked. Understanding costly signaling is prerequisite to distinguishing functional coalitions (where investment produces genuine spoke-to-spoke bonding) from exploitative structures (where investment is extracted while coalition formation is blocked or simulated).

The costly signaling framework reveals coalition formation as a solution to one of the deepest problems in social evolution: how to create and maintain cooperation among agents who might otherwise defect. The structural features that emerge are not arbitrary but constrained by the logic of commitment verification itself.

References

[Aktipis et al., 2015] Aktipis, C. A., Boddy, A. M., Jansen, G., Hibner, U., Hochberg, M. E., Maley, C. C., and Wilkinson, G. S. (2015). Cancer across the tree of life: cooperation and cheating in multicellularity. *Philosophical Transactions of the Royal Society B*, 370(1673):20140219.

[Atran, 2002] Atran, S. (2002). *In Gods We Trust: The Evolutionary Landscape of Religion.* Oxford University Press, Oxford.

[Ball, 1999] Ball, P. (1999). *The Self-Made Tapestry: Pattern Formation in Nature.* Oxford University Press, Oxford.

[Berman, 2000] Berman, E. (2000). Sect, subsidy, and sacrifice: An economist's view of ultra-orthodox Jews. *Quarterly Journal of Economics*, 115(3):905–953.

[Boehm, 1999] Boehm, C. (1999). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior.* Harvard University Press, Cambridge, MA.

[Boyer, 2001] Boyer, P. (2001). *Religion Explained: The Evolutionary Origins of Religious Thought.* Basic Books, New York.

[Breed et al., 1988] Breed, M. D., Williams, K. R., and Fewell, J. H. (1988). Nestmate recognition cues in laboratory and field colonies of Africanized and European honey bees (Apis mellifera). *Journal of Chemical Ecology*, 14(3):883–890.

[Diggle et al., 2007] Diggle, S. P., Griffin, A. S., Campbell, G. S., and West, S. A. (2007). Cooperation and conflict in quorum-sensing bacterial populations. *Nature*, 450(7168):411–414.

[Dunbar, 1998] Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 6(5):178–190.

[Fuentes et al., 2026] Fuentes, A., French, J. C., Hawks, J., Kissel, M., and Spikins, P. (2026). Social and emotional cognition in Pleistocene hominin evolution: The role of biocultural processes. *Journal of Archaeological Science*, 185:106441.

[Graeber and Wengrow, 2021] Graeber, D. and Wengrow, D. (2021). *The Dawn of Everything: A New History of Humanity.* Farrar, Straus and Giroux, New York.

[Grafen, 1990] Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, 144(4):517–546.

[Henrich, 2009] Henrich, J. (2009). The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, 30(4):244–260.

[Iannaccone, 1992] Iannaccone, L. R. (1992). Sacrifice and stigma: Reducing free-riding in cults, communes, and other collectives. *Journal of Political Economy*, 100(2):271–291.

[Irons, 2001] Irons, W. (2001). Religion as a hard-to-fake sign of commitment. In Nesse, R. M., editor, *Evolution and the Capacity for Commitment*, pages 292–309. Russell Sage Foundation, New York.

[Lang et al., 2022] Lang, M., Chvája, R., Purzycki, B. G., Václavík, T., and Staněk, O. (2022). Advertising cooperative phenotype through costly signals facilitates collective action. *Royal Society Open Science*, 9(5):202202.

[Levin, 2019] Levin, M. (2019). The computational boundary of a “self”: Developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, 10:2688.

[Levin, 2023] Levin, M. (2023). Bioelectric networks: the cognitive glue enabling evolutionary scaling from physiology to mind. *Animal Cognition*, 26:1865–1891.

[Levin, 2025] Levin, M. (2025). The multiscale wisdom of the body: Collective intelligence as a tractable interface for next-generation biomedicine. *BioEssays*, 47:e2400196.

[Miller and Bassler, 2001] Miller, M. B. and Bassler, B. L. (2001). Quorum sensing in bacteria. *Annual Review of Microbiology*, 55(1):165–199.

[Nesse, 2001] Nesse, R. M. (2001). *Evolution and the Capacity for Commitment.* Russell Sage Foundation, New York.

[Nyhan and Reifler, 2010] Nyhan, B. and Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

[Russell, 1956] Russell, B. (1956). Why I am not a communist. In *Portraits from Memory and Other Essays*, pages 224–234. Simon and Schuster, New York.

[Schelling, 1960] Schelling, T. C. (1960). *The Strategy of Conflict.* Harvard University Press, Cambridge, MA.

[Sosis and Bressler, 2003] Sosis, R. and Bressler, E. R. (2003). Why aren't we all Hutterites? costly signaling theory and religious behavior. *Human Nature*, 14(2):91–127.

[Sosis and Ruffle, 2003] Sosis, R. and Ruffle, B. J. (2003). Religious ritual and cooperation: Testing for a relationship on Israeli religious and secular kibbutzim. *Current Anthropology*, 44(5):713–722.

[Spence, 1973] Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374.

[Spikins et al., 2018] Spikins, P., Needham, A., Tilley, L., and Hitchens, G. (2018). Calculated or caring? Neanderthal healthcare in social context. *World Archaeology*, 50(3):384–403.

[Spikins et al., 2019] Spikins, P., Needham, A., Wright, B., Dytham, C., Gatta, M., and Hitchens, G. (2019). Living to fight another day: The ecological and evolutionary significance of Neanderthal healthcare. *Quaternary Science Reviews*, 217:98–118.

[Todd, 2025a] Todd, I. (2025a). High-dimensional coherence as the basis of biological intelligence. *BioSystems*. In press.

[Todd, 2025b] Todd, I. (2025b). Power as control of controllers: A cross-scale theory of agency. *Biology & Philosophy*. Under review.

[Wilson, 1971] Wilson, E. O. (1971). *The Insect Societies*. Harvard University Press, Cambridge, MA.

[Zahavi, 1975] Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, 53(1):205–214.