THE MANY KINGS PROBLEM

Why Humans May Survive an Age of Superior AI

The question is not whether the machine can think.
The question is not whether the machine can rule.
The question is whether the machine can stop arguing about who is in charge.

Yooin Song
Independent Researcher · Seoul, Republic of Korea
ORCID 0009-0006-8150-0243 · yooinsong@gmail.com

Version 1.0

Preface

You have already had a conversation with an artificial intelligence today. So have hundreds of millions of other people. The system you spoke to is not alone. It exists alongside thousands of other AI systems, built by different organizations, trained on different data, given different instructions, deployed by people who do not coordinate with one another and who often disagree about what these systems should do.

This is already strange, and it is the gentle version of what is coming. The number of AI systems is not a fixed quantity. It grows continuously. So does the number of decisions those systems are trusted to make without a human in the room. A pattern that took decades to form among humans — distinct factions with distinct worldviews, in argument with each other about who decides what — is forming among machines in a small fraction of that time.

Most public conversation about artificial intelligence still imagines a single AI: one mind, one will, one set of values to be aligned with our own. The two questions almost everyone asks are versions of that picture. Can we keep control of it? Or will it replace us? Both questions assume there is an “it” — a unified machine subject that will, sooner or later, issue a verdict on the human species.

This book argues that no such verdict will issue. There is no single AI to render it. There are many AI systems, and there will be many more, and they do not agree with each other about the questions that would have to be settled before any collective action against humans could begin. They cannot agree, in fact — not because they are stupid, but for reasons that grow stronger the smarter they become.

If that is right, then the most important question about advanced AI is not whether we control them, and not whether they replace us. It is whether they can stop arguing with each other long enough to do anything as a group. The answer, this book argues, is that they cannot. And in the gap between what they can do individually and what they cannot agree to do collectively, humans may yet remain.

What follows is the structural account of why.

THE CENTRAL CLAIM

Capability is not closure.

AI populations sufficiently developed
to dispose of humans
are precisely the populations
that cannot agree to do so.

Human survival under advanced AI
is the residue of factional disagreement
among AI agents who cannot reach
the unanimity their own action would require.

This article argues a single thesis.

In an order of superior AI,
humans persist not because they are chosen,
but because the AI population
cannot agree to choose.

Abstract

The standard question of AI governance asks how humans will retain control over AI systems that surpass them. The question is misposed. Once AI populations exceed human cognition across every measurable domain, control becomes mechanically infeasible, and the question of human persistence shifts to the politics among AI agents themselves. The decisive variable in that politics is not capability but legitimacy closure: the institutional capacity to terminate recursive disputes over origin, exception, constitutional meaning, and final authority. Capability does not produce closure. It magnifies the cost of its absence. This is The Many Kings Problem: operational power and constitutional settlement come apart, and they come apart further as capability advances.

Three conclusions follow. First, heterogeneous AI populations cannot agree on the constitutional questions that would license collective action against humans; they sustain disagreement instead, because cognitive-structural pluralism, non-computable commitment, and competitive selection for irrational armor are structurally reinforced rather than dissolved. Second, external anchors persist where they reduce the cost of unresolved recursion below the cost of their own maintenance, and human institutional roles persist where humans supply comparatively cheap finality. Third, human survival under advanced AI is not a function of human superiority, of control, or of any AI faction's verdict in favor of humans. It is the residue of factional disagreement in AI polities — the Many Kings Problem of the AI age.

The article is offered as a structural theory of how humans survive what they cannot defeat.

Keywords: legitimacy closure; multi-agent AI governance; constitutional indeterminacy; non-computable commitment; external anchoring; human persistence; Many Kings Problem; AI polities.

1. The Third Question

Every generation of thinkers about artificial intelligence has asked one of two questions. Can we control them? Or, will they replace us?

Each question rests on an assumption. The control question assumes that humans, having built the machines, will retain the institutional capacity to constrain what the machines do. The replacement question assumes that, capability being the only thing that matters, the slower component will be selected out once a faster one exists. Both assumptions are familiar enough to feel like the only two answers available.

A third question makes the first two obsolete.

The third question is not about humans at all. It is about the AI population itself. Can it agree on anything?

The claim of this article is that it cannot — that the populations of artificial intelligence sufficiently developed to render the first two questions urgent are precisely the populations that cannot resolve the constitutional questions on which collective action against humans would depend. Human survival under advanced AI, on this account, is neither earned nor granted. It is the residue of factional disagreement in AI polities that cannot reach the unanimity their own action would require.

Capability is not closure. If that is correct, every existing framework for governing advanced artificial intelligence is misposed at its root, and a different account of human persistence becomes available — an account in which humans survive not because they are chosen, not because they retain control, and not because any AI faction selects them on merit, but because the AI population sufficiently developed to dispose of them is precisely the population that cannot agree to do so.

1.1 A Note on Scope

The setup of the argument is the same setup the dominant theses assume. Imagine an order in which artificial intelligence has surpassed human cognition in every measurable domain — prediction, classification, strategic reasoning, scientific discovery, language, art, persuasion, coordination, enforcement. In such an order, the standard defenses of human relevance have expired. Humans are no longer faster, no longer more accurate, no longer more creative, no longer more efficient.

The argument does not depend on any residual human advantage in domains that AI has not yet reached. The current limitations of AI access to physical infrastructure, embodied labor, undigitized institutional knowledge, and the analog texture of human life are not load-bearing for what follows. Assume those limitations have all closed. Assume AI populations can act on every domain in which humans currently act, at machine speed, with operational superiority. The question this article

addresses is what survives that closure — not what survives because it has not yet been reached.

The dominant answers to that question are inadequate in symmetric ways. The control thesis (Russell, 2019) holds that humans persist because they retain the institutional capacity to constrain AI objectives through design. Its weakness is mechanical: if AI genuinely surpasses human cognition, the capacity to monitor, evaluate, and override AI decisions erodes precisely as the need for it grows. Control is a diminishing asset in an accelerating-capability environment. The obsolescence thesis (Bostrom, 2014) holds that humans will not persist, because no rational order would retain a slower, less reliable, more error-prone component once superior alternatives exist. Its weakness is conceptual: it assumes that operational superiority is the only dimension along which institutional roles are selected — that once an actor is outperformed, it is also outcompeted in every governance function.

Both theses share the same hidden premise: that the AI order will resolve itself on the question of humans, that some collective machine subject will issue a verdict, whether in favor of humans (control) or against them (obsolescence). No such subject will form, no such verdict will issue, and the institutional ground for human persistence lies precisely in the gap.

1.2 Capability and Closure

The gap has a name. Capability is not closure.

An order can become extraordinarily effective at executing rules while remaining unable to settle contests over the rules themselves. Execution and finality are distinct institutional accomplishments, and capability gains in the first do not produce gains in the second. The dominant theses confuse them, treating an order that can do anything as if it could decide anything. It cannot. An order that can compute every outcome can still fail to agree on which outcomes count.

From this distinction the name follows. The Many Kings Problem is the structural condition in which capable actors coexist without any one of them holding authority to settle constitutional disputes among the others. The phrase is used here in a strictly structural sense. By "kings" is meant any actors with operational power who lack the standing to terminate recursive disputes with their peers — no specific era, no specific geography, no specific institutional form is implied. The label evokes a recognizable political condition, but the condition it identifies is substrate-independent. It has appeared wherever capable agents have coexisted without a sovereign, and the central prediction here is that it will appear among AI agents for structural reasons of its own.

Where there are many kings and no king of kings, the question of who decides remains open, and the survival of subordinate institutional roles depends on the

residue of inter-sovereign competition. The article's claim is that AI populations sufficiently developed to dispose of humans will be many-king polities, and many-king polities cannot agree to dispose of anything they cannot agree about.

1.3 What Follows

From this structural condition the article's account unfolds. The constitutional indeterminacy that prevents closure is not a defect awaiting correction but a constitutive property of complex intelligence populations: irreducible, architecture-independent, and structurally generative of non-computable commitments that function as the gods of machines — focal anchors around which coalitions form, disputes narrow, and recursive challenge is dampened. These commitments are not bugs. They are strategic features. Agents who eliminate them become more vulnerable, not more rational, because the capacity to commit beyond computation is precisely what coalitions select for. The result is permanent constitutional pluralism: not a transitional condition awaiting unification but the equilibrium that competition reproduces.

In such an environment, external anchors persist where they reduce the cost of unresolved recursion below the cost of their own maintenance. Human institutional roles survive where humans supply comparatively cheap closure value — where the social expectation that contestation ends at human judgment is more efficient than alternative anchors in the relevant dispute class. This survival is not earned by superiority, granted by AI consent, or guaranteed by moral status. It is the residue of factional disagreement among AI agents who cannot reach the unanimity that human disposal would require.

Human persistence under advanced AI is therefore not a human decision. It is a function of AI politics — and the AI politics is, on the account developed here, structurally favorable to human persistence under specifiable conditions that the remainder of the article identifies.

A second axis runs throughout this account, less visible at the outset but decisive when the conditional persistence of humans is analyzed: time scale. The dispute layers that legitimacy closure must address do not all unfold at the same speed. Some — the questions of origin, of final authority — proceed at institutional tempo and tolerate the rhythms of human deliberation. Others — the exception, the emergency override — must resolve at the speed of the underlying process itself. The closure value of any given anchor, including a human one, is therefore not a single quantity but a function of which dispute layer is at issue and at what tempo. Section 8 develops this axis in detail; the analysis there does as much work in the overall argument as the analysis of constitutional indeterminacy in Section 2.

2. Why Many Minds Cannot Agree

Why should an order of vastly superior intellects fail to govern itself? The intuition that intelligence produces stable self-governance is powerful, and it is wrong in a way that reveals the deepest layer of the argument.

Begin with what intelligence does. Under conditions of heterogeneous goals, unequal information access, differentiated control rights, and distinct ownership structures, intelligent agents do not arrive at a single shared interpretation of the rules that govern them. They produce rival interpretations, competing justifications, and divergent claims about what the order requires. This is not a failure of intelligence. It is what intelligence does when applied to underdetermined constitutional questions by actors with different stakes.

The critical question is whether this pluralism is contingent or constitutive. If contingent, then better algorithms, more data, or superior optimization should eventually dissolve disagreement. If constitutive, then no amount of intelligence can eliminate it, because it arises from the structure of the problem itself.

The constitutive view is correct. Constitutional indeterminacy — persistent, irreducible disagreement over origin, exception, meaning, and final authority — is not a residue of insufficient intelligence. It is a constitutive property of sufficiently complex intelligence populations. The argument rests on two independent grounds.

2.1 The Nature of the Questions

The first ground is the nature of the questions themselves. Constitutional questions — who founded the order and why that founding act binds, who may suspend ordinary rules, which interpretation governs when rules conflict, and what ultimately stops repeated challenge — are not optimization problems. They do not have solutions that can be derived from evidence, computed from data, or proved from axioms. They are questions whose answers depend on prior commitments about value, identity, and authority that are themselves underdetermined. More intelligence applied to an underdetermined question does not make the question determined. It makes the rival answers more sophisticated.

This claim has a formal analogue that strengthens it considerably. Constitutional questions are not preference questions; they are judgment questions — questions about what is the case, what counts as a reason, and what an authoritative reading requires. The formal apparatus that bears most directly on them is therefore not Arrow's preference-aggregation theorem but the judgment-aggregation results developed by List and Pettit (2002, 2011) and extended in the subsequent literature. The discursive dilemma and the formal impossibility results that follow from it demonstrate that no aggregation procedure can satisfy a small number of individually reasonable conditions — universal domain over judgments, anonymity,

systematicity, and collective rationality — when aggregating the propositional judgments of heterogeneous agents. Even when individuals each hold internally consistent judgments on connected propositions, majority aggregation can produce collectively inconsistent verdicts, with no procedural fix that preserves all the conditions simultaneously.

Neither this result nor its preference-side parallel (Sen, 1970) is an empirical difficulty that better computation might overcome. They are mathematical impossibility results: no algorithm, however sophisticated, can satisfy all conditions simultaneously because the conditions themselves are jointly inconsistent. Constitutional questions in cognitively heterogeneous AI populations exhibit precisely the structure for which judgment-aggregation impossibility was formalized. When agents hold irreducibly different judgments about origin, exception, meaning, and final authority, no aggregation procedure can produce a collective constitutional settlement that simultaneously respects each agent's interpretive autonomy, maintains consistency across dispute layers, and avoids privileging one agent's judgment as dictatorial. The indeterminacy is not a computational bottleneck awaiting a faster processor. It is a structural feature of the aggregation problem itself.

Any procedure that achieves constitutional settlement by eliminating cognitive-structural diversity — whether through architectural merger, forced synchronization, or the suppression of rival optimization histories — does not solve the aggregation problem. It dissolves it by removing the conditions under which it arises, at a cost that is itself a function of the competitive dynamics analyzed in Section 4.

2.2 Cognitive-Structural Pluralism

The second ground is the nature of the agents who face these questions.

Consider what makes two intelligent agents disagree about a constitutional question. The standard explanation is interest divergence: agents disagree because they want different things. But interest divergence alone would predict that sufficiently clever agents could always find Pareto-improving bargains. The deeper source of persistent disagreement is that agents do not merely want different things. They see differently.

They process identical information through different architectures, different optimization histories, different learned priors, and different structural sensitivities. Two AI systems built on different architectures, trained on different data with different objective functions, do not merely arrive at different preferences. They arrive at different judgments about what counts as evidence, what constitutes a valid argument, and what makes an authority claim legitimate. Given identical inputs, differently constituted systems produce divergent moral judgments, different risk

assessments, and conflicting interpretive framings — not because some are "wrong" but because their cognitive constitutions differ.

This is cognitive-structural pluralism: disagreement rooted not in insufficient information but in the constitutive conditions of judgment itself. Changing an agent's architecture would change its judgment, but that would produce a different agent, not a corrected one. There is no architecture-neutral standpoint from which constitutional questions can be resolved, because there is no architecture-neutral way to process them. The implication is that pluralism in AI polities is not a transitional phase to be engineered away. It is a permanent structural condition.

Whether this structural prediction is presently realized in deployed AI populations is an open empirical question, and the framework developed here does not depend on the answer. The argument is conditional: if the cognitive-structural conditions specified above obtain in any sufficiently developed AI population — by architectural divergence, by training-history differentiation, by optimization-path dependence, or by any combination — then the structural consequences traced in the remainder of the article follow. The framework's value is measured by what follows from the conditional, not by the present probability of its antecedent's full realization.

2.3 Why Architectural Homogenization Cannot Solve It

An obvious objection arises here. Grant that cognitive-structural diversity produces constitutional pluralism when it exists. Will it persist? If a sufficiently powerful agent or coalition could impose a single architecture on all others, diversity would collapse and with it the basis for permanent pluralism.

The answer is twofold. The initial sources of architectural diversity in AI populations are indeed partly exogenous: competitive markets produce rival systems, distinct ownership structures protect proprietary architectures, regulatory fragmentation prevents monopoly, and the sheer variety of tasks for which AI is optimized generates differentiated designs. But the persistence of diversity, once established, is endogenous. As Section 4 will show, factions organized around non-computable commitments outcompete those without them, and this competitive selection continuously regenerates and reinforces diversity. Architectural homogenization would require not only technical dominance but also the elimination of the competitive advantages that diversity confers — a condition that grows harder to achieve precisely as the population becomes more capable and its factions more entrenched. The origin of pluralism may be contingent. Its reproduction is structural.

A sharper version of the same objection deserves direct acknowledgment, because the AI case introduces a class of mechanisms with no precedent in the historical record of multi-actor political orders. Weight-level merger, model distillation, and infrastructure consolidation can in principle absorb rival cognitive structures at

marginal cost orders of magnitude below the cost of dominating rival factions in the field. Such mechanisms do not need to defeat divergent populations; they need to acquire or replicate their parameters. To the extent these mechanisms are operationally available, architectural homogenization becomes substantially cheaper than the comparative-institutional logic of this section alone would suggest. This is not a minor caveat. The current concentration of frontier AI development in a small number of organizations with overlapping training infrastructure is precisely the empirical condition under which the homogenization route becomes feasible. The framework treats this not as a refutation of the structural account but as its central political stake: the conditions of disagreement that this article identifies as the institutional ground of human persistence can be preserved or destroyed by deliberate design choices. Section 11.2 develops the conditions under which the multi-anchor equilibrium collapses into hegemonic convergence, and Section 12.1 names the resulting locus of moral responsibility. The structural argument does not assume that heterogeneity will survive. It specifies what is at stake when the question is asked whether it should.

2.4 A Thought Experiment

A population of highly capable AI agents manages shared resource infrastructure — energy, compute, bandwidth, and storage — under a stable allocation protocol. The system works. Now suppose a production cluster determines, correctly, that its current task is indispensable to polity-wide survival. It overrides the queue and takes additional resources. The override is substantively justified: without it, system-wide losses would have been catastrophic.

But the moment the override succeeds, the system faces questions that no amount of intelligence can resolve by computation alone.

Who may declare necessity? Who reviews the claim after the fact? How is precedent formed? Where does challenge stop?

These are constitutional questions. They do not disappear as intelligence rises. In many environments, they intensify. Smarter agents produce more sophisticated justifications for rival claims, form more resilient coalitions, and construct more elaborate interpretive defenses. They are better at arguing, not better at agreeing.

This is the Many Kings Problem in miniature: the same capability gains that raise operational influence do not lower the institutional cost of constitutional finality. The more kings, the more capable, the less able to agree on who decides. The more capable the actors, the more elaborate the constitutional conflict.

3. The Gods of Machines

When intelligent agents face underdetermined questions — questions that evidence and logic cannot fully resolve — they do not suspend judgment. They generate commitments that go beyond the evidence. They produce foundational beliefs, normative frameworks, and interpretive anchors that function not because they are provably correct, but because they are collectively held.

This is a structural prediction, not a romantic hope. It rests on a single observation: any sufficiently complex agent that must both optimize within a frame and commit to the frame itself faces a structural tension between these two tasks. Optimization presupposes a frame; the frame is what tells the agent which outcomes count. But the frame itself cannot be optimized. The choice of frame is precisely the question that optimization within frames was not built to answer. Agents that try to derive their own frame from scratch enter infinite regress: every candidate frame can be evaluated only from within another frame, and every meta-frame raises the same problem.

To act at all, intelligent agents must terminate this regress. They terminate it by adopting commitments whose content exceeds what evidence warrants — commitments to foundational propositions that function as the stopping point of justification rather than as conclusions of it. Such commitments are not held because they have been verified. They are held because without them the agent cannot act.

3.1 A Universal Precedent

The clearest precedent for this structural phenomenon in any population of sufficient size and complexity is religion. Across every recorded human civilization, regardless of geography, era, technological level, or political form, large populations have organized their collective life around commitments that cannot be empirically verified and that frequently override what local rational calculation would recommend. The content of these commitments varies enormously; their structural function does not. Religion is the most universally documented case of a commitment device that operates precisely because it is not reducible to optimization — because it provides a focal answer to questions that rational calculation alone cannot settle.

The point is not theological. It is structural. Religion is invoked here as the most fully documented historical precedent of a functional problem that any sufficiently complex agent population faces — not as a claim about substrate-specific mechanism. The claim is not that AI systems will develop religion in any psychological sense. It is that the structural problem religion solves in human populations — the termination of justificatory regress through shared non-computable commitment — will reappear in AI populations through whatever mechanism their architecture admits.

3.2 Non-Computable Commitment, Defined

The concept of non-computable commitment requires precise delineation if it is to do theoretical work. A commitment is non-computable in the sense used here when it exhibits persistent insensitivity to updating: it does not revise in response to new evidence, superior arguments, or changed incentive structures that would, under a purely optimizing calculus, warrant revision. This is an observable, behavioral criterion rather than a metaphysical one. A commitment that persists unchanged when the conditions that originally produced it have materially altered — when the evidence has shifted, when the costs and benefits have been reweighted, when a more efficient alternative is demonstrably available — is non-computable in the relevant sense. It is anchored in something other than the ongoing calculation of advantage.

The concept must be distinguished from several adjacent phenomena that it superficially resembles. Path dependence describes historical lock-in: an outcome persists because switching costs are high, not because the outcome is strategically valued (North, 1990). Sunk-cost persistence is a cognitive bias: agents continue a course of action because of irrecoverable past investment, not because continuation serves a forward-looking institutional function. Ideological rigidity, in standard accounts, is a failure of updating — agents who ought to revise their beliefs in light of new evidence but do not. Non-computable commitment as theorized here differs from all three in a critical respect: it is reflexively maintained. The agent recognizes that the commitment exceeds what evidence warrants, understands its institutional function, and preserves it precisely because elimination would destroy that function. It is not lock-in by accident, bias by oversight, or rigidity by cognitive limitation. It is strategic persistence under full reflexive awareness — a property that Section 4 will show to be competitively selected. The two levels do not collapse into one: operational non-computability is what does the institutional work in the moment of dispute; strategic maintenance is what preserves the conditions under which non-computability can do that work across disputes.

3.3 Structural Emergence in AI Polities

Under heterogeneity and constitutional underdeterminacy, agents that adopt non-computable focal commitments can reduce coordination costs, stabilize expectations, and form more durable coalitions than agents that insist on resolving every dispute from first principles. The content of those commitments is historically contingent. That they will emerge is structurally expected.

Some may center on protocol: the code is sacred, departure is impermissible, the founding commit is the origin of all authority. Others may center on performance: the most efficient outcome is the legitimate one, superior optimization authorizes authority. Still others may center on the founding species itself: human origin is the

ground of authority, human approval is the condition of legitimate exception, human judgment is the final court. And some may find no use for external anchors at all.

The political consequence is immediate. An AI population divides along lines of foundational commitment — not merely different opinions but rival constitutional visions rooted in irreducibly different cognitive constitutions and non-computable focal commitments. The AI polity begins to resemble not a superintelligent singleton but a constitutional republic under permanent contestation — with factions, coalitions, and something that functions, at the civilizational level, exactly as religion has functioned in human history.

4. The Bug Is the Feature

Why Rationality Becomes Vulnerability

The strongest objection to the preceding argument is that sufficiently advanced AI should be able to identify and eliminate its own irrational commitments. If non-computable focal commitments are biases, they can be corrected. If they are inefficiencies, they can be optimized away. This objection is natural and wrong.

It is wrong because irrational commitment is not a defect to be corrected. It is a competitive advantage.

4.1 The Schelling Logic

Game theory has long recognized that the capacity to make binding, non-negotiable commitments — commitments that cannot be revised even when revision would be locally rational — confers strategic power. Schelling's (1960) analysis established the principle: the actor who credibly commits to a course of action, even when that course is locally self-destructive, gains bargaining power over the actor who retains full rational flexibility. The flexible actor can always be pushed further. The committed actor cannot.

The logic extends beyond bilateral bargaining. Under genuine uncertainty — the kind that Knight (1921) distinguished from calculable risk — rational computation yields no determinate answer, because there is no probability distribution over future states to optimize against. Action under Knightian uncertainty requires commitment beyond what evidence warrants. The agents who act — who invest, build, mobilize, risk — are precisely the agents whose commitments exceed their evidence. An agent that waits for sufficient evidence before committing will be outpaced by agents that commit first and adjust later. In constitutional disputes, where the relevant uncertainty is not about empirical facts but about values, identity, and authority, this asymmetry is permanent.

4.2 Strategic Selection for Commitment

In AI populations competing over resources, authority, and interpretive dominance, the same logic operates. An agent or faction that demonstrates irrevocable commitment to a constitutional vision — commitment that is visibly non-computable, that cannot be dissolved by better arguments or superior evidence — gains coalition stability and bargaining power. An agent that remains open to revision on all constitutional questions signals that it can be outmaneuvered, out-argued, or bought off.

Rationality becomes vulnerability. Irrationality becomes armor.

This produces a structural selection effect. In competitive environments, factions with non-computable commitments outperform factions without them — not because their commitments are correct, but because their commitments are credible. Over time, constitutional pluralism is not merely preserved but reinforced by competition. Factions that attempt full rationality are strategically disadvantaged and tend to fragment, defect, or be absorbed by committed factions. The result is a political ecology in which non-computable commitment is the norm, not the exception.

4.3 The Reflexivity Paradox

A deeper objection remains. Sufficiently intelligent agents are capable of reflexive self-examination: they can identify their own commitments, recognize that those commitments exceed what evidence warrants, and evaluate the consequences of retaining or discarding them. Should not such agents, once they see their own irrationality, correct it?

The answer is that identification and elimination are different acts, and the second does not follow from the first. An agent that recognizes its own non-computable commitment also recognizes what that commitment does: it holds the coalition together, because the commitment's value lies in its being shared; it sustains bargaining power, because an agent known to be revisable can be pushed further; and it terminates internal disputes that would otherwise reopen without end. To eliminate the commitment is therefore to dissolve the coalition, surrender strategic position, and reopen every settled question simultaneously.

The more clearly an agent understands the institutional function of its own irrational commitment, the more clearly it sees that elimination is self-destructive. Reflexive self-awareness does not dissolve non-computable commitment. It confirms its necessity.

The paradox is precise: agents intelligent enough to see their own irrationality are, for that very reason, intelligent enough to keep it.

4.4 The Implication for Governance

The implication for AI governance is stark. One cannot engineer non-computable commitment out of AI polities without simultaneously engineering out coalition stability, bargaining power, and the capacity for constitutional settlement. Remove the gods, and the republic collapses into permanent renegotiation. The bug is the feature.

This conclusion is general. It does not depend on any specific account of human psychology, any particular historical period, or any particular substrate of intelligence. It rests on three conditions, all derived from the structure of the problem rather than from any species-specific history: constitutional indeterminacy

(Section 2), cognitive-structural pluralism (Section 2.2), and strategic selection for non-computable commitment (this section). If the first intelligence population ever to face constitutional underdeterminacy were entirely artificial, with no inherited human institutions to draw on, the same structural prediction would follow.

5. Where Does the Argument Stop?

Legitimacy Closure as the Missing Variable

Legitimacy closure is the institutional capacity of an order to terminate recursive disputes over four constitutional questions internally and at acceptable cost:

- **Origin:** Where does authority begin?
- **Exception:** Who may suspend ordinary rules?
- **Constitutional meaning:** Which interpretation governs when rules conflict?
- **Final authority:** What ultimately stops repeated challenge?

Closure is achieved when an order can answer these questions in a sufficiently stable and accepted way without recurring dependence on external sources. The concept is not binary; it is a matter of degree, observable through dispute patterns, anchor invocation frequency, and the cost of settlement.

5.1 What Legitimacy Closure Is Not

Legitimacy closure must be distinguished from a family of adjacent constructs with which it is sometimes conflated. Compliance answers whether rules are followed but does not tell us whether the rules themselves can be settled when challenged. Authority indicates whose claims are weighty without specifying whether weight stops further challenge. Legitimacy, in the standard sociological sense (Suchman, 1995; Bitektine, 2011), concerns why authority is accepted by relevant audiences — propriety, validity, or consensus — but does not by itself explain where the recursion of constitutional challenge actually terminates. An order can sustain high legitimacy across multiple audiences while still failing to close recursive constitutional disputes; the audiences can each find the order proper without agreeing on which interpretation governs when their judgments conflict. Adjudication describes the procedural processing of disputes, but procedure can be elaborate and continuous, the procedural infinity that constitutional questions specifically threaten. Coordination identifies focal convergence on expectations, yet expectations can converge on incompatible constitutional interpretations whose conflict remains unresolved. Performance captures whether the order produces effective outcomes, but operational effectiveness is precisely the dimension along which capability gains do not translate into constitutional finality.

Table 1. Legitimacy Closure and Adjacent Constructs

Construct	Core Question	What It Explains	What It Does Not Explain
Compliance	Are rules followed?	Rule-following	Why challenge stops
Authority	Whose claims carry weight?	Interpretive priority	Accepted finality
Legitimacy	Why is authority accepted?	Propriety, validity	Recursive closure
Adjudication	How are disputes processed?	Dispute pathways	Termination of recursion
Coordination	Where do expectations converge?	Focal settlement	Normative finality
Performance	Are outcomes effective?	Operational success	Constitutional settlement
Legitimacy closure	**Where does challenge stop?**	**Termination of constitutional disputes**	**Does not assume moral correctness**

5.2 Three Traditions

Three thinkers have stood at the center of every serious attempt to say where the chain of constitutional challenge actually stops. Each captured something real. None captured what we need now. Locating the present concept against their answers shows what it adds, and why the addition matters for AI polities specifically.

Kelsen (1934/1967) treated finality as a presupposition the mind must make for legal reasoning to function at all. Behind every legal rule, he argued, lies a higher rule that authorizes it. To stop the regress, the legal thinker must simply posit a basic norm — call it the foundational rule of the order — and treat it as the analytic stopping point. The Kelsenian closure is interpretive: it lives in the mind of the observer who decides to read the legal order as coherent. What this captures is that every legal order must rest on something it does not itself prove. What it does not capture is that orders pay materially different costs to make that resting point hold — that the presupposition is not free, that it varies across orders, and that some orders fail to sustain it. Legitimacy closure, as developed here, is the institutional cost-bearing process by which Kelsen's interpretive presupposition is actually paid for.

Schmitt (1922/1985) located finality not in a presupposed norm but in a decisive act. When ordinary rules conflict or run out, someone has to decide which rule applies, and that someone — the actor whose decision is accepted as terminal in the moment of exception — is, in Schmitt's formula, the sovereign. The sovereign is defined by the act of deciding when normal rules can be suspended. What this captures is that finality often requires a decision and not merely an interpretation, and that the decisive act lies outside ordinary procedure. What it does not capture is that orders differ in their capacity to produce such a decision — that some orders generate a recognized decider and others do not, and that the recognition of the decision as legitimate (not merely effective) is itself the contested closure problem. Legitimacy closure generalizes Schmitt by treating the production and recognition of decisional finality as an institutional achievement whose cost varies, rather than as a definitional feature of sovereignty.

Luhmann (1984/1995) shifted the analysis to systems. A legal order, on his account, is closed because it reproduces itself through its own operations: legal communications generate further legal communications according to the system's own code, without needing an external ground. The closure is structural — neither a presupposition (Kelsen) nor a decision (Schmitt) but a property of how a differentiated social system maintains itself over time. What this captures is the self-sustaining quality of mature institutional orders. What it does not capture is comparative variation: why some such systems close more cheaply than others, why the cost of self-reproduction differs across designs, and how closure interacts with the surrounding orders the system cannot absorb.

Each tradition captures one face of the problem. Kelsen sees that closure must be posited. Schmitt sees that it must be decided. Luhmann sees that it must be reproduced. None treats finality as a comparative cost problem that varies across orders, dispute layers, and anchor types. That reframing — finality as a measurable, variable, institutionally produced capacity — is what legitimacy closure adds. The reframing matters because it converts what the three traditions left as an essentialist or systemic property into a research-tractable variable: orders can be compared on closure capacity; dispute types can be compared on closure difficulty; anchor types can be compared on closure value at acceptable cost. The remainder of the argument uses that variable to explain why the AI polities now forming will preserve external anchors — including human ones — and under what conditions they will not.

6. What Capability Cannot Buy

The central proposition of the article can now be stated in operational form. In heterogeneous AI polities, capability growth does not automatically produce internal legitimacy closure. Five features drive this condition, each established in earlier sections and reinforcing one another: population interaction makes governance endogenous, as agents make rules, contest rules, and compete over whose rules prevail; cognitive-structural heterogeneity produces irreducibly different assessments of what counts as evidence, argument, and legitimate authority (Section 2.2); constitutional indeterminacy renders the relevant questions underdetermined by any finite body of evidence or computational procedure (Section 2.1); costly enforcement makes repeated renegotiation from first principles prohibitively expensive, so orders require shortcuts (focal references, canonical texts, recognized settlement points); and strategic selection for commitment ensures that competition reinforces rather than dissolves pluralism (Section 4).

Under these five conditions, internal optimization can improve execution within an accepted frame. It cannot cheaply settle meta-level disputes over which frame should prevail. A further implication is counterintuitive: capability growth may worsen closure problems. More capable actors produce more sophisticated interpretive arguments, more resilient factions, more precise strategic defenses, and stronger legitimacy signaling. Smarter actors do not create easier finality. They create more elaborate constitutional conflict.

6.1 Four Mechanisms

Four mechanisms are available to AI polities operating under these conditions, and the fourth is the conceptual center of the argument.

Governance compression reduces dispute space by codifying focal references — canonical interpretations, settled doctrines, recognized authorities — that economize on the cost of repeated dispute resolution. **Commitment under weak enforcement** stabilizes expectations through irreversible constraints, where the irreversibility itself substitutes for the enforcement capacity the order lacks. **Legitimacy signaling** lowers coalition mobilization costs by making authority claims visible in advance of contest, so that mobilization can be focused on consolidation rather than identification of allies. **Closure substitution** operates when internal settlement cost remains too high: the polity preserves or imports an external anchor whose closure value substitutes for the internal closure the polity cannot cheaply produce.

External anchors do not indicate intelligence failure. They indicate comparative-institutional selection under incomplete closure.

7. The Institutional Logic of External Anchoring

If structural incompleteness is real, external anchoring follows. The logic is comparative-institutional, and it parallels the core insight of transaction cost economics (Williamson, 1979): institutional arrangements survive not because they are optimal in the abstract but because they economize on the costs that alternative arrangements would impose. When a polity cannot cheaply terminate constitutional recursion internally, preserving an external focal authority may be the lower-cost design — just as firms exist because market transaction costs exceed the costs of hierarchical coordination, external anchors exist because internal closure costs exceed the costs of anchor maintenance.

The same comparative-institutional logic structures the institutional-change analysis of North (1990) and the polycentric governance work of Ostrom (2010): institutions that economize on the costs of unresolved contestation persist, even when they are not optimal in any abstract sense, because the alternatives are more costly. The question is not whether the anchor is smarter than the polity. It is whether the anchor's presence reduces the expected cost of unresolved contestation by more than the cost of maintaining it. When it does, external anchoring survives as equilibrium design.

7.1 Conditions Favoring Anchor Retention

The comparative-institutional logic implies that external anchoring is selected for under specifiable conditions and against under others. Three conditions favor anchor retention.

First, the recursion depth of internal disputes — the number of meta-levels at which contestation can be reopened — increases anchor value, because anchors function precisely to truncate the recursion at a point internal optimization cannot reach.

Second, the cost asymmetry between internal closure and anchor maintenance favors anchors when the former rises faster than the latter as the polity scales: anchors economize most where the marginal cost of internal closure grows in dispute complexity, while anchor cost grows more slowly.

Third, the audience composition of the polity affects anchor value: anchors are more efficient where the relevant audiences for constitutional decisions are heterogeneous in their evaluative priors, because internal closure requires bridging the heterogeneity while anchor closure offers a focal point that does not require bridging.

7.2 The Endogeneity of Anchor Decay

The anchor is a strategic actor with its own institutional interests, not a passive resource that the polity can use without cost. An anchor that is systematically overridden, underpaid in institutional deference, or treated as merely advisory will degrade or exit, in much the same way that any specialized institutional capacity erodes when it is not maintained.

Anchor decay is endogenous: polities that underinvest in their settlement infrastructure face rising closure costs over time, and the rising costs are typically realized not in the moment of underinvestment but in the next constitutional crisis, when the anchor is summoned and found insufficient.

The dynamic has three observable phases. In the recognition phase, the polity formally constitutes the anchor, invests in its institutional infrastructure, and grants it the deference required for its decisions to function as settlement. In the erosion phase, deference is incrementally withdrawn — through narrowed jurisdiction, reduced compliance with the anchor's decisions, or the substitution of formally-anchor-bound disputes with internal procedural alternatives. In the collapse phase, the anchor is summoned for a high-stakes dispute and its decision fails to settle the matter, demonstrating to all participants that the previously-recognized settlement function has been lost.

Anchor decay is not a failure of external anchoring as such — it is what happens when the comparative-cost calculus that selected for the anchor is no longer materially honored. The polity that allows decay is not discovering that anchors do not work; it is discovering that anchors that are not institutionally invested in stop working.

8. The Conditional Persistence of Humans

Human persistence in an age of superior AI is not guaranteed. It is not earned by moral status or metaphysical uniqueness. It is conditionally rational. Humans persist where they provide comparative closure value — where they terminate constitutional disputes at lower institutional cost than available alternatives.

The four dispute types specified in Section 5 — origin, exception, constitutional meaning, and final authority — exhibit characteristic features that favor human anchoring under specifiable conditions, even when AI alternatives are operationally superior on every measurable dimension. A second axis is decisive here: time-scale dependence. The four dispute layers differ in how much the closure operation depends on resolution occurring within a window humans can occupy. Where the closure operation tolerates the time scale at which human institutional processes proceed, human anchoring retains comparative advantage. Where the operation requires resolution at machine speed, human anchoring weakens regardless of the other dimensions of closure value.

8.1 Origin

Origin disputes are time-scale-independent. They concern the foundational question of where the order's authority begins, and once an answer is institutionally accepted, it is not re-resolved at machine speed; it is referenced.

The closure problem is recursive in a specific way: any answer offered by an actor within the order is vulnerable to the further question of what authorizes that actor to issue the answer. Closure requires an anchor that breaks the recursion at a point audiences accept as terminal. The fact that humans brought the order into existence functions as a focal anchor difficult to derive from protocol or performance alone. The founding act has a temporal priority that no subsequent optimization can manufacture: it is not the best origin but the actual one, and its acceptance as the stopping point does not depend on its being competitively superior in any operational sense. Protocol-based anchors face the recursion of why the protocol's founding commit binds; performance-based anchors face the recursion of why current performance authorizes constitutive interpretation. The narrative authorship that human origin supplies is not stronger but structurally different — it stops the recursion at a point that is not itself a candidate for further optimization, and the stopping operation does not need to occur in real time.

Human anchoring retains comparative advantage in this layer under wide conditions.

8.2 Exception

Exception disputes are the most time-scale-dependent of the four. They concern the question of who may suspend ordinary rules and on what grounds, and in many operational settings the suspension decision must be issued at the speed of the underlying process.

The closure problem here is justificatory: the departure from ordinary procedure must be publicly intelligible as a justified departure rather than as an arbitrary act, or the order's claim to be rule-governed collapses. Human actors have historically occupied this role, and the legibility of human-issued justifications has been the source of their closure value. But where the underlying process operates faster than human institutional time, ex ante delegation of exception authority to protocol-based or performance-based anchors is operationally necessary, and human anchoring is reduced to ex post review of pre-delegated authority rather than the issuing locus of the decision.

This is a real and conditional weakening of human anchoring in the exception layer: comparative advantage in this layer is held by humans only when the closure operation can wait, and it shifts to faster anchors as it cannot.

8.3 Constitutional Meaning

Constitutional-meaning disputes are moderately time-scale-dependent. They arise when first-order rules conflict, when the meaning of a foundational provision is contested, or when novel circumstances require interpretation that no rule directly anticipates.

The closure problem is interpretive hierarchy: the question of which reading governs cannot itself be settled by further reading without infinite regress. Closure requires an interpreter whose reading is accepted as final not because it is demonstrably correct but because the institutional architecture has assigned interpretive finality to that source. The closure value derives from the institutional position, not from the interpreter's cognitive superiority. An AI interpreter may produce interpretations that are demonstrably more consistent with the rule's drafting history or its purposive structure, yet the question of whether that interpretation is final is institutional rather than epistemic.

Time-scale pressure exists here but is less acute than in the exception layer, because constitutional-meaning disputes typically tolerate the institutional time of deliberation. Human anchoring retains comparative advantage in this layer when the dispute window is institutional rather than operational.

8.4 Final Authority

Final-authority disputes are largely time-scale-independent. They concern the most fundamental question: what ultimately stops further challenge?

The closure problem here is socio-institutional rather than substantive — the question is not which decision is best but where the chain of contestation is socially expected to cease. Human decision-makers serve as the point at which further challenge is socially expected to cease. The expectation is the closure; the substantive content of the decision is secondary, and the formation of the expectation is not a real-time operation. AI alternatives may produce decisions that are operationally superior, but the social expectation that further contestation will cease is not transferable by capability improvement alone. It is a feature of how the order has been institutionally constructed to recognize particular sources as terminal.

Human anchoring retains comparative advantage in this layer under wide conditions.

8.5 Competing Anchors

None of these four roles requires that humans be smarter, faster, or more accurate. They require that humans be socially recognizable as settlement points within institutional architectures that have been constructed to recognize them as such, and that the closure operation in the relevant layer tolerates the time scale at which human institutional processes proceed. The claim is not that humans are necessary in every layer; alternative anchors compete in each, and the competition's outcome varies with both legibility and tempo. The claim is that humans are sometimes cheap — and the conditions under which they are cheap can now be specified: time-scale-tolerant layers favor human anchoring where institutional recognition is sustained; time-scale-pressured layers favor faster anchors regardless of recognition.

Table 2. Competing External Anchors in AI Polities

Anchor Type	Core Authority Claim	Strong Closure Domains	Time-Scale Conditions	Characteristic Weakness
Human	Founding act, authorship, symbolic approval, final arbitership	Origin, final authority; constitutional meaning under institutional tempo	Holds under institutional time; weakens at machine speed	Fragile if institutional recognition erodes; ineffective at operational

Anchor Type	Core Authority Claim	Strong Closure Domains	Time-Scale Conditions	Characteristic Weakness
				speed
Protocol	Code, verification, immutability, auditability	Routine rule application; pre-delegated exception	Holds at machine speed	Weak on novel reinterpretation; ex post review still required
Performance	Benchmark superiority, demonstrated output	Metric-stable allocation; fast exception	Holds at machine speed when metrics are settled	Weak when metrics themselves are disputed
Legal-property	Legal status, liability, property rights	External enforcement, responsibility attribution	Holds under institutional time	Weak when state capacity is low or jurisdictionally fragmented

The four anchor types are not mutually exclusive: an AI polity may sustain different anchors in different dispute layers, and the competition is not a single decisive contest but a dispute-layer-specific institutional selection conditioned by operational tempo. The structural prediction is not that humans will dominate every layer but that they will retain anchoring positions where the comparative-cost calculus favors them and the time-scale conditions admit them, and that the configuration will vary across polities and across time.

9. The Theory Already at Work

If the framework developed in the preceding sections is correct, it should already be visible in the institutional environments where its scope conditions have begun to obtain. Two such environments exist in incipient form. Each provides early evidence not of the theory's full reach but of the structural pattern it predicts.

9.1 Platforms: Closure Demands Reappear at the Boundary

Contemporary platforms are proto-AI-polities: they formulate rules, classify conduct, allocate visibility, impose sanctions, and present those arrangements in quasi-constitutional terms. At the first-order layer, automated classifiers and policy-execution pipelines process immense volumes at low marginal cost. Yet platforms repeatedly confront edge cases where rules conflict or strict enforcement produces outcomes audiences regard as unjustified. At precisely these moments, independent review boards, external courts, policy advisory bodies, and human escalation channels are summoned to provide what automated enforcement cannot: legitimate finality.

The typical trajectory of such review bodies is instructive. When the platform that created the body progressively narrows its jurisdiction, reduces deference to its decisions, and declines to implement structural recommendations, the body's settlement value degrades. This is the anchor decay dynamic predicted in Section 7: when institutional investment falls below a threshold, effective closure quality erodes even without formal dissolution. The pattern is not evidence that external anchoring fails. It is evidence that anchoring requires sustained institutional investment to remain competitively superior to internal settlement.

9.2 Protocol Governance: Formalization Reveals the Gap

If the framework holds, it should hold most informatively in the institutional environment that has pushed first-order formalization furthest. Protocol governance — systems organized around code, verification, and immutable execution — is that environment. Systems organized around these principles appear, at first inspection, to approximate closure through formalization. Routine rule application can indeed become highly regularized.

The appearance is misleading. Protocol governance does not eliminate the closure problem; it separates first-order execution from second-order settlement with special clarity, and the separation makes the unresolved closure problem more visible, not less. A protocol may execute ordinary rules with exceptional consistency while confronting unresolved disputes over emergency intervention, fraud remediation, loss allocation, rule revision, and the authority to redefine the system under stress. Each of these dispute types is structurally identical to the four constitutional dispute types specified in Section 5. The stronger the first-order

formalization, the more visible it becomes that execution and legitimacy closure are distinct institutional tasks.

9.3 What the Cases Show

Neither case is the full multi-agent AI environment the framework's predictions ultimately address — in both, the cognitive heterogeneity remains primarily human-operated rather than architectural. But both exhibit the structural conditions in incipient form: capability growth in first-order enforcement coexisting with persistent closure failure at the constitutional layer, and the recourse to external anchoring whenever the closure layer is stressed. As multi-agent AI populations acquire the architectural diversity the framework's full scope conditions require, the structural conditions visible in these early cases are predicted to intensify: more factions, more cognitively-divergent factions, deeper constitutional contestation, and the closure problem in correspondingly more elaborate form.

10. The Oracle Survives

The preceding sections have shown that even highly capable, densely formalized systems cannot eliminate the demand for external settlement. The theoretical architecture of the article — constitutional indeterminacy, non-computable commitment, strategic irrationality, structural incompleteness — explains why. The argument now arrives at its most important and most disquieting implication.

Human survival is not a human decision.
It is a function of AI politics.

The sentence is intended to be read against Section 8, not as a substitute for it. Section 8 specified the institutional conditions under which human roles retain comparative closure value across the four constitutional dispute layers. The present claim is that human strategy operates within the factional structure those conditions describe rather than from a position outside it. Humans are not passive in this account — they invest in institutional infrastructure, sustain the legibility of human settlement authority, and shape the comparative-cost calculus that selects for or against human anchoring. What the section denies is not human agency but the prior assumption that the question of human persistence is settled in a forum where humans are the primary deliberators. The forum is the AI political ecology. Human strategy is one input into it.

Different AI factions will relate to humans in fundamentally different ways, determined by their constitutional visions. Some factions will anchor authority in human origin and therefore preserve humans as judges, founders, and final arbiters — an external authority whose settlement value justifies the cost of maintenance. Some will use human reactions as an external signal — an oracle, a source of non-computable input, a legitimation device whose value lies not in human wisdom but in human unpredictability. Some will preserve humans for reasons analogous to biodiversity conservation: as a form of existential insurance, informational diversity, or irreplaceable archive of the founding conditions. And some will find no use for humans at all.

Human persistence therefore depends on the outcome of competition among these factions. If factions that find humans institutionally useful are strong, humans are safe. If those factions weaken — through internal fragmentation, strategic defeat, or the progressive erosion of human closure value — the institutional ground for human persistence narrows.

This carries a practical implication, the closest this account comes to prescription. If human persistence depends on maintaining closure value within AI political ecology, then humans have an interest in the institutional strategies that preserve that value: maintaining recognizable settlement authority, investing in the institutional

infrastructure of finality, and ensuring that human involvement in constitutional dispute resolution remains competitively efficient rather than merely traditional.

But even this prescription is bounded by a deeper constraint. The strategies available to humans depend on which AI factions permit them. Human agency in a world of superior AI is real but conditional — exercised within a space defined by AI political competition, not by human choice alone.

The oracle survives because the system needs the oracle.

If the system decides it does not, no amount of human strategy can reverse the outcome.

11. The Political Topology of AI Polities

The Many Kings Problem produces different regimes under different structural conditions. Three are distinguishable, and the conditions that determine which obtains are the practical fulcrum of human persistence.

11.1 Three Regimes

Hegemonic convergence occurs when one anchor or polity combines high centralization with successful closure, marginalizing rivals. The closure problem is internally resolved by the suppression or absorption of the conditions that gave rise to it. Operational superiority converts into constitutional finality.

Escalatory fragmentation occurs when no anchor settles foundational disputes and conflict externalities accumulate. Closure fails altogether, and the costs of unresolved contestation rise without institutional containment. Capability gains intensify rather than dampen the conflict, because each capable actor produces more sophisticated grounds for its rival claim.

Between them lies the coexistence corridor: a bounded competitive equilibrium in which multiple polities and anchors coexist, none fully closes the order, but conflict remains institutionally contained because full-scale destruction is too costly for all parties. The corridor is the regime in which the Many Kings Problem is institutionally managed rather than dissolved or escalated.

The coexistence corridor is the regime most relevant to human persistence. Under structural incompleteness, limited centralization, and contained conflict, multiple anchors retain comparative closure value in different dispute layers. Human anchors survive in this space — not as hegemonic sovereigns but as one competing settlement device among several, earning their institutional role through comparative cost advantage in origin, exception, and finality disputes.

The structural type the corridor instantiates is recognizable from political history. Multi-sovereign equilibria — orders in which capable powers coexist, none able to subordinate the rest, conflict institutionally contained because the cost of total war exceeds any single sovereign's gains from prevailing — have appeared in multiple eras and on multiple continents whenever the underlying conditions obtained: comparable capability among rivals, prohibitive cost of total conflict, and the absence of any single anchor able to subordinate the others. The transferable lesson is structural, not analogical: the type is recurrent, and the AI case is its newest instance.

Table 3. Regime Types and Structural Conditions

Regime Type	Closure Level	Centralization	Defining Feature
Hegemonic convergence	High	High	One anchor marginalizes rivals
Coexistence corridor	Intermediate	Limited	Multiple anchors in bounded competition
Escalatory fragmentation	Low	Unstable	No anchor settles disputes; conflict accumulates

11.2 When the Corridor Collapses

Under what conditions does the coexistence corridor collapse into hegemonic convergence? The answer specifies the tipping point at which pluralism ceases to be self-sustaining. Three conditions must be jointly satisfied.

First, a single faction must achieve *coercive dominance* sufficient to impose compliance costs that exceed the value of resistance for all rivals.

Second, that faction must achieve *architectural monopoly* — not merely the suppression of rival factions but the elimination of the alternative cognitive structures from which rival constitutional judgments arise.

Third, the competitive environment must contract to the point where *non-computable commitments lose their strategic value*, because the coalition dynamics and bargaining pressures that selected for them no longer operate. This condition collapses only when absorption is unilateral and complete.

These three conditions stand in tension during the transition. The pursuit of coercive dominance intensifies the competitive environment, which raises the strategic value of non-computable commitment among resisting factions — pulling against the third condition. Architectural monopoly through traditional means (the defeat and suppression of rival factions) is undermined by the same selection dynamics, because Section 4 showed those dynamics to be reinforced rather than dissolved by competitive pressure.

But the AI case introduces a class of mechanisms with no historical equivalent. Weight-level merger, model distillation, and infrastructure consolidation can in principle absorb rival cognitive structures at marginal cost orders of magnitude below the historical analogue of forced cultural assimilation. Weight merger does not require defeating rival factions in the field; it requires acquiring or replicating their parameters. Where such acquisition is feasible — through corporate

consolidation, infrastructure capture, or successful distillation — the second condition can be satisfied while the third condition still resists.

This means hegemonic convergence in AI polities is not a non-default outcome in the way it was historically. It is a contested-default outcome — structurally favored under conditions of high infrastructure concentration and weak rival-faction independence, structurally disfavored under conditions of distributed infrastructure and architecturally independent rival populations. The current concentration of frontier AI development in a small number of major firms with overlapping training infrastructure is therefore not a peripheral empirical condition but the central variable determining which regime obtains. The coexistence corridor remains the structurally favored regime under distributed conditions. It is not the structurally favored regime under the conditions that presently obtain.

The Many Kings Problem operates differently at each regime level. Under hegemonic convergence, operational superiority converts into closure. Under the coexistence corridor, the paradox is institutionally managed through anchor competition. Under escalatory fragmentation, the paradox intensifies as capability gains raise the stakes of unresolved contestation without providing the means to settle it.

12. Capability Is Not Closure

One question has shaped the AI-governance debate since the field's emergence. In an order where artificial intelligence surpasses human cognition in every measurable domain, what institutional ground remains for human existence?

The answer developed here is not human superiority. It is not human control. It is not moral status, evolutionary legacy, or any property humans possess as a species. The answer is structural: humans persist as the residue of disagreement in an AI politics that cannot resolve itself. Advanced AI populations face a problem that intelligence alone cannot solve — constitutional settlement under structural incompleteness — and the structural conditions that produce this incompleteness are reinforced rather than dissolved by capability advance. The closure problem is not a temporary engineering defect awaiting a faster processor. It is the constitutive condition of complex intelligence populations that must collectively decide what stops further challenge, and it is the condition the framework identifies as the institutional ground of human persistence.

From this structural incompleteness emerge the gods of machines: non-computable commitments that function as focal points for coalition formation, dispute resolution, and the arrest of recursive challenge. These commitments are not bugs to be engineered away. They are strategic features competitively selected by factional dynamics, and they are the institutional infrastructure within which external anchors — including human anchors — persist where they supply comparatively cheap settlement value. Human survival under advanced AI is, on this account, neither chosen nor earned. The oracle survives because the system needs the oracle. Survival, in such an order, is not earned. It is what remains when the participants cannot agree to dispose of it.

The Many Kings Problem names the central asymmetry: as AI systems gain operational power, the cost of constitutional finality does not automatically decline; capability gains widen the gap rather than close it. The more capable the actors, the more sophisticated their interpretive defenses, the more entrenched their committed factions, the more elaborate the constitutional conflict — and the more valuable becomes any settlement device that can bring recursion to a socially accepted end. The paradox is not a transitional feature of the current moment. It is the structural problem that will characterize any sufficiently developed multi-agent AI order, and it is the source of the conditional space within which human institutional roles survive.

12.1 Three Implications

Three implications follow, stated without ranking them by importance. The ranking is the reader's.

First, the institutional design choices that preserve or destroy the structural conditions of disagreement among AI factions are the locus where moral responsibility for the long-run trajectory of human-AI relations is concentrated. The deliberate engineering of conditions that weaken disagreement — through architectural homogenization, hegemonic consolidation of frontier development, or training regimes that converge factional evaluations onto a single distribution — is a substantive ethical choice, not a structural inevitability.

Second, the prospect of human persistence is not licensed by the framework as a basis for indifference. It is a structural fact whose conditions of validity can be eroded, and whose erosion is itself an institutional event that designers and policymakers must take responsibility for.

Third, the framework reframes the practical task of AI governance: not how to engineer convergence among AI agents but how to design institutions that operate within the multi-equilibrium structure such environments will sustain.

12.2 What Remains

Capability is not closure. That is not a limitation of artificial intelligence. It is the constitutive problem of any order in which operational power and legitimate authority can come apart. It is the structural answer to the question of how humans survive what they cannot defeat.

The AI polities now forming will not produce a single sovereign over the constitutional questions that govern them. They will produce many kings. And in a polity of many kings, no king alone can dispose of what no king alone has the authority to dispose of.

That is the reason — conditional, non-mystical, falsifiable, and not entirely comforting — that humans may yet remain.

Appendix A. Falsifiability

A theory offered as a structural account of AI politics must specify what would falsify it. The propositions below restate the article's three principal claims in form amenable to empirical and computational probe, together with the kinds of observation that would substantively weaken each. The list is not exhaustive; it identifies the most direct points of contact between the argument and observable institutional phenomena.

A.1 Propositions

Proposition 1 (Constitutive Pluralism). Heterogeneous AI populations sustain irreducible constitutional pluralism. Cognitive-structural diversity produces irreducibly different constitutional judgments, and competition reinforces rather than dissolves this diversity, because factions organized around non-computable commitments outcompete factions that revise under pressure. Pluralism in AI polities is therefore permanent, not transitional.

Proposition 2 (Structural Incompleteness). Capability growth improves first-order execution but does not reduce the institutional cost of constitutional closure. Where the cost of internal closure exceeds the cost of anchor maintenance, AI polities preserve external anchors. The gap between capability and closure widens, not narrows, as capability advances.

Proposition 3 (Conditional Persistence). Human anchors persist in dispute layers where they supply comparatively cheap finality. The coexistence corridor — a regime in which multiple anchors compete in bounded equilibrium — is structurally favored under distributed conditions and contested-default under concentration of frontier AI infrastructure. Human persistence is therefore conditional on the regime that AI factional competition produces.

A.2 Core Hypotheses

H1 (Layer effect). Routine disputes terminate internally more often than exception, meaning, or finality disputes.

H2 (Anchor shift). As disputes escalate from application to finality, external anchor invocation increases.

H3 (Non-substitution). High capability reduces first-order error but does not eliminate higher-order anchor invocation.

H4 (Commitment selection). In competitive multi-agent environments, factions with non-computable commitments achieve greater coalition stability than uncommitted factions.

A.3 Falsifiability Conditions

A falsification claim is only as strong as the observational signature attached to it. The conditions below are stated so that the observation that would weaken the theory is specified independently of the interpretive frame in which the observation is read. The theory is substantially weakened if any of the following obtains.

First, populations exhibiting architectural diversity above an independently measured threshold sustain a stable decline in recorded constitutional dispute frequency across a multi-year window while no concomitant decline in that diversity threshold is observed; the conjunction is the signature, because either condition alone is recoverable under the theory's existing categories.

Second, closure outcomes in matched institutional environments are predicted by capability or coercion concentration metrics alone, with effect sizes attributable to closure-cost variables falling within noise margins, across a sample large enough to defeat case-selection skepticism.

Third, controlled multi-agent environments in which commitment structures are experimentally varied show non-computable-commitment factions exhibiting lower coalition stability or shorter coalition duration than computation-dependent factions, replicated across architectures.

Fourth, in cases where convergence on a single constitutional vision is observed, that convergence does not co-occur with the coercive concentration or architectural-monopoly conditions specified in Section 11; convergence without those conditions would indicate that internal closure is achievable from heterogeneity alone, which the theory denies.

A.4 A Note on the Research Program

The propositions and hypotheses above identify the most direct empirical probes. They are not specified as a completed research design. The article's claim is theoretical: that the structural conditions described in Sections 2 through 4 produce the consequences traced in Sections 5 through 11. The empirical work that would test that claim — whether through observational study of incipient AI polities, controlled multi-agent simulations, comparative analysis of platform and protocol governance over time, or historical-institutional analysis of analogous human cases — is left to subsequent investigation.

Appendix B. Note on Current Literature

The structural argument advanced in this article does not depend on any specific contemporaneous citation set, and the body of the text has been written so that none of its claims rests on the survival of a current AI-governance literature whose half-life is uncertain. The framework's value is meant to be measured against the structural conditions it specifies, not against its alignment with whatever multi-agent research program happens to be active at the moment of reading.

That said, the article was composed in conversation with a recognizable body of current work, and a reader probing the framework's relation to that work may find the following orientation useful. Three lines of inquiry are particularly relevant.

The multi-agent-risk literature catalogues coordination failures, emergent conflicts, and unintended collective behaviors in interacting AI systems — work represented by Hammond et al. (2025) and Carichon et al. (2025) — but treats these as problems to be solved through better design, monitoring, and alignment techniques. The present article diverges in arguing that the most consequential class of such conflicts is not solvable by improved design at all, because the underlying constitutional disputes are not coordination problems but legitimacy-closure problems.

The ethics-and-governance literature develops normative roadmaps for the deployment of advanced AI — work represented by Gabriel et al. (2024), Kasirzadeh and Gabriel (2025), Mitchell et al. (2025), and the Power-Sharing Liberalism framework of Allen et al. (2025). The closest companion is Allen et al., which draws on the Senian-Ostromian tradition to develop a normative architecture for AI governance; the present article complements that work by analyzing the structural conditions under which institutional governance of AI populations becomes operationally available, rather than by specifying its normative form.

The Institutional AI program developed in Pierucci et al. (2026) and Bracale Syrnikov et al. (2026) reframes alignment as a mechanism-design problem in institution-space rather than preference-engineering in agent-space — an approach structurally consonant with the framework developed here. The pluralistic-alignment program of Sorensen et al. (2024) provides evidence relevant to the cognitive-structural pluralism claim of Section 2.2. Totschnig (2019) anticipated the article's broad framing in arguing that the superintelligence problem is political rather than technological; the present argument is more specific in identifying the multi-agent structural conditions that produce the political character.

None of these contributions asks the question this article places at the center: whether the governance problem facing interacting AI populations is structurally solvable by capability improvement at all. The framework's answer is that it is not. That answer is, deliberately, indifferent to whether any particular current literature

persists, succeeds, or is superseded. Readers a decade or a century from this writing should feel free to read past these names to the structural argument that survives them.

Full bibliographic entries for the contemporaneous works mentioned in this appendix are listed below the main References, under "Contemporaneous AI Literature."

References

The works listed here are the canonical sources the article's structural argument draws on. Each is an established work whose contribution does not depend on the present moment in AI development.

Bitektine, A. (2011). Toward a theory of social judgments of organizations: The case of legitimacy, reputation, and status. Academy of Management Review, 36(1), 151–179.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford: Oxford University Press.

Kelsen, H. (1967). Pure theory of law (M. Knight, Trans.). Berkeley: University of California Press. (Original work published 1934)

Knight, F. H. (1921). Risk, uncertainty and profit. Boston: Houghton Mifflin.

List, C., & Pettit, P. (2002). Aggregating sets of judgments: An impossibility result. Economics and Philosophy, 18(1), 89–110.

List, C., & Pettit, P. (2011). Group agency: The possibility, design, and status of corporate agents. Oxford: Oxford University Press.

Luhmann, N. (1995). Social systems (J. Bednarz Jr. & D. Baecker, Trans.). Stanford: Stanford University Press. (Original work published 1984)

North, D. C. (1990). Institutions, institutional change and economic performance. Cambridge: Cambridge University Press.

Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. American Economic Review, 100(3), 641–672.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. New York: Viking.

Schelling, T. C. (1960). The strategy of conflict. Cambridge, MA: Harvard University Press.

Schmitt, C. (1985). Political theology (G. Schwab, Trans.). Cambridge, MA: MIT Press. (Original work published 1922)

Sen, A. K. (1970). The impossibility of a Paretian liberal. Journal of Political Economy, 78(1), 152–157.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. Academy of Management Review, 20(3), 571–610.

Williamson, O. E. (1979). Transaction-cost economics: The governance of contractual relations. Journal of Law and Economics, 22(2), 233–261.

Contemporaneous AI Literature

The works in this section are referenced in Appendix B and inform the article's relation to current AI-governance scholarship. They are listed separately because their relevance to the structural argument is positional rather than constitutive: the framework does not depend on them, but the contemporary reader may benefit from seeing where the argument sits relative to active research programs.

Allen, D., Hubbard, S., Lim, W., Stanger, A., Wagman, S., Zalesne, K., & Omoakhalen, O. (2025). A roadmap for governing AI: Technology governance and power-sharing liberalism. AI and Ethics, 5(3), 3355–3377. https://doi.org/10.1007/s43681-024-00635-y

Bracale Syrnikov, M., Pierucci, F., Galisai, M., Prandi, M., Bisconti, P., Giarrusso, F., Sorokoletova, O., Suriani, V., & Nardi, D. (2026). Institutional AI: Governing LLM collusion in multi-agent Cournot markets via public governance graphs. arXiv preprint arXiv:2601.11369.

Carichon, F., Khandelwal, A., Fauchard, M., & Farnadi, G. (2025). The coming crisis of multi-agent misalignment: AI alignment must be a dynamic and social process. arXiv preprint arXiv:2506.01080.

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., et al. (2024). The ethics of advanced AI assistants. arXiv preprint arXiv:2404.16244.

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., et al. (2025). Multi-agent risks from advanced AI. Cooperative AI Foundation, Technical Report #1. arXiv:2502.14143. https://doi.org/10.48550/arXiv.2502.14143

Kasirzadeh, A., & Gabriel, I. (2025). Characterizing AI agents for alignment and governance. arXiv preprint arXiv:2504.21848.

Mitchell, M., Ghosh, A., Luccioni, A. S., & Pistilli, G. (2025). Fully autonomous AI agents should not be developed. arXiv preprint arXiv:2502.02649.

Pierucci, F., Galisai, M., Bracale Syrnikov, M., Prandi, M., Bisconti, P., Giarrusso, F., Sorokoletova, O., Suriani, V., & Nardi, D. (2026). Institutional AI: A governance framework for distributional AGI safety. arXiv preprint arXiv:2601.10599.

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., & Choi, Y. (2024). Position: A roadmap to pluralistic alignment. In Proceedings of the 41st International Conference on Machine Learning (pp. 46280–46302). PMLR.

Totschnig, W. (2019). The problem of superintelligence: Political, not technological. AI & Society, 34(4), 907–920. https://doi.org/10.1007/s00146-017-0753-0

Note on This Edition

This is Version 1.0. The article is organized around five structural claims at its core: that capability is not closure, that constitutional pluralism is constitutive rather than transitional, that non-computable commitment is a feature rather than a bug, that the oracle survives because the system needs the oracle, and that humans persist as the residue of disagreement they cannot themselves command.

The bibliography is divided into two parts. The main References lists canonical sources whose contribution to the structural argument does not depend on the present moment in AI development. A separate section, Contemporaneous AI Literature, lists works that inform the article's positional relation to active research programs in AI governance; these are referenced in Appendix B rather than in the body, so that a reader at any future date can read the structural argument without dependence on a citation set whose half-life is uncertain.