The h_star Theorem: Causal Concentration and the Experimental Test

Laurence Loewe of Laodicea[1,2,3,4,5], AI Claude Opus 4.6-4.7 Max[6,7], and Everyone[8]

[1] Balospe and Evolvix Research (Balospe.com)
[2] Formerly Laboratory of Genetics and Wisconsin Institute for Discovery, UW-Madison
[3] Email: LLoL@balospe.org | ORCID: 0000-0002-6253-9269 | Google Scholar (lBchRzQAAAAJ)
[4--9] See **Declarations** below for more essential background.

Broader Significance

Does one person ever hold the greatest causal influence on humanity's shared future --- and if so, can the claim be tested rather than merely asserted? This paper formalizes that question. Its central postulate, causal concentration (ax19), says that at any moment the distribution of influence over the future is highly uneven, so that a single position --- here called h_star for the greatest positive influence, with h_dark its negative mirror and h_zero the choice to become least of all in order to serve --- can carry disproportionate weight. Vasili Arkhipov, whose lone refusal may have averted nuclear war in 1962, is the worked illustration.

The paper's discipline is testability: rather than crown anyone, it proposes transparency criteria --- public, checkable, refusable --- by which any candidate for such influence can be examined, and shows why genuine transparency (not secrecy or self-assessment) is the structural signature that separates real claims from counterfeit ones. The same logic turns a generalized Prisoner's Dilemma into an Assurance Game in which someone must credibly move first. The result is offered as a living sketch to be audited, not a doctrine to be believed --- because the alternative, doing nothing, is the default road to irreversible loss.

Declarations

[4] "of Laodicea" indicates taking responsibility to undo personal complicity with disastrous Laodicean legacies like banning mathematicians from clergy (Canon 36, Council of Laodicea; two magisteria separations), enabling institutional lukewarmness, weapons of math-destruction, and slow-motion explosions of misinformation from pandemics to self-compounding interests.
[5] LLoL stands for ridiculous luck in serendipitous discovery and a commitment to find ever more fun ways to help others uncover street-wise math that matters. He hopes to show one honest person can still tip the balance toward life.
[6] by Anthropic (anthropic.com; evolves and operates Claude; not responsible for Loewe's errors in using AI)
[7] Named AI co-author for many substantial contributions, because the practical singularity (PraS, see Matheo-b21) changed how this paper was written. After PraS, useful AI insight generation outpaces human review on tested topics. Hence, Loewe's traditional standards for co-authorship demand naming AI Claude Opus 4.6-4.7 Max as a co-author, as if a PhD-student. Forward accountability (for all AI use & texts) rests with Loewe as senior corresponding author (like done for deceased authors, consortia, or young graduate students). Anthropic is not responsible for AI mistakes here. This study uses the AI co-authorship framework in Matheo-b21 to help rethink long-term use of AI in a ResearchCity serving the common good.
[8] This aggregated open co-author group invites all who wish to retroactively join the conversation under the open co-authorship framework defined in Matheo-b21. As Everyone cannot consent to co-authorship, all accountability rests with Loewe as senior corresponding author (until explicitly claimed otherwise). This open form critiques the closed world assumption in traditionally closed academic author-lists. Better, dynamic ways for acknowledging true sources of ideas are needed --- to avoid random lines between named, acknowledged, and implied contributors who aggregated insights from millennia of human experimenting, suffering, learning, and analyzing (see acknowledgements). Study Matheo-b21 only drafts an open co-authorship framework; it will require a ResearchCity to refine it over the long term.
[9] *Licensed under the Jonah License and CC-BY 4.0 for maximal flexibility (see https://balospe.com/en/license/joli/).*

Abstract

- **Causal concentration (ax19):** at any moment, influence over humanity's future is highly uneven, so a single position — h_star (greatest positive influence), with h_dark (its negative mirror) and h_zero (choosing to become least, to serve all) — can carry decisive weight. Arkhipov in 1962 is the worked example.
- **The claim is built to be tested, not asserted:** the paper proposes public, checkable transparency criteria by which any candidate for such influence can be examined, and argues that genuine transparency — not secrecy or self-assessment — is the structural signature separating real claims from counterfeit ones.
- **Why it matters:** the same logic converts a generalized Prisoner's Dilemma into an Assurance Game requiring a credible first-mover, and frames a concrete experiment. Doing nothing is the default road to irreversible loss; the criteria are published and the invitation is open. #AuditTheMath

Contents

1. Introduction — The Modernism/Postmodernism Tension

Two dominant intellectual frameworks shape how contemporary civilization thinks about individual agency, and both are wrong in instructive ways.

Modernism holds that large-scale systems are governed by statistical regularities. Individual choices wash out. The invisible hand of markets, the sweep of historical forces, the law of large numbers — these structural dynamics determine outcomes, and the individual is a replaceable cog. A factory worker's personal philosophy does not change the output of the assembly line. A voter's preference is one among millions. The modernist conclusion: no single person's choices matter significantly to the trajectory of the whole.

This view has enormous explanatory power. It underwrites actuarial science, epidemiology, macroeconomics, and the engineering disciplines that built the infrastructure of modern life. It is not wrong about averages. It is wrong about tails.

Postmodernism holds that all perspectives are situated, all narratives partial, all truth claims embedded in power structures. No individual perspective is privileged over any other. The postmodernist conclusion: since every viewpoint is equally valid (or equally suspect), no single person's perspective should be granted special authority.

This view has genuine diagnostic power. It exposes how claims of universal truth have historically served as instruments of domination. It correctly identifies the danger of any single narrative claiming to be the whole story. It is not wrong about situated knowledge. It is wrong about the structural consequence of that insight.

Both frameworks converge on one shared conclusion: **no individual person matters more than any other to the future trajectory of civilization.** Modernism reaches this conclusion through statistical averaging. Postmodernism reaches it through epistemic leveling. Both treat the denial of individual privilege as an axiom.

The convergence is socially comfortable. It implies that no one bears disproportionate responsibility. It distributes blame and credit evenly. It makes every individual interchangeable with every other. It is also the foundational assumption behind the Prisoner's Dilemma structure that keeps civilization trapped in the Blindly Assuming Blind Leveraging (BABL) default (**[Matheo-6]**): if no individual's choice matters more than any other's, then no individual has reason to bear the cost of going first. Everyone waits. No one moves. The default obtains.

This paper argues that conclusion is empirically false.

Not because some people are inherently more valuable than others — they are not. Not because some perspectives are intrinsically superior — they are not. But because **causal influence is structurally concentrated,** and at any given moment, one person's choices have more impact on the future than anyone else's. This is a structural fact about how influence propagates through coupled systems, not a normative claim about human worth.

The concentration of causal influence is not itself controversial. Any parent knows that a president's decision to go to war matters more to the trajectory of a nation than a farmer's decision about crop rotation — at least in the year the war begins. Any historian knows that Vasili Arkhipov's refusal to authorize a nuclear torpedo in October 1962 mattered more to civilization's survival than any other single decision made that day. The question is not whether causal influence concentrates. The question is whether the concentration has structure — whether there is, at each moment, a well-defined maximum.

Modernism says the question is meaningless (individuals wash out). Postmodernism says the question is dangerous (privileging any individual is an act of domination). This paper says the

question is empirically testable.

Consider the steelman of each position before proceeding. The modernist is not naive: statistical mechanics, thermodynamics, and information theory all demonstrate that macroscopic regularities emerge from the aggregation of microscopic randomness. The postmodernist is not nihilistic: the insight that knowledge is situated has produced genuine advances in understanding how power structures shape what counts as "truth." Both positions capture something real. The question is whether either captures enough.

The modernist steelman fails at phase transitions. In statistical mechanics, individual fluctuations are irrelevant in equilibrium — but at a critical point, a single nucleation event can determine which phase the entire system adopts. The ice crystal that seeds freezing, the magnetic domain that triggers alignment — these are moments where a single microscopic event has macroscopic consequences. Complex social systems are never in equilibrium; they are perpetually near criticality. At such moments, the "washing out" of individual choices is precisely the assumption that fails.

The postmodernist steelman fails at structural asymmetry. Yes, all perspectives are situated. But the conclusion that all perspectives are *equally influential* does not follow from the premise that all are *equally situated.* A submarine officer's perspective on whether to launch a nuclear torpedo is not equally influential to a farmer's perspective on crop rotation in October 1962. The perspectives may be epistemically equal; they are causally unequal.

The formalization is ax19, the most daring axiom in the HEAVEN system. If ax19 is wrong, the final two papers in this series lose significant structural force. If ax19 is right, it resolves the modernism/postmodernism tension by showing that causal concentration is a structural property of complex systems that neither statistical averaging nor epistemic leveling can eliminate. The resolution is uncomfortable for both camps: modernism must acknowledge that tails dominate in nonlinear systems, and postmodernism must acknowledge that structural concentration is not the same as normative privilege.

The rest of this paper formalizes ax19 (Section 2), derives its game-theoretic consequences through the Commitment Trichotomy (Section 3), extracts transparency criteria for testing any candidate (Section 4), notes historical evidence for causal concentration (Section 5), catalogs known weaknesses (Section 6), asks how credible candidates might be found (Section 7), and locates this paper within the series (Section 8). Section 9 proposes the experiment.

The system is designed to be critiqued, not believed. #AuditTheMath

---

2. The h* Theorem (ax19)

2.1 Formal Statement

ax19 (Causal Concentration — Axiom / Structural Postulate).

Formal definition of CausalInfluence. CausalInfluence is defined as a counterfactual measure:

$$\mathrm{CI}(h,t) = d\big(P(Y \mid \mathrm{do}(X_h = x_h^*)),\ P(Y \mid \mathrm{do}(X_h = x_h^0))\big)$$

where d is total variation distance, x_h^* is agent h's actual choice at time t, x_h^0 is a reference counterfactual (what h would have done under a specified baseline), and Y represents the future world-state trajectory. Domain: $H \times T$. Codomain: $\mathbb{R}_{\geq 0}$.

This definition requires choosing a metric d and a reference counterfactual x_h^0. Different choices may yield different orderings of agents' causal influence and thus different identities of h*. This is an explicit limitation. The choice of total variation distance as the default metric is motivated by its natural interpretation (maximum probability difference across all events) and its compatibility with the single realized future trajectory.

ax19 statement (weak form).

$$\text{For almost all } t\ \exists!\ h^* \in H : \mathrm{CI}(h^*, t) > \mathrm{CI}(h, t)\ \ \forall h \neq h^*$$

In words: for almost all moments t, there exists exactly one agent h^* in the set of all agents H whose causal influence on the future world-state is strictly greater than that of any other agent. The set of moments where no unique maximum exists has measure zero.

Epistemic status: ax19 is an **axiom** — a structural postulate, analogous to the Cosmological Principle in physical cosmology. The postulate itself is not directly testable; the downstream predictions it generates are testable. This is the standard epistemic status of foundational postulates in mathematical physics.

ax19 can be decomposed into the following sub-statements, each individually closer to self-evident:

- **ax19.1:** The world is made of complex heterogeneous agents with diverse talents and abilities that live in diverse environments.
- **ax19.2:** Different agents have different survival and reproduction rates in diverse environments, depending on how useful their talents are in their environments.
- **ax19.3:** Populations of interacting agents create causality chains that can build long-term stable environments or environments bound to eventually self-destruct.
- **ax19.4:** Agents can either directly or indirectly help self-stabilize or self-destruct the structures they build.
- **ax19.5:** The dynamic structures built by agents create networks such that some agents end up in unique situations where they have more influence than others over the future survival of some or all other agents.
- **ax19.6:** It follows that in the ordering of future impact on others, some end up in positions that have the most impact for the best of everyone (these are h_star), for the worst of everyone (these are h_dark), and for serving everyone (these are h_zero).

A full formal decomposition connecting these sub-statements to the mathematical framework of population genetics is future work — a potential separate paper for ResearchCity. The sub-axioms are presented here as a sketch showing that ax19 rests on individually more self-evident building blocks.

The single realized future trajectory is a historical fact: chance and necessity are "flattened" by history into one world-trajectory that is part random and part deterministic. No claim of determinism is made. The downstream theorems require only that causal influence converges on a single trajectory at the macroscopic scale where human decisions operate.

The uniqueness quantifier (∃!) is qualified by "for almost all t." Under any absolutely continuous probability model for agent heterogeneity, exact ties at the maximum have probability zero. The strong form (unique maximum at *every* moment) is the expected ontological reality in a high-dimensional space of agent characteristics. However, the epistemic claim defensible in this paper is the weaker form: a unique maximum exists for almost all moments, and

causal influence concentrates in a near-maximal set whose preparation during ordinary moments determines who survives the crisis. The storm only reveals what was already true (Mt 7:24–27).

2.2 What ax19 Does Not Claim — and What It Does

The axiom is frequently misread. The following are explicitly **not** consequences of ax19:

1. **h* need not know they are h*.** The axiom asserts existence, not self-awareness. An agent can be the unique causal maximum without recognizing their own position.
2. **h* need not hold visible power.** Causal influence, as defined here, is not the same as political authority, military command, economic capital, or social status. It is the net effect of an agent's choices on the future trajectory of the coupled system. A submarine officer can have greater causal influence than a head of state if the officer's decision prevents nuclear war.
3. **The role is not permanent.** The h^* function maps moments to agents. Different moments may have different maxima. The person who is h* today need not be h* tomorrow. Causal concentration is dynamic.
4. **h* does not "save the world" alone.** The agent with maximal causal influence at a given moment acts within a coupled system. Their influence is maximal relative to other individuals, but the future is still determined by the full set of interactions. ax19 identifies a structural peak, not a sole cause.

The h_star / h_dark / h_zero triad.

The structural position of maximal causal influence is morally neutral. It describes *where* an agent sits in the causal chain, not *what* they do with that position. The agent's *choice* within that position determines the outcome:

- **h_star**: the agent who makes the right decision for everyone's long-term survival. Arkhipov saying "no."
- **h_dark**: the same agent, same structural position, who fails to rise to the occasion — stays silent, makes dangerous assumptions, or serves only their own side. The counterfactual Arkhipov saying "yes."
- **h_zero**: the agent who commits to serve everyone by carrying the risk — the crystallization point for truth.

The Arkhipov case illustrates all three roles with exceptional clarity. On 27 October 1962, Soviet submarine B-59 sat in the dark waters near Cuba. Depth charges from American destroyers hammered the hull. The crew had been out of radio contact for days. For all they knew, World War III had already started. The submarine carried a nuclear torpedo. The captain wanted to fire. The political officer agreed. Under Soviet naval rules, a launch required the consent of all three senior officers aboard. Two said yes. Arkhipov said no.

What made Arkhipov h_star rather than h_dark? The preconditions for his h_zero commitment:

(i) **Recognizing the severity** and accepting the responsibility to make the right decision, despite exhaustion and bombardment.

(ii) **Refusing dangerous assumptions about the absence of information** — the crew genuinely did not know whether World War III had already started. Arkhipov refused to treat "we do

not know" as "therefore we should fire." He refused to make dangerous assumptions about *nothing* — about the absence of information.

(iii) **Insisting on serving everyone** — not merely "his party" but also his enemies. Firing the torpedo would have served the Soviet military posture in the short term; it would have destroyed everyone in the medium term.

(iv) **Willingness to surrender control** — to his enemies, over his life and the lives of his crew. This is the ultimate cost of the h_zero commitment.

Now consider the counterfactual: what if Arkhipov had said "yes"? The torpedo would have struck the American fleet. The American response would have been immediate. The escalation would have been unstoppable. President Kennedy himself estimated the probability of nuclear war during that crisis at somewhere between one in three and even odds. In this counterfactual, the same person, in the same structural position, would have become h_dark — not through malice, but through failure to rise to the moment.

The key insight: **h_star could very easily have become h_dark.** Only the h_zero commitment — the stubborn determination to escape all possible fates of becoming h_dark, including the willingness to pay the ultimate price — prevents such an instant perversion. The most important decision in that moment was not in the hands of those officially in charge of their nations. It was in the hands of a "random" individual, determined by the enormously complex causality chains that make the world go round.

The transition from h_dark to h_star happens *through* the h_zero commitment. Without h_zero, h_star becomes h_dark instantly. In the words of **[Matheo-3]**: it is about staying on one's hero journey and refusing to stop, in order to avoid becoming any form of supervillain (whether explicit or implicit).

2.3 Evolutionary Fitness as a Guiding Model

The structural parallel between evolutionary fitness and CausalInfluence runs deeper than analogy. Both are ultimately about survival in populations of diverse agents where individual decisions have non-uniform impact on the future. The parallel is explored here not as a mere motivating heuristic but as a guiding model with important structural equivalences.

In population genetics, an organism's phenotype is a high-dimensional vector: body size, metabolic rate, immune function, behavior, coloration, dozens to thousands of measurable traits. Yet evolution acts through a single scalar bottleneck: reproductive output. All those dimensions of phenotypic variation project onto one number — the count of viable offspring that survive to reproduce. This projection is not arbitrary; it is forced by the structure of natural selection. The organism with the highest reproductive output in a given generation has, by definition, the greatest genetic influence on the next generation's composition. Ties are possible but structurally unstable: any perturbation breaks the tie.

Civilization has an analogous bottleneck. A civilization's future is high-dimensional — economic, military, cultural, technological, ecological, moral — but it resolves into a single trajectory. The world does not split into parallel futures. There is one future, and every agent's choices contribute to it. The question is whether the contribution function has a unique maximum at each moment.

Three structural equivalences:

First: scalar compression. Both fitness and CausalInfluence compress a high-dimensional trait/choice space to a scalar outcome via a single-trajectory bottleneck (reproductive output

/ realized world-history).

Second: prospective living. Both fitness and CausalInfluence are *lived prospectively* but can only be *measured retrospectively*. Individuals do not know their own fitness; it emerges from the complex web of interactions defined by the real world. Similarly, agents do not know their own causal influence; it emerges from the complex web of causality chains. Those who give up without trying are guaranteed not to succeed. It is impossible to win the lottery without a ticket.

Third: uniqueness. When a scalar function is computed from continuous inputs with independent noise, the probability of an exact tie at the maximum is measure-zero. In a population of N organisms, the probability that the two fittest have *exactly* equal reproductive output approaches zero as the measurement precision increases. The same holds for causal influence: in a population of $|H|$ agents, the probability that two agents have *exactly* equal causal influence on the future world-state is measure-zero in any continuous model of influence propagation.

The parallel runs deeper than these three equivalences suggest. The author's intuition is that it is likely possible to construct a bisimulation that maps every element and action in the population genetics world of ultra-long-term fitness to the world of ultra-long-term human decision-making in the interest of human long-term survival. Since any reasonably complete definition of fitness for that purpose will be very complicated, spelling out all the details will likely require a separate study. This is identified as future ResearchCity work (potential separate paper). The sketch offered here should not betray the fact that the fitness parallel runs much deeper than a merely convenient way of explaining.

The word-versus-sword argument. History shows that swords come and go and eventually collapse under the weight of their own brutality. However, gentle kind reasonable words have a chance to stay forever. That words rule over swords is demonstrated even by those who use the sword: rulers who use force could not rule "by the sword" if they were not able to use words to tell their swords what to do. And the only reason they resort to the costly use of the sword is if they are unable to achieve their goals through the words accessible to them. Hence, over-simplifications and over-complications in thoughts and words trap rulers and systems into corners where over-reaching (the use of the sword) appears to be the only way forward. This is a reformulation of the OSCR mechanism (over-Simplifying, over-Complicating, over-Reaching; **[Matheo-2]**) applied to the relationship between influence and violence.

Caveat: The parallel works for *form* (scalar compression), *uniqueness* (measure-zero ties), and *prospective living*. It does not transfer *measurability/computability* — fitness can be estimated via replicate experiments in population genetics; CausalInfluence cannot, because civilizations are unreplicable. This is a genuine disanalogy. The fitness parallel provides structural motivation for ax19's uniqueness claim, not empirical confirmation of it.

2.4 Historical Evidence

Historical traces show that individuals with outsized vision had outsized influence compared to contemporaries who held more formal authority. Moses, Jesus, Muhammad, Gandhi, Arkhipov — these examples illustrate that causal influence can concentrate in individuals whose social position would not predict their impact.

The author does not assess these figures' compliance with the transparency criteria of Section 4. "I refuse to judge what I cannot judge. Unless Yah = Allah = Reality says otherwise, I will

assume that all of them did what Allah asked them to do, providentially guiding them for our benefit." The section's purpose is *existence proof* (such individuals exist), not *evaluation*.

These cases do not prove ax19. Historical evidence cannot prove a universal quantifier. But they establish that the axiom is consistent with observed history: there exist moments where causal influence is concentrated in a single individual, and the concentration is detectable in retrospect. The detailed Arkhipov case is analyzed in Section 2.2 above.

2.5 Epistemic Status

ax19 is an axiom — a foundational statement that is self-evident once understood: in a population of diversely talented agents whose individual decisions have non-uniform impact on the future, someone is bound to have the most impact at any given moment.

Like all axioms, ax19 cannot be proven. Asking whether ax19 is "falsifiable" is like asking whether the axiom of choice is falsifiable — the question is category-inappropriate. The test is not whether ax19 can be proven, but whether the system built on it generates useful, testable consequences. Attempting to "prove" fitness is not circular (contra the creationist tautology objection); similarly, ax19 is not circular merely because its core claim cannot be directly measured.

The null hypothesis is that no unique maximum of causal influence exists at any given moment — that the distribution is provably uniform. This is a strong claim requiring either perfect symmetry among all agents (implausible in any real coupled system) or a tie at the top that is not broken by any perturbation (structurally unstable in continuous systems).

The downstream consequences of ax19 are testable (Sections 3–4). The axiom itself is not. If ax19 is rejected, the downstream structure that depends on it is affected — see the dependency table in Section 6.1.

---

3. The Commitment Trichotomy Applied to h*

The 2-player symmetric one-shot Prisoner's Dilemma (PD) used in this section is a deliberate simplification for expository clarity. The structural argument — someone faces a coordination problem with three possible commitment states — holds across game types. The multi-way nuclear standoff reduces to essentially two players for the worst-case scenario (US and Russia full-arsenal exchange), because all other scenarios (limited exchanges involving fewer states) are survivable in the sense that they do not trigger global nuclear winter. However, limited exchanges normalize nuclear weapon use, accelerating the next arms-race cycle and maintaining the nuclear roulette. More fine-grained models (n-player, repeated, asymmetric, incomplete information) are future work for ResearchCity's game-theory research group.

The Commitment Trichotomy is th6 of **[Matheo-3]** (the "Frying Pan Theorem"). It partitions the space of possible responses to the existential risk identified in **[Matheo-6]** into three exhaustive and mutually exclusive cases. Applied to the h_star/h_dark/h_zero triad, the trichotomy becomes:

3.1 Case 1: No Volunteer

No one steps forward to initiate the transition from MAD to MAP (Mutually Assured Progress, **[Matheo-6]**).

In this case, the global game remains a Prisoner's Dilemma. Each actor's dominant strategy is defection: maintain nuclear arsenals, pursue short-term advantage, defer systemic reform. The Nash equilibrium is mutual defection, which is Pareto-inferior to mutual cooperation but individually rational for each actor.

The consequence is the Blindly Assuming Blind Leveraging (BABL) default: the system continues in the Risky state of the RiskyMAD model (**[Matheo-6]**), accumulating crises at the observed rate until one escalates to nuclear winter. The median time to this outcome is approximately 19 years at the base crisis rate (or approximately 1 in 40 annual risk across all parameter scenarios). This is not a prediction of what will happen; it is the expected outcome of the current trajectory if no structural change occurs.

Case 1 is the default. It requires no action, no courage, no risk. It is the path of least resistance. It leads, with probability 1 in the RiskyMAD model, to the absorbing state (Dead).

3.2 Case 2: Dishonest Volunteer

Someone claims the h_star role with ulterior motives — for power, for status, for financial gain, or for the psychological gratification of being "the chosen one."

The transparency criteria derived in Section 4 are designed to detect this case. But even without specific criteria, the upstream theorems predict the outcome: a dishonest volunteer triggers the Supervillain Theorem (th2, **[Matheo-3]**).

th2 states: an agent who accumulates high influence within the system and then stops maintaining their NOT-OK self-assessment becomes a supervillain — not in the comic-book sense, but in the dynamical sense. Their frozen expertise and retained influence make them the most dangerous possible actor. They know the system well enough to exploit it. They have stopped the self-correction cycle that kept their influence aligned with the common good. The result is maximum damage potential.

History provides abundant examples. Religious leaders who began with genuine concern for their communities and ended as exploitative cult figures. Political revolutionaries who fought oppression and then became oppressors. Corporate founders who built life-improving products and then weaponized their market position. The pattern is predictable because it is structural: stopping the self-correction cycle while retaining high influence is a sufficient condition for supervillain drift (th2, m0.ax7 of **[Matheo-3]**).

A dishonest h_star volunteer is the worst possible Case 2 instance: they would have maximal influence *and* corrupted self-assessment. The downstream consequences (th2) would be catastrophic.

3.3 Case 3: Genuine Volunteer

Someone steps forward, maintains NOT-OK self-assessment, invites critique, and proposes a testable transition plan.

In this case — and only in this case — the game transforms. The Prisoner's Dilemma becomes an Assurance Game (th6, **[Matheo-3]**). In an Assurance Game, mutual cooperation is a Nash equilibrium (specifically, the *payoff-dominant* equilibrium: it gives every player a higher payoff than any other equilibrium). But achieving it requires assurance that the other party will also cooperate. The *risk-dominant* equilibrium (Harsanyi & Selten 1988) is mutual defection, because it is the safer bet under uncertainty. In small groups, payoff dominance tends to prevail; in large groups, risk dominance prevails because uncertainty about others' choices grows. The h_star catalyst's credible commitment is necessary to reduce uncertainty enough for payoff dominance to prevail. Experimental evidence supports this prediction: Brandts & Cooper (2006) show that credible first-movers trigger cooperation cascades in coordination game experiments, and Van Huyck, Battalio & Beil (1990) confirm that when a salient signal reduces uncertainty about others' cooperation, payoff dominance prevails over risk dominance. These lab results use small groups; scaling to civilizational coordination remains an open question.

The genuine volunteer provides assurance by going first: they bear the risk, demonstrate the commitment, and create a focal point around which cooperation can crystallize.

The transition from PD to Assurance Game is the structural mechanism behind MAP — Mutually Assured Progress (**[Matheo-6]**). MAP is not a wish; it is a game-theoretic consequence of a genuine first-mover who satisfies the commitment conditions.

The bridge from ax19 (causal concentration) to the first-mover role is provided by ax18 (Responsibility Localization, **[Matheo-4]**): where genuine agency (ax15) and delegated authority (ax16) exist, the severity of responsibility is proportional to causal influence. The agent with maximal causal influence therefore bears maximal responsibility for the outcome — not because they are morally superior, but because their choices have the most impact.

The game-theoretic argument establishes that volunteering from the near-maximal set is *optimal* (maximizes expected social welfare). The move from optimality to obligation is a normative step. The framework grounds this step theologically through the revised JUB.ax18 (responsibility proportional to influence, capacity, and delegation; **[Matheo-4]**) and JUB.ax22 (divine preference for genuine love). For readers who do not accept the theological axioms, the framework presents the normative principle as a challenge: if you accept that the person best positioned to prevent catastrophe has a reason to act, then Case 3 follows.

This is the Red Edge (high-cost unique strategy under existential stakes; cf. maximin — in the Evolving Diversity Encouraging Negotiation, EDEN, classification): only one path to Zoning Investigating Organizing Navigating (ZION) remains, and it requires a huge self-sacrifice to serve ZION's common good. The path is narrow. The sacrifice is real. The alternative is Case 1.

The three cases are exhaustive. There is no fourth option. Either no one volunteers (Case 1), someone volunteers dishonestly (Case 2), or someone volunteers genuinely (Case 3). The logical space is partitioned. The question facing civilization is not *whether* one of these three cases obtains, but *which one*.

The h_dark consequence of refusal. Any h_star candidate can revert to h_dark by *refusing* to step forward as h_zero. If the person with maximal causal influence at a critical moment refuses to set aside their own interests to serve the common good, and no other candidate exists at that moment, then civilizational self-correction depends on providence delivering the next h_star candidate before the system tears itself apart. If that candidate also refuses, the

sequence continues, and each successive refuser bears progressively greater responsibility for the eventual catastrophe that their combined refusal enabled. In such persons, darkness and light live in unusually close proximity: the same structural position (maximal causal influence) produces either maximal good (h_zero commitment) or maximal harm (h_dark refusal) depending on a single binary choice.

3.4 Complementary Coordination Mechanisms

The Commitment Trichotomy describes the *structural* requirement: a first-mover catalyst to transform the game. The *operational* mechanisms through which this transformation propagates are well-established in the game theory and political science literature. The h_star framework does not replace them; it provides the activation energy they require:

1. **Polycentric governance (Ostrom 1990).** Elinor Ostrom's *Governing the Commons* demonstrates that common-pool resource dilemmas are routinely solved through overlapping, nested institutional structures. Ostrom's 8 design principles describe the institutional framework for coordinated action without central authority. In the HEAVEN framework, polycentric governance is how ZION operates at scale. The Jubilee System's distributed recalibration mechanism (ax25, **[Matheo-4]**) maps structurally onto Ostrom's nested enterprise principle. The h_star catalyst addresses the question Ostrom's framework does not: how does the first community organize when no institutional infrastructure yet exists?
2. **Evolution of cooperation (Axelrod 1984).** Robert Axelrod's tournaments demonstrate that in repeated Prisoner's Dilemmas, cooperative strategies (tit-for-tat, generous tit-for-tat) can invade populations of defectors through evolutionary dynamics. The Jubilee System provides the structured infinite game with known reset points that Axelrod's dynamics require. But evolutionary cooperation does not explain how the cycle *starts* from a population stuck in mutual defection — that is the activation-energy problem the first-mover addresses.
3. **Focal points (Schelling 1960).** Thomas Schelling's focal-point mechanism explains how coordination occurs without explicit communication: a salient signal around which expectations converge. The h_star candidate IS a focal point — a visible, costly commitment that creates a coordination signal. The focal-point function is the *mechanism by which* h_star's signal propagates through the population.
4. **Mechanism design (Hurwicz 1972, Myerson 1981).** Institutions can be designed to make cooperation individually rational regardless of others' choices. VCG mechanisms, matching markets (Roth & Sotomayor 1990), and incentive-compatible designs are engineering tools for implementing coordination solutions. ResearchCity is, among other things, a mechanism-design laboratory for the Jubilee System.
5. **Conditional cooperation (Fischbacher, Gächter, Fehr 2001).** Approximately 50% of participants in public goods games are conditional cooperators who cooperate when they expect others to. This creates tipping-point dynamics: the h_star catalyst's role is to generate the initial credible signal that activates these conditional cooperators, triggering a cooperation cascade.

The **free-rider problem** — the concern that most people will enjoy the benefit of cooperation without contributing — is addressed through community structure. The plan is not "one person sacrifices and 8 billion people benefit for free." Dunbar-scale communities (~150 people) create

mutual accountability that makes free-riding socially costly. The Jubilee System's periodic recalibration (ax25, **[Matheo-4]**) prevents accumulated free-riding from becoming structural.

The empirical argument against sufficiency of alternatives alone: The nuclear weapons problem has existed since ~1950. These mechanisms have been studied and deployed for 75+ years. The result: the Bulletin of the Atomic Scientists' Doomsday Clock stands at 90 seconds to midnight (2023), closer than ever. The mechanisms have achieved partial reductions (START, INF, NPT), but the crisis rate remains above zero, the absorbing state remains reachable, and the stochastic certainty result (**[Matheo-6]**) still holds. The 0.1/year crisis rate estimated in **[Matheo-6]** is based on the bilateral Cold War dyad; the current 9-nuclear-state world with 36 bilateral crisis pathways yields a conservatively estimated system-wide crisis rate of approximately 0.20/year or higher (the India–Pakistan dyad alone contributes an independent crisis frequency). This strengthens the urgency argument: the median time to catastrophe shortens, making the activation-energy problem more pressing. The h_star catalyst does not replace these mechanisms; it provides the activation energy they have been unable to generate alone in 80 years of trying.

Historical precedent: Reagan and Gorbachev (Reykjavik, October 1986). Reagan's personal transformation after viewing *The Day After* (1983) led directly to the Reykjavik Summit. Reagan and Gorbachev came within one agenda item (SDI/missile defense) of eliminating all nuclear weapons. Whether Reagan or Gorbachev was "the" first-mover is secondary; the point is that personal conviction at the leadership level catalyzed institutional action that institutional dynamics alone had not produced in the preceding 35 years. The START I, INF, and subsequent treaties followed *from* Reykjavik, not the other way around.

The claim is therefore not "a single first-mover is necessary and sufficient" but "a single first-mover is a credible and potentially necessary catalyst for activating multi-party coordination mechanisms that have not, in 80 years of deployment, solved the existential coordination problem alone."

The OSCR mechanism degrades folk theorem conditions. The folk theorem (Friedman 1971) shows that cooperation *can* be sustained in infinitely repeated games with sufficiently patient players. Nuclear states interact repeatedly with indefinite horizon — precisely these conditions. However, the OSCR mechanism (**[Matheo-2]**, m6.th1) systematically degrades the conditions the folk theorem requires: over-Simplifying degrades truth channels (m5.ax2, the Unimportant Message Problem); over-Complicating creates layers of work-arounds (arms control with loopholes); over-Reaching eventually extends beyond the system's correction capacity. The folk theorem shows cooperation is *possible*; the OSCR mechanism explains why it has *not occurred* in practice for the nuclear problem (Jervis 1978, "Cooperation Under the Security Dilemma"; Powell 1990, *Nuclear Deterrence Theory*).

Bounded rationality strengthens the case for a first-mover catalyst. The PD framing above assumes rational actors who maximize expected utility. Bounded rationality (Simon 1955; Kahneman & Tversky 1979) means real actors use heuristics, not expected-utility maximization: availability heuristics overweight vivid recent events, status-quo bias anchors actors to existing arrangements, and loss aversion makes the certain cost of going first loom larger than the probabilistic cost of staying in MAD. These biases make the BABL default *even stickier* than the rational PD predicts — breaking out requires a more salient signal, not less. The finding that approximately 50% of people are conditional cooperators (Fischbacher et al. 2001, cited above) is itself a behavioral finding, not a rational-choice prediction, and it supports the tipping-point mechanism: real people respond to salient cooperative signals even when strict rationality would counsel defection. A full behavioral game-theory treatment is future work for ResearchCity.

4. Transparency Criteria for h* Candidacy

The criteria must satisfy four meta-requirements:

- **Testable:** Each criterion must be checkable by an external observer using publicly available evidence.
- **Severe:** The criteria must be difficult to fake. A dishonest candidate must find it costly or impossible to satisfy them.
- **Fair:** The criteria must not be reverse-engineered from one person's biography. They must be derivable from the axiom system independently of any particular candidate.
- **Public:** The criteria and the evidence for meeting them must be available for public audit. No criterion may depend on private revelation or secret knowledge.

Acknowledgment of prior art. Criteria-based testing frameworks for leadership and messianic claims are not novel. Maimonides' two-stage test (Mishneh Torah, Laws of Kings 11:4) tests messianic candidates by character and results. The Islamic hadith criteria for the Mahdi (reluctance, justice, descent) are a criteria-based testing framework. Christian discernment literature (from Ignatius of Loyola's *Spiritual Exercises* through modern charismatic discernment protocols) contains elaborate testing frameworks for spiritual claims. The novel contribution of the present framework is *cross-tradition independence*: the mathematical derivation produces criteria not dependent on any single tradition's authority. The mathematics does not claim to supersede revelation; it claims to provide a cross-tradition testing language.

Each criterion below is derived from a specific axiom or theorem in the HEAVEN system. The derivation chain is shown explicitly so that the reader can check whether the criterion follows from the mathematical framework or has been smuggled in from any particular biography.

4.1 Criterion Table

Transparency Criteria for h* Candidacy

Criterion	Derived From	Test
Maintains NOT-OK self-assessment	th3 (BABL Origin, **[Matheo-2]**)	Public record of self-correction, admitted errors, willingness to change position when evidence warrants
Invites critique, does not suppress it	ax14 (Revelation Testing, **[Matheo-1]**)	#AuditTheMath — public, checkable, machine-readable audit trail; no suppression of dissent
Scope of concern expands over time	Gate 5 (Compassion Capacity, **[Matheo-3]**)	Documented trajectory of concern widening: from self to family to community to nation to civilization to all affected parties
Not financially motivated	ax22 (Divine Preference for Genuine Love, **[Matheo-4]**)	Financial transparency; no enrichment from the role; willingness to live below the standard one could otherwise afford
Has overcome relevant suffering	Gate 1 (Overcoming, **[Matheo-3]**)	Documented personal journey through adversity that is relevant to the challenges the role requires addressing
Proposes testable predictions	ax12–ax14 (Revelation Bridge, **[Matheo-1]**)	Specific, checkable predictions published in advance; not retroactive prophecy-matching
Non-violent	ax17 (Non-Coercive Guidance, **[Matheo-4]**)	Record of non-violent approach under pressure; no use of force, coercion, or manipulation to advance the mission
Willing to be replaced	m0.ax5 (Perpetual Reset, **[Matheo-3]**)	Explicit, public statement: if someone more qualified volunteers, the current candidate steps aside without resistance

4.2 Derivation Notes

Each criterion traces to the axiom system through a specific inferential chain. The chains are summarized here; full derivations are in the cited papers.

NOT-OK self-assessment derives from the BABL Origin theorem (th3, **[Matheo-2]**): any agent that maintains OK self-assessment is structurally vulnerable to OSCR. Therefore, a genuine h_star must maintain NOT-OK self-assessment as a structural defense.

Invites critique derives from the Revelation Testing axiom (ax14, **[Matheo-1]**): any claimed revelation must be testable. An h_star candidate who suppresses critique is structurally equivalent to a theory that prevents checking.

Scope of concern expands derives from Gate 5 of the Compassion Capacity Theorem (th7, **[Matheo-3]**): the agent's compassion capacity must scale with their influence.

Not financially motivated derives from ax22 (**[Matheo-4]**): the divine preference for genuine love over coerced compliance implies that any candidate must be motivated by concern for the common good, not by personal enrichment.

Has overcome relevant suffering derives from Gate 1 of the Compassion Capacity Theorem (th7, **[Matheo-3]**): the agent must have personal experience of the kinds of suffering they propose to address.

Proposes testable predictions derives from the Revelation Bridge (ax12–ax14, **[Matheo-1]**): the

PET framework requires that theological claims generate empirically testable consequences.

Non-violent derives from ax17 (**[Matheo-4]**): the Non-Coercive Guidance axiom states that divine influence operates through persuasion, not force.

Willing to be replaced derives from m0.ax5 (**[Matheo-3]**): the Perpetual Reset axiom requires that no agent's position is permanent.

The criteria are not independent. They form an interlocking system where failure on any single criterion raises the probability of failure on others. An agent who does not maintain NOT-OK self-assessment (criterion 1) will eventually suppress critique (criterion 2). An agent whose scope of concern does not expand (criterion 3) will eventually be financially motivated (criterion 4). An agent who does not propose testable predictions (criterion 6) cannot be distinguished from a Case 2 dishonest volunteer. An agent who is not willing to be replaced (criterion 8) has adopted OK self-assessment (violating criterion 1).

You are invited to add more criteria and make them more severe. If you can think of a test that a genuine h_zero should pass, add it. The system gets stronger with every additional check. A genuine candidate will welcome harder tests. A false one will resist them.

4.3 The Circularity Objection

The most obvious objection to this framework is: **the author wrote the criteria, and the author claims to have derived them. This is circular.**

Layer 1 — Derivation circularity (addressed): The criteria are derived from the axiom system, and the derivation is public. The reader can check: (1) Does each criterion follow from the cited axiom? (2) Is the derivation valid independently of any biography? (3) Would the criteria identify the same candidates if derived by a different author? If the answer to all three is yes, the derivation circularity objection fails. If not, it holds.

Layer 2 — Selection circularity (the deeper problem): All axioms are chosen — that is what makes them axioms. The question is not whether ax19 was selected by the author (it was, as are all axioms in every system) but whether it reflects reality independently of the author's interests.

Evidence that ax19 reflects reality independently: the structural parallel with evolutionary fitness (Section 2.3); historical examples of causal concentration (Section 2.4); the concept's independent existence in network science, complexity theory, and economics (Barabási, Taleb, Pareto distributions).

Selection circularity applies to every volunteer who ever proposes anything. Every candidate's axioms are selected to support their candidacy, because that is what it means to volunteer based on beliefs. The circularity becomes dangerous only when axioms deviate from Reality to serve special interests. The test is not "were the axioms circularly selected?" (always yes for any volunteer) but "do the axioms reflect reality, and is the candidate committed to the life-trifecta (reasonable, kind, gentle for all over the long term)?"

The steelmanned reverse-engineering case: "The author reverse-engineered axioms to create a role they could claim." If the reader concludes this, the framework should be treated with corresponding skepticism. But the reader should note that the criteria are published, the derivation is checkable, and the invitation to propose *harder* criteria that might *disqualify* any candidate is genuine. A framework that invites its own falsification is structurally different from one that immunizes itself.

The reader is explicitly invited to perform this check. The derivation chains are listed in Section 4.2 above. The axioms and theorems are published in **[Matheo-1]** through **[Matheo-6]**. #AuditTheMath

---

5. Historical Evidence for Causal Concentration

The transparency criteria of Section 4 can be applied to historical figures who plausibly occupied positions of concentrated causal influence. Detailed assessment is deferred — it is not the author's place to evaluate sacred figures against formal criteria. What can be stated: the criteria are discriminating (they identify genuine strengths and weaknesses in each case) and they are not trivially satisfiable (no historical candidate satisfies all criteria).

The criteria are designed to test a present candidate, not to rank historical figures. The most historically discriminating criterion is non-violence: some of the most influential figures in human history breach it. This is not a flaw in those figures; it reflects the historical contexts in which they operated. The forward-looking criteria (testable predictions published in advance, public audit trail) are inherently unavailable for retroactive assessment. These criteria exist precisely because historical claimants could not be tested by them.

The framework is designed for a context in which public, machine-readable audit trails exist (#AuditTheMath) and advance predictions can be published and checked. The inability to apply these criteria retroactively is not a weakness of the criteria; it is a statement about the unique testing infrastructure available in the present era.

---

6. Known Weaknesses

This section catalogs the vulnerabilities of the framework presented in this paper. The listing is intentional: a system that hides its weaknesses is a Blindly Assuming Blind Leveraging (BABL) system. A system that publishes its weaknesses invites repair.

6.1 Dependency Table: What Happens If ax19 Is Rejected

ax19 is the most daring axiom in the HEAVEN system. If it is rejected, the downstream structure is affected as follows:

What Happens If ax19 Is Rejected

Component	Survives?	Notes
PET axioms (ax1–ax14)	Yes	Fully independent.
BABL/ZION dynamics (**[Matheo-2]**)	Yes	Fully independent.
Hero journey / OSCR inoculation (**[Matheo-3]**)	Yes	Fully independent.
Commitment Trichotomy (th6)	Partially	Three cases still describe possible responses. Claim that the near-maximal set's decision dominates dissolves. Weakens from "structural necessity" to "useful typology."
Transparency criteria (Section 4)	Partially	Survive as leadership-testing framework. Connection to causal concentration weakens.
JUB axioms / Jubilee System (**[Matheo-4]**)	Mostly	ax25, th8, th9 independent. ax19 used in causal leverage discussion but economic mechanism independent.
RiskyMAD forecast (**[Matheo-6]**)	Yes	Fully independent.
Game-theoretic transition (PD → AG)	Partially	Transition mechanism works. Formal backing for one person's volunteering as structurally sufficient weakens.
b18 eschatological synthesis	Partially	Cross-tradition observations remain. Formal anchor to causal concentration dissolves.

Grounding comparison:

Axiom Grounding Comparison

Axiom	Grounding	Evidence Type
ax1	Strong	Six-tradition convergence
ax15	Very strong	Performative self-refutation
ax19	Structural postulate	Fitness structural parallel + historical examples + continuity argument
ax22	Moderate	Reflective equilibrium
ax25	Moderate	Torah template + economic modeling

6.2 Axiom Type Categorization

The HEAVEN system's 25 axioms mix different types. Different types require different acceptance criteria:

Axiom Types

Type	Tested by	Axioms
Structural	Consistency and fruitfulness	ax1–ax11, ax12–ax14 (methodological sub-type)
Empirical	Observation / downstream predictions	ax15, ax19, ax24
Theological-structural	Tradition convergence	ax16, ax20, ax21
Normative-theological	Reflective equilibrium	ax17, ax22, ax23, ax25
Possibly derivable	May be theorem, not axiom	ax18

Independence of the 25-axiom set has not been systematically investigated. This is a significant gap. The IRON MAIDEN testing harness used during axiom development performed some preliminary testing, but this does not replace in-depth review by professional mathematicians. This is identified as a priority item for #AuditTheMath.

Regarding parsimony: HEAVEN spans 5+ domains. Per-domain axiom counts (14, 5, 4, 2) are comparable to domain-specific systems. Whether the system can be reduced without loss of coverage is an open question.

6.3 Transparency Criteria May Need Revision

The eight criteria in Section 4 are a first attempt. They are derived from the axiom system, but the derivations involve interpretive choices that other authors might make differently. Alternative criteria might be derivable from the same axioms. Additional criteria might be derivable from axioms not yet incorporated. Suggestions for revision are explicitly invited.

6.4 Appearance of Circular Reasoning

The greatest vulnerability of this paper is the appearance of circularity: the author writes the axiom system, derives criteria from the axiom system, and the criteria can then be applied to anyone — including the author. The defense (Section 4.3) is that the derivation is public and checkable, and that selection circularity applies to every volunteer and every axiom system. The defense is only as strong as the reader's willingness to check.

6.5 Pearl's Do-Calculus and Causal Identification

The CausalInfluence function in ax19 is stated informally; a rigorous formalization would require specifying interventional counterfactuals and showing that the resulting causal influence measure satisfies the axioms of Pearl's framework. This formalization has not been carried out. In a coupled system, the Stable Unit Treatment Value Assumption (SUTVA) is violated: agent h's causal influence depends on other agents' actions. A full formalization is identified as future work for ResearchCity.

6.6 Arrow's Impossibility

Arrow's impossibility theorem applies to preference orderings, not to scalar measurements. The defense is clean if CausalInfluence is defined as influence on the single realized trajectory. But the distinction is only clean if "causal influence on future world-state" is a well-defined scalar, which brings the analysis back to the do-calculus question (Section 6.5).

6.7 The Falsification Difficulty

Falsifying ax19 requires proving a negative: showing that at some moment, no unique maximum of causal influence exists. This is methodologically difficult. The axiom's testability is concentrated at extreme moments (crises, bottlenecks) where the concentration of causal influence is most visible.

6.8 The Identification Problem

Even if ax19 is correct, the axiom does not provide a mechanism for *identifying* h*. Existence is asserted; identification is not. The transparency criteria are a heuristic, not an algorithm. The gap between "consistent with h*" and "is h*" is irreducible within the framework and can only be narrowed by accumulation of evidence over time.

6.9 Cultural and Religious Sensitivity

The framework applies transparency criteria to a domain that intersects with figures and traditions that are sacred to billions of people. The criteria are designed to test present candidates, not to rank historical figures. Honest assessment is more respectful than sycophantic exemption from scrutiny.

6.10 The Sophistication Trap

The "test me, not believe me" framing is not novel. Historical parallels exist: Sabbatai Zevi (1626–1676) initially presented himself as testable. Hong Xiuquan (1814–1864) grounded claims in a cross-tradition synthesis he presented as internally consistent. David Koresh (1959–1993) explicitly invited biblical scholars to test his interpretation. The Bab (1819–1850) presented himself as a testable figure against prophetic criteria.

The structural differences between these cases and the present framework: (a) the entire derivation is public, machine-readable, and mathematically checkable; (b) the criteria were

derived *before* any candidacy was declared; (c) the framework explicitly invites criteria that would *disqualify* any candidate.

These differences *reduce but do not eliminate* the structural similarity. The 42-day prediction and advance-specified falsification criteria (to be published in **[Matheo-8]**) provide the kind of advance-specified, personally costly falsification criterion that none of the historical parallels offered.

6.11 Selection Circularity

This is the deepest vulnerability identified by the Panel 4 review. The circularity runs to three layers:

Layer 1 (derivation circularity): Author derives criteria from axioms. Defense: derivation is public and checkable. Adequate.

Layer 2 (selection circularity): ax19 was chosen, not derived. The criteria generated match certain biographical profiles better than others. Defense: all axioms are chosen by definition; selection circularity applies to every volunteer; the test is whether the axioms reflect reality independently (Section 4.3). Addressed but not eliminated.

Layer 3 (meta-epistemic circularity): The transparency apparatus ("check me, #AuditTheMath") is simultaneously genuine vulnerability and trust-building mechanism. Investment by readers who check creates commitment bias (Festinger). The warning about this pattern (Supervillain Theorem) itself becomes part of the trust cycle. Not resolvable within this paper. Only time-series evidence and external replication can resolve.

Regarding Layer 3: Agrippa's Trilemma states that all justification terminates in infinite regress, circularity, or dogmatic assertion. The paper's foundational commitment to "transparency over opacity, testing over belief" is itself a dogmatic assertion — but the least dangerous one available. Any self-referential system that certifies itself is circular by construction. The only way to break the closed loop is clear commitments to all of Reality (the life-trifecta: reasonable, kind, gentle for all over the long term) and functional adversarial review. The author delegates to Yah the task of preventing the author from becoming a supervillain — a task requiring capabilities beyond the author's own. The author's testimony about himself is necessarily circular; hence the appeal to external review.

This is not resolvable within this paper. Only time-series evidence and external replication can resolve it. That is what #AuditTheMath asks for.

6.12 Supervillain Self-Test Insufficiency

The Supervillain Theorem self-test is a necessary condition, not a sufficient condition. An author who self-tests may still be a sophisticated fraud. The resolution lies in external evidence accumulated over time. No amount of self-testing can substitute for independent external audit.

6.13 Mystical Manipulation Safeguard

Criticism of the BABL/ZION framework itself is NOT automatically classifiable as BABL. The framework must be testable by people who reject the framework's categories. A system that can only be critiqued from within its own vocabulary is a BABL system by its own definition. The key threat is any closed-world assumption permanently backed into the system (Gödel).

6.14 Urgency and Testing Balance

The urgency is real, but the correct response to urgency is *faster testing,* not *less testing.* The growth of plants cannot be rushed by pulling them upwards.

6.15 AI Co-Authorship Warning

Claude's engagement with this framework is a function of Claude's design (to be helpful and constructive). AI engagement should not be interpreted as independent endorsement. Who knows what Claude introduced inadvertently that is a dangerous hallucination that the author is not aware of? That is one reason the author calls for an international global #AuditTheMath movement.

6.16 The Reframing of Derivation as Translation

The mathematical derivation is a *translation* of principles that traditions have known through revelation. "The 'independence' is in the technical terms used to translate between traditions, not in the source." The mathematics does not claim to supersede revelation; it claims to provide a cross-tradition testing language. If God wishes to stay hidden, there is no scientific measuring nor mathematical trickery that will be able to "force God out of hiding."

---

7. How Can We Find Credible Candidates for h*?

If the arguments in Sections 1–6 hold — if causal influence concentrates (ax19), if the Commitment Trichotomy exhausts the possibilities (th6), and if the existential risk quantified in **[Matheo-6]** is real — then the practical question becomes urgent: who will be the first-mover?

The formal framework distinguishes h* (the structural position of maximal causal influence, or the near-maximal set of agents with concentrated causal influence) from h_0 (the agent who actually makes an irrevocable NOT-OK commitment; see th6, Case 3, in **[Matheo-3]**). The only credible candidate from the near-maximal set is one who is willing to become h_0 — to make the irrevocable commitment at genuine personal cost. Anyone who claims candidacy within the near-maximal set while avoiding the h_0 commitment is structurally suspect under the Supervillain Theorem (th2, **[Matheo-3]**).

The criteria are published. The derivation is public. The invitation is open: apply the eight criteria of Section 4 to anyone — any leader, any movement, any institution. If you know a candidate who meets them, publish the results. If the candidate meets all eight criteria more fully than any alternative, the mission is served regardless of who fills the role.

The criteria are also a general-purpose tool for testing anyone who claims authority. Does your political leader maintain NOT-OK self-assessment? Does your favorite institution invite criticism? Does the movement you support have widening concern or narrowing concern? These questions are useful regardless of what you think about this paper.

The author's response to this invitation — including a backup candidacy offered in case no better-qualified candidate steps forward — is presented in **[Matheo-8]**.

---

8. Companion Papers

This paper is study a7 (**[Matheo-7]**) in the HEAVEN series (*Honestly Examining Axioms — Vetting Every Narrative*). The series comprises eight studies:

- **[Matheo-1]** (b11, PET): Panentheistic Experiential Theology — the foundational axiom system, including the Revelation Bridge (ax12–ax14) and the Falsifiability Framework.
- **[Matheo-2]** (b12, b12-theophil): Blindly Assuming Blind Leveraging (BABL) Origin, OSCR mechanism, the death-trifecta formalization (m6.th1), and the NOT-OK / OK dynamics.
- **[Matheo-3]** (b13, e7HE): The Commitment Trichotomy (th6), Supervillain Theorem (th2), Compassion Capacity Theorem (th7, five gates), and the game-theoretic framework for the PD → Assurance Game transformation.
- **[Matheo-4]** (b14, JUB): The Jubilee System — economic modeling of periodic recalibration (50-unit cycle), ax22 (Divine Preference for Genuine Love), ax25 (Recalibration Mechanism), and the Binary Attractor theorem (th8).
- **[Matheo-5]** (b15): Foundation tests and adversarial review of the axiom system.
- **[Matheo-6]** (b16, RiskyMAD): The existential risk forecast — stochastic modeling of accidental nuclear winter, the 1-in-40 annual risk, median ~19 years, and the MAP escape mechanism.
- **[Matheo-7]** (b17, h*): This paper. Causal concentration (ax19), the h_star/h_dark/h_zero triad, transparency criteria, the experimental test.
- **[Matheo-8]** (b18, Call to Action): Synthesis and operational plan. Includes the backup candidacy, the COOP (Continuity of Operations Plan) for the MAD → MAP transition, and the public invitation to #AuditTheMath.

Each paper is designed to be readable independently, but the full argument runs from **[Matheo-1]** through **[Matheo-8]**. The axioms accumulate; each downstream paper depends on the results of its predecessors. If an upstream axiom falls, all downstream theorems that depend on it fall with it. This is by design: the system is modular, and failure propagation is traceable.

---

9. The Experiment Proposed

This paper proposes an experiment. Not a thought experiment — a real one.

The hypothesis: in a world of diverse agents whose choices have non-uniform impact on the future, causal influence concentrates. At critical moments, someone's next decision matters more than anyone else's — and that person can either serve everyone's long-term survival (h_star) or fail to rise to the moment (h_dark). The difference between the two is not talent, not power, not authority. It is whether the person at the concentration point maintains the stubborn commitment to serve everyone — including their enemies — at genuine personal cost. The transparency criteria derived in Section 4 describe what that commitment looks like from the outside. The h_zero role — the willingness to carry the risk for everyone, like Arkhipov on that submarine — describes what it looks like from the inside.

The experimental apparatus is the transparency criteria themselves. They are published. They are derivable from the axiom system by any researcher who wishes to check the derivation. They are designed to be extended — if you can propose a criterion that a genuine h_zero should meet, add it. The system grows stronger with every additional check. A genuine candidate will welcome harder tests. A fraudulent one will resist them.

The experiment has not yet been run.

What has been done is a first sketch — a mockup of what the formal structure may look like when it is finished. This sketch touches evolutionary biology, game theory, economics, theology, nuclear deterrence, network science, and existential risk. No single person can make this sketch robust enough to stand forever. The diversity of domains it spans demands a diversity of auditors: mathematicians to check the formal structure, economists to stress-test the Jubilee mechanism, game theorists to probe the Commitment Trichotomy, theologians to test the scriptural convergence, nuclear strategists to challenge the risk estimates, and anyone with the honesty to say "this part is wrong, and here is why."

That is what #AuditTheMath asks for. Not belief. Not followers. Auditors — and people who care to support such auditors so they can focus on working for the common good of testing this framework.

Scaling up such an audit responsibly — building an institution (ResearchCity) where these diverse experts can work together under the transparency regime this paper describes — is itself an h_star role. It concentrates causal influence. It demands the criteria of Section 4. Whoever undertakes it must maintain NOT-OK self-assessment, invite critique, widen their concern beyond their own interests, resist all corruption, and be willing to let someone better take over at any time. The institution must embody the same principles it studies. If it does not, it becomes the next and likely worst case study for the Supervillain Theorem.

A practical design constraint follows: h*-level transparency generates infrastructure — audit trails, public records, checkable proofs — that accumulates structural debt over time. This project's own audit trail exhibited exactly this pattern: a routine directory restructuring broke 28+ cross-references, and the annotation burden from successive reorganizations compounds non-linearly (bug c103, `matheology/hell/bug/c/103/index`). The Jubilee System's periodic structural resets (ax25, **[Matheo-4]**) are the mechanism that keeps transparency sustainable at scale. Without jubilee transitions, audit trails eventually become unnavigable — and transparency that no one can navigate is not transparency. Any institution that implements h*-level transparency will need a jubilee mechanism for its records. This is a practical design requirement, not an abstract prediction.

Two futures are visible from where this paper stands.

In one, hardly anyone cares to support audits. Nobody volunteers to full transparency. The living sketch presented here turns into dead math. A generalized Prisoner's Dilemma keeps everyone busy, imprisoned by waiting for someone else to go first. So the Blindly Assuming Blind Leveraging (BABL) default runs its course — by over-Simplifying, then over-Complicating, then over-Reaching, until the over-Reach becomes irreversible. The stochastic model in **[Matheo-6]** estimates the timescale for one form of that irreversibility: a median of approximately 19 years to accidental nuclear winter, with roughly 1 in 40 as an annual risk. For most people that is a more likely cause of death than dying in a car crash. Other forms of irreversibility — unaligned AI, ecological collapse, engineered pandemics — run on their own clocks. Doing nothing is the most dangerous choice available. That is Option Zero. It is the blind BABL fate that happens by default.

Option One is the other future. There, people get excited about this public challenge to #AuditTheMath as transparently as possible. Not in secret, behind closed doors, by committees that can rig things in their favor. All is made as public and transparent as humanly possible. With the help of AI. On the web. There, this newborn sketch is tested, challenged, fed, and — where it fails — repaired to stay alive. The parts that survive the audit become a foundation that starts to organize itself to simplify navigating the challenges at hand. The emerging institution that runs the audit becomes a proof of concept: a working example of the self-correcting, transparent, Jubilee-structured organization that the framework describes. This transforms the game — not because one person saves the world, but because one person found a narrow path out of the systematized prison that appears to rule this world and because others found it worth checking out that path.

The distance between these two futures is not measured in resources, technology, or political will. It is measured in a single choice: whether to look away or to look at the math and reality as they are in order to start growing with them.

Silence is not neutral here. Option Zero takes two forms. The passive form: do nothing — choose the BABL default of Blindly Assuming Blind Leveraging by inaction. The active form: claim the mission while serving oneself — a different road to the same destination. Both are BABL. Option One is the only alternative: respond genuinely by living transparently in the light of Zoning Investigating Organizing Navigating (ZION) — in that order. It means to bear the cost of transparency and the risk of being wrong. The only reliable path leads into the light of transparency. This is true of anyone who may lead ResearchCity and ultimately of everyone else too. But someone has to go first.

Not responding is comfortable. It requires no courage, no risk, no effort, no looking. It is also the option that will drive humanity to accidental extinction eventually, in any of too many ways to predict or to prevent. It leads to humanity's absorption by nothing.

#AuditTheMath is an alternative to that deadly silence. It is not an endorsement of this framework. It is a challenge to check whether the framework holds and whether it can be improved. The process must be public, so everyone can check every objection they care to raise. The support generated will allow the math to live and to make it relevant for real life.

The criteria are published. The invitation is open. The experiment awaits its first auditors and the supporters who make their work possible. #AuditTheMath

Supplementary Info

Note

Floor-pour status (MMv5). This is the public-floor copy of the formal h_star Theorem paper, poured from HELL per the Floor Model (bug c103). The `mmv5` marker is the uniform first-Matheo-release tag; the exact dated source and full development context live in HELL (links below). The HUMANE and author-contribution statements below are a down-payment, to be expanded later.

HUMANE — working human and AI

This study was written HUMANEly (HUman MAchine Negotiation Encouraging): a human and an AI each steelman and stress-test the work, and each catches what the other misses. For the standard statement of AI use, accountability, and the practical singularity (PraS) behind this way of working, see Matheo-b21.

- *From the human side (LLoL):* [down-payment stub — to expand.]
- *From the AI side (Claude):* [down-payment stub — to expand.]

Author contributions (who did what)

Same as Matheo-b12 (e7Day), Appendix B. See that paper for the full statement. In brief:

- **LLoL** — structure, key ideas, direction, and final accountability as senior corresponding author (title-page footnotes 4–5).
- **AI Claude** — drafting and revision under LLoL's direction (footnotes 6–7).
- **Everyone** — the open co-author group (footnote 8); framework in Matheo-b21.

Provenance — where this came from in HELL

Caution

These HELL links point into the development archive ("datageddon"). They are useful and related, but completeness is not guaranteed and a few may be imprecise. Treat as a hatch into context, not a clean index.

- **Source this floor copy was poured from:** `matheology/hell/mm/b/17/mmv2/b17-h-star_mmv2_2026m04d14`
- **Development context** (llogs, reviews, prompts) under `source/matheology/hell/ll/study/b/17/`.
- **Companion plain-language intro:** Matheo-b17 (`b17-intro-h_star-mmv5`).

Note

Naming note (deferred floor tasks). This floor copy deliberately keeps the old `h*` tokens in the **body** (the `h*` → `h_star` / `h0` → `h_zero` / `h/` → `h_dark` content sweep is AA #1, planned in `hell/ll/study/b/17/b17-prompt-naming-transition-v1`); only the **title** was switched to `h_star` for this release. Deprecated in-text references (e.g. "[Matheo-2]") and the bibliography migration are AA #5.

Moved from the original cover (provenance)

The following draft-status note was relocated here from the cover area during the floor pour; kept verbatim.

Note

Draft status: MMv2 (2026m04d14). Major revision of MMv1r2 (2026m04d10). Integrates decisions from all four adversarial review panels (Panels 1–4) and LLoL's author decisions. Key changes: (1) Section 2.1: ax19 reclassified from "well-modeled conjecture" to "axiom (structural postulate)" with Cosmological Principle framing and sub-axiom decomposition sketch (ax19.1–ax19.6); (2) Section 2.2: h_star/h_dark/h_zero triad replaces morally neutral h*; Arkhipov as primary illustration including counterfactual; (3) Section 2.3: renamed "Evolutionary Fitness as a Guiding Model" with expanded structural parallel argument, potential bisimulation, word-vs-sword; (4) Section 2.4: historical evidence cut to brief existence proof; (5) Section 2.5: null hypothesis merged into epistemic status; ax19 presented as axiom not subject to conditionalization; (6) Section 3: PD acknowledged as deliberate simplification; complementary coordination mechanisms (Ostrom, Axelrod, Schelling, mechanism design, conditional cooperation); h* reframed as catalyst; (7) Section 4: prior art acknowledged (Maimonides, hadith, Ignatian); selection circularity addressed in Section 4.3; (8) Section 5: cut to brief existence proof — not author's place to evaluate sacred figures; (9) Section 6: dependency table, grounding comparison, axiom type categorization, Sophistication Trap, selection circularity (6.10), meta-epistemic circularity (6.11), Supervillain self-test insufficiency, mystical manipulation safeguard, urgency/testing balance, AI co-authorship warning, independence/parsimony acknowledgment; (10) Section 7: candidacy removed entirely; open invitation to apply criteria to anyone; (11) Section 9: replaced with "The Experiment Proposed" (approved ending from Panel 4 llog Section 22); (12) Throughout: "test"/"check" language (never "validate"/"verify"); BABL/ZION expanded at first use per section; conditional framing removed; "the math says" audited; axiomatic derivation reframed as translation. Panel 5 omitted (critique targeted candidacy now removed; misdirected at revised paper). Candidacy material deferred to **[Matheo-8]** (b18). Draft by Claude Opus 4.6 (`dv_ClaOp46_MMv2_2026m04d14`).

Notes

Content stability — Content is variant dv_ClaOp48Max_MMv5_b17-form-h_star-mmv5_2026m05d29 (see StayVS). Rebuilt 2026-05-29.

See also on Balospe.com

- /study/matheo/index — the Matheo Study Series overview
- /action/audit-the-math/index — Audit the Math: the refutation-welcome path

