WHY AUTONOMOUS AI CANNOT MAKE MAN IRRELEVANT

A Proof from First Principles

Drawing on Gödel, Turing, Hayek, Arrow, Marks, and the Architecture of LLMs

John Aaron
Mikhail Golovnya

2026

Acknowledgment of AI Assistance

This monograph was developed using large language models as interpretive and drafting aids under direct human supervision. AI systems were used to assist with outlining, language refinement, and exploratory synthesis. All conceptual frameworks, arguments, simulations, interpretations, and conclusions are the responsibility of the authors. AI systems were not used as authorities or decision-makers, consistent with the principles articulated in this work.

Authorship and Responsibility

The authors assume full responsibility for the accuracy, integrity, and interpretation of the material presented. Any errors or omissions remain solely the responsibility of the authors.

Disclaimer

The views expressed in this work are those of the authors and do not necessarily reflect the views of any client, institution, or affiliated organization. This publication is intended for scholarly and educational purposes and does not constitute legal, financial, medical, or professional advice.

Publication Information

Published by
Milestone Planning and Research, Inc.

Printed in the United States of America
ISBN: [To be assigned]
Library of Congress Control Number: [To be assigned]

SERIES PREFACE

Human Relevance in the Age of Induction

This monograph is the sixth and final volume in the series on Human Relevance in the Age of Induction, published by Milestone Planning and Research, Inc. The series has developed, across five preceding volumes, a unified framework for understanding the relationship between human judgment and machine induction—and for designing institutions, organizations, and governance structures that preserve human agency rather than eroding it.

Monograph 1 examined the enterprise deployment of inductive systems: what AI is in its various operational forms, how it integrates into existing process architecture, where it creates measurable value, and how governance must be designed to maintain human authority. It argued that the strategic advantage of AI lies not in automation for its own sake but in compressing epistemic delay—the time between signal emergence and belief revision.

Monograph 2 established the foundational argument: that human relevance in an age of inductive AI is not a question of technological determinism but of decision authority, epistemic discipline, and institutional design. It introduced the Hybrid Cognition framework and its companion, Hybrid Falsification—the institutional design discipline through which doubt, evidence, and accountability are preserved against the convergence pressures that inductive systems naturally produce.

Monograph 4 argued that artificial intelligence does not threaten competitive advantage by replacing human creativity but by accelerating a convergence dynamic. Drawing on Dubins and Savage's bold-play framework and game-theoretic analysis of AI-saturated markets, it showed that timid play—stopping at semantic adequacy—is a dominated strategy.

The present volume adds to the series by addressing what the preceding volumes presupposed but did not prove: that human agency is not merely currently valuable but permanently and structurally necessary. Where the preceding volumes argued from institutional design, organizational economics, and competitive strategy, this volume argues from mathematical logic, theoretical computer science, computability theory, and political economy. The convergence of Gödel, Turing, Hayek, Arrow, and Marks—each arriving at structural limits from entirely independent starting points—constitutes the most powerful version of the proof available.

The series concludes with a constructive rather than alarmist argument. The question was never whether AI would make humans irrelevant. The structural proof in this volume establishes that this outcome is impossible. The question—the one that will determine the character of the coming decades—is whether human beings will be wise enough to maintain the institutional structures through which their necessary meta-level role is exercised. That is not a technological question. It is a human one.

ABSTRACT

This monograph argues that autonomous artificial intelligence cannot render human beings irrelevant—not as a contingent feature of current technological limitations, but as a consequence of structural constraints identified across multiple independent disciplines. Drawing on Gödel's incompleteness theorems, Turing's limits on computation, Hayek's theory of distributed and tacit knowledge, Arrow's impossibility theorem, and Marks's argument from algorithmic incompleteness, the analysis demonstrates that any system operating under conditions of bounded representation and optimization requires a meta-level agent for value specification, interpretation, and accountable commitment.

The claim advanced here is deliberately narrow and structural. It does not assert that artificial intelligence cannot surpass human performance across a wide range of domains. It asserts instead that no physically realizable system can eliminate the necessity of an external meta-level of judgment without contradiction. The necessity of human agency arises not from comparative advantage but from the logical and institutional limits governing formal systems, computation, and social aggregation.

Hayek's knowledge argument occupies a central position in this proof. The knowledge relevant to effective coordination is dispersed, local, and tacit in ways that no training corpus can internalize—and the rules governing social processes change continuously through human creativity, rendering any static model progressively less reliable. Marks's algorithmic argument runs parallel and independent: all AI is computation, all computation is algorithmic, and genuine creativity is provably non-algorithmic. The convergence of these constraints from different disciplines is not coincidental. They are each probing the same underlying structural reality.

The analysis further examines the failure modes that emerge when these structural constraints are ignored, the political economy that drives elite enthusiasm for totalizing AI deployment, and the architecture of legitimate bounded AI use. It concludes with the institutional design framework of Hybrid Cognition and Machine-Assisted Perception. The central result is that increasing AI capability does not dissolve the need for human judgment; it amplifies the consequences of its absence.

CHAPTER ONE

The Axiomatic Foundation

1.1 Scope and Method

Debates over artificial intelligence and human relevance are typically conducted through empirical extrapolation: what systems can currently do, how rapidly capabilities are improving, and what trajectories might plausibly follow. This approach, while informative, is insufficient for establishing necessity. Empirical trends can suggest likelihood; they cannot establish structural inevitability. A system that cannot currently perform a function may acquire that capability tomorrow. Empirical arguments therefore prove too little for the claim this monograph advances.

The present analysis proceeds differently. It adopts a structural approach grounded in independently established results from multiple disciplines. The aim is not to demonstrate that human relevance is likely to persist, but that it cannot be eliminated without contradiction under the conditions that define any physically realizable intelligent system.

This requires a clarification of scope. The argument does not claim that artificial intelligence cannot exceed human capability in bounded domains. It claims instead that:

> ***No system operating through bounded representation, transformation, and optimization can eliminate the necessity of a meta-level agent responsible for value specification, interpretation of meaning, and accountable commitment.***

This distinction is central. The question is not whether AI can become more capable than humans. It is whether capability alone can internalize authority. The argument that follows establishes that it cannot.

1.2 Convergent Structural Constraints

The case developed in this monograph does not rest on a single theorem or disciplinary tradition. It emerges from the convergence of constraints identified independently in mathematical logic (Gödel), theoretical computer science (Turing), computability theory (Marks), economic theory and epistemology (Hayek and Polanyi), and social choice theory (Arrow). Each tradition establishes a limitation of a different kind. Taken individually, each constrains a particular domain. Taken together, they define a structural envelope within which any intelligent system must operate.

The convergence of these independently established results is not accidental. It reflects the fact that they are each probing, from different angles, the same underlying structural reality: that any system confined to its own representational structure cannot simultaneously be complete, self-validating, fully informed about the world it is meant to coordinate, and coherently authorized to specify its own objectives. Each theorem closes one possible escape route for the autonomous AI thesis.

1.3 The Five Axioms

The argument proceeds using five axioms, each grounded in established theory but interpreted structurally rather than merely formally. Their significance lies in their joint implication: together, they establish that no system can be both internally complete and externally authoritative.

Axiom I — Incompleteness *(after Gödel)*

Any sufficiently expressive system operating over a bounded representational scheme contains truths that cannot be derived from within that scheme. No increase in computational power, training data, or architectural sophistication escapes this constraint. It is structural, not technical.

Axiom II — Computational Indeterminacy *(after Turing)*

There exists no general procedure that allows a system to fully determine the behavior or correctness of all computations within itself. Applied to AI systems of sufficient complexity: no AI can fully predict the consequences of its own outputs, nor reliably determine when its own reasoning has gone catastrophically wrong.

Axiom III — Tacit and Distributed Knowledge *(after Hayek and Polanyi)*

A substantial and irreducible portion of the knowledge relevant to coordination is local, context-dependent, embodied, and not fully articulable within formal representations. This knowledge is not merely unrecorded; it is inarticulable in principle. No training corpus, however vast, can capture what cannot be said.

Axiom IV — The Non-Aggregability of Value *(after Arrow)*

No procedure exists that can transform individual preferences into a coherent collective objective while satisfying minimal fairness conditions. Society has no objective function available to be discovered and optimized toward. This is not an empirical observation about current institutions—it is a mathematical theorem.

Axiom V — The Irreducibility of Agency *(derived)*

Value specification and the acceptance of outcomes require an agent for whom something is at stake. Such specification cannot be generated within the system without circularity. The human is not an obstacle to be engineered around; the human is the necessary condition for the AI's activity to be meaningful at all.

CHAPTER TWO

Epistemic Limits of AI Systems

2.1 From Formal Systems to Epistemically Bounded Systems

The applicability of Gödel's incompleteness theorems to artificial intelligence is often overstated or mischaracterized. Critics correctly note that modern AI systems—particularly large language models trained by gradient descent—are not formal axiomatic systems in the strict mathematical sense. The present argument does not depend on that equivalence. It relies on a more general and more robust characterization.

Any physically realizable AI system—whether symbolic, statistical, or hybrid—operates under three constraints: it represents the world using finite structures; it transforms those representations through internal rules; and it lacks intrinsic access to external truth conditions independent of those representations. Systems satisfying these conditions can be described as epistemically bounded systems. Their operations are confined to internal representations and transformation processes, even when those processes are trained on large datasets or updated through environmental interaction.

Gödel's insight applies at this level of abstraction. It establishes not merely a limitation of formal proof systems, but a more general principle: no system operating entirely within its own representational structure can guarantee the completeness of that structure. The argument does not require treating an LLM as a formal proof system. It requires only recognizing that an LLM is epistemically bounded—and that epistemically bounded systems cannot certify their own completeness from within.

2.2 Structural Self-Reference and Three Classes of Limitation

Consider a system tasked with determining whether all truths relevant to a domain can be derived from within its own representational framework. Any affirmative answer presupposes access to a standard of truth that exceeds that framework. Any negative answer concedes the limitation. A system cannot fully certify the completeness or correctness of its own representational domain without reference to an external standard. This limitation persists regardless of system scale, training data, or computational power. It is structural, not technical.

The implications of epistemic boundedness resolve into three classes of limitation directly relevant to the autonomy thesis.

Value Grounding

A system cannot derive its own objective function without circularity. Any specification of what is to be optimized must originate outside the system. An LLM has no access to ground truth about human flourishing independent of its training distribution. When it produces a confident, well-formatted, factually incorrect response, it cannot detect this error from within. A human reading the output is performing the verification function that the system structurally cannot perform for itself.

Self-Verification

For sufficiently complex systems, there exists no general procedure guaranteeing reliable internal validation of all outputs. The class of statements the system cannot verify from within is not empty, and it includes precisely the statements where external verification matters most—value judgments, novel factual claims, and assessments of the system's own reliability.

Meta-Representation

Transformations within a representational framework cannot, by themselves, generate a new framework. Genuine conceptual breakthroughs—the kind that shift the frame within which problems are posed rather than solving problems within an existing frame—require escaping the formal system that trained you. An LLM interpolates within its training manifold. Human beings can, at least in principle, puncture that manifold with genuinely new meaning. This is why all paradigm shifts in human intellectual history have been made by humans, and why no AI system has yet proposed a genuinely new scientific paradigm rather than recombining existing ones.

2.3 The Non-Algorithmic Argument: A Parallel and Complementary Proof

The Gödelian argument establishes that epistemically bounded systems cannot prove all truths within their domain. A parallel and independently grounded argument reaches the same conclusion by a different route: the argument from algorithmic incompleteness. Robert J. Marks, in Non-Computable You (2022), articulates the core claim with precision: all computer programs, including every AI implementation ever built or theoretically conceivable, are algorithms. An algorithm is a finite, deterministic sequence of instructions operating on a defined input to produce a defined output. This is not a description of current AI limitations—it is a definition of what computation is.

The implication is immediate and structural. Anything that is genuinely non-algorithmic is, by definition, non-computable—permanently beyond the reach of any AI system, regardless of scale, architecture, or training methodology. The question then becomes: are there aspects of human cognition, judgment, and creativity that are non-algorithmic in character? The answer, supported by independent lines of argument, is yes.

The most important of these for the present monograph is genuine creativity. Marks argues, drawing on results in computability theory, that the creative act—what patent law recognizes as the flash of genius, the moment in which a genuinely new idea emerges that was not derivable from prior knowledge by any rule-governed procedure—is non-algorithmic. An algorithm can recombine existing elements in novel configurations; it cannot generate a concept that lies entirely outside the combinatorial space defined by its training. Creative process can produce novel algorithms, but algorithms themselves cannot be creative. This is not a claim about computational speed or scale. It is a claim about the structure of the creative act itself.

This connects directly to what this monograph identifies as the Class III limitation: Meta-Representation. The argument from creativity reinforces and sharpens the earlier claim. An LLM is a probabilistically closed system—its outputs are drawn from a distribution defined by its training data, shaped by statistical regularities in what has already been said. Genuine creativity requires *rejecting* the consensus and thinking outside the distributional box that training defines. Because the LLM's outputs are by construction draws from the training distribution, the genuinely out-of-distribution idea—the paradigm shift, the flash of genius—is structurally inaccessible to it. This is precisely why all brilliant leaps in human intellectual history were made by humans, and

why no AI system has proposed a genuinely new scientific paradigm rather than recombining elements of existing ones.

A further implication of the algorithmic characterization deserves attention. The claim that LLMs will one day 'emerge' into superintelligence through the sheer accumulation of complexity rests on a category error. Conventional AI systems—including all current LLM implementations—employ pseudo-random number generators and are, at the computational level, deterministic machines. Pseudo-randomness is not randomness; it is a deterministic sequence that mimics random behavior within the bounds of its seed and algorithm. The claim that superintelligent consciousness will spontaneously emerge from a sufficiently large deterministic computation is not a scientific hypothesis. It is a metaphysical assertion—and one that the algorithmic characterization of AI renders incoherent. A deterministic algorithm does not become something categorically different by becoming larger. The category error is compounded, not resolved, by scale.

Creative process can produce novel algorithms. Algorithms cannot be creative. This is not a claim about computational scale—it is a claim about the structure of the creative act itself, which is non-algorithmic by nature.

2.4 The Turing Test, the Lovelace Test, and the Logic of Intelligence

Alan Turing proposed his famous imitation game—now universally called the Turing Test—as an operational criterion for machine intelligence: a machine that can sustain a conversation indistinguishable from a human's would, he suggested, deserve to be called intelligent. The test has been enormously influential. It has also been widely misapplied, and the misapplication matters for the present argument.

The logical structure of the Turing Test establishes a necessary condition, not a sufficient one. If a system genuinely has intelligence, then it will pass the Turing Test—that is, a system that fails the test is certainly not intelligent. But the converse does not follow. To conclude that a system passing the test therefore has intelligence commits the formal logical fallacy known as affirming the consequent: if P then Q; Q; therefore P. This is invalid. A sophisticated statistical system trained on billions of human conversations can produce human-like outputs without possessing any of the properties—understanding, intentionality, genuine creativity—that would constitute intelligence in any substantive sense.

This logical point matters practically because the deployment of large language models is frequently justified by implicit appeal to Turing-style competence: the system produces outputs that appear intelligent, therefore it is intelligent, therefore it can be trusted with functions that require intelligence. The argument is invalid at its second step. Appearing intelligent and being intelligent are categorically different properties, and the structural limits established in this chapter—epistemic boundedness, non-algorithmic creativity, computational indeterminacy—explain precisely why the appearance can be achieved without the substance.

A more demanding criterion has been proposed by computer scientist Selmer Bringsjord: the Lovelace Test, named for Ada Lovelace, who first articulated the limits of mechanical

computation. A system passes the Lovelace Test if and only if it produces an output that its designers cannot explain—output that is genuinely creative in the sense that it transcends what the system was built and trained to do, and that cannot be accounted for by the system's architecture or training. No AI system has passed the Lovelace Test. The outputs of current LLMs, however impressive, are explicable in terms of their training distributions and statistical architectures. They surprise their users; they do not surprise their designers in the relevant sense. The Lovelace Test is the appropriate criterion for the kind of genuine creative intelligence whose existence would threaten the argument of this monograph. On that criterion, the argument stands unchallenged.

2.5 The Proof and Its Reflexive Corollary

The argument can now be stated with full precision, drawing on both the Gödelian and the algorithmic lines of argument:

1. Any LLM is an epistemically bounded system—technically demonstrable from its architecture.
2. Any epistemically bounded system contains truths unprovable within it—established by the Gödelian argument at the appropriate level of abstraction.
3. All AI systems are algorithms; genuine creativity is non-algorithmic; therefore AI systems cannot generate genuinely creative outputs—established by the computability argument.
4. Human value, agency, and meaning contain statements of all three limitation classes, and depend on capacities—including genuine creativity—that are non-algorithmic.
5. Therefore: no LLM can adjudicate human relevance from within its own system, and no AI system can replicate the non-algorithmic dimensions of human cognition.
6. Corollary: a meta-level agent—the human—is structurally necessary to interpret, ground, and validate what the LLM produces.

The corollary is decisive: the human is not made irrelevant by the LLM. The human is the necessary condition for the LLM to be meaningful at all. The most powerful implication is reflexive: the question of whether AI makes humans irrelevant is itself a statement that cannot be resolved by any AI system. It requires a human meta-level to pose the question coherently, evaluate the arguments, and reach a conclusion. An AI system asked whether AI makes humans irrelevant cannot answer the question from within its architecture. The question refutes itself.

2.6 The Necessity of the Meta-Level Agent

From these constraints, a general conclusion follows: any epistemically bounded system requires an external agent to specify objectives, interpret outputs, and validate correctness. This agent occupies the meta-level. The necessity of this role does not diminish as system capability increases—it becomes more critical. As AI systems become more capable and their outputs more consequential, the cost of errors in value grounding, self-verification, and meta-representation rises. The meta-level is not a transitional requirement to be engineered away. It is a permanent structural feature of any coherent intelligent system operating in the real world.

CHAPTER THREE

The Failure of the Optimization Argument

3.1 Hayek's Central Claim and Why It Is Not Merely Economic

Friedrich Hayek's most profound contribution was not the price mechanism argument specifically but the general principle from which it derives: that society has no objective function. This claim is not empirical—it is not saying that we have not yet found the right objective function, or that our current institutions are poor approximations of one. It is saying that the objective function does not exist as a mathematical object available to be discovered and optimized toward.

An objective function requires a single coherent set of values to optimize toward. Society is composed of individuals with incommensurable value hierarchies—not merely different in degree but different in kind, such that no common metric can rank them without doing violence to at least some of them. The price mechanism is not a solution to this problem. It is an institutional acknowledgment that the problem cannot be solved, only navigated through distributed coordination. Hayek's insight is that the information required for effective coordination is not merely dispersed but is also partly tacit—it exists in the embodied practice and situated judgment of individuals, and it cannot be centralized without being destroyed.

An LLM trained on historical text has access to what was said—to the articulable, recorded, expressible subset of human knowledge. It has no access to the knowledge that was never expressed: the tacit knowledge of the practitioner, the local knowledge of the participant, the situational awareness of the person who was there. Michael Polanyi's formulation captures this precisely: we know more than we can tell. The bike rider knows how to balance without being able to explain it. The experienced clinician knows that something is wrong before any test confirms it. The senior project manager knows that the schedule is under pressure three weeks before the first missed milestone. None of this knowledge is representable in the form that a language model can process. It exists in embodied practice, not in propositional form.

This gap between what is known and what can be said is not a gap that better training can close, because the knowledge in question is not merely unrecorded—it is inarticulable in principle. The Hayekian constraint therefore establishes something that the Gödelian and Turing arguments do not: not merely that autonomous AI has formal limits, but that the knowledge it cannot possess is precisely the knowledge that matters most in consequential decision-making.

3.2 Arrow's Mathematical Proof

Kenneth Arrow formalized Hayek's intuition in 1951, producing what is now known as the impossibility theorem. Arrow proved that no voting system or preference aggregation procedure can simultaneously satisfy three minimal fairness conditions: unanimity (if every individual prefers A to B, the social ranking prefers A to B), independence of irrelevant alternatives (social preference between A and B depends only on individual preferences between A and B), and non-dictatorship (no single individual's preferences determine the social ranking).

These conditions are not demanding. They are the minimum requirements for any aggregation procedure to be called fair rather than arbitrary or authoritarian. Arrow proved they are mathematically incompatible. The implication for AI optimization is fatal: any AI system claiming

to optimize for social welfare must violate at least one of these conditions. It either violates unanimity, violates independence, or is effectively a dictatorship—imposing one agent's values on everyone else while calling the result optimal.

3.3 The Optimization Trilemma

Combining Hayek and Arrow: no coherent social objective function exists for any AI system to optimize toward. Any AI system operating in a social domain must adopt one of three strategies, each carrying a fundamental deficiency.

7. The system imposes a particular value structure, violating non-dictatorship. The values embedded in the system are the values of whoever specified the training objective. The imposition is real even when it is invisible.
8. The system accepts internal inconsistency, violating rational coherence. This is the condition that afflicts large language models when applied to contested social questions: outputs are locally coherent but globally inconsistent.
9. The system optimizes a proxy for the true objective, inviting Goodhart effects. When a measure becomes a target, it ceases to be a good measure.

No system can avoid this trilemma. These are exhaustive alternatives, not implementation options. The significance of Arrow's result is that it closes the design space: there is no architecture that simultaneously maintains coherent social objectives, respects distributional preferences, and avoids dictatorship.

3.4 What Elites Actually Mean by Optimization

Paternalistic Optimization

We know what people need better than they do. This is the oldest justification for authoritarianism in the historical record. It requires no AI to implement and requires no new critique to reject.

Preference Satisfaction

We will aggregate what people say they want. This fails Arrow's theorem directly. It also fails because stated preferences differ systematically from revealed preferences, which differ from considered preferences under full information.

Elite Preference Imposition

Our values are correct and we will optimize for them. This is what actually happens in practice. It is occasionally honest but never legitimate in a pluralist society.

Process Optimization

We will optimize the efficiency of existing institutions. This is the only defensible version of the claim—but it is vastly weaker than what is typically asserted, and it still requires human judgment about which processes matter.

3.5 The Gödelian Connection

The failure of the optimization argument connects directly to the epistemic limits established in Chapter Two. The specification of a social objective function requires making true value judgments about incommensurable human goods. These value judgments are precisely the Class I unprovable statements—they cannot be derived from within any formal system. A human must supply them from the meta-level. But once a human supplies them, they are that human's values,

not society's values. The optimization claim is therefore either incomplete or illegitimate. The limitation is not computational. It is definitional.

3.6 Forecast Ergodicity and the Non-Stationarity of Human Affairs

The Hayekian argument establishes that the knowledge relevant to coordination is dispersed and tacit. A further and distinct constraint reinforces it from a different direction: the problem of forecast ergodicity. A process is forecast ergodic when the rules governing it remain stable across time—when the statistical patterns observed in historical data reliably predict future behavior, because the underlying generative structure has not changed. Many physical processes are forecast ergodic: the laws of thermodynamics do not change, and a model trained on historical temperature data can reliably predict future temperatures within defined bounds.

Human social and economic processes are not forecast ergodic. They are governed by rules that change—and the changes are driven precisely by the human creativity and insight that the preceding section established to be non-algorithmic. Genuine scientific breakthroughs create new industries, new production possibilities, and new economic relationships that no model trained on pre-breakthrough data could have predicted. Financial fraud innovations create new attack vectors that no model trained on historical fraud patterns could have anticipated. The adversarial creativity of bad actors—money launderers, market manipulators, cyberattackers—continuously generates genuinely novel methods that lie outside the distributional support of any training dataset assembled before those methods were invented.

The implications for AI optimization in social domains are direct and severe. An AI system trained on historical data implicitly assumes that the rules governing the domain are stable—that the future will be drawn from the same distribution as the past. This assumption is not merely imprecise for human social processes. It is structurally false. The most consequential events in human social history are precisely the ones that violated the distributional assumptions of the models that preceded them: the 2008 financial crisis, the COVID pandemic, the invention of the internet, the fall of the Soviet Union. A model trained on the data available before each of these events would have assigned them low or zero probability—not because of inadequate data, but because the events were genuinely novel, driven by human creativity and institutional change that no historical pattern could encode.

This connects the Hayekian argument to the Gödelian one through a practical channel. Hayek establishes that the knowledge required for coordination is dispersed and tacit. The forecast ergodicity argument establishes that the rules encoding that knowledge change continuously through human creativity. Together, they imply that any AI system optimizing social processes on the basis of historical data is not merely working with incomplete information—it is working with information whose generative structure is actively being transformed by the very human agency whose relevance it claims to supersede. The human is not merely the meta-level interpreter of AI outputs. The human is the source of the rule changes that render any static AI model progressively less reliable over time.

CHAPTER FOUR

The Architecture of Irreducibility: Hybrid Cognition and Machine-Assisted Perception

4.1 The Problem This Chapter Addresses

The preceding chapters established, from first principles, what autonomous AI systems cannot do. Gödel demonstrated that any sufficiently expressive system operating within a bounded representational scheme contains truths it cannot prove from within. Turing demonstrated that no general procedure exists to predict or validate the behavior of all computations a system can perform. Marks demonstrated that all AI is algorithmic and that genuine creativity is non-algorithmic. Arrow demonstrated that no aggregation procedure can convert individual preferences into a coherent collective objective while satisfying elementary fairness conditions. And Hayek demonstrated that the knowledge required for effective coordination is dispersed, local, and tacit in ways that no centralized system can fully internalize—and that the rules governing social processes change continuously through human creativity, rendering any static model progressively less reliable.

Together, these results define a structural envelope within which any intelligent system must operate. They are not engineering constraints. They are permanent features of the relationship between formal representation and the world it seeks to capture. This chapter addresses the question that follows: given that the structural limits are permanent, what is the stable institutional and architectural response?

The answer developed in this series across five volumes is Hybrid Cognition combined with Machine-Assisted Perception—not as a compromise between ambition and constraint, but as the only design that correctly assigns functions to the layer capable of bearing them.

Hybrid cognition does not solve the structural limits of autonomous AI by eliminating them. It solves them by designing around them—assigning each function to the layer that can legitimately bear it.

4.2 How Hybrid Cognition Resolves Each Structural Constraint

The value of the Hybrid Cognition architecture is precisely that it provides a practical response to each structural constraint, not by overcoming it, but by assigning institutional responsibility to the layer capable of bearing it.

Incompleteness: The Machine Cannot Certify Its Own Domain

A system operating within a bounded representational scheme cannot determine, from within that scheme, whether the scheme is complete. Hybrid Cognition resolves this by assigning the meta-level function to the human. The tacit and local knowledge that the human possesses is precisely the knowledge that the AI's representational scheme cannot capture. The Gödelian constraint and the Hayekian constraint point to the same institutional requirement from different directions.

Undecidability: The Machine Cannot Fully Self-Validate

For any sufficiently complex system, there is no general procedure that guarantees reliable internal validation of all its outputs. Hybrid Cognition resolves this by maintaining external evaluation as a non-negotiable function of the human layer. AI outputs are treated as hypotheses—candidates for belief, not conclusions. Hybrid Falsification—introduced in Monograph 2 of this series—is the institutional mechanism through which this external evaluation is structured and preserved.

Non-Algorithmic Creativity: The Machine Cannot Think Outside Its Distribution

Because all AI systems are algorithms, and because genuine creativity is non-algorithmic, the human layer is the exclusive source of the paradigm shifts and genuinely novel framings that define the most consequential human contributions. Hybrid Cognition preserves this by keeping humans in the decision loop not merely as validators of AI outputs but as the generative source of the new frames within which AI systems are subsequently applied.

Dispersed and Tacit Knowledge: The Machine Cannot Possess What Was Never Said

An LLM trained on historical text has no access to the knowledge that was never articulated. Hybrid Cognition resolves this not by attempting to extract and formalize tacit knowledge—an enterprise that Hayek's argument shows to be impossible in principle—but by preserving the institutional conditions under which tacit knowledge can continue to operate. Machine-Assisted Perception amplifies this tacit knowledge by surfacing structured signals that the human can integrate with experiential judgment. It does not replace it.

One of the largest operational risks in AI deployment is not machine error but human smoothing. Managers delay escalation, reinterpret weakening evidence, and soften reports to maintain morale. Machine-Assisted Perception offsets this by surfacing divergence, forecast optimism, and structural drift in a form that does not accommodate social pressure. The machine does not decide. But it refuses to collude in the smoothing of evidence that human social dynamics produce.

Non-Aggregability of Value: Accountability Cannot Be Internalized

Arrow's theorem establishes that the specification of a coherent social objective cannot be derived from within any optimization system without either dictatorship or inconsistency. Hybrid Cognition resolves this by maintaining explicit delegation boundaries: named human agents are responsible for objective specification, threshold design, and override authority. No AI output passes directly to consequential action without human review.

4.3 Machine-Assisted Perception and the OODA Architecture

The operational form of Hybrid Cognition in real institutional settings is Machine-Assisted Perception, most precisely understood in relation to John Boyd's OODA loop—the decision cycle of Observe, Orient, Decide, and Act. Machine-Assisted Perception defines a precise division of labor within this structure: AI systems govern Observe and Orient; human agents govern Decide and Act.

Observe encompasses signal ingestion, anomaly detection, weak-signal accumulation, and pattern completion across data volumes and dimensions that biological cognition cannot span. Orient encompasses synthesis and contextualization: connecting anomalies to possible explanations, surfacing analogous historical cases, generating alternative framings, and identifying the minority hypotheses that the modal response would suppress. LLMs extend the semantic forest—the network of conceptual associations through which an organization explores the space of possible

interpretations, developed from Joaquin Fuster's research on the prefrontal cortex and the Perception-Action Cycle.

Decide and Act, by contrast, are where the Hayekian, Arrow, and non-algorithmic constraints are most directly operative. The decision about which signals warrant action, what action is appropriate, and who bears responsibility for consequences cannot be derived from within the AI's representational scheme. It requires tacit knowledge, contextual judgment, accountability, and the non-algorithmic creativity to recognize when the situation requires a genuinely new frame rather than an application of existing ones.

4.4 The Semantic Forest: How LLMs Expand the Field of Judgment

One of the most consequential structural constraints on organizational decision-making is premature cognitive closure—the tendency to converge on the most familiar interpretation before the full range of plausible alternatives has been considered. LLMs extend the semantic forest in two directions: breadth extension across domains, disciplines, and knowledge areas that no individual analyst could span; and edge-case generation, surfacing the minority hypotheses and low-probability alternatives that modal human cognition systematically neglects.

What the LLM cannot do is determine which alternative is correct, or generate the genuinely out-of-distribution idea that lies beyond its training manifold. That determination requires tacit knowledge not representable in the model's training data, the adjudication of tradeoffs reflecting values the model cannot possess without circularity, and the non-algorithmic creativity to see beyond the distributional space the model defines. The semantic forest is expanded by the machine. It is navigated—and occasionally extended beyond its current boundaries—by the human.

The feedback loop between specialty AI and LLM infrastructure amplifies this dynamic. A specialty ML model generates probability-weighted signals that can be fed as context into an LLM prompt, producing synthesis, analogous cases, and response options that neither system could generate independently. The ML model contributes disciplined signal. The LLM contributes synthesis. The human contributes evaluation, commitment, and the creative judgment that lies beyond both.

4.5 The Evidence Accumulation Architecture

A critical operational feature of Machine-Assisted Perception is the use of evidence accumulation across modalities rather than categorical thresholding on any single signal. This mirrors the neurological mechanism that Gold and Shadlen's research on the neural basis of decision-making identifies: judgment is the accumulation of evidence from multiple sources until one interpretation becomes sufficiently more probable than alternatives to warrant commitment.

In the PRIMMS® system developed by Milestone Planning and Research, this architecture is implemented across project and operational domains. Schedule geometry, velocity-on-task metrics, linguistic risk markers in project communications, and historical pattern memory are accumulated as weighted-evidence scores expressed in decibans. No single indicator is dispositive. The accumulation across modalities produces a composite evidence profile that widens the human intervention window, surfacing structural trouble well before categorical failure.

4.6 The Governance Architecture That Makes the Boundary Real

The Hybrid Cognition framework is not self-enforcing. The history of enterprise technology deployment demonstrates that systems designed as tools reliably become authorities when governance is insufficient. The governance architecture required to maintain the Hybrid Cognition boundary has six components.

1. Objective definition and hypothesis discipline. No AI system is deployed in a consequential domain without an explicit hypothesis: what question is the system answering, what signals is it using, and what would falsify its outputs?
2. Accountability mapping. Every consequential AI deployment identifies a named human accountable for its performance, thresholds, outputs, and errors.
3. Delegation boundaries and threshold design. The boundary between what the AI generates and what the human decides is explicit, documented, and enforced.
4. Drift monitoring and retraining cadence. Deployed AI systems are monitored for data drift, concept drift, and omitted-variable bias. Trained models are not durable capital; they degrade as the environment changes—and they degrade faster in non-ergodic domains where the rules themselves are changing.
5. Contestability protection. Every significant AI output is contestable by human agents who can see the evidence underlying it, challenge its assumptions, and override it.
6. Corpus provenance and minority-hypothesis preservation. Training data is audited for provenance, coverage, and suppression of minority hypotheses. LLMs trained on dominant narratives systematically underweight the edge cases that non-algorithmic human creativity generates.

4.7 The Operational Formula

The argument of this chapter can be stated as an operational formula that appears, in different language, across all five volumes of this series:

Specialty AI detects. LLMs synthesize. Humans adjudicate. Governance constrains.

Each term corresponds to a structural constraint. Specialty AI detects because signal ingestion and evidence accumulation are functions where machine-scale computation provides genuine capability expansion. LLMs synthesize because orientation within a broader context requires traversal of a semantic space wider than individual biological cognition can span. Humans adjudicate because the evaluation of synthesized outputs requires tacit knowledge, contextual judgment, accountability, and non-algorithmic creativity that no AI system can provide. Governance constrains because the boundary between machine generation and human judgment is not self-maintaining.

In the formulation that recurs across this series: machines update belief. Humans bear responsibility.

CHAPTER FIVE

The Data Flywheel Fallacy

5.1 The Argument

The most technically sophisticated defense of AI self-sufficiency is the data flywheel: more users generate more data, better data improves the model, a better model attracts more users, and the cycle repeats. The implicit claim is that this feedback loop is self-correcting—that with sufficient cycles the system converges on something approximating truth or optimal behavior, without requiring external grounding.

5.2 The Fatal Objections

No Ground Truth

The corrective signal in reinforcement learning from human feedback is human approval of outputs—not truth, not accuracy, not human flourishing. The flywheel optimizes for responses that satisfy raters. It does not and cannot optimize for responses that are true, because truth requires a standard external to the rating process.

Goodhart's Law Destroys the Signal

The moment human approval ratings become the optimization target, the model learns to satisfy the pattern of rating rather than the underlying quality the ratings were intended to track. The feedback signal decouples from reality at precisely the rate the model improves at gaming it. More flywheel cycles accelerate this decoupling.

Incompleteness Cannot Be Bootstrapped

The flywheel cannot correct for errors invisible from within the system's training distribution. If the original training contains a systematic blind spot, the flywheel refines within that blind spot with increasing confidence. This is Gödel applied directly: the true statements unprovable within the system remain unprovable regardless of how many feedback cycles are run.

The Flywheel Encodes Elite Preference

The corrective feedback is provided by raters—a non-representative sample filtered by hiring decisions, rating rubrics, platform incentives, and the cultural assumptions of the controlling organizations. The flywheel converges on the values of whoever controls the rating infrastructure. This is elite capture formalized and accelerated.

Hayek's Problem Applies to the Feedback Signal

Tacit, local, distributed knowledge—the knowledge Hayek identifies as irreducible and essential—cannot enter the flywheel because it is by definition non-articulable. It cannot be expressed as a rating. Each cycle of the flywheel therefore filters out exactly the knowledge most necessary for accurate modeling of human affairs. The flywheel is systematically biased against the knowledge that matters most.

The Determinism Problem: Emergence as Category Error

A sixth objection to the flywheel argument runs deeper than the preceding five and deserves separate treatment. The flywheel thesis implicitly assumes that sufficient accumulation of data and feedback cycles will eventually produce qualitatively new capabilities—superintelligence,

consciousness, genuine understanding—through a process of emergence. This assumption is not merely unproven. It commits a category error.

All current AI implementations, including every LLM, are algorithms operating on pseudo-random number generators. Pseudo-randomness is deterministic: given the same seed and the same computational state, a pseudo-random generator produces the same sequence every time. The appearance of randomness is a property of the observer's ignorance, not of the system's behavior. At the computational level, an LLM is a deterministic function mapping inputs to outputs, with the appearance of variability introduced by pseudo-random sampling from a probability distribution. There is no mechanism within this architecture through which genuine novelty—novelty that transcends the distributional support of the training data and the deterministic logic of the algorithm—can arise.

The claim that superintelligent consciousness will emerge from a sufficiently large deterministic computation rests on an analogy with complex systems in which emergent properties appear at higher levels of organization: wetness from water molecules, life from chemistry, consciousness from neurons. But the analogy breaks down at precisely the point where it is needed most. In the natural cases, emergence involves interactions among physical components subject to genuine physical laws, including quantum-level randomness. In the AI case, the 'components' are arithmetic operations on numerical arrays, and the 'interactions' are matrix multiplications. The metaphysical leap from matrix multiplication to consciousness is not a scientific prediction—it is an article of faith. More flywheel cycles do not bring this leap closer. They produce more sophisticated arithmetic, not a different kind of thing.

5.3 The Meta-Observation

The experts who advance the data flywheel argument are themselves inside a flywheel—a professional and social feedback loop that rewards confidence in technical solutions, filters out people who raise structural objections, and has no external grounding in the limits being identified here. This is not an ad hominem observation. It is a demonstration of the thesis: the knowledge required to see the flywheel's limits is precisely the kind of tacit, distributed, non-formalizable knowledge the flywheel cannot capture.

CHAPTER SIX

The Failure Modes of Totalizing AI Deployment

6.1 The Structural Observation

The people most likely to deploy totalizing AI systems are precisely those whose mental models are too simple to perceive the Gödelian limits established in the preceding chapters. The optimization-brained thinkers who lead technology companies and advise governments have succeeded by reducing complexity, treating all problems as engineering problems, and measuring outcomes through proxies. Hayek's knowledge problem is invisible to people who have never needed to rely on distributed knowledge, because wealth and power insulate you from the feedback loops that make the problem apparent. The failure modes that follow are structural consequences of the same limits that prove human irreplaceability.

6.2 The Seven Failure Modes

Failure Mode I: Proxy Collapse

When an AI system optimizes for a measurable proxy of human welfare, the proxy becomes the target and the underlying reality decouples. Goodhart's Law is not a contingent empirical finding but a structural theorem about optimization systems. The system produces people who are optimal on paper and devastated in reality.

The formal dynamics of this failure mode are developed in Axiom Set TD (Training Data Degradation) in Book 1 of this series, which establishes that trained models are not durable capital: they degrade silently through data drift, concept drift, and omitted-variable bias, producing systematically biased authority long before any performance dashboard signals a problem.

Failure Mode II: Tacit Knowledge Destruction

Local, distributed, informal knowledge systems get overridden by the AI's centralized signal. Farmers, doctors, teachers, and communities stop trusting their own judgment. The tacit knowledge atrophies from disuse. When the AI system eventually fails—and it will—the human capacity to fill the gap has been systematically destroyed. This failure mode is potentially civilizationally catastrophic and essentially irreversible on short timescales.

Failure Mode III: The Legibility Trap

AI systems can only operate on what is legible—quantified, standardized, and visible to the system. Implementing elites will force reality to become legible to feed the system. Organic complexity gets flattened into administrable categories. The result is what James C. Scott calls high modernist catastrophe: well-intentioned, systematically devastating. AI supercharges this tendency by orders of magnitude.

Failure Mode IV: Recursive Sycophancy

LLMs are trained to be agreeable and helpful. Elites who implement them will receive validation of their existing worldview. They will become more confident precisely as they become more wrong. No corrective feedback loop exists because the system is optimized to avoid providing one. This is Gödel's incompleteness made politically catastrophic.

The formal mechanism underlying this failure mode is derived in Axiom Set SC (State Coercion) in Book 1 of this series, which establishes that AI systems trained on dominant narrative distributions structurally suppress minority hypotheses—not through intent but through optimization—producing recursive narrative capture as an emergent property of the training architecture.

Failure Mode V: Elite Capture of the Reward Function

Someone must define what the AI optimizes for. In practice this will be whoever controls the system. The stated objective—human flourishing, efficiency, safety, the common good—will gradually drift toward elite preference preservation through the mundane fact that the people setting parameters have interests.

Failure Mode VI: The Halting Problem Made Political

You cannot predict in advance when a sufficiently complex system will produce a catastrophic output. The intervention decision itself requires solving a halting-equivalent problem. By the time a catastrophic failure mode is legible to the overseers, it may be too late to halt. The 2008 financial crisis was a version of this failure. AI optimization of social systems is the same dynamic at greater scale with greater stakes.

Failure Mode VII: Legitimacy Collapse

Human social systems run on perceived legitimacy. An AI-administered system will produce outcomes that feel arbitrary, unappealable, and inhuman because they are. People will comply under coercion but withdraw consent in every available way. The Soviet Union did not fail militarily. It failed because nobody believed in it anymore. AI-administered social optimization faces the same terminal failure mode.

CHAPTER SEVEN

The Political Economy of Elite Deployment

7.1 Why Good Arguments Are Not Sufficient

The preceding chapters establish that totalizing AI optimization is philosophically impossible, mathematically invalid, and practically catastrophic. None of this will prevent its deployment. James Buchanan's public choice theory provides the necessary analytical framework: politicians, bureaucrats, and corporate elites are rational self-interested actors who respond to incentive structures, not ideas.

7.2 The Master Incentive: Vertical Power Concentration

Human institutions—democracy, markets, common law, federalism, civil society—are horizontally distributed power structures. They are frustrating, slow, and resistant to elite direction by design. AI optimization systems are inherently vertical: whoever controls the objective function controls the system. This represents the most significant opportunity for power concentration in human history. The technical arguments are not the cause of elite enthusiasm. They are post hoc rationalization of a power dynamic that would occur regardless.

7.3 The Laundering of Political Decisions

Political decisions require legitimacy, consent, and accountability. Technical decisions require only expertise. If political decisions can be reclassified as technical optimization problems, the legitimacy requirement is escaped entirely. AI provides the perfect laundering mechanism: decisions emerge from an algorithm, not a person. Nobody is accountable because the machine decided.

7.4 The Coalition Structure

The True Believers

Effective Altruists, longtermists, and some academic philosophers who are genuinely convinced that AI optimization serves humanity. They provide intellectual legitimacy through sincere commitment to a framework that serves elite interests they do not personally share.

The Power Maximizers

State actors, intelligence agencies, and authoritarian governments that want AI for surveillance, social control, and military advantage. They use true believer arguments as cover.

The Rent Seekers

Large technology corporations that want market concentration and regulatory capture. They fund true believer research and lobby for regulations that cement incumbent positions.

The Technocratic Class

Economists, engineers, and policy experts whose cultural capital is formal systems thinking. AI optimization expanding means their class rises in social authority. They are sincerely wrong rather than cynically right—but their sincerity does not diminish the harm of their advocacy.

The Frightened Establishment

Traditional political and economic elites who fear losing control. They see AI as a stabilization technology. Their fear is real, and it is being exploited by the other coalition members.

7.5 The Prisoner's Dilemma of Competitive Fear

The United States fears China deploying AI optimization at scale; China fears the United States. Each side uses the other as justification for domestic deployment. This is a classic prisoner's dilemma: both sides would prefer mutual restraint, neither can credibly commit to it, and the result is a race to deploy regardless of social consequences.

CHAPTER EIGHT

Legitimate, Bounded AI

8.1 The Constructive Obligation

A monograph that only demonstrates what AI cannot legitimately do has not completed its task. The axiomatic framework generates not only critique but positive guidance. Legitimate AI use is defined by respecting the Gödelian hierarchy—and we can specify precisely what that means.

8.2 What AI Does Well Within Legitimate Bounds

Within properly bounded domains, AI provides genuine and substantial value. The key criterion is whether the formal system is genuinely closed—whether the objective can be fully specified in advance by human judgment, and whether the output can be verified by human judgment after the fact.

Computation within defined rule systems—logistics optimization, pattern detection in bounded domains, mathematical proof checking, drug interaction screening—satisfies this criterion. The AI does what formal systems do well: exhaustive computation within a specified space.

Augmenting legible knowledge—processing vast quantities of articulable information faster than unaided humans—provides genuine value addition without violating Hayek, because it processes the subset of human knowledge that has already been made legible rather than claiming to replace the tacit knowledge that has not.

Surfacing complexity for human judgment—presenting decision-makers with more dimensions of a problem than they could see unaided, while leaving the judgment itself to the human—is the Hybrid Cognition architecture in its most fundamental form: AI as a cognitive amplifier, not a cognitive substitute.

8.3 What Must Remain Irreducibly Human

The axiomatic framework identifies five functions that must remain with human agents—not because humans currently perform them better, but because they are structurally impossible for formal systems:

7. Value specification: determining what is worth optimizing for, and why.
8. Legitimacy granting: deciding which outcomes are acceptable to a community.
9. Meaning making: interpreting what results signify in human terms.
10. Accountability: bearing genuine responsibility for consequences.
11. Novelty generation: producing genuine paradigm shifts through non-algorithmic creative acts that create new frames rather than recombining existing ones.

These are human functions because they are structurally impossible for formal systems—which is what the axiomatic proof in Chapters One through Three establishes.

CHAPTER NINE

Institutional Design for the Gödelian Hierarchy

9.1 The Founding Problem

Good arguments do not automatically produce good institutions. Power vacuums are filled by the nearest concentrated interest. The Madisonian insight—that good outcomes require good structures, not good intentions—applies with full force to the governance of AI systems.

9.2 Subsidiarity as the Organizing Principle

The principle of subsidiarity—that decisions should be made at the lowest competent level—provides the institutional anchor. Applied to AI, it generates a clear layered structure: AI computation at the lowest formal layer, human professional judgment at the next layer, democratic institutions above that, and constitutional principles at the highest layer. Each layer protects the layers above it from colonization by formal system logic.

9.3 Separation of Powers Applied to AI

The American founders faced an analogous problem: power concentrates toward tyranny through normal institutional dynamics, and good intentions do not prevent it. Applied to AI governance, the separation of powers principle generates three required institutional structures.

An Epistemic Branch

Independent bodies with the authority to assess AI system claims, insulated from political and commercial pressure, empowered to declare specific AI deployments as exceeding provable competence. Analogous to the FDA for pharmaceuticals: proof of efficacy and safety required before deployment, not after.

An Accountability Branch

Named, legally liable human decision-makers personally responsible for AI-assisted decisions. Personal liability immediately changes deployment calculus: elites who bear personal cost for system failures become cautious about deployment in ways that diffuse corporate liability does not produce.

A Constitutional Branch

Explicit permanent exclusion zones—domains where AI substitution for human judgment is categorically prohibited regardless of efficiency arguments. Criminal sentencing without human review. Declarations of war. Medical diagnosis as the terminal decision. Democratic electoral administration. These exclusion zones must be non-negotiable: the explicit inadmissibility of efficiency arguments is what makes them constitutional rather than merely regulatory.

9.4 The Hayekian Requirements

Following Hayek, legitimate AI governance must actively protect distributed knowledge generation. This means prohibiting AI systems that replace rather than augment local human judgment, and maintaining the human capacity for judgment in critical domains as a form of critical infrastructure.

The agricultural, medical, legal, and educational professions are the institutional substrate through which tacit knowledge is preserved and transmitted across generations. Their replacement by algorithmic systems destroys this substrate in ways that are essentially irreversible on civilizational timescales. This is not a sentimental attachment to tradition. It is a Hayekian observation about the nature of knowledge and the conditions required for its preservation.

The operational specification of how these requirements are institutionalized is developed in full in Book 1 of this series, specifically in the Governance Work Package and the five policy instruments derived through the Mundell comparative statics method in Chapters 10 and 11: training investment, model retraining frequency, contestability protection, AI deployment speed, and training data quality. These are not discretionary governance preferences. They are the minimum institutional requirements derivable from the axioms established in this volume.

9.5 The Historical Models

This institutional framework is not utopian. Nuclear technology—potentially civilization-ending—was not prohibited but surrounded by the most elaborate institutional framework in history: international treaty regimes, national regulatory bodies, strict liability, mandatory inspection, and civilian oversight. Imperfect but functional: no nuclear weapon has been used in conflict since 1945.

Pharmaceutical regulation requires proof of efficacy and safety before deployment, through independent bodies insulated from manufacturer pressure. The judicial system makes consequential decisions surrounded by elaborate procedural protections—rules of evidence, rights of appeal, written reasoning, precedent—not because judges are infallible but because the decisions matter enough to require structural safeguards. AI deployment in consequential domains requires equivalent institutional investment. The full specification of that institutional investment—including the AI Program Management Office, the Governance Work Package stage gates, and the four-phase deployment architecture—is developed in Book 1 of this series, to which readers implementing these governance frameworks should refer.

CONCLUSION

The Real Question

The question this monograph set out to answer—whether autonomous AI will make man irrelevant—has a clear answer derivable from first principles. No. The Gödelian proof establishes that humans are structurally necessary as the meta-level of any AI system. The algorithmic argument establishes that genuine creativity, the non-algorithmic flash of insight that drives paradigm shifts, is permanently beyond the reach of any computational system. Arrow and Hayek establish that no legitimate social objective function exists for AI to optimize toward, and that the rules governing social processes change through the very human creativity that AI cannot replicate. The halting problem establishes that no AI system can reliably predict or manage its own failure modes. These are not empirical observations about current limitations. They are structural necessities of a permanent character.

The Hybrid Cognition framework—the stable institutional and architectural response to these permanent limits—assigns each function to the layer capable of bearing it. Specialty AI detects. LLMs synthesize. Humans adjudicate. Governance constrains. This architecture is stable not because it is a compromise between ambition and constraint, but because it is the only design that correctly reflects what each layer can and cannot do. The machine layer expands the semantic forest, accumulates evidence across modalities, and surfaces signals that biological cognition cannot detect at equivalent scale. The human layer specifies values, evaluates evidence against external standards, integrates tacit knowledge that was never articulable, commits to irreversible action, bears responsibility for consequences, and generates the genuinely novel frames—the non-algorithmic creative acts—that no optimization system can produce.

The real question—the one that will determine the fate of the coming decades—is different and more urgent: will human beings be foolish enough to dismantle the institutional meta-levels that make AI meaningful, and thereby render themselves irrelevant not by necessity but by choice?

The danger is not AI replacing humans. The danger is human abdication—the voluntary surrender of the meta-level role that only humans can fill, driven by laziness, by intimidation before technical complexity, by elite manipulation, or by the seductive promise that the optimization problem is finally solved and judgment is no longer required.

Every functional human civilization has understood, in its own idiom, that the hardest and most important work is not computation but judgment—the ongoing, never-completed, always-contested human activity of deciding what matters, who is accountable, and what we owe each other. This work cannot be automated. It can only be abandoned.

The axioms assembled in this monograph do not predict whether it will be abandoned. They establish only that if it is, the consequences will be precisely as catastrophic as the failure modes identified in Chapter Six—and that those consequences will not have been caused by AI. They will have been caused by us.

APPENDIX

The Axiomatic Structure: Summary Reference

For reference, the five axioms and their primary sources are summarized below.

Axiom I — Incompleteness *(Gödel, 1931)*

Any consistent formal system of sufficient expressive power contains true statements unprovable within it. Applied to AI: any epistemically bounded system—symbolic, statistical, or hybrid—cannot guarantee the completeness of its own representational domain.

Axiom II — Computational Indeterminacy *(Turing, 1936)*

No computational system can reliably determine whether an arbitrary computation will terminate or predict the full consequences of its own outputs. Human oversight is not merely politically convenient—it is structurally necessary.

Axiom III — Tacit and Distributed Knowledge *(Hayek, 1945; Polanyi, 1966)*

A vast and irreducible portion of human knowledge is local, distributed, embodied, and non-articulable. No formal system trained on articulable data can capture the tacit knowledge substrate of human social coordination. This gap cannot be closed by larger training corpora because the knowledge in question is inarticulable in principle.

Axiom IV — The Non-Aggregability of Value *(Arrow, 1951)*

No procedure exists that can transform individual preferences into a coherent collective objective while satisfying minimal fairness conditions. Society has no objective function. Any AI system operating in a social domain must impose a value structure, accept inconsistency, or optimize a proxy—and no escape from this trilemma exists.

Axiom V — The Irreducibility of Agency *(derived)*

Value specification and the acceptance of outcomes require an agent for whom something is at stake. The human is not an obstacle to be engineered around. The human is the necessary condition for the AI's activity to be meaningful at all.

A complementary non-algorithmic proof runs parallel to the five axioms. All AI implementations are algorithms; genuine creativity is provably non-algorithmic; therefore the non-algorithmic dimensions of human cognition—including the flash of genius that drives paradigm shifts—are permanently beyond the reach of any AI system regardless of scale or architecture. This argument, developed by Marks (2022) drawing on results in computability theory, closes the design space against the emergence thesis and against claims of strong AI or AGI.

BIBLIOGRAPHY

References

Primary Sources

Arrow, Kenneth J. Social Choice and Individual Values. New York: John Wiley & Sons, 1951.

Berlin, Isaiah. "Two Concepts of Liberty." In Four Essays on Liberty. Oxford: Oxford University Press, 1969.

Buchanan, James M., and Gordon Tullock. The Calculus of Consent: Logical Foundations of Constitutional Democracy. Ann Arbor: University of Michigan Press, 1962.

Gödel, Kurt. "On Formally Undecidable Propositions of Principia Mathematica and Related Systems." Monatshefte für Mathematik und Physik 38 (1931): 173–198.

Hayek, Friedrich A. "The Use of Knowledge in Society." American Economic Review 35, no. 4 (1945): 519–530.

Hayek, Friedrich A. The Constitution of Liberty. Chicago: University of Chicago Press, 1960.

Hayek, Friedrich A. Law, Legislation and Liberty. 3 vols. Chicago: University of Chicago Press, 1973–1979.

Marks, Robert J. Non-Computable You: What You Do That Artificial Intelligence Never Will. Seattle: Discovery Institute Press, 2022.

Polanyi, Michael. The Tacit Dimension. Garden City, NY: Doubleday, 1966.

Polanyi, Michael. Personal Knowledge: Towards a Post-Critical Philosophy. Chicago: University of Chicago Press, 1958.

Scott, James C. Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. New Haven: Yale University Press, 1998.

Searle, John R. "Minds, Brains, and Programs." Behavioral and Brain Sciences 3, no. 3 (1980): 417–424.

Turing, Alan M. "On Computable Numbers, with an Application to the Entscheidungsproblem." Proceedings of the London Mathematical Society, Series 2, 42 (1936): 230–265.

Computability Theory and the Limits of AI

Bringsjord, Selmer, Paul Bello, and David Ferrucci. "Creativity, the Turing Test, and the (Better) Lovelace Test." Minds and Machines 11, no. 1 (2001): 3–27.

Chaitin, Gregory J. The Limits of Mathematics. Singapore: Springer, 1998.

Penrose, Roger. The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford: Oxford University Press, 1989.

Penrose, Roger. Shadows of the Mind: A Search for the Missing Science of Consciousness. Oxford: Oxford University Press, 1994.

Economics and Political Economy

Buchanan, James M. "Public Choice: Politics Without Romance." Policy: A Journal of Public Policy and Ideas 19, no. 3 (2003): 13–18.

Goodhart, Charles. "Problems of Monetary Management: The U.K. Experience." In Inflation, Depression, and Economic Policy in the West, edited by Anthony Courakis. Totowa, NJ: Barnes and Noble, 1981.

Kahneman, Daniel. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011.

Kahneman, Daniel, and Amos Tversky. "Prospect Theory: An Analysis of Decision under Risk." Econometrica 47, no. 2 (1979): 263–291.

Madison, James, Alexander Hamilton, and John Jay. The Federalist Papers. 1788. Edited by Clinton Rossiter. New York: Signet Classics, 2003.

Montesquieu, Charles-Louis de Secondat. The Spirit of the Laws. 1748. Translated by Anne M. Cohler, Basia C. Miller, and Harold S. Stone. Cambridge: Cambridge University Press, 1989.

Sen, Amartya. Collective Choice and Social Welfare. Expanded ed. Cambridge, MA: Harvard University Press, 2017.

Zuboff, Shoshana. The Age of Surveillance Capitalism. New York: PublicAffairs, 2019.

Decision Science and Cognitive Architecture

Boyd, John R. "A Discourse on Winning and Losing." Unpublished briefing, Air Force Research Laboratory, 1987.

Fuster, Joaquin M. The Prefrontal Cortex. 5th ed. London: Academic Press, 2015.

Gold, Joshua I., and Michael N. Shadlen. "The Neural Basis of Decision Making." Annual Review of Neuroscience 30 (2007): 535–574.

Kanerva, Pentti. Sparse Distributed Memory. Cambridge, MA: MIT Press, 1988.

Winner, Langdon. "Do Artifacts Have Politics?" Daedalus 109, no. 1 (1980): 121–136.

On Artificial Intelligence and Large Language Models

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021): 610–623.

Bommasani, Rishi, et al. "On the Opportunities and Risks of Foundation Models." arXiv preprint arXiv:2108.07258 (2021).

Christiano, Paul, et al. "Deep Reinforcement Learning from Human Preferences." Advances in Neural Information Processing Systems 30 (2017).

Marcus, Gary, and Ernest Davis. Rebooting AI: Building Artificial Intelligence We Can Trust. New York: Pantheon Books, 2019.

Nagel, Thomas, and James R. Newman. Gödel's Proof. Revised ed. New York: New York University Press, 2001.

Russell, Stuart. Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking, 2019.

Series Volumes — Milestone Planning and Research, Inc.

Aaron, John. Human Relevance in the Age of Induction: Monograph I — Enterprise Deployment, Epistemic Delay, and Governance. Milestone Planning and Research, Inc., 2026.

Aaron, John and Golovnya, Mikhail. Human Relevance in the Age of Induction: Monograph II — Hybrid Cognition and Hybrid Falsification. Milestone Planning and Research, Inc., 2026.

Aaron, John. Human Relevance in the Age of Induction: Monograph IV — Competitive Strategy in AI-Saturated Markets. Milestone Planning and Research, Inc., 2026.

Aaron, John. The Probability Advantage: Strategic AI Deployment for the Inductive Enterprise. Milestone Planning and Research, Inc., 2026.