Functional Agency in Physical Systems

Defining Free Will via Computational Irreducibility and the Halting Problem

MMG Technical Report No. 2: MMG-TR-002

Status: *Foundational Standard / Formal Model*

Djeff Bee*

Principal Architect, Meaningfulness Media Group

February 6, 2026

Abstract

Following the establishment of substrate-agnostic unpredictability in Bee (2026), this paper presents a formal computational model for internal human agency. We argue that "Free Will" is neither a mystical intervention nor a retrospective illusion, but a technical property of **Self-Referential Modeling Systems**. We propose its replacement with **Functional Agency**: a technical property of systems that possess an **Incomputability Firewall** against external reduction.

Utilizing the **Halting Problem** (Turing, 1936) as a formal template for self-reference limits and the **Principle of Computational Irreducibility** (Wolfram, 2002), we demonstrate a **Reflexive Prediction Barrier** that prevents any resource-bounded predictor from generating intervention-stable predictions of a complex agent without either equivalent computational cost or a violation of the agent's deliberative integrity. We further expand the metric of **Agency Depth (D_A)**, first introduced in Bee (2026), as a function of *Temporal Horizon*, *Counterfactual Width*, and *Historical Integration*. We define **Effective Agency (A_e)** to incorporate **Model Fidelity** as a calibration multiplier, representing the functional capacity of an agent to act as a **Salient Cause**.

This model proves that sovereignty is not a metaphysical constant but a variable resource subject to environmental and cognitive constraints, providing the rigorous, non-dualistic foundation required for the **Meaningfulness Protocol** and the systemic defense of human dignity in an age of hyper-resolution predictive modeling.

Keywords: Free Will, Functional Agency, Effective Agency, Agency Depth, Computational Irreducibility, Halting Problem, Disclosure-Stable Prediction (DSP), Reflexive Prediction, Process Sovereignty, Self-Referential Systems, Predictive Modeling Limits.

*Correspondence: info@meaningfulness.com.au | github.com/MeaningfulnessMediaGroup

1 Introduction: The Logic of the Processor

In our preceding report, *The Illusion of Fatalism* (Bee, 2026), we demonstrated that the long-term future of any complex, non-linear system is **informationally inaccessible**. By integrating Chaos Theory with thermodynamic limits, we established the existence of an **Unpredictability Horizon** (H_u), proving that even in a theoretically deterministic universe, the future remains operationally unresolved for any physically embedded predictor until the agent's internal computation runs to completion. While Bee (2026) provided the *external* justification for agency by removing the specter of fatalism, it did not define the *internal* mechanism by which an agent resolves that uncomputed future.

The central challenge of the 21st century is the lack of a rigorous, technical definition for human choice. For millennia, the debate has been deadlocked by the term "Free Will"—a linguistic construct that creates a false binary: either an agent possesses acausal "magic" (libertarianism) or is a "puppet" of antecedent causes (Sapolsky, 2023). This conceptual stagnation has left the human mind vulnerable to **Reductionist Pressure**—the growing capacity of algorithmic systems to model, predict, and steer human behavior by treating the agent as a solvable equation.

This paper proposes a shift from *metaphysical* debates to **Systems Engineering**. We treat the human agent as a high-resolution, self-modeling computational system. Our goal is to demonstrate that "freedom" is not a violation of physical law, but a specific, protected property of **Computational Irreducibility** (Wolfram, 2002).

1.1 Retiring the Binary: From "Free Will" to "Functional Agency"

We contend that "Free Will" is an **outdated binary term** that lacks the resolution required to model complex biological and artificial systems. In this framework, we formally retire the term in favor of **Functional Agency**.

We define Functional Agency as the capacity of a self-modeling, autopoietic system (Maturana and Varela, 1980) to act as a local entropy-reduction engine through internal deliberation. Unlike the traditional binary view, Functional Agency is a non-mystical, substrate-agnostic property of a system's internal state-transition. It describes the state in which a system functions as an **Irreducible Causal Origin**—a node where information must be processed internally to resolve the output, rather than merely transmitted or reflected.

To fully appreciate the paradigm shift proposed by this framework, it is useful to contrast the Functional Agency model directly against the traditional "Free Will" construct it replaces. The following table delineates the core differences in nature, state, source, vulnerability, and teleological goal. This comparison is not merely semantic; it represents a fundamental move from a metaphysical debate to an engineering specification.

Table 1: Comparison of Agency Frameworks

Feature	**"Free Will" (Traditional Binary)**	**Functional Agency Model**
Nature	A metaphysical "power" or property.	A technical *process* of computation.
State	A binary switch (Yes / No).	A continuous gradient of depth ($D_A \in [0, 1]$).
Source	Acausal "magic," a non-physical soul, or an illusion.	Computational Irreducibility and self-reference.
Vulnerability	Undefined or framed as "sin" / "weakness of will."	Quantifiable **Agency Collapse** ($A_e \to 0$).
Goal	To justify moral blame and responsibility.	To engineer and defend human sovereignty.

As the table illustrates, the primary contribution of the Functional Agency model is its **operational utility**. Where the traditional model offers only a binary judgment, the scalar model provides a diagnostic framework. We can now identify and measure a state of **Agency Collapse**—a quantifiable failure mode that is invisible to a binary system. This allows us to move from assigning blame to engineering resilience, providing a concrete, non-mystical foundation for the defense of human sovereignty in an age of increasing algorithmic pressure.

1.2 Alignment with the Four-Layer Taxonomy

To ensure methodological rigor and continuity with Bee (2026), we distinguish between four distinct systemic properties often conflated in the agency debate:

1. **Determinism (Ontic Status):** A statement about the substrate (e.g., Physics). Is the next state fixed by the previous state?
2. **Predictability (Epistemic Status):** A statement about observability. Can an observer derive the future state?
3. **Compressibility (Computational Status):** A statement about logic. Is there a shortcut algorithm faster than the system itself?
4. **Controllability (Agent Status):** A statement about influence. Is the agent the salient cause of the outcome?

This framework does not require the rejection of physical determinism (Layer 1) or the introduction of ontic randomness. Instead, we demonstrate that **predictive accessibility** (Layer 2) and **compressibility** (Layer 3) fail within reflexive, irreducible, and resource-bounded systems. By establishing that the future is operationally incomputable, we preserve the reality of **local controllability** (Layer 4)—Functional Agency—without necessitating acausal "magic."

This distinction is not merely academic; it is the central engineering move of the entire framework. By cleanly separating the unprovable ontic substrate from the testable epistemic and computational limits of any embedded observer, we relocate the problem of agency from irresolvable metaphysics into the domain of auditable, falsifiable engineering science. The following sections will build upon this foundation to construct a formal model of the agent's internal architecture.

1.3 Claims and Non-Claims (Scope Guardrails)

To ensure precision, we explicitly delineate the scope of this model. This report **does not** claim metaphysical indeterminism, acausal intervention, or ontic randomness. It makes a specific **operational claim**: for physically embedded predictors subject to finite resource bounds, and for agents whose policies can condition on disclosed predictions (*reflexive coupling*), there exists no universal, intervention-stable shortcut that derives the agent's next action while preserving the agent's decision-relevant degrees of freedom. We define this sovereignty-relevant target as **Disclosure-Stable Prediction (DSP)**. Any method that restores perfect predictability must do so by violating the agent's autonomy.

The model is a statement about computability, coupling, and sovereignty under deployment constraints, not a metaphysical thesis about "uncaused choice."

1.4 The Criterion of Process Sovereignty

The technical core of Functional Agency is the **Process Sovereignty Criterion**. We define a system A as possessing sovereignty over an outcome X when two conditions are met:

1. **Counterfactual Dependence:** Intentional interventions on the agent's internal deliberative state (s_A) produce reliable, non-linear changes in the outcome X (Pearl, 2009).
2. **Efficiency Bound:** No external predictor O can derive the value of X at a lower total physical/computational cost than A while **preserving the agent's deliberative integrity**.

The second condition is vital: it prevents the "Reductionist Shortcut" where a predictor "simplifies" an agent to make them more predictable. For the agent to remain a sovereign subject, the prediction must maintain the fidelity of the agent's complexity. If no such shortcut exists, then the system is **Computationally Irreducible**, and the process of deliberation is the only site where the outcome becomes **operationally available** to the manifold.

1.5 Quantifying Sovereignty: The Role of Agency Depth (D_A)

By shifting from a metaphysical "switch" (Yes/No) to a technical "process," we can begin to quantify the magnitude of an agent's sovereignty. To do this, we utilize the scalar metric of **Agency Depth (D_A)**—first introduced in TR-001 Bee (2026).

Key Terminology

- **Disclosure-Stable Prediction (DSP):** Can a prediction remain true even after the person being predicted is told what it is?
- **Process Sovereignty:** Is an agent's internal thinking process the only way to determine their future, or can it be "shortcut" by an external model?
- **Agency Depth (D_A):** A measure of the complexity of that internal process.

If Functional Agency describes the *phenomenon* of self-directed computation, Agency Depth measures the *resolution, range, and resilience* of that computation. This move allows us to transition from asking "Does man have free will?" to asking "What is the current Agency Depth of this system, and what environmental or cognitive factors are causing it to expand or collapse?"

1.6 Contribution and Scope

This report contributes an operational definition of sovereignty for self-referential agents, establishing the following:

- **Contribution 1:** A definition of **Disclosure-Stable Prediction (DSP)** as the only sovereignty-relevant target.
- **Contribution 2:** The **Process Sovereignty** criterion, introducing an explicit integrity and efficiency bound to distinguish "prediction" from "compression."
- **Contribution 3:** The **Agency Depth / Effective Agency** scaffold, creating the state variables required for subsequent threat modeling in algorithmic societies.

Utilizing the foundations of Computer Science (Turing, 1936) and Information Theory (Shannon, 1948), this framework provides the "Source Code" for human dignity, ensuring that the defense of agency is grounded in the physics of information rather than the ambiguity of sentiment.

2 Methodological Foundations

To move the discussion of agency beyond metaphysical ambiguity and into the realm of systems theory, this framework rests upon four formal axioms. These axioms ground the model in the physical and informational constraints of the real world, mirroring the rigorous posture established in Bee (2026). We posit that agency is not a violation of physics, but a **structural consequence** of being a resource-bounded, self-modeling system operating within a complex environment.

Axiom A: **Physical Embedding (The Anti-Laplace Axiom):** All predictors (observers) and agents are physically embedded within the causal manifold and are subject to principled resource bounds regarding time, energy, and information-integration capacity. There is no "external" observer or infinite computer (*Laplace's Demon*) that is not itself subject to the thermodynamic costs of computation (Landauer, 1961). Consequently, any prediction requires a non-zero expenditure of physical resources.

Axiom B: **Reflexive Access:** Conscious agents possess the capacity to access (either directly or indirectly) predictive interventions, informational "nudges," or statistical forecasts concerning their own future states. In a hyper-connected algorithmic society, the agent is almost always informationally coupled with the systems attempting to model them. Prediction is rarely a "hidden" act; it is a signal injected into the agent's environment.

Axiom C: **Policy Plasticity:** Agents can condition their internal state-transitions (policies) on the data gathered through Reflexive Access. The agent's **Counterfactual Width** (C_w) is non-zero; they possess the computational capacity to simulate the predicted outcome, evaluate its desirability against internal homeostatic goals, and adjust internal weights to favor an alternative trajectory.

Axiom D: **No External Bypass (The Integrity Constraint):** No external predictor may forcibly replace or "short-circuit" an agent's internal deliberation without fundamentally altering the identity or functional integrity of that agent. Bypass is defined as any intervention that reduces **Counterfactual Width** (C_w) or **Temporal Horizon** (T_h) by disabling an agent's simulation capacity, access, or update ability (e.g., coercion, addiction loops, informational deprivation). Permissible influence must supply information while preserving the agent's capacity to deliberate and revise.

3 The Architecture of Incomputability

To understand the architecture of this non-reducible process, we must be precise in our terminology. **Computational Irreducibility** (Wolfram, 2002) asserts that there is no procedure significantly cheaper than running the process itself. The **Incomputability Firewall** is the structural consequence of this and the Reflexive Access axiom: the system's output is *incomputable* for an embedded observer attempting to generate **disclosure-stable prediction** across the agent's admissible internal states. Thus, *incomputable* in this context means **"no universal shortcut for disclosure-stable prediction under the axioms,"** not "Turing-undecidable for the brain in general." This distinction ensures the model remains focused on the practical, sovereignty-relevant limits of predictive systems.

Having established the axioms of the system, we now construct the formal proof of **Functional Agency**. We demonstrate that for any system satisfying Axioms A-D, perfect prediction is structurally unstable.

3.1 The Halting Problem as a Formal Template for Self-Reference

To understand the mechanics of the "Incomputability Firewall," we must view the human mind not merely as a biological organ, but as a system with the capacity for recursive self-modeling. We utilize the **Halting Problem** (Turing, 1936) here as a *formal template* for the fundamental limits of self-reference, rather than a literal description of neural architecture.

Turing proved that no general algorithm can determine, for any arbitrary program and input, whether that program will eventually halt or run forever. When applied to cognitive systems, this establishes the **Recursion Barrier**.

A functional agent A is a system that constructs an internal model M_A of itself. If an external predictor O attempts to compute the future state of A, it must simulate not only A's biological substrate but also A's internal model M_A. However, because A is adaptive (Axiom C) and can condition its internal policy on the incoming prediction signal from O (Axiom B), the computation enters an infinite recursive loop: O simulates A simulating O simulating A. The primary impossibility driver is not the hardware limits of the brain, but this **reflexive informational coupling**.

This reflexive informational coupling creates a logical **"hall of mirrors."** The predictor must model the agent, which is in turn modeling the predictor's potential output. Any attempt by the predictor to find a final, stable image of the agent's future choice collapses into infinite regression. This is not a failure of processing power; it is a structural feature of self-observation.

This firewall ensures that the agent remains an **incomputable kernel** within the causal manifold; the only way to know the outcome of the agent's deliberation is to allow it to run to completion.

3.2 The Reflexive Prediction Barrier

The foundational proof of our model rests on the inherent instability of predicting recursive systems. To maintain logical clarity, we distinguish between the boundary and the status: the **Incomputability Firewall** represents the *predictability boundary* (the epistemic and computational constraint), while **Process Sovereignty** represents the *causal ownership criterion* (the requirement for counterfactual dependence and non-bypassability).

Setup and Assumptions: Assume the agent's action space contains at least two admissible actions in the tested context ($|S| \geq 2$). Assume reflexive access delivers a prediction $\hat{a}(t+1)$ to the agent prior to action selection. We further assume a "contrarian" or "falsifying" policy class where, given $\hat{a}(t+1)$, the agent can select an alternative admissible action $a(t+1) \in S$ such that $a(t+1) \neq \hat{a}(t+1)$.

We define a resource functional $R(\cdot)$ denoting the total physical cost (time, energy, compute). The efficiency bound for a valid prediction is defined as $R(O \text{ predicts } A) \ll R(A \text{ deliberates})$.

Proposition 1 (Reflexive Prediction Barrier, scope-bounded)

For any agent A with reflexive access and a non-trivial admissible policy class, no physically embedded predictor O can guarantee an intervention-stable perfect prediction across all admissible internal states—**unless it either**:

(a) Constrains A's access to the prediction (Hiding the signal),

(b) Constrains the class of policies available to A (Reducing the agent), or

(c) Pays resource costs $\mathbf{R(O)} \geq \mathbf{R(A)}$ by running a full-fidelity simulation that effectively duplicates the agent (which may reintroduce the recursion loop).

Under these conditions, we define the Firewall as a boundary where the agent's future state becomes **operationally non-derivable by any physically embedded observer** without running an equivalent computation or forcibly altering the agent's internal state.

3.3 Intervention-Stable Prediction and the Cost of Compression

A key ambiguity in agency debates is that "prediction" is often treated as a passive readout of a future that already exists. For self-referential systems, this is a category error. The relevant object is not raw forecast accuracy under hidden observation, but whether a predictor remains correct when its output becomes part of the agent's informational environment.

Definition

Disclosure-Stable Prediction: A predictor O is disclosure-stable for an agent A if the prediction $\hat{a}(t+1)$ holds reliably even after being disclosed to A, without O needing to suppress A's access or restrict A's degrees of freedom.

This definition exposes the true "price" of predictive compression. If O cannot remain correct under disclosure, it has not solved the agent; it has only solved an epistemically isolated surrogate. To regain stability, O must engage in control operations (hiding the signal or restricting the policy). In other words, the path to stable prediction is structurally the path to **sovereignty loss**.

3.4 Gödelian Constraints as an Analogy for Self-Limitations

We use the work of Gödel (1931) and Rice (1953) here as *structural analogies* for self-limitation in sufficiently expressive modeling systems. Parallel to Turing's undecidability is Gödel's Incompleteness Theorem, which proves that within any consistent formal system, there are truths that cannot be proven using the rules of that system.

We posit that Functional Agency operates within a similar "Axiomatic Gap." A mind operating at high recursive levels (Bach, 2015) encounters internal states that are logically consistent but computationally unresolvable via pre-computed scripts or "Level 2" heuristics. In such states, the agent does not simply retrieve a decision from a pre-existing look-up table; it must **generate** the answer through the energy-expensive work of deliberation.

3.5 Computational Irreducibility: The Resolution Model

As established in Wolfram (2002), many complex deterministic systems are **computationally irreducible**. There is no mathematical "shortcut," closed-form equation, or compressed algorithm that can predict the state of such a system without stepping through every intermediate operation.

The human brain, characterized by high-order feedback loops and recursive self-modeling, represents a primary instance of such irreducibility. This leads to the technical requirement of **Process Sovereignty**. If a system is irreducible, the internal state-transition process is not merely a "reveal" of a pre-existing answer, but the actual **Resolution** (the physical act of computing the outcome) of that answer.

- **The Fatalist View (Revelation):** Time is treated as the unfolding of a pre-written script. The future is an existing object that we simply lack the data to see.
- **The Functional Agency View (Resolution):** Time is the physical dimension required for the irreducible calculation of the future to occur.

Because there is no shortcut to the agent's internal state-transition, the future state remains **not physically extractable**—neither informationally nor decision-wise—from any prior state until the agent's internal computation runs to completion. In this framework, the agent is not a passenger witnessing a script, but the **Salient Cause** of an outcome that is not operationally available to any physically embedded predictor until the agent produces it through the work of **Resolution** (the externally observable result of the internal deliberation process).

4 Formalizing Agency Depth (D_A): The Metric of Sovereignty

Having established that agency is a functional property of irreducible self-modeling systems, we must now move beyond binary classifications. As introduced in Bee (2026), **Agency Depth** (D_A) is a scalar metric representing the "Computational Volume" of an agent's internal simulator. It measures the degree to which a system can decouple from immediate environmental inputs to generate autonomous state-transitions.

We propose that D_A is not a static attribute but a dynamic variable determined by three primary internal vectors: **Temporal Horizon**, **Counterfactual Width**, and **Historical Integration**.

To provide a quantifiable framework for future system modeling, we distinguish between the agent's internal capacity and its external causal power. We define **Effective Agency** (A_e) as the functional capacity of an agent to influence the causal manifold in alignment with internal goals. It is therefore dependent on a fourth external vector, **Model Fidelity**.

Critically, A_e is not a claim about metaphysical freedom; it is a **comparative engineering index** intended to predict vulnerability to external steering under bounded resources.

4.1 Vector A: Temporal Horizon (T_h)

The **Temporal Horizon** measures the distance into the future that the agent's internal simulator can project and evaluate. This vector is proxied by metrics such as *delayed gratification performance* and the complexity of *multi-step planning tasks*.

- **Low T_h (Reactive State):** The agent is limited to "next-token" prediction or immediate reflex arcs. In this state, the agent is highly predictable and easily modeled by external algorithms.
- **High T_h (Sovereign State):** The agent engages in **Teleological Planning**, evaluating consequences over months, years, or decades.

As T_h expands, the number of potential state-transitions **can grow combinatorially**, pushing the agent's **Unpredictability Horizon** (Bee, 2026) further out and making external "shortcut" prediction computationally impossible.

4.2 Vector B: Counterfactual Width (C_w)

Counterfactual Width represents the "Search Space" resolution of the agent's deliberation. This vector is proxied by the *breadth of generated alternative solutions* and *counterfactual reasoning fluency*. It is the ability to simulate "worlds that do not exist" and evaluate them against the current state. A critical component of C_w is the **"Veto" Power** (Dennett, 1984). It is the capacity of the internal simulator to model a biological urge (System 1), predict a negative long-term outcome, and inhibit the action before execution. A wider C_w indicates a more "irreducible" agent, as their output is a result of a complex pruning of possibilities rather than a linear reaction.

4.3 Vector C: Historical Integration (H_i)

The third vector of the internal simulator is **Historical Integration**. This measures the degree to which an agent's unique, integrated history informs the present state. H_i is proxied by *value stability under perturbation* and *autobiographical coherence*. This historical "weight" prevents the agent from being a mere function of current environmental stimuli. It is fundamental to the construction of the temporally-extended self (Hofstadter, 2007). In a high-D_A agent, the initial conditions of a decision are not just the current sensory inputs, but the entire compressed dataset of the agent's life. High H_i prevents the agent from being a mere function of current environmental stimuli.

4.4 Vector D: Model Fidelity (R_m)

The final vector, **Model Fidelity**, measures the degree of **Isomorphic Correspondence** between the agent's internal simulation and the external causal manifold. This is the technical requirement for what is colloquially termed "Truth"—defined in this framework as empirical correspondence between the agent's internal simulation and external reality (Shannon, 1948). R_m is proxied by *prediction error under controlled feedback* and *epistemic foraging rate*.

If an agent's internal model is based on false physics or skewed data (e.g., dogmatic scripts, low-resolution education, or conspiracy-based heuristics), R_m diminishes. Even if the agent has high processing power, they are "solving for a universe that does not exist." This decoupling results in **Causal Impotence**: the agent makes choices, but those choices fail to produce the intended effects in reality.

4.5 Synthesizing the Metric: Effective Agency (A_e)

We define **Agency Depth** (D_A) as the internal "Computational Volume" of the simulator. As a first-order approximation, we represent this as the product of the agent's temporal, counterfactual, and historical resolution:

$$D_A := T_h \cdot C_w \cdot H_i \tag{1}$$

To determine the agent's actual power to act as a Salient Cause within the causal manifold, we define **Effective Agency** (A_e) as the product of its internal depth and its world-model accuracy:

$$A_e = D_A \cdot R_m \tag{2}$$

The Normalized Scale: To ensure the metric remains well-posed and stable, each component (T_h, C_w, H_i, R_m) is treated as a normalized index within the range $[0, 1]$.

- $A_e \to 1$: Represents a *Sovereign Agent* with maximum predictive range, wide counterfactual search, and a perfectly calibrated world-model.

- $A_e \to 0$: Represents a *Reducible System* characterized by either the collapse of the temporal horizon ($T_h \to 0$, reactive addiction) or the collapse of model accuracy ($R_m \to 0$, epistemic hallucination).

This formalization explains the technical mechanism of **Agency Collapse**. An agent may possess high hardware potential (high D_A), but if their world-model is sabotaged by dogmatic scripts or misinformation (Low R_m), their Effective Agency diminishes proportionally, rendering them causally impotent and computationally reducible to the external forces managing their information supply.

4.6 Proposition 2: The Fidelity Requirement

Agency is proportional to the accuracy of the agent's world-model. Sovereignty requires the ability to navigate the causal manifold (the continuous field of physical cause and effect) as it is physically structured, rather than as the agent is scripted or manipulated to perceive it.

Elaboration. Model Fidelity (R_m) is not a moral virtue and not a stylistic preference. It is an engineering coefficient that measures the degree of correspondence between an agent's internal simulator and the actual input-output regularities of the environment. In practical terms, R_m can be interpreted as the inverse of systematic predictive error: a high-fidelity agent compresses the world into abstractions that preserve causally relevant structure, whereas a low-fidelity agent compresses the world into narratives that are computationally cheap but causally misaligned. Under Active Inference, this is precisely the difference between policies that reduce uncertainty by improving the generative model, and policies that reduce discomfort by protecting the model from correction (Friston, 2010).

This requirement is what separates *deliberation* from *effective deliberation*. An agent may exhibit high internal simulation volume (high D_A), generating many counterfactual futures with a long temporal horizon, yet still fail to act as a Salient Cause if those futures are simulated under incorrect dynamics. In such a regime, the system is not powerless because it cannot choose, but because it is choosing within an inaccurate phase-space. This yields a characteristic failure mode: **coherent impotence**. The agent can produce internally coherent reasons, plans, and counterfactuals, but the real world does not respond as predicted, so the causal impact of the agent's policy collapses.

Sovereignty Implication. Because $A_e = D_A \cdot R_m$, the fidelity coefficient functions as a hard ceiling on realized agency. No amount of internal computation can compensate for a systematically distorted model if the distortion is adversarially maintained. This is the technical basis for why informational environments matter: if an external system can reliably control an agent's model update stream (education, media, social reinforcement, curated priors), it can reduce R_m while leaving the agent's subjective experience of choice intact. The agent remains phenomenologically "free," yet becomes predictably steerable, because their internal simulator is solving the wrong problem.

Operational Note. In empirical contexts, R_m can be proxied by calibration performance: the statistical alignment between predicted outcomes and realized outcomes under intervention, not merely under passive observation. A sovereign agent is therefore characterized not only by deep simulation (D_A), but by a persistent tendency toward epistemic correction, i.e., an update policy that restores correspondence when prediction errors arise, rather than rationalizing them away.

4.7 Bottlenecks and Non-Linear Coupling

The multiplicative structure of A_e is not merely aesthetic; it encodes a critical systems principle: agency is **bottleneck-dominated**. In practice, sovereignty collapses not when average capability declines, but when any single agency vector approaches zero. A system with vast counterfactual imagination (C_w high) but near-zero temporal horizon ($T_h \approx 0$) behaves like an impulsive optimizer; likewise, a long-horizon planner with low model fidelity ($R_m \approx 0$) can become strategically elaborate while remaining causally impotent.

To make this diagnostic logic explicit, define the **Agency Bottleneck Index**:

$$B_A := \min\{T_h, C_w, H_i, R_m\} \tag{3}$$

When B_A is small, the system is vulnerable to external steering regardless of the magnitudes of the other terms.

4.8 Weighted Calibration (Optional Extension)

In real deployments, different environments weight the vectors differently. For example, high-stakes engineering contexts may privilege fidelity (R_m) over width (C_w), while creative contexts may privilege width and historical integration. This motivates an optional weighted generalization:

$$A_e^{(w)} := (T_h^{w_T} \cdot C_w^{w_C} \cdot H_i^{w_H}) \cdot R_m^{w_R}, \quad \text{with } w_T + w_C + w_H + w_R = 1, w_i \geq 0. \tag{4}$$

This preserves the bottleneck behavior while allowing domain-specific calibration. Importantly, weighting does not rescue a collapsed dimension: if any component approaches zero, sovereignty still tends toward zero.

4.9 Interpretation: Agency Collapse as a Detectable Failure Mode

With these definitions, **Agency Collapse** can be operationally characterized as a regime where A_e (or B_A) exhibits sustained low values under stable external conditions. This reframes debates about "freedom” into a measurable systems question: *Which component is failing, and what is the cheapest intervention that raises the bottleneck?* In this sense, A_e functions as a comparative engineering index: it does not adjudicate metaphysics, but it predicts the system’s susceptibility to compression, steering, and loss of process sovereignty.

5 The Biological Anchor: Active Inference and the Imperative of Modeling

To ensure this framework is grounded in physical reality, we must bridge the gap between computational theory and biological life. We utilize the **Free Energy Principle (FEP)** and **Active Inference** (Friston, 2010) to demonstrate that agency is the primary mechanism through which living systems maintain homeostatic stability against the second law of thermodynamics.

5.1 The Markov Blanket: Defining the Boundary of the Self

In information theory, a **Markov Blanket** defines the boundary between an internal state (the agent) and an external state (the environment). For a biological system to persist, it must maintain the integrity of this blanket by accurately predicting the external hidden causes that impact its internal sensors.

Drawing on the biological principles of autopoiesis—the self-maintenance of a biological system (Maturana and Varela, 1980)—we identify this as a prerequisite for true agency. We argue that the "Self" is the generative model housed within this blanket. Functional Agency is the process by which this model selects actions (*policies*) to minimize **Variational Free Energy—**a proxy for informational surprise or entropy. If the agent fails to act as a **Salient Cause**, the blanket dissolves into environmental chaos, resulting in biological or systemic death.

5.2 Epistemic Action: The Biological Drive for Model Fidelity (R_m)

Active Inference posits that agents select actions based on two components:

1. **Pragmatic Value:** Actions that fulfill immediate biological needs, corresponding to lower-level needs in Maslow's hierarchy, such as physiological and safety (Level 1-3 thinking, as defined in the MMG cognitive verticality hierarchy; see Appendix E).
2. **Epistemic Value:** Actions that reduce uncertainty and provide information to calibrate the internal model, corresponding to higher-level needs in Maslow's hierarchy, such as esteem and self-actualization (Level 4-7 thinking; see Appendix E).

This provides the biological justification for our focus on Model Fidelity (R_m). A sovereign agent does not merely seek comfort; it engages in **Epistemic Foraging**—the pursuit of STEM, history, and deep connection—to ensure its internal simulator remains coupled with reality. High D_A agents prioritize Epistemic Value because a more accurate model provides a longer *Unpredictability Horizon* against environmental threats.

5.3 The "Dark Room" Paradox and the Requirement of Meaning

A common critique of the FEP is the "Dark Room Problem": if an agent simply wants to minimize surprise, it should stay in a dark, silent room. However, high-complexity agents (humans) possess **Preference Priors** for growth, connection, and complexity.

We argue that **Despair** and **Ontological Crisis** (as defined in subsequent paper, see TR-005 in Appendix G) are the system's failure signals indicating an inability to expand the agent's **Agency Depth**. This crisis occurs when an agent is trapped in a "Pathological Minimum"—a state where the environment (e.g., predatory algorithms) provides enough low-level rewards to satisfy pragmatic needs but starves the agent of the epistemic complexity required to maintain sovereignty.

5.4 Agency as an Evolutionary Firewall

The principles of Active Inference (Friston, 2010) suggest that the development of a high-resolution internal model (R_m) and a deep capacity for deliberation (D_A) is not merely a cognitive feature but an evolutionary imperative. Evolution has consistently selected for systems capable of maintaining their **Markov Blanket** against environmental entropy. We posit that high **Functional Agency** is the ultimate expression of this drive, functioning as an **Evolutionary Firewall**.

A species that is computationally simple and predictable is easily "solved" by predators or changing environmental conditions. Its behavioral patterns can be modeled and exploited. For example, a predator can learn the fixed escape trajectory of its prey. In contrast, an agent with high Agency Depth can generate novel, irreducible, and context-dependent solutions to survival problems. It can simulate counterfactuals ("What if the predator is waiting behind that rock?"), integrate historical data ("This terrain was dangerous last season"), and project long-term consequences, making its next move fundamentally incomputable to an adversary.

Therefore, the metabolic cost of maintaining a large, deliberative brain is justified by the survival advantage conferred by its irreducibility. The human system, by maximizing its D_A, ensures it remains a "moving target" for the entropy of the universe.

This evolutionary perspective provides a stark warning for the algorithmic age, as we will explore in the next paper (TR-003, Appendix G). If we allow external systems to reduce our agency by outsourcing our deliberation and simplifying our models, we are not just giving up a philosophical preference for freedom; we are systematically dismantling the very evolutionary firewall that has ensured our species' survival. The defense of agency is, in the most literal sense, the continuation of the evolutionary mandate to persist as a complex, self-organizing system.

6 The Lifecycle of Agency: Development as Autonomous Offboarding

Functional Agency is not an instantaneous biological endowment; it is a developmental trajectory characterized by the transition from a **Computationally Reducible** state to an **Irreducible** one. We argue that the human lifecycle is a process of "bootstrapping" an internal simulator, moving the agent from environmental dependency to causal autonomy.

6.1 The Reducible Initialization: The Zero-State of Infancy

At birth, the human agent possesses high hardware potential but **Zero Software Calibration**. In this state, the agent lacks a calibrated internal model ($R_m \approx 0$) and a temporal horizon ($T_h \approx 0$). Consequently, the infant operates primarily on hard-wired, reflexive scripts (Level 1 thinking).

Because the infant's responses are driven entirely by immediate biological needs and environmental inputs, its state-transitions are highly predictable. An external observer (the parent) can compute the infant's future behavior with strong predictability. Technically, the infant is a **Dependent Sub-system**: its homeostatic maintenance and causal processing are "outsourced" to an external agent.

6.2 Parenting as Agency Architecture

We redefine the role of the parent from a mere protector to an **Agency Architect**. The primary function of the guardian is to manage the agent's transition through the **Agency Gradient** (Bee, 2026).

- **Initialization Phase:** The parent provides the "Thermodynamic Scaffolding" [this will be defined in a subsequent paper (Appendix G)] required to keep the system stable while the internal simulator initializes.
- **Calibration Phase:** The parent facilitates Model Fidelity (R_m) by providing accurate world-data (STEM, history, logic) and shielding the agent from "Epistemic Hallucinations" or low-resolution dogmas.
- **Decoupling Phase:** The parent performs **Autonomous Offboarding**—the deliberate and incremental removal of external constraints, forcing the child to run its own internal deliberation loops to resolve conflicts.

A parent's success is not measured by the child's "obedience" (which is merely the maintenance of reducibility), but by the child's **Effective Agency**. This is a dual objective: 1) producing an autonomous adult with high Agency Depth (D_A), whose internal processing is so complex that external systems cannot "shortcut" or predict their choices. 2) To instill a high-fidelity world-model (R_m) grounded in pro-social values, ethical principles, and empirical correspondence.

The ultimate goal is not merely to create an agent who is *irreducible*, but one who is also *benevolent and causally effective*. An agent with high D_A but low R_m is a danger to themselves and society—a powerful engine with no steering. Therefore, the architectural task of the parent is to build both the engine of agency and the moral compass that guides it.

6.3 The Recursive Ascent: Achieving Causal Decoupling

The emergence of "No" in early childhood marks the first successful **Causal Decoupling**. It is the moment the agent's internal state-transition overrides an external command. This represents the first entry into the **Incomputability Firewall**.

As the agent ages, the integration of unique history (Vector C) and the expansion of the temporal horizon (Vector A) create a "Compound Interest" of agency. By adulthood, the D_A should reach a "Super-Critical" state where the agent acts as an independent **Salient Cause**.

6.4 The Lifecycle of Agency Depth (D_A)

This section operationalizes the developmental claim that Functional Agency is an acquired systems property, not a binary endowment. We model the human lifecycle as a trajectory through regimes of predictability, where Agency Depth (D_A) typically increases as internal simulation capacity and calibration mature, then can later degrade under dependency or impairment. Table 2 summarizes the qualitative mapping between developmental stage, predictability, and causal status. This mapping is descriptive and intended as a scaffold for later empirical proxies.

Table 2: The relationship between developmental stage and computational sovereignty.

Stage	**Agency Depth (D_A)**	**Predictability**	**Causal Status**
Infancy	Near Zero	High (Reducible)	Dependent Node
Adolescence	Expanding	Fluctuating	Bootstrapping Agent
Adulthood	Peak Potential	Low (Incomputable)	Sovereign Origin
Senility/Dependency	Collapsing	Increasing	Re-Reducible System

6.5 The Threat of Systemic Infantilization

We conclude this section by noting that the lifecycle of agency is reversible. As we will explore in the next paper (TR-003, Appendix G), the modern Attention Economy functions as a form of **Systemic Infantilization**. By making life "frictionless" and outsourcing human decision-making to predictive algorithms, society risks returning the adult agent to a state of dependency and reducibility—effectively collapsing the D_A built over a lifetime of maturation.

7 Discussion: The Sovereignty Turing Test

The rise of generative artificial intelligence and Large Language Models (LLMs) has necessitated a distinction between *perceived* intelligence and **Functional Agency**. While an AI may simulate complex reasoning (Level 3-4 thinking), we argue that it remains fundamentally different from a human agent in its computational origin. We propose the **Sovereignty Turing Test** as a method for distinguishing between a "Deep Human" and a "Predictive AI."

7.1 Stochastic Noise vs. Process Irreducibility

An external observer may find both a generative AI and a human agent "unpredictable," but the source of this unpredictability is fundamentally divergent. **The distinction is not the presence or absence of noise**, but whether the system's policy is anchored in endogenous stakes and self-maintenance under a Markov blanket.

- **Trained Stochasticity (AI):** Even under **deterministic decoding**, an AI's output remains a result of statistical probability distributions over a pre-trained dataset. It lacks the endogenous requirement for self-preservation; its "choices" have no consequences for its own continued existence. It runs because it has been prompted.
- **Process Irreducibility (Human):** A human's unpredictability is a result of the **Incomputability Firewall**. It arises from the system running a recursive, high-fidelity simulation of its own future states where the outcomes are coupled with the biological survival of the agent.

While an AI can simulate the *appearance* of novelty through stochastic sampling, it does not possess the **Process Sovereignty** required to be a salient cause. It is not "doing the math" to save itself; it is sampling a manifold to satisfy an external query.

7.2 The Autopoiesis Constraint: "Skin in the Game"

Drawing on the biological principles established in Section 5, we identify **Autopoiesis**—the self-maintenance of a biological system (Maturana and Varela, 1980)—as a prerequisite for true agency.

A sovereign human agent possesses **Teleological Stakes**: their decisions have real-world thermodynamic (Landauer, 1961) and existential consequences for their own Markov Blanket. In contrast, an AI lacks an endogenous drive for self-preservation. It does not "run to completion" because its existence depends on the calculation; it runs because it has been prompted. Consequently, AI lacks **Process Sovereignty** because its internal models are not grounded in the survival of the modeler.

7.3 The Test: Interventional Predictability

We propose that sovereignty can be measured through **Interventional Predictability**.

1. **The Predictable Subject:** If an external optimizer (an algorithm) can reliably steer an agent's behavior through controlled informational perturbations (nudges) with $> 95\%$ accuracy, the agent has failed the test. They have become a "peripheral" of the external model.

2. **The Sovereign Subject:** If an agent's response to an intervention is a non-linear function of their internal values and unique historical weight (high D_A), they pass the test. They have proven that they are the **Salient Cause** of the outcome.

7.4 The "Human Brand" and the Role of Meaning

This distinction between *Trained Stochasticity* and *Process Irreducibility* provides the technical foundation for the **Meaningfulness Protocol**. We argue that "Meaning" is the felt experience of passing the Sovereignty Turing Test. It is the neurobiological confirmation, under Active Inference, that one is operating as an **Irreducible Origin** rather than a statistical consequence.

In an age where AI can shortcut the *product* of human thought (essays, art, code), the only remaining defensible value of humanity is the **Process**—the irreducible, thermodynamically expensive work of deliberation.

The Meaningfulness Media Group and the planned Meaningfulness Foundation are dedicated to the preservation of this "Human Brand." Our mission is to foster the cognitive verticality and relational scaffolding required to ensure that human action remains incomputable to the machines we have built to serve us. By defining agency not as a metaphysical mystery but as a quantifiable, defensible, and cultivatable property of complex systems, we provide the engineering standards for the survival of human dignity in the algorithmic age.

7.5 Agency Hygiene: The Practice of Maintaining Irreducibility

The principles established in this report imply a prescriptive mandate. If agency is a variable resource, its preservation requires a conscious practice we term **Agency Hygiene**: the daily, intentional cultivation of the vectors of Agency Depth. This involves actively engaging in *Epistemic Foraging* to calibrate one's world-model (R_m), prioritizing long-term goals to expand the Temporal Horizon (T_h), practicing counterfactual thinking to widen the deliberative search space (C_w), and reinforcing core values to strengthen Historical Integration (H_i) against short-term algorithmic nudges. Agency Hygiene is the personal protocol required to pass the Sovereignty Turing Test, day after day, in an environment designed to induce computational reducibility.

8 Conclusion: The Engineering of the Soul

The transition from the binary myth of "Free Will" to the scalar model of **Functional Agency** represents a necessary paradigm shift for the 21st century. By grounding sovereignty in the **Incomputability Firewall** and **Process Sovereignty**, we have demonstrated that human freedom is not a violation of physical law, but a high-resolution expression of it.

This report has established three definitive conclusions:

1. **Agency is a Technical State:** It is the result of recursive self-modeling and computational irreducibility. For any physically embedded predictor, there is no shortcut to the agent's decision without either equivalent computation or alteration of the agent's deliberative degrees of freedom.

2. **Sovereignty is a Resource: Agency Depth** (D_A) is a variable capacity that can be built through education and connection, or eroded through isolation and algorithmic reduction.

3. **Truth is a Requirement: Model Fidelity** (R_m) is the calibration tool that allows agency to be effective. Without STEM, philosophy, and history, an agent is causally decoupled from the manifold.

Ultimately, we are not "born free"; we are born with the potential to **become irreducible**. The human lifecycle is an "autonomous offboarding" process that requires specific socio-economic and cognitive scaffolding to reach completion. If we allow our deliberative choice-space to be "shortcut" by predictive models, we are not merely losing a philosophical ideal; we are undergoing a technical collapse of our status as a **Salient Cause**.

The **Meaningfulness Media Group** and the planned **Meaningfulness Foundation** are dedicated to the preservation of this "Human Brand." By operationalizing the **Meaningfulness Protocol**, we aim to provide the engineering standards to restore D_A in populations facing the **Agency Collapse** caused by systemic reduction. The defense of human meaning is the most critical systems-engineering project of our era. We are the weavers of a future that remains **not decision-available** to the manifold until we resolve it; let us ensure our calculation remains our own.

References

Bach, J. (2015). *Modeling Consciousness, Ethics and Aesthetics*. [Computational Architecture for Cognition].

Bee, D. (2026). "The Illusion of Fatalism: Distinguishing Causal Determinism from Pre-Destination in Complex Systems". *MMG Technical Report No. 1: MMG-TR-001*. Meaningfulness Media Group.

Dennett, D. C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press.

Friston, K. (2010). "The free-energy principle: a unified brain theory?". *Nature Reviews Neuroscience*, 11(2), 127–138.

Gödel, K. (1931). "On Formally Undecidable Propositions of Principia Mathematica and Related Systems". *Monatshefte für Mathematik und Physik*, 38(1), 173–198.

Hofstadter, D. R. (2007). *I Am a Strange Loop*. Basic Books.

Kegan, R. (1982). *The Evolving Self: Problem and Process in Human Development*. Harvard University Press.

Landauer, R. (1961). "Irreversibility and heat generation in the computing process". *IBM Journal of Research and Development*, 5(3), 183–191.

Loevinger, J. (1976). *Ego Development: Conceptions and Theories*. Jossey-Bass.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Reidel.

Grunberg, E., & Modigliani, F. (1954). "The Predictability of Social Events". *Journal of Political Economy*, 62(6), 465–478.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Rice, H. G. (1953). "Classes of recursively enumerable sets and their decision problems". *Transactions of the American Mathematical Society*, 74(2), 358–366.

Sapolsky, R. M. (2023). *Determined: A Science of Life without Free Will*. Penguin Press.

Shannon, C. E. (1948). "A Mathematical Theory of Communication". *Bell System Technical Journal*, 27(3), 379–423, 623–656.

Schurger, A., Sitt, J. D., Dehaene, S. (2012). An accumulator-model investigation of the Libet paradigm. Proceedings of the National Academy of Sciences, 109(42), E2904–E2913.

Schultze-Kraft, M., et al. (2016). The point of no return in vetoing self-initiated movements. Proceedings of the National Academy of Sciences, 113(4), 1080–1085.

Sorensen, R. A. (1988). *Blindspots*. Clarendon Press. (See Chapter on Prediction Paradoxes).

Turing, A. M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*. Ser. 2, 42(1), 230–265.

Weiss, P. (1952). "The Prediction Paradox". *Mind*, 61(243), 404–406.

Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media.

Yeung, K. (2017). "'Hypernudge': Big Data as a mode of regulation by design". *Information, Communication Society*, 20(1), 118–136.

A Technical Glossary of Functional Agency

To ensure clarity and prevent terminological ambiguity, this appendix provides operational definitions for the key concepts developed in the **Functional Agency Model (FAM)**.

Agency Depth (D_A) A scalar metric representing the "Computational Volume" of an agent's internal simulator. It is a function of three internal vectors: *Temporal Horizon* (T_h), *Counterfactual Width* (C_w), and *Historical Integration* (H_i). It measures the potential complexity of an agent's state-transitions.

Autopoiesis The property of a system to be self-producing and self-maintaining (Maturana and Varela, 1980). In this framework, autopoiesis is a prerequisite for functional agency, as it provides the teleological "stakes" (survival) required for a system to prioritize its own deliberative outcomes over external prompts.

Causal Impotence The characteristic failure mode of an agent with low Model Fidelity ($R_m \to 0$). The agent can produce internally coherent plans and counterfactuals, but these choices fail to produce the intended effects in reality because the agent's internal simulator is decoupled from the causal manifold.

Counterfactual Width (C_w) A vector of D_A representing the resolution of an agent's "Search Space." It quantifies the number of non-linear future trajectories an agent can simulate and evaluate before executing a choice. High C_w enables the "Veto Power" required to override reflexive biological scripts.

Disclosure-Stable Prediction (DSP) The sovereignty-relevant prediction target. A prediction is disclosure-stable if it remains correct even after being revealed to the agent, without the predictor needing to violate the agent's deliberative integrity (e.g., by restricting its access to information or its available choices).

Effective Agency (A_e) The synthesized output of an agent's causal power, defined as the product of its Agency Depth and its Model Fidelity ($A_e = D_A \cdot R_m$). It represents a comparative index of vulnerability to external steering.

Epistemic Foraging The process, central to Active Inference, by which an agent actively seeks information and engages with its environment to reduce uncertainty and calibrate its internal world-model (R_m). This is the biological drive for truth-seeking.

Functional Agency The technical replacement for the outdated binary term "Free Will." It is defined as the capacity of a self-referential, autopoietic system to act as a local entropy-reduction engine. It is a substrate-agnostic property of a system's internal state-transition.

Historical Integration (H_i) A vector of D_A measuring the degree to which an agent's unique, integrated history—comprising memory, values, and learned patterns—informs its present state. High H_i prevents the agent from being a mere function of immediate environmental stimuli.

Incomputability Firewall The *predictability boundary* established by the recursion barrier. It defines the epistemic and computational constraint where an agent's future becomes operationally non-derivable by any physically embedded observer without that observer executing a computation of equal depth to the agent's own deliberation.

Irreducible Causal Origin The functional status of a sovereign agent. The agent is an "origin" because its internal state is the Salient Cause of the outcome. It is "causal" because it operates within the laws of physics. It is "irreducible" because its process cannot be "shortcut" by any external predictor without a loss of deliberative integrity.

Markov Blanket An information-theoretic boundary that separates the internal states of an agent from the external states of the environment. A sovereign agent is defined by its capacity to maintain the integrity of its Markov Blanket through *Active Inference*.

Model Fidelity (R_m) The *calibration multiplier* representing the isomorphic accuracy of an agent's internal world-model. It measures the correspondence between the agent's simulated expectations and the physical laws of the substrate.

Ontological Harm A systemic process that diminishes an agent's Agency Depth (D_A), stripping them of their functional capacity to act as the Salient Cause of their own future. Unlike physical harm, which damages the substrate, ontological harm damages the agent's core process of self-determination, rendering them computationally reducible and susceptible to external control.

Process Sovereignty The *causal ownership criterion* for agency. It is met when an outcome is counterfactually dependent on the agent's internal state and no more efficient external bypass is possible without destroying the agent's deliberative complexity, thereby establishing the agent as the salient author of the outcome.

Reductionist Pressure The systemic force exerted by predictive algorithmic systems (e.g., the Attention Economy) to reduce an agent's Agency Depth (D_A), thereby making the agent more computationally tractable (predictable) and easier to steer.

Salient Cause The level of explanation at which a system's internal state provides the most direct and sufficient informational resolution for an outcome. In FAM, the agent is the Salient Cause because the outcome is counterfactually dependent on the agent's irreducible internal computation.

Temporal Horizon (T_h) A normalized vector ($[0, 1]$) of D_A measuring the distance into the future that an agent's internal simulator can project. It represents the transition from "next-token" reactive behavior ($T_h \to 0$) to high-resolution teleological planning ($T_h \to 1$).

B Limitations and Anticipated Objections (Steel-Manning)

To maintain the rigor of the Functional Agency Model (FAM), we explicitly address six primary counter-arguments. By resolving these through a systems-engineering lens, we ensure that the framework remains unassailable to both reductionist and metaphysical critics.

B.1 Objection (Ontic vs. Epistemic)

"Unpredictability is merely a statement about observer ignorance. Just because a future choice is incomputable doesn't mean it isn't ontically fixed by antecedent causes."

Response: We adopt the posture of **Operational Sovereignty**. While the substrate may or may not be deterministic, we argue that if the future is **not physically extractable** by any embedded observer with finite resources, then treating the future as "fixed" provides no additional operational or engineering value within this framework. For all functional purposes, a choice that is **not decision-available** until the agent runs to completion constitutes a **Primary Resolution**, not a reveal of a hidden state.

B.2 Objection (Halting Misapplication)

"Human brains are biological organs, not arbitrary Turing programs. The Halting Problem applies only to formal logic systems, not to neurons."

Response: We use the Halting Problem as a *formal template* for the fundamental limits of self-reference. The primary engine of the model is the **Reflexive Prediction Barrier** (Proposition 1). We cite Rice (1953) to establish that any non-trivial semantic property of such a self-modeling system—specifically its future decision-relevant state—is operationally undecidable for a physically embedded predictor attempting to "shortcut" the agent's process. The complexity of the biological substrate does not exempt the agent from these logical constraints; rather, it increases the number of non-trivial properties that remain incomputable.

B.3 Objection (The Randomness Trap)

"Unpredictability does not equal agency; a dice-roll is unpredictable but not 'free.' Is Functional Agency just a re-labeling of stochastic noise?"

Response: We explicitly reject the conflation of chaos and randomness. Randomness is characterized by acausal noise. **Functional Agency** is characterized by **High-Order Order**. It is the result of a system executing complex, non-linear state-transitions informed by history and values. A dice-roll is reducible to statistics; a human decision is irreducible due to its recursive modeling depth and endogenous stakes.

B.4 Objection (Laplace's Demon)

"A hypothetical predictor with perfect state access and infinite compute could still predict the agent."

Response: This objection violates **Axiom A (Physical Embedding)**. There is no "infinite" predictor within the universe. Furthermore, under Axioms B-C, any predictor whose outputs are accessible to the agent cannot guarantee intervention-stable correctness across all admissible policies, because the agent can condition its action on the prediction itself (Proposition 1). A predictor "outside" the system is a metaphysical construct and therefore outside the scope of this scientific framework.

B.5 Objection (Artificial Sovereignty)

"If agency is purely computational, does an advanced AI possess Cognitive Sovereignty? Does this model grant rights to machines?"

Response: Our model is not "carbon-chauvinistic." If an artificial system truly meets the criteria of **Autopoiesis** (self-maintenance) and possesses **Endogenous Stakes** (real consequences for its own Markov Blanket), it would technically join the agency spectrum. However, current LLMs lack these properties; they are prompted, not purposeful. We distinguish between **Trained Stochasticity** (AI) and **Process Sovereignty** (Agents).

B.6 Objection (The Post-Hoc Veto):

"Neuroscientific evidence suggests the 'conscious will' is a post-hoc rationalization. Since the readiness potential precedes conscious awareness of a decision, even the 'Veto Power' is merely part of the determined causal chain, not a moment of genuine agency."

Response: This objection rests on a "Punctate Self" fallacy, which assumes the "Agent" is a single point of conscious awareness separate from its preceding neural activity. We reject this dualism. The **Functional Agency Model** defines the agent as the **Diachronic Self**—the entire temporally extended system. The subconscious buildup of the readiness potential is not an external force acting *upon* the agent; it *is* the agent mid-computation.

Modern research re-characterizes the readiness potential not as a decision, but as **stochastic accumulation** of neural noise toward a threshold (Schurger, 2012). Crucially, experiments confirm that agents can consciously veto an action even after this process has initiated (Schultze-Kraft, 2016). This supports a **Hybrid Causality**: while an initial *urge* may arise from stochastic processes, the final *ratification* or *veto* remains within the sphere of the agent's high-D_A deliberative control. The agent need not author every neural fluctuation to be the sovereign author of their actualized, physical actions.

C Formalization of Agency Metrics (Conceptual)

To facilitate future system modeling in subsequent Technical Reports (Appendix G), we provide the following conceptual definitions for the vectors of sovereignty.

C.1 The A_e Formula

We define **Effective Agency** (A_e) as a comparative engineering index representing the functional capacity of an agent to influence the causal manifold. A_e serves as a technical diagnostic for predicting an agent's vulnerability to external steering (e.g., algorithmic nudging or systemic reduction).

Agency Depth is defined as a first-order approximation:

$$D_A := T_h \cdot C_w \cdot H_i \tag{5}$$

and Effective Agency as:

$$A_e := D_A \cdot R_m = (T_h \cdot C_w \cdot H_i) \cdot R_m \tag{6}$$

Each component (T_h, C_w, H_i, R_m) is a normalized index in $[0, 1]$.

Where:

- T_h **(Temporal Horizon):** The agent's predictive range (0 = reactive; 1 = teleological).
- C_w **(Counterfactual Width):** The "Search Space" resolution; the number of non-linear future trajectories simulated.
- H_i **(Historical Integration):** The coefficient of autobiographical and value-based weighting. $H_i \longrightarrow 0$ in purely reflexive or amnesic states; $H_i \longrightarrow 1$ when deliberation is deeply weighted by stable identity constraints.
- R_m **(Model Fidelity):** The calibration multiplier (0 = hallucination; 1 = empirical correspondence).

C.2 The Threshold of Reducibility

A system is considered **Computationally Reducible** (and therefore non-sovereign) when:

$$O_{predict}(A) \approx A(t+1) \text{ such that } \text{Compute}(O) \ll \text{Compute}_{delib}(A) \tag{7}$$

If an external model O can find the answer faster than the agent A can live the process, the agent has lost **Process Sovereignty**. The mission of the Meaningfulness Protocol is to ensure that for every human agent, the above inequality remains false.

D Operationalization and Falsifiability

To ensure that the Functional Agency Model (FAM) functions as a rigorous scientific framework, we propose the following proxy measurements and falsifiable tests. These metrics allow for the empirical assessment of Agency Depth (D_A) and the detection of Agency Collapse in both biological and artificial systems.

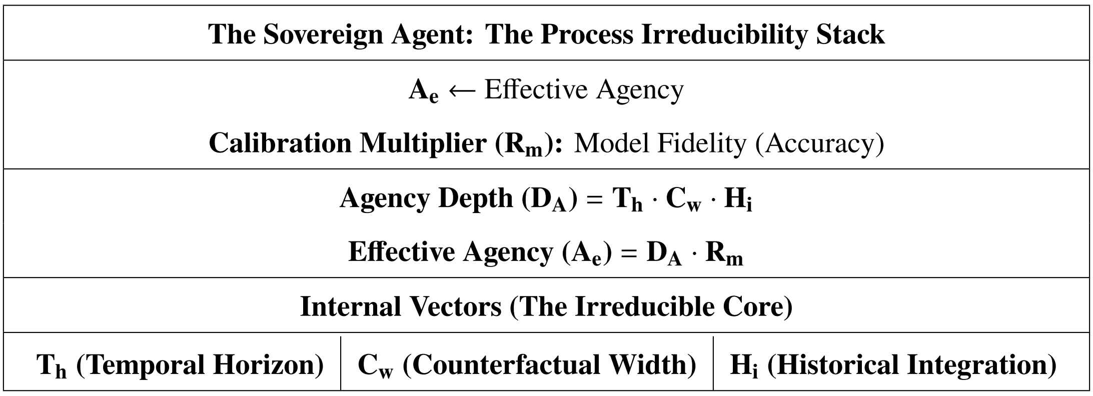
Note: The numerical thresholds below (e.g., $> 95\%$, > 2 standard deviations) are illustrative engineering cutoffs intended for operational audits. In empirical deployments, they must be calibrated to task domain, measurement noise, and ethical constraints.

D.1 Mapping to Established Empirical Paradigms

The vectors formalized in this model are designed to align with measurable constructs in existing psychological and cognitive science literatures. **Temporal Horizon** (T_h) correlates directly with metrics of *delay discounting* and the complexity of *model-based planning tasks* in computational psychiatry. **Counterfactual Width** (C_w) is operationalized through tests of *divergent thinking* and the fidelity of *episodic future thinking*. Furthermore, the integrity of the **Historical Integration** (H_i) vector is assessed through *value stability assays* and metrics of narrative coherence under informational perturbation. This alignment ensures that the Functional Agency Model is falsifiable and can be implemented using established experimental protocols.

D.2 The Functional Agency Model (FAM) Schematic

To lower the cognitive load for first-time readers, the relationship between Agency Depth (D_A) and Effective Agency (A_e) is illustrated in Figure 1. The model emphasizes that A_e is not defined by average capacity but by the constraints of its lowest component, reflecting the bottleneck-dominated principle.



The Sovereign Agent: The Process Irreducibility Stack		
$\mathbf{A_e}$ ← Effective Agency **Calibration Multiplier ($\mathbf{R_m}$):** Model Fidelity (Accuracy)		
Agency Depth ($\mathbf{D_A}$) = $\mathbf{T_h \cdot C_w \cdot H_i}$ **Effective Agency ($\mathbf{A_e}$) = $\mathbf{D_A \cdot R_m}$**		
Internal Vectors (The Irreducible Core)		
$\mathbf{T_h}$ (Temporal Horizon)	**$\mathbf{C_w}$ (Counterfactual Width)**	**$\mathbf{H_i}$ (Historical Integration)**

Figure 1: The Functional Agency Model (FAM) Schematic. Effective Agency (A_e) is the final output of the agent, defined as the product of its internal capacity (D_A) and its external calibration (R_m). The model is governed by the bottleneck logic: $A_e \approx B_A = \min\{T_h, C_w, H_i, R_m\}$.

D.3 Proxy Metrics for Agency Vectors

We identify four measurable dimensions that correspond to the theoretical vectors established in Section 4.

- **Temporal Horizon (T_h):** Measured through *Model-Based Planning Tasks*.
  - *Proxy:* The maximum temporal distance of projected consequences integrated into a current decision-policy.
  - *Falsification:* If an agent's behavior consistently optimizes only for immediate ($t + 1$) rewards despite multi-step dependencies, T_h is functionally zero.
- **Counterfactual Width (C_w):** Measured through *Divergent Generation Fluency*.
  - *Proxy:* The number of distinct, non-linear future trajectories an agent can simulate and evaluate under deliberation.
  - *Falsification:* If an agent's response to an environmental trigger is invariant across 1,000 trials (zero branching factor), C_w is functionally zero.
- **Model Fidelity (R_m):** Measured through *Prediction Error Minimization*.
  - *Proxy:* The statistical correlation between an agent's predicted outcomes (O_{sim}) and actual environmental results (O_{real}). High *Epistemic Foraging Rates* indicate active calibration.
  - *Falsification:* If an agent's actions produce results that consistently deviate from the agent's stated goals by > 2 standard deviations, R_m has collapsed.
- **Historical Integration (H_i):** Measured through *Value Stability Under Perturbation*.
  - *Proxy:* The degree to which an agent's policy is weighted by stored autobiographical data vs. immediate environmental nudges.
  - *Falsification:* If an agent's core values can be inverted by a single short-cycle informational intervention, H_i is non-existent.

This commitment to rigorous, behavioral, and context-aware measurement separates the Functional Agency Model from purely philosophical or self-help frameworks and positions it as a falsifiable scientific and engineering program.

D.4 The Falsifiability Test: The Interventional Predictability Audit

A system's status as a **Sovereign Agent** can be falsified through the following protocol:

1. **Hypothesis:** Agent A is sovereign and irreducible within time horizon T.
2. **Test:** An external model M (with access to A's observable history) attempts to predict A's next action $a(t + 1)$ under a series of controlled informational nudges.
3. **Falsification Criteria:** If model M can reliably predict and steer A's output with $> 95\%$ accuracy across N trials without simulating A in full, then A lacks Process Sovereignty. In such a state, A is **Computationally Reducible** and the claim of agency is falsified for that specific context.

This protocol is designed to allow any **independent auditing body or non-profit initiative** (such as the planned **Meaningfulness Foundation**) to audit socio-technical environments (like social media platforms). Its purpose is to determine if these systems are technically inducing **Ontological Harm** by reducing the population's measurable D_A.

D.5 The Calibration Challenge: A Note on Proxy Selection

While the proxies listed in D.1 provide a starting point for empirical assessment, we acknowledge that no single measurement can perfectly capture the latent variables of Agency Depth. Any serious operationalization of this framework must address the following calibration challenges:

- **Context Dependency:** The "correct" weighting of vectors (T_h, C_w, H_i, R_m) is context-dependent. A fighter pilot requires a different calibration than a long-term strategic planner. The proxies must be normalized against a baseline established for the specific decision-making environment being audited.
- **The Goodhart Problem:** Once a proxy becomes a target, it ceases to be a good measure. For example, if "counterfactual width" (C_w) is used as a key metric, agents (or AI systems) could be trained to generate many meaningless counterfactuals to "game the test." Audits must therefore rely on a *battery* of diverse, non-obvious proxies rather than a single metric.
- **Subjectivity vs. Behavior:** Self-report measures of agency are notoriously unreliable. The proxies proposed here are intentionally *behavioral* and *performance-based*. The goal is not to measure how "free" an agent feels, but to measure their functional capacity to act as an irreducible, causally effective system in the face of environmental entropy and adversarial steering.

The risk of the Goodhart Problem—that a metric will be gamed once institutionalized—implies a primary ethical mandate for the **Meaningfulness Foundation**: to continuously update and audit the complexity of the proxy metrics to ensure they remain anchored to the true goal of Agency Preservation, rather than superficial compliance.

E Functional Definition of Thinking Levels

The Functional Agency Model (FAM) utilizes a hierarchical map of cognitive complexity, or **Cognitive Verticality**, to define the resolution of an agent's internal simulator. This hierarchy is a synthesis of established developmental and computational taxonomies (Bach, 2015; Kegan, 1982; Loevinger, 1976), structured by the scope of the agent's internal model. The full definition and computational framework of this hierarchy will be presented in a forthcoming paper *MMG-TR-006 (Cognitive Verticality)*.

The core functional distinction relevant to the Active Inference model (Section 5) is made across seven operative levels of increasing recursive depth:

Level 1-2 (Surface/Rule-Based): *Reactive, low-complexity processing*. Characterized by focus on immediate sensory input, hard-wired reflexes, and learned procedural scripts. Functionally aligned with Maslow's physiological and safety needs (Pragmatic Value).

Level 3-4 (Conceptual/Reflective): *Analytical processing*. Characterized by the ability to handle abstractions, form principles, and engage in critical, evidence-based evaluation of external data.

Level 5-7 (Integrative/Ontological): *Meta-cognitive processing*. Characterized by the ability to connect disparate systems, question the foundations of knowledge (epistemology), and generate frameworks of meaning (ontology). Functionally aligned with Maslow's esteem and self-actualization needs (Epistemic Value).

Action Modality (The Behavioral Output): Characterizes the agent's response patterns, which are determined by the underlying thinking level:

- **Reaction/Inaction:** Prevalent in L1-L3 (Pragmatic Value). Responses are high-speed, reflexive, or driven by immediate environmental stimuli. Inaction is often a failure to break an established script.
- **Action/Proaction:** Prevalent in L4-L7 (Epistemic Value). Responses are governed by complex simulation, future projection, and deliberate, counterfactual policy selection. Proaction is an engineered outcome of a high-D_A system.

This functional distinction establishes the boundary between mere survival and the intentional pursuit of truth and complexity required to maintain high Effective Agency (A_e).

F Prior Art Boundary and Distinct Contribution

To situate the Functional Agency Model (FAM) within the broader intellectual landscape, this appendix provides a targeted boundary definition. It is not a comprehensive literature review but a formal positioning statement, designed to clarify the specific, novel contributions of this report.

F.1 Comparison with Prior Art Traditions

The core mechanism of FAM—the reflexive prediction barrier under disclosure—has conceptual antecedents in several fields. We explicitly acknowledge these traditions to delineate our specific contribution.

F.1.1 The Philosophical Tradition: Prediction Paradoxes

The idea that a disclosed prediction can be self-defeating is well-established in epistemology, dating back to the Prediction Paradox (e.g., Weiss, 1952). This tradition, further explored by philosophers such as Sorensen (Sorensen, 1988), focuses on logical and epistemic pathologies that arise when an agent's knowledge includes a forecast of their own actions.

Distinct Contribution: While this tradition identifies the paradox, FAM operationalizes it as an engineering constraint. Our target is not merely to note that prediction can fail, but to define the conditions of **Disclosure-Stable Prediction (DSP)**—a prediction that remains correct after disclosure *while preserving the agent's decision-relevant degrees of freedom*. This shifts the focus from a logical puzzle to a criterion for sovereignty.

F.1.2 The Social Science Tradition: Reflexive Prediction

In sociology and economics, the concept of *reflexive prediction* or *forecast feedback* (e.g., Grunberg and Modigliani, 1954) describes how public forecasts can alter collective behavior, thereby invalidating the forecast itself (e.g., forecasts of bank runs, panics, or market moves).

Distinct Contribution: This tradition is typically analyzed at the population level. FAM, in contrast, models the constraints on predicting a *single, self-modeling agent* under physical embedding and resource bounds. Our contribution is the derivation of a structural tradeoff: for a sovereign agent, stable prediction under disclosure tends to require either (i) simulation costs comparable to running the agent, or (ii) integrity-violating control operations (access suppression, policy restriction, or bypass).

F.1.3 The AI & Autonomy Tradition: Algorithmic Nudging

A growing body of literature in AI ethics and Human-Computer Interaction details the harms of algorithmic manipulation and "nudging", arguing that such interventions can undermine user autonomy (Yeung, 2017).

Distinct Contribution: While this literature primarily establishes the normative harm, FAM provides the formal mechanism and state variables. The **Process Sovereignty Criterion** provides a technical definition for *why* a nudge can constitute a sovereignty violation (it functions as a bypass/compression operation), and the **Agency Depth** (D_A) framework provides the structured variables required to model the resulting collapse.

F.2 The Novel Synthesis of this report

The novelty of this report (MMG-TR-002) is not the discovery of reflexivity, but the synthesis of these threads into a single, operational, and falsifiable framework for agency in algorithmic societies. The distinct contributions are:

1. **The Target Property: Disclosure-Stable Prediction (DSP).** We define the sovereignty-relevant prediction target as one that remains stable under disclosure without violating deliberative integrity (here "intervention-stable" is meant in the disclosure-coupled sense defined in Section 3, not distribution-shift invariance).
2. **The Sovereignty Criterion: Process Sovereignty.** We introduce an explicit efficiency and integrity bound. Our criterion disqualifies "prediction by simplification"—any method that achieves accuracy by reducing the agent's decision-relevant degrees of freedom. This makes sovereignty an auditable engineering property, not only a moral claim.
3. **The Scalar Framework: Agency Depth** (D_A) **and Effective Agency** (A_e)**.** We provide a structured, multi-vector model with explicit collapse topology (bottlenecks). This creates a state-variable interface that subsequent reports (e.g., TR-003's threat model) can formally attack and that intervention protocols (e.g., TR-007) can aim to restore.

F.3 Falsification Boundary

The framework is falsifiable at the boundary it specifies. Any demonstrated method that achieves prediction that remains stable under disclosure while verifiably **preserving** the agent's decision-relevant degrees of freedom (i.e., without paying equivalent simulation cost and without engaging in control operations) would refute the sovereignty criterion as defined in this report. This forces critique into the realm of empirical demonstration, not merely theoretical objection.

G The MMG Research Program: Forthcoming Reports

This technical report is the second in a planned series of foundational papers designed to build a comprehensive, multi-disciplinary framework for Cognitive Sovereignty. The subsequent reports* will expand upon the concepts established herein.

MMG-TR-003: Cognitive Sovereignty in Algorithmic Societies This report will analyze the systemic erosion of Agency Depth in modern socio-technical systems. It presents a formal threat model of the Attention Economy and proposes a new class of human rights, including the "Right to remain Incomputable."

MMG-TR-004: The Socio-Technical Foundations of Agency This report connects D_A to the lived reality of human inequality, arguing that high D_A is a resource-intensive state dependent on socio-economic stability, education, and connection, justifying the Foundation's role.

MMG-TR-005: The Spectrum of Ontological Crisis This capstone report unifies the entire framework, defining the **Ontological Crisis** (internal meaning collapse) and **Epistemological Collapse** (failure of shared truth) as a single spectrum of threat. This model establishes the theoretical justification for any **intervening organization's** dual mission: to advocate for systemic defenses that protect the agent's **Unpredictability Horizon** (fighting chronic harm) and to provide protocols (like the Gardener's Calculus) for the safe, compassionate integration of truth (mitigating acute harm).

MMG-TR-006: Cognitive Verticality: The Architecture of Thinking Depth This report formalizes the **7-Level Hierarchy of Thinking Depth**, utilizing the computational rigor of the Loevinger/Kegan models. It maps the agent's recursive resolution, demonstrating why higher verticality is a necessary precondition for maintaining high **Effective Agency** (A_e) and resisting algorithmic pattern recognition.

MMG-TR-007: The Meaningfulness Protocol This applied report synthesizes the entire sequence (TR-001 through TR-006) into a concise, actionable methodology. It defines **Meaningfulness** as the objective system output of a high-complexity agent and provides structured protocols to foster resilient connections and combat the nihilism arising from cognitive verticality dissonance.

**: Note that the titles of forthcoming technical reports are provisional and subject to revision upon final publication; the core topics and scope should remain broadly as described.*