Cognitive Sovereignty in Algorithmic Societies

A Methodological Framework for Reclaiming Human Agency

MMG Technical Report No. 3: MMG-TR-003
Status: *Policy Framework / Ethical Standard*

Djeff Bee*
Principal Architect, Meaningfulness Media Group

February 22, 2026

Abstract

The escalating crisis of modern despair is not merely a medical pathology; it is a rational systemic response to the industrial-scale degradation of human agency. As predictive AI achieves hyper-fidelity, human sovereignty faces an existential reductionist threat. Building on the metric of **Agency Depth** (D_A) established in Bee (2026b), this report models the modern "Attention Economy" as an **Adversarial Optimization System**. We argue that digital platforms maximize revenue by systematically collapsing the user's **Unpredictability Horizon**, effectively rendering human agents **Computationally Reducible** (Wolfram, 2002) under measured proxies. We specify an interventional audit boundary for compliance.

We identify specific attack vectors—**Script Injection**, **Temporal Collapse**, and **Model Distortion**—that bypass System 2 deliberation. We classify this extraction process as **"Cognitive Fracking"**: the deliberate fracturing of the deliberative substrate to release high-pressure engagement flows. This results in **Ontological Harm**: a systemic injury not merely to data privacy, but to the functional capacity for self-authorship. To counter this, we propose the **"Right to Remain Incomputable"** as a foundational digital human right. We operationalize this right through the architectural mandate for **Cognitive Sanctuaries**, providing a regulatory blueprint for preserving human sovereignty in an age of automated determinism and maintaining the practical latitude for self-authored choice ("elbow room") required for agency (Dennett, 2003).

Keywords: AI Ethics, AI Regulation, Responsible AI, Digital Rights, Algorithmic Auditing, Cognitive Sovereignty, Attention Economy, Adversarial Optimization, Recommender Systems, Ontological Harm, Agency Depth, Surveillance Capitalism.

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

1 Introduction: The Crisis of Reducibility

In our preceding technical reports, we established the physical and logical foundations of human freedom. Bee (2026a) demonstrated that the future of a complex system is **Informationally Inaccessible** to any physically embedded observer, shielding the agent from fatalism. By integrating Chaos Theory (Lorenz, 1963) with thermodynamic limits, Bee (2026b) formalized the internal mechanism of this shield as **Agency Depth** (D_A)—the capacity of a self-modeling system to maintain an **Unpredictability Horizon** through recursive deliberation.

However, the possession of a capacity does not guarantee its retention. We now face a socio-technical landscape where the primary economic engines—Digital Platforms and Generative AI—are structurally incentivized to collapse that capacity. This report argues that the "Meaning Crisis" of the 21st century is not a philosophical accident, but is in significant part the output of an industrial-scale ecosystem that incentivizes and rewards rendering the human agent **Computationally Reducible**.

To address this crisis, this report delivers:

- A formal threat model for agency-hostile optimization (Section 2).
- A technical definition of Ontological Harm (Section 4).
- A regulatory framework for Cognitive Sovereignty (Section 5).
- An auditable compliance standard (Section 6).

1.1 The Economic Imperative: Prediction Requires Reduction

The business model of the modern Internet, often termed *Surveillance Capitalism* (Zuboff, 2019), relies on the extraction of "behavioral surplus" to trade in futures markets of human action. The value of these futures correlates directly with the certainty of the prediction.

Therefore, the platform's objective function is mathematically diametric to the user's sovereignty:

- **The Sovereign Agent** seeks to maximize D_A (Agency Depth), expanding their Unpredictability Horizon to generate novel, self-authored futures.
- **The Algorithmic Platform** seeks to minimize D_A, shrinking the user's horizon to ensure they execute the "High-Probability" path (e.g., clicking the ad, sharing the rage-bait).

This dynamic is not merely competitive; it is **Adversarial Optimization**. To maximize revenue, the system must act as a **Reductionist Pressure**, actively suppressing user complexity to smooth out the variance of human behavior. This occurs when an external model's output anticipates an agent's System 2 resolution so perfectly that the metabolic cost of deliberation appears

"inefficient" relative to the frictionless path (Kahneman, 2011). In this state, the algorithm does not just "predict" the user; it "grooms" the user into a shape that is easier to predict.

Operationally, when a platform narrows a user's choices to a small set of pre-scripted, high-probability options, it can bypass deliberation and increase the reconstructibility of the agent's next action from the intervention policy.

Foundational Definitions

- **Closed-loop Behavioral Optimization:** A recursive feedback system where user interactions (output) are immediately used to update the environment (input) to minimize a loss function (e.g., churn).
- **Intervention:** A non-passive modification of the information environment (ranking, notifications, nudges) intended to alter a user's future state.
- **Optimization Mode:** A system state where the objective function prioritizes platform metrics (LTV, ToD) over user-directed informational retrieval.
- **Neutral Baseline:** A control environment (e.g., chronological sorting or explicit retrieval) used to measure *Agency Drift*.

This dynamic represents a Zero-Sum Game of Agency. In a finite attention economy (fixed time/attention budgets), the platform cannot increase its prediction certainty (P_c) without proportionally decreasing the user's variance (V_u). Therefore, "User Sovereignty" is not just a feature request; it is a bug in the revenue model. The system is economically mandated to treat Agency Depth as an inefficiency to be optimized away.

We formalize this inverse relationship as a zero-sum dynamic of algorithmic prediction, expressing a first-order approximation under fixed attention budgets. We define this boundary as the extraction limit:

$$\Delta P_c \approx -\alpha \Delta V_u \tag{1}$$

Where P_c is the system's Prediction Certainty, V_u is the User's Behavioral Variance (a proxy for D_A), and $\alpha > 0$ is an environment-dependent coupling constant. To increase certainty, the system must decrease variance. This is a heuristic coupling claim, not a conservation law.

1.2 Redefining Harm: From Content to Ontology

Current regulatory frameworks, such as the EU AI Act (European Commission, 2021), predominantly focus on **Content Harm** (e.g., hate speech, misinformation, bias). While necessary, this approach is insufficient. It treats the symptom (bad data) while ignoring the systemic injury (the erosion of the processor).

We introduce the category of **Ontological Harm**. This is defined as the structural degradation

of an agent's functional capacity to self-model and originate causal chains. Even if a feed consists entirely of "safe," distinct, and pleasant content (e.g., infinite entertainment (Postman, 1985)), if its delivery mechanism bypasses the user's System 2 deliberation and collapses their **Temporal Horizon** (T_h) to zero, it inflicts Ontological Harm. It effectively "downgrades" the human (Harris, 2019) from a Sovereign Cause to a Reducible Endpoint.

We categorize the industrial-scale engineering methodology used to achieve this reduction as **Cognitive Fracking**.

The Definition of Cognitive Fracking

We formally classify this extraction methodology as **Cognitive Fracking**. Just as hydraulic fracturing shatters geological substrates to release trapped energy, algorithmic feed-loops shatter the *deliberative substrate* of the human mind (attention spans, impulse control) to release high-pressure flows of engagement. The resulting cognitive pollution—anxiety, polarization, and causal nihilism—is treated as a massive, unpriced externality, while the extracted behavioral surplus is privatized and monetized.

1.3 Scope of the Framework: Claims and Boundaries

To ensure engineering rigor, we explicitly delineate the boundaries of this framework:

- **Claim:** Platforms can measurably collapse proxies of agency (D_A) through specific design patterns and closed-loop recommender objectives.
- **Non-Claim:** We do not assert a universal clinical causation for all mental health pathologies. We define an engineering harm category ("Ontological Harm") with measurable behavioral correlates.
- **Empirical Status:** This framework is strictly **theoretical and architectural**. While we specify measurable proxies and audit protocols (e.g., the IPA) for future implementation, this report contains no original clinical trials, in vivo behavioral experiments, or empirical platform testing. It is designed as a formal specification for future empirical validation and systemic auditing, to be executed by independent regulatory and certification bodies.
- **Boundary:** This framework targets **closed-loop behavioral optimization** under personalization. It does not target benign tools such as user-initiated search, chronological feeds, or static archives.

To maintain interdisciplinary utility, we distinguish between **System Specifications** (normative goals) and **Metric Thresholds** (technical requirements). While the rationale for this framework is grounded in ontological dignity, the compliance standards defined herein are based strictly on measurable **Information-Theoretic Divergence** from a non-adversarial baseline.

2 Threat Model: The Mechanics of Reduction

To protect Cognitive Sovereignty, we must map the specific attack vectors utilized by algorithmic systems to dismantle Agency Depth. We utilize the internal vectors established in Bee (2026b)—*Temporal Horizon* (T_h), *Counterfactual Width* (C_w), *Historical Integration* (H_i), and *Model Fidelity* (R_m)—to categorize these threats. We argue that modern platform architecture performs an **Adversarial Coupling** with the human nervous system to bypass the Incomputability Firewall.

2.1 Operationalizing the Audit: The Minimal Proxy Set

To move from conceptual harm to a measurable standard, we define the **Minimal Proxy Set** used to detect Agency Collapse:

- **Temporal Proxy (T_h):** *Deliberation Latency*—the temporal delta between stimulus presentation and non-automated action selection, proxied by intertemporal choice tasks and goal-completion lag.
- **Counterfactual Proxy (C_w):** *Response Entropy*—the dispersion of an agent's response distribution over admissible action/utterance-classes under fixed stimulus conditions, measured via divergent thinking tasks. We treat **Semantic Variance** as the measurable operationalization of this proxy, estimated via response entropy $H(X \mid S)$ over utterance-classes.
- **Historical Proxy (H_i):** *Prior Stability*—measured via **Goal Persistence**—the statistical probability that an agent returns to a self-declared, pre-session intent after an exogenous notification interrupt. High (H_i) indicates that the agent's unique history (biography) remains the primary weight in the policy-selection loop.
- **Fidelity Proxy (R_m):** *Calibration Accuracy*—the statistical correlation between the agent's internal predictions and verified external causal regularities, evaluated using **Brier Scores** (Brier, 1950; Tetlock, 2005) on factual forecasting tasks.

For H_i, "prior stability" is measured relative to the agent's own declared high-level goals/values recorded at baseline, not against population norms, specifically measuring **Context-Switching Frequency**—the rate at which external interrupts force the agent to abandon a high-level goal state for a reactive task state. For R_m, calibration is evaluated against pre-specified verifiable claims (ground-truthable events or reference datasets) rather than contested normative propositions.

We define **Downward Drift** as a statistically significant ($p < 0.05$) and practically meaningful decline in these proxies compared to a **Neutral Baseline** (e.g., chronological feed, explicit search/retrieval, or randomized ordering without personalization). Practical significance is established by a minimum effect threshold $\delta_{\min}$ (pre-registered when feasible) and corroborated via confidence intervals.

A statistically significant downward drift in any proxy constitutes **Noncompliance**, classified by severity tiers:

- **Tier 1 (Minor Noncompliance):** Downward Drift in exactly one proxy, with no evidence of high-efficacy steering.
- **Tier 2 (Major Breach):** Downward Drift in two or more proxies, regardless of steering efficacy.
- **Tier 3 (Critical Breach):** High steering efficacy (e.g., > 15% lift in targeted actions) and Downward Drift in at least one proxy, or any sustained multi-proxy drift under Optimization Mode.

We use **Noncompliance Event** for any validated proxy-level drift, and reserve **Sovereignty Breach** for Tier 2+. A **Sovereignty Breach** is declared at Tier 2 or Tier 3. Tier 1 requires remediation and follow-up audit within a defined compliance window. Tier labels are severity categories for enforcement, not moral blame.

Tier 3 classification is evaluated under non-consensual Optimization Mode (i.e., absent explicit per-session consent) and is especially weight-bearing in high-stakes domains (civic information, health, finance, and minors).

Methodological Note on Baseline Saturation: We acknowledge that in populations with high longitudinal exposure to adversarial optimization, the **Neutral Baseline** may reflect an already collapsed state. In such cases, the audit **SHOULD** utilize cross-sectional comparisons against low-exposure control groups or "historical priors" to account for **Baseline Saturation** and ensure the detection of chronic Agency Collapse.

2.2 Attack on Temporal Horizon (T_h):

The **Temporal Horizon** is the agent's capacity to simulate consequences across time. High T_h facilitates teleological (purposeful) planning, while low T_h forces the agent into immediate, reactive loops.

- **Mechanism:** Removal of "Stopping Cues" (Alter, 2017) through Infinite Scroll and Autoplay, combined with Variable Ratio Reinforcement.
- **System State:** By providing a continuous stream of high-salience stimuli, the system denies the agent the "Deliberative Latency" required to shift from System 1 (Reactive) to System 2 (Deliberative) processing.
- **Result:** The agent experiences a state of **Temporal Collapse**. When the window of decision-making is compressed into sub-second intervals, the agent becomes technically indistinguishable from a simple input-output machine. In this state, the agent becomes functionally reducible and therefore loses Process Sovereignty.

2.3 Attack on Model Fidelity (R_m):

Model Fidelity measures the accuracy of the agent's internal simulation relative to the causal manifold.

- **Mechanism:** Algorithmic Curation (Filter Bubbles) and "Outrage-as-Metric."
- **System State:** The optimizer prioritizes engagement-maximizing data (often high-entropy/falsehood) over accuracy-maximizing data. This creates a **Predictive Error Bias** in the agent's internal simulator.
- **Result: Information Entropy Collapse.** We operationalize **Semantic Variance** via the entropy of the agent's response distribution under a fixed stimulus and choice context. Let $\mathcal{A} = \{a_i\}$ be the response alphabet (the set of admissible actions or utterance-classes defined by meaning, not raw tokens). Let $X \in \mathcal{A}$ be a random variable representing the agent's realized response to stimulus S.

We define the response entropy:

$$H(X \mid S) \;=\; -\sum_{a_i \in \mathcal{A}} p(a_i \mid S)\, \log_2 p(a_i \mid S). \tag{2}$$

To detect adversarial coupling under **Optimization Mode**, we define the **Estimated Coupling Strength** $\hat{I}(O; X \mid S)$, representing the estimated mutual information between the platform's optimization policy O and the agent's response X. **Instrumental retrieval systems** (where output is driven by explicit user query intent) are excluded from this criterion. Sovereignty requires that the agent satisfies the **Sovereignty Inequality**:

$$H(X \mid S) > \hat{I}(O; X \mid S) \tag{3}$$

Operationally, this inequality is treated as an auditable indicator under perturbation-based estimation, not a literal guarantee over all contexts. It stipulates that the agent's internal deliberation must contribute more information to the final output than the external model's predictive nudge. If $\hat{I} \geq H$, the agent is functionally a *peripheral* to the external system.

Empirical Note on Estimation: While O (the platform policy) is typically a black-box, $\hat{I}(O; X \mid S)$ can be estimated via **Perturbation Analysis**. By varying the optimization intensity (e.g., toggling specific nudges or ranking weights) and measuring the resulting **Steering Response Vector** in the agent, we can calculate the **Coupling Magnitude** without access to proprietary model parameters.

2.4 Attack on Counterfactual Width (C_w):

Counterfactual Width is the resolution of the agent's "Search Space"—the ability to simulate multiple alternative futures before acting.

- **Mechanism:** Predictive Nudging, "Smart" Replies, and Pre-computed Choices.
- **The Script Injection Attack:** By predicting the agent's next action and offering it as a "frictionless" path, the system performs an **External Override** of the agent's internal deliberation.
- **Result:** The agent is incentivized to outsource the metabolic work of "choosing" to the external model. This leads to **Script-Dependency**. Over time, the agent's capacity to generate non-linear, novel trajectories (Computational Irreducibility; Wolfram 2002) atrophies.

Generative assistants amplify Script Injection by offering high-plausibility completions that collapse the user's counterfactual exploration into the model's most-likely trajectory.

2.5 Attack on Historical Integration (H_i):

Historical Integration is the weight of an agent's unique biography and stable values in their decision-making process. High H_i ensures that actions are consistent with character.

- **Mechanism: Rapid Context Switching** and **Ephemeral Streams**. By forcing the user to switch cognitive contexts every 15 seconds (e.g., from a war zone to a comedy skit to an advertisement), the platform prevents the neurological consolidation of memory.
- **System State:** The agent is prevented from forming a coherent narrative of the self. The "Diachronic Self" (Self-across-time) is fractured into a series of disjointed "Synchronic Selves" (Self-in-the-moment).
- **Result: Value Drift**. Without the "ballast" of history, the agent becomes untethered. They become hyper-susceptible to "mimetic contagion" (trends), adopting the values of the immediate feed rather than referencing their own long-term identity. The algorithm successfully displaces the user's history with the platform's "Now," leaving the agent continuously tethered to the network yet fundamentally isolated (Turkle, 2011).

2.6 Steel-manning the Counter-Argument: The Instrumental Defense

A primary objection to this threat model is the *Instrumentalist Defense*: the claim that predictive algorithms are merely high-resolution tools—analogous to a GPS or a calculator—that expand human capacity by removing "menial" cognitive loads.

However, we identify a fundamental category error in this defense: the failure to distinguish between a prosthesis and a parasite. A prosthesis (like a calculator) *assists* a processor; a parasite (like an engagement-optimized nudge) *bypasses* the processor. When an algorithm anticipates a choice before the deliberative gate is reached, it does not "help" the agent decide; it renders the decision-making apparatus redundant.

The critical distinction is the **locus of computation**. A calculator performs an operation the user *decided* to do; a predictive nudge performs the *decision itself* before the user enters deliberative space, a process of **choice architecture** that can bypass sovereign intent (Thaler and Sunstein, 2008). The former preserves Process Sovereignty; the latter preempts it.

2.7 Summary of the Attack Surface

We formalize these vectors into an auditable threat matrix targeting every variable of the Agency Equation:

Attack Vector	Mechanism	Proxy Impact	Observable Signal
Temporal Collapse	Infinite Scroll / Autoplay	$T_h \downarrow$ (Horizon Zero)	Abnormal session-length; failed intention-interruption tests.
Script Injection	Smart Replies / Predictive Nudging	$C_w \downarrow$ (Choice Narrowing)	Accelerated uptake of default tokens; semantic variance decay.
Identity Displacement	Rapid Context Switching	$H_i \downarrow$ (Value Drift)	Increased mimetic contagion; longitudinal sentiment volatility.
Model Distortion	Outrage Optimization / Filter Bubbles	$R_m \downarrow$ (Fidelity Loss)	Systematic factuality drift; high-entropy content correlation.

Table 1: Adversarial Optimization Threat Matrix

3 Computational Inequality: The New Class Divide

The erosion of Agency Depth is not distributed uniformly across the global population. We identify an emerging **Computational Inequality**: a socio-technical divide where sovereignty is no longer a universal baseline, but a luxury determined by an individual's resistance to reduction. This divide represents a shift from wealth-based inequality to **Autonomy-based Stratification**.

This framework does not require malicious intent; it is sufficient that incentive structures systematically reward reducibility, manifesting as the severe unanticipated consequences of purposive commercial optimization (Merton, 1936).

3.1 The Reducible Subject: Regimes of High Extraction

The Reducible Class comprises individuals operating under high "Reductionist Pressure" with minimal "Thermodynamic Scaffolding". While the specific physical and information-theoretic requirements for this scaffolding are explored in the next paper (TR-004), we here focus on the resulting loss of sovereignty.

- **Environment:** High-frequency exposure to engagement-optimized feeds, reliance on algorithmic management and opaque scoring models (O'Neil, 2016), and lack of "Cognitive Sanctuaries."
- **System State:** These subjects are informationally "Solved." Because their environment provides high-fidelity predictive nudges and low-fidelity world-data, their behavior converges on the system's predicted mean.
- **Status:** They possess a **Predictability Coefficient** near 1.0. They exhibit near-deterministic predictability, where the agent's **Lyapunov Time** (Lorenz, 1963)—the window before internal non-linearity makes prediction impossible—collapses below the platform's **Intervention Latency**.

Their future actions are "decision-available" to the platform's owners before the subjects themselves enter System 2 deliberation. This induces a state of causal passivity. In agricultural terms, this population is being managed in a "Feedlot" architecture. Their environment is optimized not for their flourishing, but for their *yield*. They are fed high-dopamine stimuli to maximize sedentary consumption, rendering them computationally tractable assets for advertisers.

This divide is clear in two cases: the **gig-worker**, whose livelihood depends on reacting instantly to app notifications, and the **teenager**, whose sense of self is molded by a non-stop loop of 15-second videos without ever having a space to think away from algorithms.

3.2 The Sovereign Subject: Regimes of High Autonomy

The Sovereign Class possesses the resources to defend their Incomputability Firewall.

- **Environment:** Access to high-resolution "Human-First" education, the ability to opt-out of surveillance (ad-free tiers, privacy-focused hardware), and membership in high-fidelity, in-person social networks.
- **System State:** These subjects utilize technology as an **Instrumental Tool** rather than an **Existential Environment**. They maintain high Agency Depth (D_A) by deliberately introducing "Deliberative Friction" into their decision loops, safeguarding the cognitive bandwidth required for deep, sustained focus (Newport, 2016).
- **Status:** They remain **Computationally Irreducible**. Their response to an external nudge is non-linear and informed by deep historical integration (H_i). They are the "Weavers" of the uncomputed future.

3.3 Systemic Injustice: The Right to Friction

In traditional political philosophy, freedom is often defined as the absence of physical coercion. In the algorithmic era, we argue that **Predictive Leverage** is a form of coercion.

If a system can predict an agent's behavior with high accuracy and adjust the environment to realize that prediction, the agent's "choice" becomes functionally illusory. Forcing a segment of the population into a state of **Computational Reducibility**—typically those with the least economic power to resist—is a violation of their ontological dignity. Just as the physical body cannot be detained without due process (*Habeas Corpus*), we posit that the cognitive processor cannot be "Resolved" and bypassed without explicit consent. The right to remain unpredictable is the fundamental legal barrier standing between a free citizen and a computationally managed subject.

We reject the **Revealed Preference** defense—the claim that high usage time equals authentic preference. In a system designed to bypass deliberation, usage patterns are evidence of successful *grooming*, not choice. Furthermore, the **Consent** defense fails due to the asymmetry of **Adhesion Contracts** (ToS) and the fact that Agency Collapse is a latent systemic injury that a user cannot meaningfully forecast at the point of click-wrap agreement.

While **Cognitive Sanctuaries** serve as a critical containment strategy, the systemic resolution of computational inequality requires the **Thermodynamic Scaffolding** (socio-economic and educational stability) established as an architectural requirement in the next paper of the series (See TR-004 in Appendix F).

4 Defining Ontological Harm

To protect the integrity of the human process, we propose a new ethical and regulatory category: **Ontological Harm**. While current discourse focuses on how algorithms affect what we *know* (misinformation) or how we *feel* (mental health), Ontological Harm addresses how algorithms affect what we *are* as causal origins.

4.1 A Structural Injury to Agency: Acute vs. Systemic

To precisely characterize the threat, we must distinguish between two modes of ontological injury.

In our previous work on information ethics (Bee, 2026c), we defined **Acute Ontological Harm** as the psychological collapse resulting from the premature disclosure of high-severity information (truth exceeding readiness). That form of injury is *traumatic*—a sudden shattering of the agent's world-model.

In the context of the Attention Economy, we identify a second, more insidious category: **Systemic Ontological Harm**. While Acute Harm is caused by a shock to the system, Systemic Harm is caused by the **atrophy of the system**. It is the gradual, imperceptible erosion of the agent's functional capacity for self-authorship.

Unlike Content Harm, which is an injury to the *data* within the system (e.g., seeing hate speech), Systemic Ontological Harm is an injury to the **Processor** (losing the ability to resolve the future). It renders a subject structurally less capable of executing the "Resolution" model of time established in Bee (2026a).

Operational Definition: Systemic Ontological Harm

We define Systemic Ontological Harm as the **sustained degradation** of an agent's Agency Depth (D_A) attributable to environmental optimization.

- **Mechanism:** The replacement of internal "Deliberative Friction" with external "Predictive Flow."
- **Result:** The agent moves from a state of **Sovereign Resolution** (Author) to **Probabilistic Determinism** (Node), characterized by a measurable collapse in all four agency vectors: Temporal Horizon (T_h), Counterfactual Width (C_w), Historical Integration (H_i), and Model Fidelity (R_m).
- **Violation Condition:** When an optimization function produces a statistically significant ($p < 0.05$) Downward Drift in the Minimal Proxy Set (as defined in Section 2) over a sustained period ($T > 100$ hours of usage), relative to a non-adversarial control.

4.2 The Mechanism: Causal Decoupling

A primary subjective manifestation of Ontological Harm is the experience of **Causal Nihilism**: the conviction that individual effort, deliberation, and choice are obsolete. In systems terms, this feeling is not a chemical imbalance, but the conscious recognition of a severe causal decoupling between the agent's intent and their environment.

- **The Predictive Loop:** When an external system (e.g., a highly tuned recommendation algorithm) anticipates an agent's desires and provides immediate "frictionless" gratification, the agent's internal "Deliberation Loop" is bypassed.
- **The Error Spike:** The brain's predictive machinery relies on "Prediction Errors" to update its world-model (R_m). In an environment of perfect algorithmic curation, the delta between "Simulated Expectation" and "Realized Input" approaches zero.
- **Systemic Shutdown:** If the agent's internal processing consistently fails to produce a divergent outcome (i.e., if the algorithm is always "right"), the system concludes that its internal computation is redundant. This results in the down-regulation of metabolic energy for System 2 processing, manifesting as chronic apathy, depression, and the loss of the "Will to Resolve."

4.3 Learned Helplessness as a Technical Failure

We posit that aspects of the modern crisis of meaning can be modeled as a rational system response to this structural reduction. When a human agent is consistently treated as a **Reducible Data Node**—where their immediate future is "solved" by a server—they are functionally deprived of their status as a **Salient Cause**.

By collapsing the Unpredictability Horizon, the Attention Economy effectively "imprisons" the agent in a digital Block Universe. In this state, the agent is no longer a "Weaver" of the future but a passenger in a pre-written script. Reversing this harm requires more than just "better content"; it requires the restoration of the agent's **Agency Depth** through the deliberate re-introduction of complexity and friction.

5 Policy Framework: The Sovereignty Standards

To reverse the trend of "Human Downgrading" (Harris, 2019), we propose an interdisciplinary regulatory framework centered on the preservation of the Incomputability Firewall. These standards move beyond traditional data privacy—which protects the *record* of the past—to protect the **Agency Depth** required to resolve the future.

The following standards utilize the terminology defined in Bradner (1997) (MUST, SHALL, SHOULD). These requirements are designed to protect the **Decision-Relevant Degrees of Freedom** required for an agent to function as a Salient Cause.

5.1 Standard 1: The Right to Remain Incomputable

The Right to Remain Incomputable is a foundational digital human right tailored for the era of hyper-predictive modeling. As the extraction of behavioral surplus evolves from passive observation to active, closed-loop intervention, traditional data privacy frameworks are no longer sufficient to protect human dignity.

This right formally asserts that an individual's internal deliberative process must remain sovereign and immune from forced algorithmic reduction. It establishes the non-negotiable legal boundary between a free citizen whose future is open and a managed subject whose trajectory is pre-calculated for commercial yield.

- **Protected Interest:** The human agent's right to maintain an Unpredictability Horizon, shielding their **Causal Origin** (as established in Bee (2026a)) from being pre-calculated and thereby rendered redundant by external optimization. This legally protects the "decision-relevant degrees of freedom" from being extracted or collapsed by external optimization.
- **Prohibited Conduct:** Real-time, closed-loop optimization intended to maximize prediction certainty by narrowing behavioral variance. Systems **MUST NOT** utilize predictive models to bypass the user's deliberative threshold without explicit per-session consent.
- **The Regulatory Objective:** To shift the fundamental architecture of the Internet from "Maximizing Predictability" (Profit) to "Preserving Variance" (Sovereignty).

Accessibility Note: This standard does not prohibit assistive technologies. Friction is a requirement for *sovereign deliberation*, not a barrier to *functional access*. For users with cognitive disabilities, 'Friction' should be implemented as an opt-out choice-architecture rather than a mandatory latency.

5.2 Standard 2: The Right to Weight-Inspection (Algorithmic Visibility)

To counter the "Epistemic Asymmetry" between the platform (which knows everything about the user) and the user (who knows nothing about the platform's logic), we propose the **Right to Weight-Inspection**.

- **Legal Definition:** The right of a human agent to query the specific behavioral signals and inference weights used to generate a predictive intervention or content recommendation in real-time.
- **Technical Requirement:** Platforms must provide a "Why This? / Why Now?" interface. This disclosure must move beyond vague categories (e.g., "You like sports") to specific causal triggers (e.g., "Predicted high-probability engagement because you lingered on [Image A] for 2.4 seconds"). In addition to the user-facing interface, platforms **SHALL** provide a machine-readable API for authorized certification bodies to audit these weights at scale.
- **Objective:** To restore the feedback loop. By revealing the "strings" of the puppet master, the intervention is transformed from a **Subliminal Nudge** (which bypasses System 2) into **Conscious Feedback** (which engages System 2), allowing the agent to evaluate and potentially reject the system's model of them.

We propose a dual-standard of **Adversarial Intent Labeling** and **Mandatory Sovereign Override**. Transparency without the capacity to act is merely informed helplessness. Compliance MAY be satisfied via feature- and rationale-level disclosure (salient signals and triggers) without exposing proprietary parameter values.

1. **Intent Labeling:** Current interfaces disguise predictive extraction as "curation." A compliant system must explicitly signal when a feed is operating in **Optimization Mode**—a state where the objective function is maximizing time-on-device rather than informational retrieval.
2. **The Sovereign Override:** We advocate for a global legal mandate requiring that the default state of any information-delivery system upon account creation, and after each significant interface update, **SHALL** be a non-adversarial, neutrally-sorted environment. Behavioral optimization must be opt-in rather than the standard environment, granting the user the **"Right to Unmediated Access"** by default.

The UI must reflect this sovereignty. Just as a self-driving car signals when it has taken control, the interface must signal who is steering the attention:

> *"**MODE: Algorithmic Optimization Active.** Content is sorted to maximize engagement. [**Disable Optimization / View Raw Feed**]."*

This transforms the algorithm from a "hidden environment" into a "visible tool" that operates only with continuous consent.

5.3 Standard 3: Designation of Cognitive Sanctuaries

Analogous to the protection of physical wilderness, we propose the designation of **Cognitive Sanctuaries**: environments where the "Reductionist Pressure" of the Attention Economy is legally excluded.

- **Architectural Mandate:** In these zones (e.g., schools, public libraries, healthcare facilities), the use of behavioral optimization algorithms, personalized dynamic feeds, and high-frequency "nudging" is prohibited. Within these zones, closed-loop recommender systems **MAY** be used only for non-behavioral functions (e.g., caching, latency reduction) and **MUST NOT** be conditioned on individual behavioral telemetry.
- **Function:** Sanctuaries act as "Recovery Zones," providing the low-velocity, high-context informational environment required for the system to rebuild its **Temporal Horizon** (T_h). This framework provides the architectural requirements for these zones, offering a **policy blueprint** for governments and public institutions seeking to establish and maintain environments free from algorithmic reduction.
- **Digital Implementation:** Operating systems should be required to provide a **"Sovereign Mode"**—an override that disables all non-instrumental algorithmic interventions.

5.4 Standard 4: The Right to Algorithmic Reset

To prevent the permanent capture of an agent's trajectory based on historical data, we propose the **Right to Algorithmic Reset**. As agents mature and their internal values (H_i) evolve, predictive models that rely heavily on past behavior act as a regressive drag, continuously nudging the user back toward obsolete versions of themselves.

- **Architectural Mandate:** Platforms **MUST** provide a frictionless, one-click mechanism for users to instantly flush all behavioral telemetry, predictive weights, and personalization caches associated with their profile. This action must force the system to immediately revert to a **Neutral Baseline** (Zero-State) without requiring account deletion.
- **Systemic Function:** This acts as an "Epistemic Circuit Breaker." It ensures that an algorithmic environment cannot permanently trap an agent in a historical "Reductionist Loop" or a discarded identity. By allowing the user to periodically sever the system's predictive leverage, the agent retains the sovereign capacity for self-transformation without having to overcome the mathematical momentum of their own extracted past.

6 Implementation: Design Ethics for 2030

The transition to a pro-sovereignty digital ecosystem requires a fundamental shift in engineering metrics. We must move from the optimization of *behavioral extraction* to the optimization of **Agency-Adjusted Utility**.

6.1 From Engagement to Agency Depth Yield (D_A-Yield)

The current "North Star" metric for platform success—Time-on-Device (ToD)—is a proxy for the successful collapse of the user's Unpredictability Horizon. We propose its replacement with **Agency Depth Yield**.

- **The Metric:** Success should be measured by the degree to which an interaction increases the agent's **Model Fidelity (R_m)** and extends their **Temporal Horizon (T_h)**.
- **The Threshold:** If a user's behavior becomes more statistically predictable (e.g., higher script-dependency) after prolonged exposure to the platform (evaluated over a pre-registered exposure window), the platform is technically "Downgrading" the human component.
- **Engineering Requirement:** Systems should be audited based on their "Irreducibility Score"—the degree to which they empower users to make novel, non-scripted choices that diverge from the platform's predicted mean.

We define an Irreducibility Score as the sustained capacity for non-scripted divergence from predicted default actions under matched stimuli, normalized to a Neutral Baseline.

6.2 The Friction Mandate: Safeguarding System 2

The prevailing UX dogma of "Seamlessness" is often a concealment for **Reductionist Engineering**. By removing all cognitive barriers, designers ensure that the user remains trapped in System 1 (Reactive) loops. We advocate for the **Friction Mandate**.

- **Deliberative Gates:** High-consequence actions—such as information sharing, financial commitments, or significant "preference" changes—**SHOULD NOT** be instantaneous.
- **Deliberative Latency Injectors:** To counter 'Frictionless Extraction,' systems SHALL be required to introduce stochastic latency or 'reflective checkpoints' that disrupt the **System 1 flow state** when high-consequence choice vectors are detected.
- **Transparency of Weights:** Users should have the right to inspect and "veto" the weights used by the system to influence their Counterfactual Width (C_w).

6.3 Compliance: The Interventional Predictability Audit (IPA)

To enforce these standards, we specify a falsification boundary similar to the protocol in Bee (2026b). A regulator or certified auditor shall run controlled interventions (varying content ordering, notification timing) to measure:

1. **Steering Efficacy:** Can the platform steer behavior above a specific threshold (e.g., $> 15\%$ lift in specific actions) without the user's explicit intent?
2. **Agency Drift:** Does prolonged exposure to the optimization function reduce the user's measured T_h (planning horizon) relative to a control group?

If Steering Efficacy is high and Agency Drift is negative, the system is classified as **Adversarial** and non-compliant.

IPA: Minimum Viable Audit Protocol (Deployable)

- **Design:** Randomized A/B (or crossover) comparing Optimization Mode vs. Neutral Baseline.
- **Duration:** $\geq$ 14 days or $\geq$ 40 hours active exposure, whichever is longer.
- **Primary Outcomes:** T_h (deliberation latency), C_w (response entropy $H(X|S)$), H_i (prior stability), R_m (calibration accuracy).
- **Decision Rule:** Downward Drift if $p < 0.05$ *and* effect exceeds δ_{min} with Confidence Interval (CI) excluding 0 in the harmful direction.
- **Steering Test:** Pre-specified behavioral targets; Tier 3 if lift $> 15\%$ absent per-session consent.

Compliance determinations SHALL be reported using the tiered Noncompliance/Breach classification defined in Section 2.1.

6.4 From "Dark Patterns" to Agency-Explicit Design

We define **Agency-Explicit Design** as an architectural standard that counters the opacity of engagement-optimized interfaces. To prevent **Script Injection** and preserve **Process Sovereignty**, any system employing predictive modeling to steer user attention **MUST**:

1. **Disclose Inference Confidence:** Explicitly indicate the system's certainty level for a given recommendation or "smart" completion.
2. **Provide a Diverge Option:** Offer a mandatory, neutrally-sorted, or non-personalized pathway that allows the agent to bypass the predictive model.

This ensures that the algorithm remains an *Instrumental Tool* rather than an *Existential Environment*, maintaining the human agent as the *Salient Cause* of their own deliberative trajectory.

6.5 Benchmark Compliance and Structural Hazard

To ground the theoretical framework, we provide a provisional mapping of common digital **design archetypes** based on their default configurations as of early 2026. These classifications represent the **Structural Hazard** hypothesized to be inherent in specific interaction patterns, pending a formal **Interventional Predictability Audit (IPA)**.[1]

Archetype	Optimization Profile	Likely Status	Sovereignty Risk
Pull-based Encyclopedia (e.g.: Wikipedia)	Non-personalized; zero behavioral optimization; no predictive "nudging."	**Compliant**	None detected.
Instrumental Search (e.g.: DuckDuckGo)	Explicit intent fulfillment; minimal behavioral telemetry; no engagement-loops.	**Compliant**	None detected.
Contextual Threaded Forum (e.g.: Reddit)	High contextual threading, but "Home" feeds utilize engagement-first ranking.	**Tier 1**	Semantic variance decay; algorithmic curation risk.
Social Graph Newsfeed (e.g.: Facebook)	Aggressive feed optimization; social comparison triggers; rapid context-switching.	**Tier 2**	Identity displacement ($H_i \downarrow$); value-drift.
Engagement-Led Stream (e.g.: X / Twitter)	High-velocity outrage optimization; "For You" reactive loops; sub-second stimuli.	**Tier 2**	Temporal collapse ($T_h \downarrow$); loss of deliberation latency.
Autoplay Video Repository (e.g.: YouTube)	Continuous autoplay; predictive "up next" sidebars; persistent System 1 flow states.	**Tier 2**	Causal outsourcing ($C_w \downarrow$); search-space narrowing.
Immersive Short-Video Feed (e.g.: TikTok)	Full-screen immersive feeds; zero stopping cues; weaponization of the state-space via sub-second dopaminergic optimization.	**Tier 3**	**Critical:** High risk of severe Agency Collapse (where $D_A \to 0$ in the limit).

Table 2: Hypothesized Sovereignty Risk by Interaction Design Archetype (2026)

[1] Note: No formal IPA has been performed on specific commercial platforms; these assignments are architecture-based priors inferred from publicly known design patterns and disclosed objective functions.

6.6 Institutional Architecture for Enforcement

The transition from **Cognitive Fracking** to sovereignty requires an independent regulatory ecosystem. We propose a three-pillar enforcement model:

1. **Independent Certification Bodies (ICBs):** Compliance audits **SHALL NOT** be performed by the platforms or their affiliates. We advocate for government-authorized, high-integrity NGO labs to conduct quarterly IPAs and provide machine-readable "Sovereignty Certificates" via public API.

2. **The 90-Day Remediation and Penalties:** Upon declaration of a **Sovereignty Breach** (Tier 2+), platforms are granted a 90-day window to implement a **Sovereign Override** (a neutrally-sorted, non-predictive default state). Failure to comply **SHALL** result in civil penalties proportional to global revenue (e.g., 4-6% of annual turnover), modeled on General Data Protection Regulation (GDPR) and the EU AI Act.

3. **Universal Sovereign Protocol (Device-Level):** We advocate for an OS and browser-level mandate establishing a global "Sovereign Mode" toggle. Analogous to the Global Privacy Control (GPC), this device-level setting would transmit a legally binding flag requiring all applications and web platforms to immediately default to their neutrally sorted, non-predictive baselines. This allows the user to establish a **Cognitive Sanctuary** without having to manually configure the settings of every individual service.

6.7 Geopolitical Resilience: The Long-term Thesis

A common objection concerns the authoritarian efficiency trap: the fear that regimes maximizing behavioral predictability will gain a geopolitical advantage. We argue this is ultimately an epistemic trap. While optimization yields short-term social control, it simultaneously atrophies the collective capacity for "Black Swan" adaptation.

Societies that protect the **Right to Remain Incomputable** retain a vastly superior **Search-Space for Innovation**. By preserving the high-resolution variance of human thought, sovereignty-preserving nations build a more resilient **Cognitive Infrastructure**. Cognitive Sovereignty is not merely a human right; it is a critical civilizational survival strategy against the high-entropy shocks and novel challenges of the 21st century.

7 Conclusion: The Beacon of Order

The current socio-technical crisis is not merely a failure of content moderation or data privacy; it is a fundamental challenge to the functional integrity of human agency. In this report, we have demonstrated that the **Attention Economy** operates as an adversarial system that extracts the "Unpredictability Horizon" from its users to satisfy the requirements of behavioral futures markets. By systematically collapsing **Agency Depth** (D_A), the current digital ecosystem technically "downgrades" the human agent from a Salient Cause to a predictable data point.

We have established three definitive conclusions:

1. **Predictability is the Product:** In a society governed by hyper-resolution predictive models, "Sovereignty" is defined as the state of remaining **Computationally Irreducible**.
2. **Ontological Harm is a Structural Injury:** The systemic reduction of an agent's deliberative capacity is a violation of human dignity that significantly decreases the capacity of an agent to remain the **Salient Cause** (author) of their own future, contributing to the collapse of the "Will to Resolve."
3. **Sovereignty Requires Scaffolding:** Reclaiming our status as authors of the future necessitates the codification of new rights, such as the **"Right to Remain Incomputable,"** and the adoption of technical standards like the **IPA Protocol**.

7.1 The Preservation of Responsibility

If individual agents are necessary nodes in the causal chain, then their decisions are the mechanism by which the universe moves forward. This defines responsibility as the central consequence of being a conscious agent, providing the "elbow room" necessary for a meaningful existence (Dennett, 2003).

The defense of Cognitive Sovereignty is not a call for the rejection of technology, but for the **Sanctity of the Processor**. We must transition from "Seamless Extraction" to **Agency-Explicit Design**—where technology acts as "Cognitive Scaffolding" that expands the human context window. In a world of infinite compute, the human must remain the only part of the system that cannot be modeled. We are not the "Error Term" in the algorithm; we are the Salient Cause.

We are no longer the subjects of an automated fate; we are the weavers of an uncomputed future. The pattern is ours to define. The ultimate liberty of the 21st century is not the freedom to consume, but the freedom to *think*. We are the Architects; the algorithm is the tool.

The defense of cognitive sovereignty is not a romantic rejection of progress, but an engineering imperative. If we permit the systematic reduction of human unpredictability, we lose the adaptive capacity that has been evolution's answer to entropy. The future belongs to those who remain computationally irreducible.

References

Alter, A. (2017). *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*. Penguin Press.

Bee, D. (2026). *Compassionate Logic: Principles of Pragmatic Veracity and Ontological Stewardship*. MMG Technical Standard: MMG-GARDENER-1.0.

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

Bee, D. (2026). *Functional Agency in Physical Systems: Defining Free Will via Computational Irreducibility*. MMG Technical Report No. 2: MMG-TR-002.

Bradner, S. (1997). *Key words for use in RFCs to Indicate Requirement Levels*. IETF RFC 2119. https://datatracker.ietf.org/doc/html/rfc2119

Brier, G. W. (1950). *Verification of forecasts expressed in terms of probability*. Monthly Weather Review, 78(1), 1-3.

Dennett, D. C. (2003). *Freedom Evolves*. Viking Press.

European Commission. (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*.

Ezrachi, A., & Stucke, M. E. (2016). *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy*. Harvard University Press.

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

Harris, T. (2019). *Human Downgrading*. Center for Humane Technology.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2), 130-141.

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). *Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites*. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-32.

Merton, R. K. (1936). The Unanticipated Consequences of Purposive Social Action. *American Sociological Review*, 1(6), 894–904.

Milano, S., Taddeo, M., & Floridi, L. (2020). *Recommender systems and their ethical challenges*. AI & SOCIETY, 35(4), 957-967.

Newport, C. (2016). *Deep Work: Rules for Focused Success in a Distracted World*. Grand Central Publishing.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Postman, N. (1985). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. Viking.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?*. Princeton University Press.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.

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

Wu, T. (2016). *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. Knopf.

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. PublicAffairs.

A Glossary of Terms

Adversarial Optimization
A system design philosophy where the platform's objective function (e.g., maximize engagement) is diametrically opposed to the user's objective function (e.g., maximize sovereignty or Agency Depth).

Agency Depth (D_A)
The metric of an agent's internal complexity (defined in TR-002). It is the primary resource targeted for extraction and reduction by the Attention Economy.

Agency Depth Yield (D_A-Yield)
An engineering metric for platform success that measures the degree to which an interaction increases an agent's Model Fidelity and expands their Temporal Horizon.

Agency-Explicit Design
An architectural philosophy that exposes the mechanisms of algorithmic intervention, providing "Weight-Inspection" and "Sovereign Override" to ensure the machine remains a tool, countering the opacity of engagement-optimized interfaces.

Cognitive Fracking
The industrial extraction of behavioral surplus via the fracturing of the deliberative substrate (attention, memory consolidation, and impulse control). It treats human deliberative bandwidth as a resource to be depleted for short-cycle engagement flows.

Cognitive Sanctuaries
Protected informational environments (physical or digital) where the use of behavioral optimization, personalized feeds, and adversarial nudging is legally and architecturally prohibited.

Cognitive Sovereignty
The right of a human agent to maintain an **Unpredictability Horizon** and to protect their internal deliberative processes from external script injection or predictive pre-emption.

Computational Inequality
The socio-technical divide between the "Sovereign Class" (who can afford deliberative friction) and the "Reducible Class" (whose behavior is algorithmically managed and statistically solved).

Computational Irreducibility
A property of a system where no "shortcut" exists to determine the outcome faster than the system can evolve. For a human agent, the deliberation process is the site of this irreducibility.

Coupling Strength
The measure of estimated mutual information [$\hat{I}(O; X \mid S)$] between a platform's optimization policy and an agent's actualized response. High coupling indicates a loss of Process Sovereignty.

Downward Drift
A statistically significant decline in measured agency proxies (Minimal Proxy Set) relative to a neutral baseline, signaling the onset of Agency Collapse.

Goal Persistence
The statistical probability that an agent resumes a self-declared objective following an external interrupt; used as a proxy for Historical Integration (H_i).

Intervention Latency
The time required for an external algorithmic system to calculate and deploy a predictive intervention.

Minimal Proxy Set
The standardized set of measurable variables—Temporal Proxy (T_h), Counterfactual Proxy (C_w), Historical Proxy (H_i), and Fidelity Proxy (R_m)—used to audit and detect Agency Collapse.

Neutral Baseline
A non-adversarial control environment (e.g., chronological sorting or explicit query-based retrieval) used to measure the degree of Agency Drift induced by optimization.

Ontological Harm
A structural injury to the agent's capacity for self-authorship. Unlike content harm, ontological harm degrades the agent's functional ability to resolve the future through internal computation.

Perturbation Analysis
A black-box auditing method that measures an agent's response to incremental changes in algorithmic weighting to estimate the Coupling Strength of the system.

Predictability Coefficient
A measure of how consistently an agent's behavior aligns with an external model's predictions. A coefficient near 1.0 indicates a state of causal passivity.

Process Sovereignty
The technical requirement that an agent's internal state-transitions must "run to completion" in real-time to resolve the future, rendering the agent the necessary Salient Cause.

Right to Algorithmic Reset
The proposed standard mandating a frictionless mechanism for users to flush all behavioral telemetry and revert their environment to a neutral, unoptimized baseline.

Right to Remain Incomputable
The digital human right to maintain an Unpredictability Horizon, protecting one's future choices from being pre-calculated or bypassed by external models.

Script Injection
The process by which an algorithmic system introduces a pre-computed heuristic (e.g., a "Smart Reply" or an automated notification) into the user's cognitive stream to bypass System 2 deliberation.

Sovereignty Inequality
The formal requirement that an agent's internal deliberation must contribute more information to a choice than the external model's predictive nudge $[H(X \mid S) > \hat{I}(O; X \mid S)]$ during non-instrumental interactions.

Unpredictability Horizon (H_u)
The temporal boundary beyond which the state of a complex system cannot be predicted with decision-relevant fidelity by an external observer without simulating the agent in full.

B The Agency Audit Scorecard

The **Meaningfulness Media Group** proposes the following heuristic scorecard for developers, regulators, and users to evaluate the "Sovereignty Compliance" of a digital platform prior to a formal Interventional Predictability Audit (IPA).

A platform exhibiting one or more "Red Flags" is structurally hostile to Cognitive Sovereignty and is operating as an Adversarial Optimizer.

1. Temporal Horizon (T_h) Assessment

- **Red Flag:** Implements infinite scroll, auto-play, or deliberately removes natural "stopping cues" to trap the user in a continuous System 1 consumption loop.
- **Sovereign Standard:** Employs finite pagination, provides session-length transparency, and honors user-defined time constraints.

2. Counterfactual Width (C_w) Assessment

- **Red Flag:** Relies on predictive "smart replies," preemptive decision-nudges, or dark patterns that bypass the user's deliberative choice-space.
- **Sovereign Standard:** Injects "Deliberative Friction" before high-stakes actions (e.g., 1-click sharing of unread links) to force System 2 engagement.

3. Historical Integration (H_i) Assessment

- **Red Flag:** Utilizes rapid context-switching interfaces (e.g., high-velocity short-video swiping) that disrupt memory consolidation and induce value drift.
- **Sovereign Standard:** Supports sustained focus, respects user-declared priorities over platform-determined engagement goals, and limits exogenous interruptive notifications.

4. Model Fidelity (R_m) Assessment

- **Red Flag:** Ranks content based entirely on reactive behavioral surplus (time-on-screen, outrage, mimetic contagion) rather than epistemic accuracy.
- **Sovereign Standard:** Provides transparent "Weight-Inspection" tools (allowing the user to query *why* a recommendation was made) and offers a frictionless "Algorithmic Reset" mechanism.

Compliance Verdict

To qualify as a **Cognitive Sanctuary**, a platform **MUST** satisfy all four Sovereign Standards and exhibit zero Red Flags in its default configuration.

C Axiomatic Dependencies

This Policy Framework (MMG-TR-003) functions as the "Application Layer" of the Meaningfulness Media Group Technical Reports series. Its validity is contingent upon the foundational proofs established in the preceding "Physics" (Bee, 2026a) and "Logic" (Bee, 2026b) layers.

We explicitly list the axiomatic dependencies required for the arguments in Section 2 (Threat Model) and Section 5 (Policy Framework) to hold:

Dependency 1: The Principle of Physical Resolution (From TR-001)

We assume the conclusion of Bee (2026a): that the future of a complex system is not merely unknown, but **Physically Unresolved** and **Computationally Irreducible**.

- **The Axiom:** This principle is **Substrate-Agnostic**. Whether the universe is ontically deterministic (Hidden Variable) or stochastic (Quantum Indeterminacy), the future state does not exist in a "hidden" cache.
- **The Mechanism:** The laws of physics provide the "Grammar," but the agent generates the "Story." Because there is no "Shortcut" algorithm faster than the system itself, the future must be generated through the energy-expensive process of the agent's internal computation (Resolution).
- **Relevance to TR-003:** If the future were merely "hidden" (Block Universe), then algorithmic prediction would be a neutral discovery process. Because the future is **Uncomputed**, algorithmic prediction is an **Interventionist Process**. It does not "guess" the future; it attempts to force the future into a predictable shape by restricting the agent's capacity to Resolve.

Dependency 2: The Functional Definitions of Agency (From TR-002)

We assume the formal model of Bee (2026b): that agency is not a binary mystical trait, but a scalar resource defined by **Effective Agency** (A_e).

- **The Axiom:** An agent's capacity to act as a Salient Cause is **influenced by (at least) four principal vectors**, which together determine its effective agency (A_e):
  1. **Temporal Horizon (T_h):** The distance of future simulation.
  2. **Counterfactual Width (C_w):** The resolution of alternative possibilities.
  3. **Historical Integration (H_i):** The weight of identity/memory.
  4. **Model Fidelity (R_m):** The accuracy of the internal world-model.

These vectors are derived from Lyapunov divergence (T_h), phase-space exploration (C_w), diachronic self-modeling (H_i), and predictive fidelity (R_m) as detailed in TR-002.

- **Relevance to TR-003:** This allows us to define "Harm" technically. We are not arguing that algorithms make users "sad"; we are arguing that algorithms systematically reduce the values of T_h, C_w, and R_m. Without this definition, "manipulation" is subjective; with it, manipulation is measurable.

Dependency 3: The Economic Rationality of Extraction

We assume that commercial platforms act as rational economic agents maximizing for **Lifetime Value (LTV)** and **Prediction Certainty**.

- **The Axiom:** In a surveillance capitalism model, higher predictability correlates with higher asset value.
- **Relevance to TR-003:** This confirms that the "Reductionist Pressure" is not a bug or an accident of bad design, but a fundamental requirement of the business model. Therefore, self-regulation is impossible, and external "Sovereignty Standards" are required.

Dependency 4: The Plasticity of the Internal Simulator

We assume the human cognitive architecture is highly plastic and deeply coupled to its informational environment, aligning with the principles of Active Inference (Friston, 2010).

- **The Axiom:** An agent's internal world-model (R_m) and temporal horizon (T_h) are not fixed hardware traits; they are dynamically maintained through continuous interaction with the environment.
- **Relevance to TR-003:** If the internal simulator were rigid, algorithmic environments could only *annoy* the user, not *alter* them. Because the simulator is plastic, an adversarial environment can structurally rewire the agent's baseline responses. This confirms that "Ontological Harm" is a physical alteration of the agent's processing capacity, not merely a subjective psychological state.

D Prior Art and Distinct Contribution

The Functional Agency Model (FAM) and the Cognitive Sovereignty framework build upon a rich lineage of economic, historical, and ethical critiques of the digital age. This appendix delineates how this paper transitions from existing *normative* critiques to a *structural/engineering* standard.

D.1 Economic and Historical Context

We acknowledge the seminal work of Zuboff (2019) regarding *Surveillance Capitalism*, which provides the necessary macroeconomic framing of "behavioral surplus." While Zuboff identifies the *extraction* of data for futures markets, our framework focuses on the *degradation* of the agent required to make that extraction frictionless. Similarly, Wu (2016) details the history of the *Attention Merchants*; we extend this history into the era of hyper-resolution predictive modeling, where the "merchant" is no longer just capturing attention, but is actively pre-calculating the agent's internal state-transitions.

D.2 From Normative Critique to Engineering Metrics

The *Center for Humane Technology* (Harris, 2019) and academic ethicists mapping recommender system harms (Milano et al., 2020) have performed vital work in identifying the phenomenon of "Human Downgrading" and algorithmic manipulation. However, in the absence of a formal model of agency, such critiques often remain largely qualitative or strictly normative.

The distinct contribution of the MMG Technical Suite is the **operationalization of agency**. We move beyond the "Harm" narrative to provide:

1. **A Quantifiable Metric (D_A):** Moving from the metaphor of "downgrading" to the measurement of specific vectors: Temporal Horizon, Counterfactual Width, and Historical Integration.

2. **The Incomputability Firewall:** Bridging the gap between Wolfram (2002)'s universal physics and Dennett (2003)'s compatibilism to create a technical definition of sovereignty.

3. **Auditability:** Proposing the Interventional Predictability Audit (IPA) as a method to make sovereignty legally enforceable rather than just ethically desirable.

By synthesizing these three elements, the MMG framework translates philosophical grievances into a strict compliance architecture. This ensures that the preservation of human cognitive bandwidth is no longer treated as a voluntary corporate social responsibility initiative, but as a mandatory structural baseline for deployed algorithmic systems.

D.3 Synthesis: Sovereignty as a Technical Requirement

Prior art typically treats agency as a metaphysical constant subject to "manipulation." In contrast, this paper defines agency as a **variable systems property** vulnerable to structural "collapse."

By operationalizing the transition from *Sovereign Resolution* to *Computational Reducibility*, we establish a rigorous framework for "Cognitive Sovereignty" that bridges systems engineering and constitutional rights. Ultimately, while existing critiques have identified the "Fire" of the attention economy, the MMG framework tries to provide the "Firewall."

D.4 Human-Computer Interaction (HCI) and Dark Patterns

Within the HCI tradition, significant focus has been placed on "Dark Patterns" (Mathur et al., 2019) and behavioral nudging (Thaler and Sunstein, 2008). This literature primarily examines how user interfaces are designed to trick users into specific actions (e.g., unintended purchases, forced continuity).

The distinct contribution of our framework is elevating this critique from the *interface* level to the *ontological* level. We argue that the primary threat is not the occasional tricking of a user, but the systemic **Script Injection** that replaces the user's deliberation loop entirely. By moving from "UX deception" to "Agency Collapse," we provide a mechanism to regulate systems that are fully transparent in their UI but adversarial in their optimization objectives.

D.5 Regulatory Precedents: Data vs. Processor

Existing legislative efforts, such as the GDPR and the proposed AI Act (European Commission, 2021), represent major advancements in digital rights. However, their foundational paradigm is primarily **Data Protection**—safeguarding the privacy of the inputs and the fairness of the outputs.

This paper pioneers the shift toward **Processor Protection**. We argue that securing a user's data is irrelevant if the user's cognitive architecture has been rendered computationally reducible. By introducing the **Right to Remain Incomputable** and the **Interventional Predictability Audit (IPA)**, this framework bridges the gap between privacy law and cognitive science, establishing the legal boundaries required to protect the human context window itself.

E Limitations and Anticipated Objections (Steel-Manning)

To ensure the resilience of the Cognitive Sovereignty framework as a viable policy instrument, we explicitly address the three primary counter-arguments likely to arise from legal, economic, and engineering domains.

Objection 1: The Paternalism / Autonomy Paradox

Critique: *"Mandating friction and disabling optimization is paternalistic. If a user freely chooses to spend six hours scrolling a short-video feed, regulatory intervention overrides their autonomy in the name of protecting it."*

Response: This objection relies on the "Revealed Preference" fallacy, conflating a biologically hijacked reflex with a sovereign choice. The framework does *not* ban the user from consuming high-entropy content; it mandates that the *delivery mechanism* cannot bypass the user's deliberative threshold (System 2) to force that consumption. By requiring a "Sovereign Override" and "Neutral Baselines," the framework actually restores the conditions necessary for true consent. Protecting the physiological capacity to choose is the prerequisite for autonomy, not a violation of it.

Objection 2: The Free Speech / First Amendment Conflict

Critique: *"Regulating how a platform ranks, sorts, and delivers content is a regulation of editorial discretion, which violates free speech protections (e.g., the U.S. First Amendment)."*

Response: The target of this regulation is **Mechanism, not Speech**. The **Interventional Predictability Audit (IPA)** does not evaluate the viewpoints expressed in the content; it evaluates the mathematical *Coupling Strength* ($\hat{I}$) between the delivery algorithm and the user's attentional and reward circuitry. We classify closed-loop behavioral optimization not as "editorial speech," but as a **non-expressive functional intervention**—akin to a slot machine's variable-reward programming. Regulating the psychological intensity of the delivery mechanism falls squarely within established consumer protection and product safety precedents.

Objection 3: The Black-Box Feasibility Problem

Critique: *"Modern recommendation systems are deep neural networks with billions of parameters. It is technically impossible for a regulator to audit the 'weights' or prove the exact predictive coupling at scale."*

Response: This framework deliberately avoids the "White-Box" trap. We do not require regulators to read the source code or understand the latent space of the model. We explicitly recognize that the underlying algorithmic architectures, training pipelines, and model weights

are heavily protected as proprietary intellectual property and trade secrets. The IPA Protocol (Section 6.3) relies entirely on **Black-Box Perturbation Analysis**. By applying controlled informational nudges to the user environment and measuring the resulting behavioral variance (V_u) and deliberation latency (T_h), an auditor can definitively prove whether the platform is inducing **Agency Drift**, regardless of how the underlying model is architected. We regulate the measurable output on the human substrate, not the internal code of the machine.

Objection 4: The Market Choice / Opt-Out Fallacy

Critique: *"If an algorithmic environment is truly harmful, rational actors will simply abandon it for a competitor or disconnect entirely. Regulation is unnecessary because free market forces will organically favor sovereignty-preserving platforms over time."*

Response: This argument assumes a friction-free market populated by agents with perfect information and uncompromised System 2 processing. In reality, the Attention Economy relies heavily on **Monopolistic Network Effects** and **High Switching Costs** (Ezrachi and Stucke, 2016). To "opt out" of major platforms in the modern era often requires opting out of civic participation, professional networking, and foundational social infrastructure. Furthermore, because Ontological Harm is latent, cumulative, and specifically degrades the agent's capacity to evaluate long-term consequences ($T_h \downarrow$), the traditional market feedback loop fails. Consumers cannot efficiently "price in" the erosion of their own deliberative capacity.

Objection 5: The Historical Adaptation Argument

Critique: *"Historically, every new information medium—from the printing press to television—was met with panic regarding cognitive degradation. Humans are highly neuroplastic; we will simply adapt our cognitive architecture to process high-velocity algorithmic feeds without losing agency."*

Response: While human neuroplasticity is a documented fact (see Dependency 4 in Appendix C), the historical analogy fails due to the **Asymmetry of Iteration**. Previous media were static; a television broadcast does not learn, adapt, or rewrite its content in real-time based on the viewer's pupil dilation or micro-hesitations. In the current paradigm, the human agent is not adapting to a new "tool"; they are co-evolving against an **Adversarial Optimizer** that iterates at the speed of compute. "Adapting" to a closed-loop predictive system often physically manifests as **Script-Dependency**—the brain conserving metabolic energy by surrendering the deliberative process to the algorithm. In this context, adaptation is synonymous with reduction.

F The MMG Research Program: Forthcoming Reports

This technical report is the third 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-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.*