AEGIS Protocol: A Forensic Methodological Audit and Architectural Reconstruction

1. Introduction: The Ergodicity Crisis and the imperative for N-of-1 Precision

The paradigm of Evidence-Based Medicine (EBM), the cornerstone of clinical practice for the last half-century, is currently facing an epistemic crisis precipitated by the digital transformation of healthcare. Traditional EBM relies fundamentally on the Randomized Controlled Trial (RCT) to generate the Average Treatment Effect (ATE). The statistical validity of applying a population-derived ATE to a specific individual patient rests on a largely unexamined assumption from statistical mechanics: **ergodicity**. This assumption posits that the ensemble average (the mean outcome of a population at a single point in time) is equivalent to the time average (the mean outcome of a single individual over an infinite time horizon). In complex biological systems—characterized by feedback loops, non-stationarity, hysteresis, and path dependence—this assumption is demonstrably false. A drug with a positive ATE may be inert or even toxic for a specific patient due to distinct genetic, environmental, or physiological boundary conditions.[1]

The transition from population-based statistics to precision N-of-1 therapeutics represents the defining computational challenge of modern healthcare. It requires a shift from "learning from N subjects" to "learning from T time points for $N=1$." This inversion moves the problem from the domain of "Big Data" (where the Law of Large Numbers suppresses variance) to the domain of "Small Data" (where variance, autocorrelation, and unobserved confounding dominate). The attached "AEGIS Protocol" (Adaptive Engineering for General Individualized Safety) proposes a next-generation architecture designed to synthesize the disparate strengths of recent advancements—specifically the CHRONOS, STAR, and SCANT protocols—into a cohesive "Grey Box" cyber-physical system. It aims to solve the "Small Data Paradox" not by training neural networks from scratch, but by embedding mechanistic priors into Universal Differential Equations (UDEs), and to solve the safety problem via Formal Verification rather than probabilistic penalties.

However, a rigorous forensic audit of the AEGIS specification reveals that while the architectural intuition is sound, the specific methodological choices contain latent theoretical flaws that threaten the validity, safety, and identifiability of the system. This manuscript serves two purposes: first, to perform a forensic academic audit of the AEGIS protocol, identifying critical vulnerabilities such as the misappropriation of Natural Language Processing (NLP) performance metrics, the divergence risks of the Unscented Kalman Filter (UKF) in chaotic

physiological regimes, and the violation of exchangeability in Micro-Randomized Trials (MRTs) due to circadian confounding. Second, to reconstruct the protocol into **AEGIS 2.0**, a peer-review-ready specification that integrates **Action-Centered Contextual Bandits** with **Double-Robust G-Estimation**, an **Adaptive Constrained UKF (AC-UKF)**, and a **Simplex Safety Architecture** grounded in Signal Temporal Logic (STL). This reconstruction provides the rigorous mathematical guarantees required for regulatory acceptance in high-stakes digital health interventions.

---

2. Forensic Audit of Layer 1: Data Ingestion and The Sentinel System

The foundation of any N-of-1 causal inference engine is its data layer. Unlike traditional clinical trials where data is collected by trained coordinators under controlled conditions, N-of-1 digital trials rely on noisy, heterogeneous data streams: patient-reported outcomes (PROs) via mobile diaries, continuous sensor data from wearables, and sporadic clinical notes. The AEGIS protocol correctly identifies that "garbage in, garbage out" is the Achilles' heel of these systems and proposes a "Sentinel System" to ingest, adjudicate, and standardize this data. However, our audit reveals a critical reliance on misappropriated performance metrics and a misunderstanding of the semantic entropy inherent in patient-generated text.

2.1 The "Hybrid Adjudication" Fallacy: A Category Error in NLP Validation

The AEGIS specification proposes a "Confidence-Based Adjudication" workflow where a Large Language Model (LLM) extracts structured state variables (S_t) from unstructured text (e.g., patient diaries). The protocol explicitly claims:

> *"Recent research indicates that this 'hybrid' adjudication (AI + Human for uncertain cases) achieves **91% agreement** with gold-standard committees while reducing human workload by **84%**."* [1]

Forensic Analysis of the Citation:
A forensic trace of this statistic reveals it does not originate from an N-of-1 PRO context but rather from the DELIVER trial and related analyses of the Dapagliflozin heart failure studies.2 In those studies, NLP models were deployed to adjudicate Heart Failure Hospitalizations based on clinical notes (discharge summaries, physician narratives) stored in Electronic Health Records (EHRs). The "91% agreement" refers to the concordance between the NLP model's binary classification (Event/No Event) and a Clinical Events Committee (CEC) of human experts reviewing the same professional documentation.
Theoretical Vulnerability: The Domain Shift Problem
Applying performance metrics from EHR-based population trials to N-of-1 patient diaries

constitutes a severe category error and a threat to internal validity. The linguistic distribution of clinical notes differs fundamentally from patient-generated text:

1. **Lexical Consistency:** Clinical notes are written by professionals using a standardized, albeit complex, ontology (e.g., "dyspnea," "edema," "orthopnea"). Patient diaries are highly idiosyncratic, employing colloquialisms, metaphors ("felt like a truck hit me," "brain fog"), and ambiguous temporal markers ("after the show," "a while ago"). An NLP model tuned or validated on EHR data (as implied by the cited statistics) will suffer significant performance degradation—known as **domain shift**—when applied to the high-entropy distribution of PROs.
2. **Adjudication Target:** The DELIVER study adjudicated a binary outcome (Event/No Event). AEGIS requires the extraction of **continuous state variables** (e.g., "pain intensity," "dietary adherence") and, crucially, **precise timestamps** for causal inference. The error surface for continuous variable extraction is fundamentally different. A minor timestamp error (e.g., mapping "this morning" to 11:00 AM instead of 08:00 AM) can invert the causal ordering of Treatment (A_t) and Outcome (Y_{t+1}), leading to causal feedback loops where the effect is mistaken for the cause.
3. **Confidence Calibration:** The protocol relies on the LLM's "Self-Confidence Score" ($C \in $) to trigger human review. However, modern LLMs are notoriously **uncalibrated**, often assigning high confidence to "hallucinated" extractions. Relying on raw confidence scores without calibration or entropy-based uncertainty quantification exposes the system to silent data corruption.

2.2 Semantic Sentinels and The Collider Bias Risk

The AEGIS protocol uses "Semantic Sentinels" (S_{sem}) to flag "Event Markers" (e.g., "started yoga") without adding nodes to the Causal DAG to avoid "graph instability".[1] The rationale is that N-of-1 data is too sparse to support continuous structure learning (the "Phoenix" failure mode).

Methodological Critique:
While avoiding graph instability is crucial, the proposed solution of simply "flagging" an event without mathematically adjusting for it introduces Unobserved Confounding. If an external event U_t (e.g., "started yoga") causes both a reduction in the outcome Y_t (e.g., "Stress") and a change in the treatment assignment A_t (e.g., "Medication Adherence"), failing to include U_t in the causal adjustment set violates the Sequential Ignorability assumption (or "No Unmeasured Confounders" assumption) required for valid G-estimation.

$$Y_t \not\perp A_t | H_t \quad \text{if } U_t \in \text{Parents}(Y_t) \cap \text{Parents}(A_t) \text{ and } U_t \notin H_t$$
The Semantic Sentinel essentially acts as a sensor for a confounder but fails to pass that information to the estimator in a usable format. By excluding these "flagged" variables from the adjustment set, the system allows the treatment effect estimate to become biased. Conversely, if the system adapts treatment probability p_t based on U_t but does not

adjust for it in the outcome model, it introduces bias. If U_t is a common effect of treatment and outcome (a collider), conditioning on it (even implicitly via selection) opens a backdoor path, distorting the causal signal.

2.3 Standardization Gaps: The Ontology Problem

The protocol mentions using **FHIR** (Fast Healthcare Interoperability Resources) and **OMOP** (Observational Medical Outcomes Partnership) for standardization.[1] While these are appropriate standards for *storage*, they are insufficient for *inference* without a semantic reasoning layer. The protocol lacks a mechanism to resolve semantic ambiguity. For instance, mapping "drowsiness" (PRO) and "somnolence" (EHR) to the same causal node requires a formal ontology like **SNOMED CT** with is_a relationships. The current specification treats this as a formatting exercise rather than a semantic prerequisites for causal identification.

---

3. Forensic Audit of Layer 2: The Physiologic Digital Twin (State Estimation)

AEGIS proposes a "Grey Box" approach using **Universal Differential Equations (UDEs)** and the **Unscented Kalman Filter (UKF)** to estimate hidden physiological states.[1] This layer is tasked with solving the "Small Data Paradox" by enforcing physical priors.

3.1 Universal Differential Equations (UDEs) in N=1

The protocol utilizes UDEs defined as:

$$\frac{dx}{dt} = f_{mech}(x, u; \theta_{fixed}) + NN(x, u; \theta_{learn})$$

This relies on the work of Rackauckas et al. 5 regarding Scientific Machine Learning.
Theoretical Assessment:
This is the strongest theoretical component of the AEGIS architecture. In N-of-1 trials, data is sparse (T is small). A pure neural network (NN) requires thousands of samples to learn basic physics (e.g., "insulin clears from the blood over time"). By embedding f_{mech} (e.g., the Bergman Minimal Model for glucose dynamics), the NN only needs to learn the residual dynamics—the patient-specific deviation from the textbook model. This reduces the sample complexity by orders of magnitude, effectively making "Small Data" sufficient for robust estimation. The fixed prior f_{mech} acts as a regularizer, constraining the hypothesis space to physically plausible trajectories.

3.2 The Unscented Kalman Filter (UKF) Divergence Risk

AEGIS selects the UKF over the Extended Kalman Filter (EKF) to avoid Jacobian instability in non-linear biological systems.[1] It cites standard benefits: 2nd-order accuracy in capturing the

mean and covariance of the transformed distribution, and derivative-free implementation.[8]

Forensic Analysis: The Chaos & Divergence Problem
While UKF is generally superior to EKF for non-linear systems, the protocol fails to account for UKF Divergence in systems with specific structural properties often found in physiology (e.g., chaotic attractors in heart rate variability, limit cycles in circadian rhythms). Research by Feng, Tse, and others 10 demonstrates that in chaotic regimes (positive Lyapunov exponents), the standard UKF error covariance (P_k) can become inconsistent.

- **The Mechanism of Failure:** In a chaotic system, small uncertainties in the state estimate grow exponentially over time (the "butterfly effect"). If the filter's update frequency is lower than the timescale of the chaotic divergence, the predicted covariance P_k^- (calculated via the Unscented Transform) may underestimate the true dispersion of the state. This leads to the **"Smug Filter"** problem: the filter believes it is precise (P_k is small) and therefore ignores new measurements (the Kalman Gain K_k approaches zero). The state estimate $\hat{x}_k$ diverges from the true state x_k while the filter reports high confidence.
- **Sigma Point Collapse:** If the physiological process noise covariance Q is underestimated (a common occurrence when learning from sparse N-of-1 data), the sigma points may collapse towards the mean. This effectively reduces the UKF to a biased estimator that fails to capture the non-linear spread of the distribution.
- **Positivity Violation:** Standard UKF assumes Gaussian distributions for the state variables. However, biological variables (e.g., Cortisol, Glucose, Drug Concentration) are strictly non-negative. A standard UKF can generate sigma points in the negative domain (e.g., -5 mg/dL glucose). When these negative sigma points are propagated through the biological ODE f_{mech}, they can cause the solver to crash (e.g., taking the square root of a negative number) or return physical nonsense, destabilizing the entire Digital Twin.

3.3 The Absence of Adaptive Noise Estimation

The AEGIS 1.0 specification assumes fixed noise covariances (Q for process noise, R for measurement noise). In an N-of-1 setting, the patient's physiology is non-stationary. A patient fighting an infection will have much higher physiological volatility (higher Q) than when healthy. A fixed Q leads to suboptimal filtering: if Q is too low during a high-volatility period, the filter lags; if Q is too high during a stable period, the filter chases noise. The lack of **Adaptive Covariance Matching** is a significant oversight that limits the robustness of the Digital Twin.

---

4. Forensic Audit of Layer 3: Causal Inference Engine

The AEGIS causal engine uses **Micro-Randomized Trials (MRTs)**, **Structural Nested Mean Models (SNMMs)**, and **G-Estimation** to answer the counterfactual question "Why?" and

estimate the Individual Treatment Effect (ITE).[1]

4.1 Micro-Randomized Trials (MRTs) and Exchangeability

The MRT is the correct experimental design choice for developing Just-In-Time Adaptive Interventions (JITAIs).[13] By randomizing the intervention A_t at hundreds of decision points k, it maximizes the effective sample size and power to detect proximal effects.

Theoretical Flaw: The Circadian Confounding Trap
The protocol assumes that randomization p_k (the probability of treatment) guarantees exchangeability (the independence of treatment assignment and potential outcomes). However, in human physiology, Time of Day is a massive, cyclical confounder.

- **The Scenario:** Consider an MRT for physical activity. The protocol randomizes a "Walk Prompt" (A_t) with equal probability ($p=0.5$) at 9:00 AM and 9:00 PM. The outcome Y_{t+1} (Step Count) is naturally higher in the morning due to circadian rhythms, regardless of the prompt.
- **The Bias:** A naive estimator might calculate the effect as $E - E$. If the prompts are delivered uniformly, this is unbiased. However, if the patient's *availability* (a context variable S_t) varies—e.g., they are only available to receive prompts in the evenings when they are tired—the "treated" moments will systematically sample from the "low activity" circadian phase. If the causal model does not explicitly adjust for the **cyclical nature of the baseline outcome**, the treatment effect will be confounded by the time of day. The AEGIS protocol mentions "Sentinel-Triggered Regimes" but does not explicitly enforce **Stratified Randomization** or **Time-Varying Centering** to handle this deterministic confounding.

4.2 G-Estimation and The Blip Function

AEGIS uses G-estimation to estimate the parameters ψ of the Blip Function $\gamma(H_t, A_t; \psi)$.[1]

$$E = \gamma(H_t, a_t; \psi)$$

The Blip function represents the causal effect of a "blip" of treatment at time t on the outcome, removing the effect of all future treatments.
Audit of Identifiability:
The document claims this solves "time-varying confounding." This is theoretically sound.15 G-estimation is superior to standard regression (which introduces bias when adjusting for colliders or mediators) and Inverse Probability Weighting (IPW) (which is numerically unstable with high-variance weights in long time series). However, the protocol does not detail the estimating equation or the conditions for Double Robustness in the N-of-1 context. Without linking the G-estimator to the Digital Twin's predictions, the estimator remains inefficient, relying solely on the randomization for consistency.

4.3 Martingale Confidence Sequences (MCS)

AEGIS uses MCS for "Anytime Validity" to allow continuous monitoring of the trial without inflating Type I error rates (the "Peeking" problem).[1]

Audit:
The protocol cites the STAR protocol for MCS but relies on vague descriptions. Standard Hoeffding-based MCS are often too wide (conservative) to be practically useful in the short duration of an N-of-1 trial (e.g., 2-4 weeks). If the confidence sequence does not shrink fast enough, the trial will never conclude efficacy, rendering the "Anytime Validity" feature theoretically nice but clinically useless. The protocol lacks a specification for Mixture Martingales or Betting Martingales, which are necessary to achieve tight bounds for small sample sizes.17

---

5. Forensic Audit of Layer 4: The Optimiser (Decision & Policy)

Once the causal effect is estimated, AEGIS must select the optimal action. This is the domain of the Optimiser, which balances exploration (learning) with exploitation (healing). AEGIS employs **Action-Centered Contextual Bandits** and **Thompson Sampling**.[1]

5.1 Action-Centered Bandits: The Variance Reduction Key

The choice of Action-Centered Bandits is methodologically excellent.[19]

$$R_t = f(S_t) + A_t \cdot \tau(S_t) + \epsilon_t$$

By decomposing the reward into a baseline $f(S_t)$ and a treatment effect $\tau(S_t)$, the bandit focuses solely on learning $\tau(S_t)$. Since $f(S_t)$ (the patient's baseline health trajectory) accounts for the vast majority of the variance in the outcome, subtracting it out (Variance Reduction) leads to much faster learning rates. The regret bounds scale with the dimension of τ rather than the full state space dimension.
Critique: The "Seldonian Bottleneck"
The protocol fails to address Posterior Collapse in Thompson Sampling when safety constraints are active. If the Safety Layer (Layer 5) blocks the "optimal" arm repeatedly because it is near a constraint boundary, the posterior distribution for that arm may not update (as the action is never taken). This can lead to a "deadlock" where the bandit remains uncertain about a potentially highly effective treatment because it is never allowed to try it. Alternatively, if the safety layer forces a "safe" action, the bandit might update its policy based on the safe action's outcome, potentially learning a suboptimal policy that is "safe but ineffective." The protocol lacks a mechanism for Off-Policy Safety Evaluation or Safe

Exploration Padding.

---

6. Forensic Audit of Layer 5: The Dual-Safety Supervisor

AEGIS employs a "Swiss Cheese" safety model: **Signal Temporal Logic (STL)** (Hard Constraints) and **Seldonian Constraints** (Probabilistic Constraints).[1]

6.1 Signal Temporal Logic (STL) and The Circularity Flaw

STL provides a rigorous mathematical language for specifying physiological boundaries (e.g., "Glucose must not drop below 70 mg/dL for more than 15 minutes").[21] The "Robustness Degree" ρ quantifies the spatial and temporal distance to a violation.

Forensic Analysis: The Dependence on the Twin
The protocol relies on the Digital Twin (UKF/UDE) to predict the future trajectory $x_{t:t+k}$ and compute the robustness $\rho(x_{t:t+k}, \phi)$.

- **The Circularity:** The safety of the system depends entirely on the accuracy of the Digital Twin. If the UKF diverges (as identified in Section 3) or if the UDE model is misspecified, the predicted trajectory might look "safe" ($\rho > 0$) while the patient is actually on a trajectory toward a critical event. The "Hard" safety layer is effectively "Soft" because it relies on a probabilistic model (the Twin). A safety layer that trusts the model it is supposed to police is a fundamental architectural violation of safety engineering principles (e.g., simplex architecture).

6.2 Seldonian Constraints in a Cold Start

Seldonian algorithms ensure $P(g(\theta) \le 0) \ge 1-\alpha$ with high confidence.[22]

Audit:
The protocol correctly identifies this as the solution to "probabilistic safety" (e.g., preventing the probability of nausea from exceeding 5%). However, Seldonian algorithms typically require a batch of training data (D_{train}) to compute the safety test before deployment. In an N-of-1 trial, we start with $N=0$ data points. The protocol offers no solution for the "Cold Start" problem. How does the system guarantee safety on Day 1 when it has no data to compute the concentration inequalities required for the Seldonian test?

---

7. AEGIS 2.0: The Reconstructed Architecture

Based on the forensic audit, we present the reconstructed **AEGIS 2.0 Protocol**. This architecture resolves the identified flaws: divergence of UKF, circadian confounding in MRT,

citation errors in adjudication, and circularity in safety logic.

7.1 Table of Architectural Reconstructions

Layer	Component	AEGIS 1.0 (Flawed)	AEGIS 2.0 (Reconstructed)	Mathematical Justification
L1: Data	Adjudication	"Hybrid" NLP (91% acc - Population metrics)	**Ontology-Constrained Sentinel** with Entropy Thresholding	Prevents domain shift error; quantifies uncertainty via $H(S_t)$.
L1: Data	Standardization	FHIR/OMOP	FHIR/OMOP + **Propensity Augmentation**	Satisfies Sequential Ignorability by passing confounders to Causal Engine.
L2: State	Model	UDE + Standard UKF	UDE + **Adaptive Constrained UKF (AC-UKF)**	Prevents divergence in chaotic regimes; enforces physiological non-negativity.
L3: Causal	Design	Standard MRT	**Contextual Stratified MRT**	Removes circadian confounding via time-varying randomization $p_t(S_t)$.
L3: Causal	Inference	G-Estimation	**Double-Robust G-Estimation**	Consistent if *either* Twin *or* Randomization

			(linked to Bandit)	is correct.
L4: Control	Policy	Action-Centered Bandit	Action-Centered Bandit with **Constrained Thompson Sampling**	Solves "Seldonian Bottleneck" via rejection sampling on the posterior.
L5: Safety	Supervisor	STL (Twin-Dependent)	STL + **Model-Free Simplex Reflex**	Breaks circularity; guarantees safety even if Digital Twin diverges.

7.2 Detailed Reconstructed Workflow

Layer 1: The Ontology-Constrained Sentinel System

To address the "Hybrid Adjudication" fallacy, AEGIS 2.0 enforces strict semantic constraints.

1. Entropy-Based Thresholding: The confidence threshold τ_{auto} is dynamic. Let $P(\hat{S}_t)$ be the probability distribution over extracted tokens. We compute the entropy $H(\hat{S}_t) = -\sum p_i \log p_i$.

   $$\text{Trigger HITL if } H(\hat{S}_t) > \delta_{entropy}$$

   This captures ambiguity that raw confidence scores miss.
2. **SNOMED CT Mapping:** The LLM does not output strings; it maps to **SNOMED CT** concept IDs using a constrained decoding grammar (e.g., JSON-Schema).
   - *Constraint:* extract(text) -> {concept_id: "SNOMED:22253000", value: float, unit: "mg"}
   - This ensures that "Drowsiness" and "Sleepiness" map to the same node in the Causal DAG, preserving data density.
3. **Propensity Augmentation:** When the Semantic Sentinel detects an event U_t (e.g., "Yoga"), it does not edit the DAG. Instead, it appends U_t to the **Propensity Set** $\mathcal{H}_t$ used by Layer 3. This ensures U_t is adjusted for in the G-estimation without requiring structural learning.

Layer 2: The Adaptive Constrained UKF (AC-UKF)

To prevent divergence and ensure physical plausibility, we implement the **AC-UKF**.

1. **Constraint Enforcement:** We use **Projected Sigma Points**.
    - Let $\mathcal{X}_{sigma}$ be the set of sigma points.
    - Apply projection operator $\Pi_{\mathcal{C}}(x) = \max(0, x)$ (or specific physiological bounds) to every point before propagating through f_{mech}.
    - This prevents the ODE solver from encountering invalid states (e.g., negative glucose).
2. **Adaptive Noise Estimation:** We estimate the process noise covariance Q_k online to match the observed residuals.[24]
    - Calculate residual: $\epsilon_k = y_k - h(\hat{x}_k^-)$.
    - Estimate theoretical covariance: $S_k = H P_k^- H^T + R_k$.
    - Adaptation Law:

      $$\Delta Q_k = K_k (\epsilon_k \epsilon_k^T - S_k) K_k^T$$
      $$Q_{k+1} = \text{smoothing}(Q_k + \lambda \Delta Q_k)$$
    - If the model mismatch increases (high residuals), Q_k inflates. This increases the Kalman Gain K_k, forcing the filter to rely more on measurements than the potentially incorrect UDE physics. This prevents the "Smug Filter" divergence.

Layer 3: Double-Robust Causal Inference Engine

We link the Bandit (Layer 4) and the G-Estimator (Layer 3) to achieve **Double Robustness**.

- The Estimating Equation: We estimate the Blip parameter ψ by solving:

  $$\sum_{t=1}^T \left(Y_{t+1} - \hat{\mu}(S_t) - \psi A_t S_t \right) (A_t - p_t) = 0$$
    - $\hat{\mu}(S_t)$ is the baseline prediction from the **Digital Twin** (Layer 2).
    - $(A_t - p_t)$ is the centered treatment assignment (from the MRT).
- **Proof of Double Robustness:**
    - If the **Digital Twin is correct** ($\hat{\mu} \approx E$), the residual variance is minimized, and the estimator is highly efficient.
    - If the **Digital Twin is wrong** (UDE misspecified), the term $(A_t - p_t)$ has mean zero (guaranteed by randomization). Therefore, the expectation of the estimating equation is still zero at the true ψ.
    - **Result:** The estimator $\hat{\psi}$ is consistent even if the physics model is completely wrong, provided the randomization probabilities are known. This is the "Fail-Safe" for causal inference.

Layer 4: Action-Centered Bandits with Constrained Sampling

To handle the "Seldonian Bottleneck," we employ **Constrained Thompson Sampling**.

1. **Posterior Sampling:** At time t, draw a candidate parameter vector $\tilde{\beta}$ from the posterior $N(\hat{\beta}_t, \Sigma_t)$.
2. **Safety Rejection:** Check if the policy induced by $\tilde{\beta}$ violates the Seldonian

constraint using the *current* safety data buffer.
    - If **Safe**: Accept $\tilde{\beta}$ and select action $A_t = \arg\max E$.
    - If **Unsafe**: Reject $\tilde{\beta}$ and resample.
3. **Result:** The bandit explores the parameter space but is constrained to the "Safe Set." This allows it to find the optimal *safe* treatment without violating the "Do-No-Harm" principle.

Layer 5: The Simplex Safety Architecture

We replace the Twin-dependent STL monitor with a **Simplex Architecture**.

1. **Complex Controller (AEGIS):** The UDE/Bandit system proposes an action $A_{complex}$.
2. **Safety Monitor (STL):** Checks $A_{complex}$ against the Digital Twin's prediction. If $\rho < 0$, block.
3. **Reflex Controller (The Failsafe):** A simple, model-free logic layer (e.g., "If $HR > 140$, stop").
    - This layer ignores the Twin and looks directly at the sensors.
    - **Decision Logic:** $A_{final} = (Reflex(S_t) == Safe)? A_{complex} : A_{reflex}$.
    - This breaks the circularity. Even if the Twin hallucinates that the patient is safe, the Reflex layer (which has no AI) will catch the violation based on raw sensor data.

---

8. Theoretical Validation: Simulation Scenarios

To validate the superiority of AEGIS 2.0, we consider two theoretical scenarios where AEGIS 1.0 would likely fail.

Scenario A: The "Flu" Shock (Non-Stationarity)

- **Context:** A patient in a hypertension trial develops the flu. Their baseline heart rate (HR) spikes, and variability (Q) increases.
- **AEGIS 1.0 Failure:** The standard UKF assumes fixed Q. It interprets the high HR residuals as measurement noise, keeping the state estimate smooth but wrong (lagging). The Causal Engine attributes the high HR to the drug (Treatment A) or lack thereof, biasing the effect estimate.
- **AEGIS 2.0 Success:** The **AC-UKF** detects the spike in residuals. The adaptation law inflates Q_k. The filter opens up, tracking the new volatile state accurately. The **Supremum Wald Test** (running on the residuals) detects the structural break.[25] The system declares a "Regime Shift," resets the Bandit's covariance (increasing exploration), and effectively "re-learns" the patient's dynamics in the new physiological context.

Scenario B: The "Circadian" Trap

- **Context:** An activity promotion trial. The patient is naturally more active in the morning.

- **AEGIS 1.0 Failure:** The MRT randomizes prompts uniformly. The bandit learns that "Prompts in the morning" yield high steps, attributing the effect to the prompt.
- **AEGIS 2.0 Success:** The **Contextual Stratified MRT** adjusts randomization or, more importantly, the **Double-Robust G-Estimator** includes the cyclical baseline $\mu(t)$ (Fourier terms) in the adjustment. The estimator subtracts the morning "lift" from the outcome. It reveals that the *marginal* effect of the prompt is actually zero (or lower) in the morning, correctly identifying that the prompt is only effective in the afternoon "slump."

---

9. Conclusion

The original AEGIS Protocol represented a significant theoretical ambition: to unify Causal Inference, Control Theory, and Formal Verification for N-of-1 medicine. However, our forensic audit revealed that the specific implementations—specifically the reliance on population-level NLP metrics, standard UKF, and naive MRTs—introduced critical vulnerabilities that threatened the system's validity and safety.

The **AEGIS 2.0** reconstruction presented here addresses these flaws through rigorous mathematical re-engineering. By enforcing **Ontology-Constrained Adjudication**, implementing the **Adaptive Constrained UKF**, and establishing **Double Robustness** through the linkage of Action-Centered Bandits and G-estimation, we provide a blueprint that is not only theoretically elegant but clinically robust. The addition of the **Simplex Safety Architecture** ensures that the system respects the "Do-No-Harm" imperative even in the face of AI model failure. This reconstructed specification offers a definitive path forward for the regulatory approval and clinical deployment of autonomous precision therapeutic systems.

Mathematical Appendix: Derivations

A.1 AC-UKF Covariance Adaptation

The measurement residual is $\epsilon_k = z_k - h(\hat{x}_k^-)$.
The theoretical covariance of the residual is $S_k = H_k P_k^- H_k^T + R_k$.
We seek to match the actual covariance $C_{\epsilon} \approx \epsilon_k \epsilon_k^T$ with the theoretical covariance S_k.
If $\epsilon_k \epsilon_k^T > S_k$, the model is overconfident. We inflate Q_k.
The update law (simplified):

$$Q_{k+1} = Q_k + \alpha K_k (\epsilon_k \epsilon_k^T - H_k P_k^- H_k^T - R_k) K_k^T$$

where α is a forgetting factor. This ensures asymptotic convergence of the covariance.24

A.2 Double Robustness of the G-Estimator

Let the estimating function be $U(\psi, \eta) = (Y - \mu(S; \eta) - \psi A S)(A - p)$.
We want $E[U(\psi^*, \hat{\eta})] = 0$.

$$E[U] = E_S$$

$$= E_S E + Cov(Y, A | S)]$$

Since randomization ensures $E = 0$ and conditional independence (ignoring the blip), the first term vanishes.
The Double Robustness comes from the fact that we model $\mu(S)$ (the nuisance parameter). If $\mu(S)$ is correct, variance is low. If $\mu(S)$ is wrong, the term $(A-p)$ (which is mean zero) ensures the integral is still zero, preserving consistency.15

A.3 Action-Centered Regret Bound

For the linear reward model $r_t = f(S_t) + A_t \tau(S_t) + \epsilon_t$, the Action-Centered Bandit achieves a regret bound:

$$R(T) = \tilde{O}(d_{eff} \sqrt{T})$$

where d_{eff} is the dimension of the treatment effect parameters τ, not the baseline f. Standard Contextual Bandits (e.g., LinUCB) would have regret scaling with $dim(f) + dim(\tau)$. Since f (baseline physiology) is complex and high-dimensional, while τ (treatment effect) is often sparse, this represents a massive gain in sample efficiency.19

Works cited

1. HYPERION Architecture_ Flaws and Solutions.pdf
2. Clinical Events Classification (CEC) in Clinical Trials: Report on the Current Landscape and Future Directions — Proceedings from the CEC Summit 2018 | Request PDF - ResearchGate, accessed December 13, 2025, https://www.researchgate.net/publication/357453902_Clinical_Events_Classification_CEC_in_Clinical_Trials_Report_on_the_Current_Landscape_and_Future_Directions_-_Proceedings_from_the_CEC_Summit_2018
3. Natural Language Processing to Adjudicate Heart Failure Hospitalizations in Global Clinical Trials | Circulation, accessed December 13, 2025, https://www.ahajournals.org/doi/10.1161/CIRCHEARTFAILURE.124.012514
4. Natural Language Processing for Adjudication of Heart Failure Hospitalizations in a Multi-Center Clinical Trial - ResearchGate, accessed December 13, 2025, https://www.researchgate.net/publication/373346839_Natural_Language_Processing_for_Adjudication_of_Heart_Failure_Hospitalizations_in_a_Multi-Center_Clinica

l_Trial
5. Universal Differential Equations for Scientific Machine Learning - Chris Rackauckas MIT, accessed December 13, 2025, https://www.youtube.com/watch?v=bBH8HVEr0-A
6. (PDF) Universal Differential Equations for Scientific Machine Learning - ResearchGate, accessed December 13, 2025, https://www.researchgate.net/publication/338569581_Universal_Differential_Equations_for_Scientific_Machine_Learning
7. [2001.04385] Universal Differential Equations for Scientific Machine Learning - arXiv, accessed December 13, 2025, https://arxiv.org/abs/2001.04385
8. The Unscented Kalman Filter and Particle Filter Methods for Nonlinear Structural System Identification with Non-Collocated Heter - Columbia University, accessed December 13, 2025, http://www.columbia.edu/cu/civileng/smyth/papers/stcdoc_rev_4.pdf
9. Unscented Filtering and Nonlinear Estimation - UBC Computer Science, accessed December 13, 2025, https://www.cs.ubc.ca/~murphyk/Papers/Julier_Uhlmann_mar04.pdf
10. Convergence Analysis of the Unscented Kalman Filter for Filtering Noisy Chaotic Signals, accessed December 13, 2025, https://ira.lib.polyu.edu.hk/bitstream/10397/740/1/noisy-chaotic_07.pdf
11. (PDF) Convergence Analysis of the Unscented Kalman Filter for Filtering Noisy Chaotic Signals - ResearchGate, accessed December 13, 2025, https://www.researchgate.net/publication/224714618_Convergence_Analysis_of_the_Unscented_Kalman_Filter_for_Filtering_Noisy_Chaotic_Signals
12. Convergence Analysis of The Unscented Kalman | PDF - Scribd, accessed December 13, 2025, https://www.scribd.com/document/857287494/Convergence-Analysis-of-the-Unscented-Kalman
13. Micro-Randomized Trials: An Experimental Design for Developing Just-in-Time Adaptive Interventions - NIH, accessed December 13, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC4732571/
14. Microrandomized Trials: An Experimental Design for Developing Just-in-Time Adaptive Interventions - Ambuj Tewari, accessed December 13, 2025, https://www.ambujtewari.com/research/klasnja15microrandomized.pdf
15. Penalized G-estimation for effect modifier selection in a structural nested mean model for repeated outcomes | Biometrics | Oxford Academic, accessed December 13, 2025, https://academic.oup.com/biometrics/article/81/1/ujae165/7954699
16. Doubly Robust Estimation of Optimal Dynamic Treatment Regimes - PMC - NIH, accessed December 13, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC4245503/
17. [1810.08240] Time-uniform, nonparametric, nonasymptotic confidence sequences - arXiv, accessed December 13, 2025, https://arxiv.org/abs/1810.08240
18. Time-uniform, nonparametric, nonasymptotic confidence sequences, accessed December 13, 2025, https://par.nsf.gov/servlets/purl/10251927
19. Action Centered Contextual Bandits - Ambuj Tewari, accessed December 13,

2025, https://www.ambujtewari.com/research/greenewald17action.pdf
20. [1711.03596] Action Centered Contextual Bandits - arXiv, accessed December 13, 2025, https://arxiv.org/abs/1711.03596
21. Signal Temporal Logic - Moodle@Units, accessed December 13, 2025, https://moodle2.units.it/pluginfile.php/560752/mod_resource/content/1/SignalTemporalLogic.pdf
22. Security Analysis of Safe and Seldonian Reinforcement Learning Algorithms, accessed December 13, 2025, https://proceedings.neurips.cc/paper/2020/file/65ae450c5536606c266f49f1c08321f2-Paper.pdf
23. AI Safety · Tutorial 1, accessed December 13, 2025, https://aisafety.cs.umass.edu/seldonian/tutorial1.html
24. A novel adaptive unscented Kalman filter for nonlinear estimation - ResearchGate, accessed December 13, 2025, https://www.researchgate.net/publication/224303740_A_novel_adaptive_unscented_Kalman_filter_for_nonlinear_estimation
25. Lecture 12: Structural Breaks and Threshold Model, accessed December 13, 2025, https://www.fsb.miamioh.edu/lij14/672_s12.pdf