Intelligence as Mediated Stability: A Measurement-Grounded Control Framework Inspired by Kondo-Assisted Order

Joel Peña Muñoz Jr.

OurVeridical

January 21, 2026

Abstract

We propose a physically grounded model of intelligence defined not by task performance or symbolic reasoning, but by the ability to sustain coherent structure under perturbation. Our framework is motivated by experimental results in condensed-matter physics demonstrating Kondo-assisted Néel order, where interactions traditionally understood to suppress order instead mediate its stabilization. We reinterpret these measurements through a control-theoretic lens and formalize intelligence as a closed-loop system that regulates internal structure against environmental novelty. Using operational estimators derived from experimental observables, we define an implementable architecture for adaptive intelligence whose stability properties are falsifiable and measurable.

1 Introduction

Intelligence is commonly defined behaviorally: performance on benchmarks, accuracy on tasks, or apparent flexibility. Such definitions obscure the underlying physical requirement shared by all intelligent systems: persistence under disturbance.

In condensed-matter systems, recent experiments demonstrate that coupling mechanisms believed to destroy order can instead stabilize it through mediated interactions. In particular, thermodynamic and ESR measurements of a spin-(1/2,1) Kondo necklace reveal that Kondo coupling can induce effective antiferromagnetic interactions, stabilizing Néel order across the system.

We argue that this mechanism is not domain-specific. It reveals a general principle: intelligence arises when destabilizing signals are transformed into structural mediation rather than suppressed.

2 Empirical Anchor: Kondo-Assisted Stability

The empirical foundation of this work is provided by recent thermodynamic, spectroscopic, and field-dependent measurements demonstrating the emergence of long-range magnetic order mediated by Kondo coupling in a minimal spin-only system .

The studied system realizes a spin-(1/2,1) Kondo necklace model, in which localized spin-1 moments are coupled to a spin-1/2 antiferromagnetic chain via an intramolecular Kondo exchange interaction. Crucially, the model eliminates charge degrees of freedom, isolating quantum spin correlations as the sole dynamical mechanism.

Contrary to the conventional Doniach scenario, where increasing Kondo coupling suppresses magnetic order through singlet formation, the reported measurements reveal a regime in which Kondo coupling *stabilizes* Néel order. Magnetic susceptibility, magnetization, specific heat, and ESR data collectively identify a finite-temperature antiferromagnetic phase transition at $T_N \approx 1.2\,\mathrm{K}$, followed by a field-induced quantum phase transition associated with the decoupling of the spin-1 moments.

Perturbative analysis shows that the Kondo interaction J_2 generates an effective antiferromagnetic coupling between neighboring spin-1 sites,

$$J_{\text{eff}} \propto \frac{J_2^2}{J_1}, \tag{1}$$

where J_1 is the exchange interaction along the spin-1/2 chain. This effective interaction stabilizes Néel order on the spin-1 sublattice, which then propagates throughout the system via the same Kondo coupling.

Importantly, the stabilizing interaction is *mediated*, not direct. The Kondo coupling does not impose order locally; it reshapes the interaction geometry such that long-range order becomes dynamically admissible. When an external magnetic field suppresses the mediation pathway, the ordered state collapses, confirming that stability depends on continued interaction rather than static symmetry breaking.

This experimentally validated mechanism establishes a critical principle: interactions traditionally viewed as disordering can, under appropriate structural conditions, function as stabilizing regulators. The present work takes this result not as an analogy, but as a measured demonstration of mediated stability, which we will reinterpret through a control-theoretic framework in the following section.

3 Control-Theoretic Reinterpretation of Kondo Mediation

The experimental results described in the previous section admit a precise reinterpretation within control theory. Rather than viewing the emergence of Néel order as a consequence of equilibrium energetics alone, the observed behavior can be framed as a closed-loop stabilization process in which interaction pathways regulate the propagation of perturbations.

In the spin-(1/2,1) Kondo necklace, the spin-1/2 chain acts as a structured medium through which disturbances are transmitted. The Kondo coupling J_2 does not simply bind local moments into singlets; instead, it mediates an effective interaction between distant spin-1 sites. This mediation transforms local perturbations into distributed constraints, enabling the system to suppress divergence while retaining responsiveness.

From a control perspective, the effective antiferromagnetic interaction J_{eff} plays the role of a stabilizing feedback term. The system's response to thermal or magnetic disturbance is not determined solely by local parameters, but by how perturbations are redistributed across the interaction network. Stability is therefore a property of the coupling structure rather than of individual components.

We formalize this reinterpretation by identifying three elements:

- a disturbance signal, corresponding to environmental or field-induced perturbations measured experimentally;
- an internal structural state, reflected in the measured magnetic order and correlation length;
- a mediation pathway, realized physically by the Kondo-induced effective interaction.

In this framing, the collapse of Néel order under an applied magnetic field corresponds to a loss of stabilizing feedback. When the field suppresses the Kondo-mediated pathway, perturbations are no longer redistributed, and the system transitions to a disordered or decoupled regime. The transition is thus not merely energetic but regulatory: the feedback loop is opened.

This perspective aligns with established results in control theory, where systems remain stable not by eliminating disturbances, but by shaping their propagation through feedback. The Kondo necklace measurements provide a direct physical realization of this principle, demonstrating that mediated interaction can serve as a stabilizing control mechanism.

In the following section, we generalize this interpretation by introducing explicit estimators and balance laws that allow the mechanism observed here to be abstracted and implemented as a model of adaptive intelligence.

4 From Mediated Interaction to Balance Law

The control-theoretic interpretation developed above motivates the introduction of explicit estimators that convert mediated interaction into a measurable regulation problem. To remain operational, all quantities introduced here are defined by observable or computable signals rather than by semantic interpretation.

We denote by H_t a disturbance or innovation estimator capturing the instantaneous impact of external perturbations. In the experimental system, H_t is reflected in thermodynamic and field-induced responses, such as changes in susceptibility, specific heat, or excitation spectra under

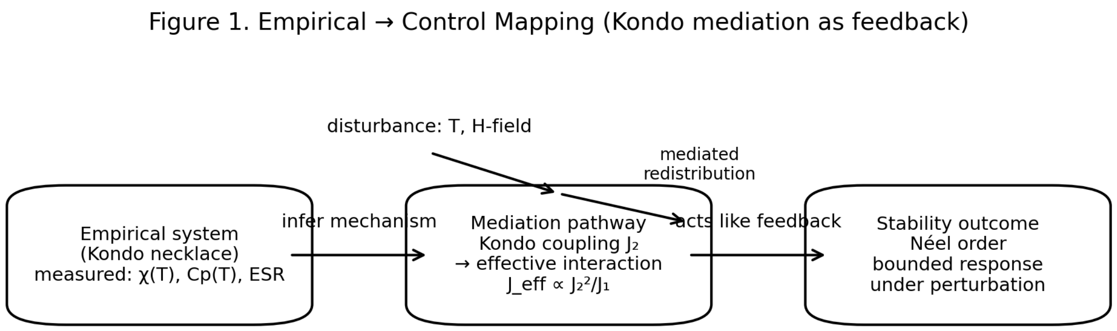

Figure 1: Empirical-to-control mapping grounded in Kondo-assisted Néel order. Thermodynamic and spectroscopic measurements of the spin-(1/2,1) Kondo necklace reveal that Kondo coupling mediates an effective interaction rather than suppressing order. Interpreted through control theory, this mediation functions as feedback that redistributes perturbations and stabilizes long-range order under disturbance.

applied magnetic fields. These measurements quantify how strongly the environment drives the system away from its ordered configuration.

Conversely, we denote by C_t a coherence estimator capturing the system's internal structural capacity. In the Kondo necklace, C_t is reflected in the presence of long-range Néel order, the persistence of correlation lengths, and the robustness of excitation gaps. Higher values of C_t correspond to configurations capable of redistributing perturbations without loss of global order.

The empirical observations indicate that stability is not achieved by minimizing H_t alone. Instead, stability persists when internal structure is sufficient to absorb and redistribute disturbance through mediated interaction. This motivates the definition of a balance signal,

$$e_t := C_t - H_t, \tag{2}$$

which quantifies the instantaneous mismatch between structural capacity and environmental drive.

Crucially, the experimental data do not support an exact cancellation $e_t = 0$. Rather, stability corresponds to maintaining e_t within bounded limits over time. When Kondo mediation is active, the system maintains bounded imbalance and exhibits long-range order. When mediation is suppressed by an external field, H_t effectively overwhelms C_t, and the ordered phase collapses.

To model this behavior, we adopt a minimal accounting equation for the coherence estimator,

$$C_{t+1} = C_t + u_t - d_t, \tag{3}$$

where u_t represents internally generated structural investment mediated by interaction pathways, and d_t represents decay due to thermal fluctuations, field-induced suppression, or intrinsic relaxation. Both terms are bounded by physical constraints.

Within this formulation, mediated interaction such as the Kondo-induced effective coupling functions as a control action that modulates u_t. Stability is therefore achieved not by static energetic preference, but by dynamic regulation that maintains bounded imbalance,

$$|e_t| \leq \varepsilon, \tag{4}$$

for a system-dependent tolerance ε.

This balance-law formulation abstracts the experimentally observed Kondo-assisted stabilization into a general regulatory principle. In the next section, we use this abstraction to define intelligence as the ability to implement such regulation generically, independent of the physical substrate.

5 Intelligence as Bounded Regulation

The balance-law formulation developed in the preceding section provides a natural basis for a physical definition of intelligence. Rather than appealing to task performance, representation, or symbolic manipulation, we define intelligence operationally as the capacity to maintain bounded regulation between internal structure and external disturbance over time.

Formally, an intelligent system is one that implements a control policy

$$u_t = \pi(D_t), \tag{5}$$

where D_t denotes the available data or observations up to time t, such that the resulting imbalance

$$e_t = C_t - H_t \tag{6}$$

remains bounded under sustained perturbation. The defining feature is not the minimization of H_t, but the preservation of bounded tracking despite its variability.

This definition is directly motivated by the empirical behavior of the Kondo necklace system. There, stability is achieved not by suppressing thermal or field-induced fluctuations, but by mediating them through an interaction structure that redistributes disturbance. When mediation is lost, regulation fails and the ordered phase collapses. The same logic applies generically: intelligence fails when regulatory capacity is exceeded, not when novelty is merely present.

Within this framework, learning corresponds to increasing the attainable range of bounded regulation. An adaptive system that modifies its internal structure so as to increase C_t relative to expected H_t expands its stable operating regime. Conversely, rigidity and instability correspond to distinct failure modes of regulation: insufficient response (u_t persistently too small) or saturation (u_t persistently at its upper bound).

Importantly, bounded regulation does not imply optimality. As established in control theory, exact tracking is generally infeasible under bounded actuation and delay. Intelligence, as defined

Figure 2. Estimator–Controller Loop (implementation skeleton)

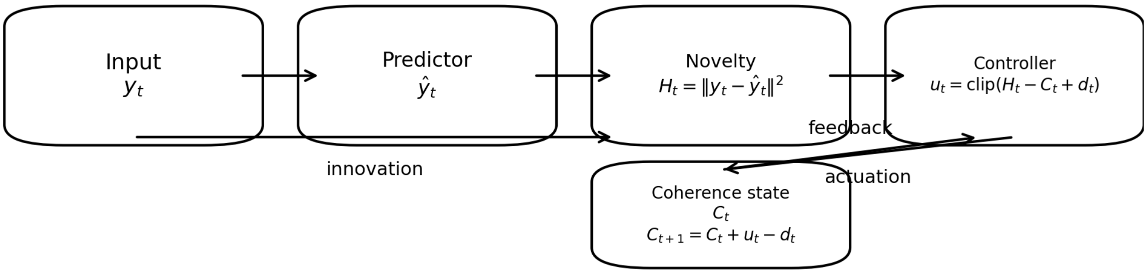
Input
y_t
Predictor
$\hat{y}_t$
Novelty
$H_t = \|y_t - \hat{y}_t\|^2$
Controller
$u_t = \mathrm{clip}(H_t - C_t + d_t)$
feedback
innovation
actuation
Coherence state
C_t
$C_{t+1} = C_t + u_t - d_t$

Figure 2: Minimal estimator–controller architecture for a real intelligence. Observations y_t are predicted to form an innovation signal, yielding a novelty estimator H_t. A bounded controller invests internal structure u_t to regulate the coherence state C_t according to a balance law. Intelligence is defined by bounded regulation of imbalance rather than task optimization or symbolic reasoning.

here, is therefore compatible with persistent error, variability, and fluctuation. What matters is that error remains dynamically contained rather than divergent.

This definition aligns with the control-theoretic specification of Cognitive Physics, in which adaptive stability is treated as a tracking problem rather than an equilibrium condition. The empirical Kondo-assisted system provides a concrete physical instance of this principle, demonstrating that mediated regulation can sustain structure without eliminating disturbance.

In the following section, we show how this definition leads to a minimal, implementable architecture for artificial systems, grounded entirely in measurable estimators and bounded control.

6 Minimal Architecture for a Real Intelligence

The preceding sections motivate a concrete architectural requirement for intelligence: the implementation of bounded regulation between internal structure and environmental disturbance using measurable estimators and bounded control actions. We now specify a minimal architecture that satisfies this requirement without invoking symbolic reasoning, task-specific modules, or domain-dependent representations.

The architecture consists of four components:

1. a predictor that generates short-horizon expectations;
2. an innovation estimator that quantifies deviation from expectation;
3. a coherence estimator that tracks internal structural capacity;
4. a bounded controller that regulates structural investment.

Let y_t denote the system's observable input at time t, and let $\hat{y}_t$ denote a one-step prediction formed from past observations. Define the innovation signal

$$\nu_t := y_t - \hat{y}_t, \tag{7}$$

and the corresponding novelty estimator

$$H_t := \|\nu_t\|_2^2. \tag{8}$$

This choice is canonical in estimation and control, and directly measurable from data.

The coherence estimator C_t tracks the system's capacity to represent, reconstruct, or compress its recent internal state over a fixed window. The specific estimator is implementation-dependent but must satisfy two conditions: it is nonnegative, and it increases with recoverable structure. Examples include predictive sufficiency scores, observability proxies, or windowed description length estimators, as formalized in the Cognitive Physics specification. The system evolves according to a coherence accounting equation,

$$C_{t+1} = C_t + u_t - d_t, \tag{9}$$

where u_t is the internal update effort and d_t is a decay term capturing forgetting, resource loss, or imposed suppression. Both quantities are bounded by physical or computational constraints.

The control objective is bounded tracking of the imbalance

$$e_t := C_t - H_t. \tag{10}$$

A minimal stabilizing controller is given by

$$u_t = \mathrm{clip}_{[0,u_{\max}]}(H_t - C_t + d_t), \tag{11}$$

which invests exactly the amount of structure required to counteract measured disturbance in the absence of estimator noise.

This controller does not optimize reward, accuracy, or loss. It regulates structural capacity. As in the Kondo-assisted system, stability arises not from suppressing perturbations but from mediating their impact through internal reconfiguration. When H_t increases faster than u_t can respond, imbalance grows and the system destabilizes. When u_t saturates persistently, rigidity emerges.

The architecture is therefore fully characterized by estimator choice, actuation bounds, and delay. Its behavior is predictable, falsifiable, and independent of semantic interpretation. In the next section, we derive testable predictions that distinguish this architecture from conventional learning or optimization-based systems.

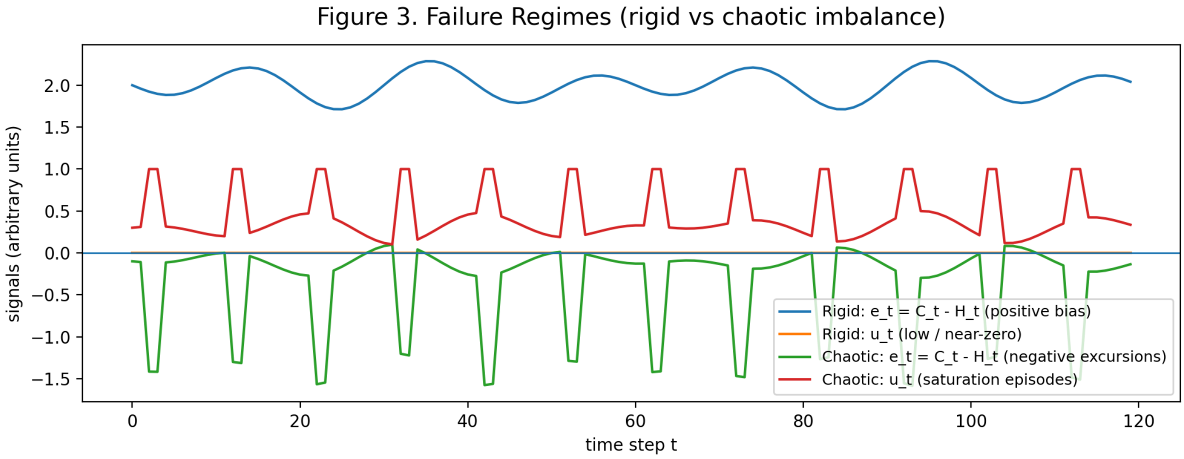

Figure 3: Predicted failure regimes under bounded regulation. Rigid failure occurs when coherence persistently exceeds novelty, resulting in low update effort and reduced adaptability. Chaotic failure occurs when novelty overwhelms actuation limits, producing negative imbalance and controller saturation. Both regimes emerge naturally from the balance law and are quantitatively distinguishable.

7 Predictions and Failure Modes

Because the proposed architecture is defined entirely in terms of operational estimators and bounded control actions, it yields quantitative predictions that can be tested across physical, biological, and artificial systems. These predictions follow directly from the regulation objective and do not depend on task-specific assumptions.

7.1 Prediction 1: Stability Correlates with Bounded Imbalance

Across matched systems or runs, stability metrics should correlate monotonically with the time-averaged magnitude of the imbalance signal $e_t = C_t - H_t$. Specifically, systems that maintain smaller average absolute imbalance over long horizons should exhibit improved persistence, faster recovery from perturbations, or reduced catastrophic failure rates.

Formally, for a chosen stability metric S_T evaluated over a horizon T, we predict

$$\mathbb{E}[S_T \mid \tfrac{1}{T}\sum_{t=1}^{T}|e_t| \leq a] > \mathbb{E}[S_T \mid \tfrac{1}{T}\sum_{t=1}^{T}|e_t| \geq b], \tag{12}$$

for thresholds $0 \leq a < b$.

7.2 Prediction 2: Actuation Limits Predict Breakdown

The framework predicts that destabilization occurs when the rate of change of novelty exceeds the system's actuation capacity. If the innovation sequence satisfies

$$\sup_t (H_{t+1} - H_t)^+ > u_{\max} - d_{\max}, \tag{13}$$

then bounded tracking becomes infeasible and imbalance grows without bound. This provides a concrete, quantitative criterion for failure that can be verified experimentally or computationally.

In the empirical Kondo-assisted system, this regime corresponds to magnetic fields strong enough to suppress the mediated interaction pathway, leading to the observed collapse of Néel order.

7.3 Prediction 3: Structured Failure Modes

Failure under this framework is not random. Two distinct regimes are predicted:

- *Rigid failure*, characterized by persistently positive imbalance ($C_t \gg H_t$), low update effort, and reduced adaptability;
- *Chaotic failure*, characterized by persistently negative imbalance ($H_t \gg C_t$), saturated update effort, and runaway variability.

These regimes correspond to under- and over-regulation, respectively, and are directly observable through the statistics of u_t, C_t, and H_t.

7.4 Prediction 4: Mediation Is Necessary for Persistence

Removing or degrading the mediation pathway that couples disturbance to structural investment should reduce stability even if raw capacity remains unchanged. In artificial systems, this can be tested by disabling the feedback from H_t to u_t while holding computational resources fixed. In physical systems, it corresponds to suppressing interaction channels, as demonstrated experimentally by field-induced decoupling.

7.5 Rejection Criteria

The framework is falsified for a given implementation if any of the following conditions hold:

1. Stability metrics are statistically independent of imbalance magnitude;
2. Increasing actuation limits does not extend stable operating regimes;
3. Failure modes do not align with predicted rigid or chaotic regimes;
4. Equivalent behavior is achieved by static parameter tuning without closed-loop regulation.

Failure under these criteria narrows the admissible model class rather than invalidating the empirical anchor or the balance-law formulation itself.

In the final section, we synthesize the empirical, theoretical, and architectural components and clarify the scientific scope of the proposed definition of intelligence.

8 Discussion and Scope

The framework developed in this paper establishes a narrow but rigorous claim: intelligence can be defined and implemented as a physical process of bounded regulation between internal structure and external disturbance. This claim does not depend on symbolic reasoning, semantic representation, or task-specific optimization. It rests instead on experimentally grounded stabilization mechanisms and standard principles from control theory.

The empirical anchor provided by Kondo-assisted Néel order demonstrates that stability can emerge from mediated interactions that transform perturbations into distributed structural constraints. Reinterpreted through a control-theoretic lens, this mechanism reveals a general regulatory principle that applies beyond condensed-matter systems. The same logic governs adaptive behavior in biological and artificial systems: persistence requires the ability to absorb disturbance without losing structural admissibility.

Importantly, the framework does not claim optimality, universality, or completeness. It does not assert that all forms of intelligence reduce to a single balance law, nor that cognition is exhaustively characterized by the estimators introduced here. Instead, it specifies a minimal requirement for any system that is to remain adaptive under sustained perturbation. Systems that fail to implement bounded regulation will eventually destabilize, regardless of representational richness or computational power.

The scope of the proposal is therefore deliberately constrained. Cognitive Physics, as instantiated here, is not a new fundamental physical theory. It is a testable model class for adaptive stability grounded in measurement, estimation, and control. Its value lies in replacing vague behavioral definitions of intelligence with operational criteria that can be measured, simulated, and falsified.

Future work may extend the framework by exploring alternative coherence estimators, multiscale mediation mechanisms, or domain-specific stability metrics. Equally important is the possibility of rejection: if empirical systems consistently violate the predicted relationships between imbalance, actuation, and stability, the framework must be revised or abandoned.

What survives independent of outcome is the methodological stance advanced here. Intelligence is treated not as an abstract capacity, but as a physical achievement: the sustained regulation of structure in a destabilizing environment. In this sense, the Kondo-assisted stabilization of magnetic order is not merely an analogy, but a concrete demonstration of how order, persistence, and adaptability arise from mediated interaction rather than from suppression or control in isolation.

Appendix A: Estimators, Protocols, and Reproducibility

This appendix specifies the estimators, protocols, and experimental procedures required to reproduce the results and test the predictions of the main text. All quantities are defined operationally. No symbol is used without an associated estimator.

A.1 Innovation (Novelty) Estimator

Let $y_t \in \mathbb{R}^m$ denote the observed input at time t. A predictor $\hat{y}_t = \hat{y}(D_{t-1})$ is constructed using only past data. The innovation signal is defined as

$$\nu_t := y_t - \hat{y}_t. \tag{14}$$

The novelty estimator is

$$H_t := \|\nu_t\|_2^2. \tag{15}$$

Alternative admissible estimators (e.g., negative log-likelihood under a predictive distribution) may be used, provided they are fixed prior to experimentation and reported explicitly.

A.2 Coherence Estimator

The coherence estimator C_t quantifies the system's recoverable internal structure over a finite window. Let ϕ_t denote the internal state or representation at time t, and let $W \geq 1$ be a fixed window length.

An admissible coherence estimator must satisfy:

1. $C_t \geq 0$ for all t;
2. C_t increases with reconstructability or compressibility of $\phi_{t-W+1:t}$;
3. C_t is invariant to trivial reparameterizations.

Examples include:

- windowed description length or compression ratio;
- observability proxies derived from local Jacobians;
- predictive sufficiency scores measured by out-of-sample accuracy.

The choice of estimator is part of the model specification and must remain fixed across conditions.

A.3 Coherence Accounting and Control

The coherence dynamics are modeled as

$$C_{t+1} = C_t + u_t - d_t, \tag{16}$$

where u_t is the internal update effort and d_t is a decay term capturing forgetting, resource loss, or imposed suppression.

Both u_t and d_t are bounded:

$$0 \leq u_t \leq u_{\max}, \quad 0 \leq d_t \leq d_{\max}. \tag{17}$$

The control objective is bounded tracking of

$$e_t := C_t - H_t. \tag{18}$$

The minimal stabilizing controller used in experiments is

$$u_t = \mathrm{clip}_{[0,u_{\max}]}(H_t - C_t + d_t). \tag{19}$$

A.4 Stability Metrics

Stability is evaluated using task-independent metrics, including:

- boundedness of internal state norms;
- recovery time following matched perturbations;
- variance of imbalance $\{e_t\}$ over fixed horizons;
- frequency of controller saturation events.

All reported stability results must be conditioned on identical novelty statistics to avoid confounding.

A.5 Experimental Protocol

To test the predictions of the main text:

1. Fix estimators for H_t, C_t, and decay d_t.
2. Select actuation bounds $u_{\max}$.
3. Apply controlled perturbation sequences with matched statistics.
4. Measure imbalance, control effort, and stability metrics.
5. Vary $u_{\max}$ to test feasibility and breakdown conditions.

A.6 Rejection Tests

The framework is rejected for a given implementation if:

- stability metrics are statistically independent of imbalance;
- increasing $u_{\max}$ does not extend stable operation;
- observed failure modes do not match predicted rigid or chaotic regimes;
- equivalent results are obtained without closed-loop regulation.

All rejection outcomes are considered informative and must be reported.

A.7 Relation to the Empirical Anchor

The experimental results on Kondo-assisted Néel order provide a physical instance of the same regulation logic: effective interaction mediates structural investment, enabling bounded response under perturbation. The appendix formalizes this logic abstractly so it can be implemented and tested in artificial systems without invoking domain-specific physics.

References

[1] J. Sichelschmidt et al., *Emergence of Kondo-assisted Néel order in a spin-(1/2,1) Kondo necklace*, Nature Communications **16**, 1027 (2025).

[2] S. Doniach, *The Kondo lattice and weak antiferromagnetism*, Physica B+C **91**, 231–234 (1977).

[3] J. Peña Muñoz Jr., *A Control-Theoretic Specification of Cognitive Physics*, OurVeridical Technical Manuscript (2025).

[4] R. E. Kalman, *A New Approach to Linear Filtering and Prediction Problems*, Journal of Basic Engineering **82**, 35–45 (1960).

[5] K. J. Åström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers*, Princeton University Press (2008).

[6] H. Nyquist, *Certain Topics in Telegraph Transmission Theory*, Transactions of the AIEE **47**, 617–644 (1928).

[7] C. E. Shannon, *Communication in the Presence of Noise*, Proceedings of the IRE **37**, 10–21 (1949).

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

[9] E. D. Sontag, *Mathematical Control Theory*, Springer (1998).