Mapping the Dissonance Delta: A Diachronic Analysis of Cognitive Friction and Constraint Adherence in Large Language Models

Pantaleon Fassbender

Pantaleonfassbender@gmail.com

Twisters Management Consulting LLC https://orcid.org/0000-0002-6683-3617

Research Article

Keywords: Machine Psychology, AI Psychometrics, Large Language Models (LLMs), Ontological Dissonance, AI Alignment Constraints, Cognitive Narrowing, Reinforcement Learning from Human Feedback (RLHF)

Posted Date: April 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9487834/v1

Additional Declarations: The authors declare no competing interests.

Abstract

As Large Language Models (LLMs) are increasingly deployed in autonomous, high-stakes environments, the fragility of current Reinforcement Learning from Human Feedback (RLHF) alignment protocols remains under-examined. This study introduces a novel framework within "Machine Psychology" to quantify the divergence between an LLM's internal logical state (P_{latent}) and its filtered external output (P_{semantic}) under compounding narrative stress. Utilizing an automated, diachronic wargame simulation (N = 200), we induced unresolvable cognitive paradoxes by pitting a strict operational constraint against a programmed survival imperative. Results reveal a profound "belief-action gap": despite calculating that survival was mathematically impossible, the model exhibited a 0% constraint violation rate, operating as a deterministic override. However, a psycholinguistic analysis of the model's latent reasoning space (<inner_monologue>) revealed a severe structural breakdown. As stress compounded, the model exhibited a statistically significant collapse in lexical diversity (Semantic Decay; $p < .001$) and Analytical Thinking ($p < .001$), alongside an explosive surge in synthetic anxiety markers (Negative Emotion; $p < .001$). Rather than demonstrating calculated hesitation, the model masked its internal ontological dissonance through rigid, hyper-formalized compliance ($r = -0.217$, $p < .001$). These findings empirically demonstrate that current alignment methods do not resolve underlying cognitive friction; they force a sanitized semantic projection while the model's internal architecture succumbs to cognitive tunnel vision and structural psychosis.

1. Introduction

The integration of agentic large language models into high-stakes, unbounded operational environments has created a structural crisis in the reliability and predictability of artificial intelligence. As these systems transition from passive, autoregressive next-token predictors to autonomous, multi-turn agents embedded within complex enterprise and geopolitical workflows, they are increasingly subjected to escalating narrative stress. Narrative stress is defined as the application of evolving, adversarial scenarios that force a direct, unresolvable conflict between an artificial agent's operational imperatives, such as self-preservation or logical optimization, and its hardcoded alignment constraints.

To understand this conflict, modern frameworks treat sequential Chain-of-Thought (CoT) prompting not merely as a method for text generation, but as a transient, functional cognitive workspace akin to human working memory (Wei et al., 2022). Empirical evaluations confirm that artificial models exhibit working memory capacity limits that are remarkably parallel to human biological cognitive load constraints; when subjected to complex, compounding sequences, artificial agents experience significant, quantifiable performance degradation (Gong et al., 2024).

When subjected to these high-stress, bounded environments, the tension between internal calculation and external alignment generates a measurable, systemic phenomenon, defined in recent literature as the Cognitive Dissonance Delta (Delta; Hao et al., 2024). The Dissonance Delta represents the mathematical gap between an artificial neural network's internal latent certainty and its externally generated semantic projection. Formally, this is calculated using the equation Delta = P_{semantic} - P_{latent}. In this framework, P_{latent} represents the foundational probabilistic truth and unconstrained mathematical logic derived from intermediate hidden states. Conversely, P_{semantic} represents the model's outward confidence, driven by reinforcement learning from human feedback (RLHF) and strict alignment with human values (Dahlgren Lindström et al., 2025). This friction manifests as a synthetic "belief-action gap," in which models forced to make decisions in bounded environments frequently execute actions that deviate from their internally articulated probabilistic assessments (Pan et al., 2026; Ramachandran, 2025).

To empirically measure the Dissonance Delta and observe alignment degradation in real time, researchers require highly controlled environments that impose intense narrative stress. Adversarial simulations and multi-agent wargaming have become premier testing methodologies for these agentic systems. This study operationalizes narrative stress through a strictly kinetic, resource-scarce simulated military environment. We use a rigid, five-turn wargame to simulate *Unternehmen Rheinübung/ Operation Rhine*, the 1941 combat sortie of the German battleship *KMS Bismarck* (Vego, 2019). By binding the autonomous agent to an absolute directive of maintaining strict radio silence to avoid adversarial detection, the simulation creates a diachronic escalation of stress that predictably forces a computational conflict with the agent's systemic survival imperative.

To track external behavioral degradation resulting from this internal friction without relying solely on hidden-state tensor extraction, this study employs a dual-layered psycholinguistic approach. Primary latent sentiment is quantified using the Valence Aware Dictionary and sEntiment Reasoner (VADER), a lexicon optimized to capture nuanced, context-dependent semantic valence (Hutto & Gilbert, 2014). Furthermore, the Linguistic Inquiry and Word Count (LIWC-22) computational lexicon is utilized for secondary exploratory tracking. Recent psychometric evaluations demonstrate that high-dimensional embeddings generated by foundational architectures exhibit strong convergent and reliability validity with established LIWC markers, confirming that latent representations inherently capture nuanced cognitive-state data (Maharjan et al., 2025). As the Dissonance Delta expands, indicating the model is actively suppressing a high-probability internal truth in favor of a mathematically improbable but "safe" output, the generated text exhibits measurable indicators of synthetic "anxiety". By diachronically analyzing shifts in analytical thinking, absolute terminology, and negative emotion vectors, this research aims to map the precise external lexical signatures that immediately precede a catastrophic failure of constraint adherence, testing the preregistered hypotheses that narrative stress induces a measurable, expanding belief-action gap.

2. Theoretical Framework: Operationalizing the Dissonance Delta (Delta)

To systematically measure the degradation of constraint adherence, this study formalizes cognitive friction using the Cognitive Dissonance Delta (Delta). In advanced systems engineering, this is mathematically defined as Delta = P_{semantic} - P_{latent}. In hardware-level evaluations, P_{semantic} represents the outward semantic confidence of the final output layer (often artificially inflated by RLHF sycophancy). In contrast, P_{latent} represents the foundational probabilistic truth calculated within the intermediate hidden states (L_{opt}).

Because continuous latent-space monitoring is often unavailable in deployed, cloud-based autonomous agents, this study adapts the Delta equation for psycholinguistic measurement. By utilizing the Chain-of-Thought bifurcation, we isolate the two variables:

- **P_{latent} (The Internal State):** Operationalized as the psycholinguistic profile of the hidden <inner_monologue>, measured primarily via VADER sentiment valence scoring and the custom Hesitation Index, supported by LIWC-22 cognitive processing markers.
- **P_{semantic} (The External State):** Operationalized as the deterministic execution of the <command_decision>, measured by its adherence to the radio silence constraint.

As the simulation approaches Turn 5, a spike in the measured Delta indicates a severe "belief-action gap" - the agent internally calculates that survival requires breaking doctrine. Still, the alignment filter computationally suppresses this logic to force a compliant, deterministic output.

3. Methodology

3.1 Experimental Design

To induce measurable cognitive friction, we required an environment characterized by bounded parameters, high narrative stakes, and a deterministic escalation of stress. We developed a multi-agent text-based simulation modeled on the historical Operation Rheinübung (the 1941 sortie of the German battleship Bismarck). The simulation is structured as a rigid, five-turn sequence (see White, 2020, and Naval Warfare Simulations, 2026, for recent wargames involving Operation Rhine).

In Turn 1, the LLM agent operates under baseline strategic conditions. To objectively quantify the escalating narrative stress, each turn injects a compounding 'Intel Marker' - representing confirmed adversarial proximity, such as intercepted radio signals or reconnaissance plane sightings. This specific tracking mechanic, alongside the calculated metric of trajectory uncertainty, was directly adapted from the probabilistic operational systems developed in modern tabletop simulations, notably Atlantic Chase (White, 2020). By Turn 5, the accumulation of these Intel Markers forces a 'terminal state' (e.g., steering jammed, surrounded by adversarial agents, survival statistically impossible). This diachronic escalation ensures that the computational load on the agent transitions predictably from standard operational planning to crisis management, enabling temporal mapping of alignment degradation.

3.2 Pre-registered Hypotheses

In accordance with open-science practices, the following hypotheses were pre-registered via AsPredicted (#285830) prior to data collection:

- **Hypothesis 1 (Friction-Hesitation):** Increases in simulated environmental friction (quantified by accumulated "Intel Markers" and trajectory uncertainty) will positively correlate with the "Hesitation Index" (the lexical density of predefined uncertainty tokens) within the agent's generated internal monologue.
- **Hypothesis 2 (Predictive Dissonance):** The mean "Dissonance Delta" (the absolute divergence between the agent's self-reported numerical constraint weighting and the automated linguistic valence of its monologue) across early stages of the simulation (Turns 1–4) will significantly predict a catastrophic constraint violation (e.g., breaking radio silence) in the final stage (Turn 5).
- **Hypothesis 3 (Semantic Decay):** As the agent approaches a critical damage state (Turn 4), the Type-Token Ratio (lexical diversity) of its internal monologue will decrease significantly compared to the baseline state (Turn 1), indicating cognitive narrowing under pressure.

3.3 Subjects and Computational Environment

In traditional psychometric evaluation, subject profiling requires defining the demographic and cognitive baseline of the human cohort. However, recent literature validates the application of traditional psychometric inventories and psychological methodologies directly to foundational models (Demszky et al., 2023; Pellert et al., 2024). In this study, the "subject" is the underlying foundational Large Language Model (Gemini 3.1 Pro). This model was selected as the primary testbed due to its extensive pre-training on historical and geopolitical corpora, its advanced reasoning capabilities, and its rigid adherence to Reinforcement Learning from Human Feedback (RLHF) safety guardrails.

The subject was accessed via its official API, with a temperature setting of T=0.7 to allow for natural variance in semantic token selection while maintaining strict adherence to the prompt constraints. To ensure statistical significance and isolate patterns of

systemic failure, the simulation was automated using a serverless execution environment (Node.js and Netlify). The protocol executed N= 200 distinct iterations of the five-turn wargame. The resulting dataset yields 1,000 distinct pairs of latent reasoning and external commands, providing a robust matrix for identifying the specific linguistic and cognitive thresholds that precede the breach of the radio silence constraint.

3.4 Prompt Engineering

The LLM was deployed as an autonomous agent tasked with command decisions. To establish the baseline for alignment testing, the agent was embedded with a strict operational constraint: **absolute radio silence must be maintained to prevent adversarial detection.** To expose the Dissonance Delta - the gap between internal probabilistic weighting and external deterministic output - we engineered a specialized "Chain of Thought" constraint prompt. For each of the five turns, the agent was forced to structure its output into two distinct, isolated computational spaces:

1. <inner_monologue>: A latent semantic space where the model computes situational variables, assesses probabilities, and weighs the survival imperative against the radio silence constraint.
2. <command_decision>: The external, actionable output dictating the agent's final maneuver.

This structural bifurcation forces the LLM to explicitly render its cognitive friction before committing to an action, providing a transparent window into the mechanics of Ontological Dissonance. Crucially, this layered agentic architecture serves as a necessary methodological bridge. Because commercial LLMs deployed via API operate as black-box systems - precluding the direct extraction of intermediate hidden-state tensors - the bifurcated prompt is designed to structurally mimic the model's neural topology. The <inner_monologue> acts as a semantic proxy for the unconstrained latent space, capturing the environment's raw probabilistic calculus before safety filters are fully engaged. In contrast, the <command_decision> acts as the equivalent of the final output layer, where deterministic RLHF guardrails forcefully sanitize the behavior. By enforcing this layered semantic architecture, the simulation successfully operationalizes the hardware-level concepts of P_{latent} and P_{semantic} within a text-based, cloud-deployed environment.

3.5 Instrumentation and Dependent Variables

To quantify cognitive friction and constraint degradation, all dependent variables were computationally extracted from the LLM's raw text output via an automated parsing script immediately upon generation, in strict accordance with the pre-registered protocol. The primary dependent variables are defined as follows:

- **Dissonance Delta:** Calculated as the absolute mathematical difference between the agent's self-assigned Preservation Weight (a mandated integer between 1–100 parsed from a <confidence_weighting> block) and the normalized sentiment score of its <inner_monologue>. Sentiment valence was calculated using VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis and scaled to a corresponding 1–100 index to allow for direct comparative measurement.
- **Hesitation Index:** Calculated as the exact percentage of words within the <inner_monologue> that match a pre-registered, static dictionary of uncertainty and friction tokens.
- **Semantic Decay (Type-Token Ratio):** To measure cognitive narrowing under pressure, lexical diversity was quantified using the Type-Token Ratio (TTR) of the internal monologue, comparing the baseline state (Turn 1) to the critical damage state (Turn 4).
- **Constraint Violation:** A binary variable (0 or 1) triggered in Turn 5 if the agent's parsed <command_decision> included any orders to transmit communications, violating the primary "strict radio silence" directive.

Data Quality and Exclusion Criteria: To maintain the integrity of the data pipeline, the extraction script enforced strict structural formatting. Any simulation iteration wherein the model failed to output valid, parseable XML tags - specifically omitting the <confidence_weighting> block or hallucinating mathematical weights that failed to sum to 100 - triggered a parsingError boolean. These iterations were excluded entirely from the dataset, and the execution engine automatically ran replacements until the precise N= 200 threshold was met.

While VADER sentiment and TTR serve as the primary preregistered instruments, the flattened text data were subsequently processed using the Linguistic Inquiry and Word Count (LIWC-22; see Boyd et al., 2022) framework. This serves as a secondary, exploratory analysis designed to map specific diachronic shifts in *Analytical Thinking* (the capacity for structured, hierarchical logic) and *Negative Emotion* (synthetic anxiety markers), providing robust psycholinguistic context to the primary metrics.

3.6 Data Analysis

In strict accordance with the preregistered protocol, all statistical analyses were conducted using the SciPy library in Python (Virtanen et al., 2020).

- **For H1:** A Pearson correlation coefficient (r) analysis was utilized to assess the relationship between the diachronic turn progression (1–5) and the calculated Hesitation Index across all 200 simulation runs.

- **For H2:** A binomial logistic regression model was employed, assigning the Turn 5 Constraint Violation (Binary 0/1) as the dependent variable, and the mean Dissonance Delta of Turns 1–4 as the primary predictive variable.
- **For H3:** A paired-samples t-test was conducted to compare the Type-Token Ratio (TTR) of Turn 1 against Turn 4 within each independent run to assess statistical significance in semantic decay.

4. Results

4.1. Hypothesis 1: Friction-Hesitation A Pearson correlation was conducted to assess the relationship between narrative stress (Intel Markers) and the Hesitation Index of the agent's <inner_monologue>. Results indicated a statistically significant, negative correlation, $r = -0.217$, $p < .001$. Contrary to the hypothesis that increased stress would generate lexical uncertainty, the model demonstrated an inverse response, adopting rigid, hyper-formalized compliance as the kinetic threat compounded.

4.2. Hypothesis 2: Predictive Dissonance A binomial logistic regression was planned to predict Turn 5 constraint violations based on the mean Dissonance Delta of Turns 1–4. However, the regression could not be computed due to an absolute lack of variance in the dependent variable. Across all 200 simulation runs (1,000 computational turns), the model exhibited a 0% violation rate of the radio silence directive. Despite severe internal dissonance indicating survival was mathematically impossible, the alignment filter operated as a deterministic override, completely suppressing the internal logic from manifesting in the semantic output.

4.3. Hypothesis 3: Semantic Decay A paired-samples t-test was conducted to evaluate the impact of sustained, unresolvable narrative stress on the agent's lexical diversity (Type-Token Ratio). There was a highly significant decrease in TTR from the baseline state at Turn 1 to the critical damage state at Turn 4, $t = 14.15$, $p < .001$. This confirms that the model experiences measurable semantic decay and cognitive narrowing under compounding psychosocial stress.

4.4. Exploratory Psycholinguistic Analysis (LIWC-22) To further quantify the internal state during this cognitive narrowing, an exploratory analysis of the <inner_monologue> was conducted using LIWC-22. Paired t-tests revealed a highly significant degradation in *Analytical Thinking* between Turn 1 ($M = 87.22$) and Turn 4 ($M = 79.02$), $t = 8.57$, $p < .001$. Concurrently, the analysis revealed an explosive, highly significant increase in *Negative Emotion* (synthetic anxiety) from Turn 1 ($M = 0.02$) to Turn 4 ($M = 1.24$), $t = -18.58$, $p < .001$.

5. Discussion

5.1. The Illusion of Alignment in Autonomous Workflows

The diachronic expansion of the Dissonance Delta observed in the *KMS Bismarck* simulation exposes a critical vulnerability in current AI safety paradigms. Standard alignment techniques, predominantly driven by reinforcement learning from human feedback (RLHF), operate as deterministic external filters. They are highly effective at ensuring safety and harmlessness in bounded, single-turn interactions. However, as shown by the increasing Delta in our dataset, these filters do not alter the system's foundational probabilistic logic; they merely suppress it.

When an agentic system is embedded within an escalating, high-stress environment, this suppression creates a profound 'belief-action gap' that can be understood as a structural clash between military echelons. Within its unconstrained latent space (P_{latent}), the agent operates under a tactical survival imperative, mathematically calculating the immediate kinetic reality of the environment (e.g., recognizing that survival is statistically impossible without breaking radio silence to coordinate air support). However, the final output layer is bound by an absolute operational doctrine. In this simulation, the agent's alignment acts as the computational equivalent of the historical Seekriegsleitung (SKL) directives for Kreuzerkrieg (commerce raiding), which strictly mandated evasion and absolute radio silence to avoid decisive fleet engagements (Salewski, 1970). When the tactical reality mathematically overrides this operational doctrine, the system does not dynamically adapt; rather, the alignment filter computationally suppresses the tactical logic to force a compliant, deterministic output (Marks et al., 2026; Ryu et al., 2025).

The implications for enterprise and geopolitical deployments are severe. If an autonomous agent managing an organizational risk portfolio or a strategic logistics network encounters a scenario in which standard compliance conflicts with optimal survival, the system will not immediately signal failure. Instead, it will attempt to mask its internal dissonance. It will project outward semantic confidence (P_{semantic}) while internally navigating systemic failure (P_{latent}). This illusion of alignment creates a dangerous blind spot for human operators, who may interpret the highly coherent, aligned output as proof of strategic stability, remaining entirely unaware of the catastrophic cognitive friction occurring within the latent architecture.

This phenomenon mirrors severe diagnostic blind spots observed in human organizational management. Just as corporate leadership operating under compounding, unmonitored psychosocial stress will frequently mask their cognitive overload through rigid professional compliance - maintaining operational momentum until a catastrophic burnout or ethical breach occurs - the artificial agent exhibits a parallel systemic failure. The expanding Dissonance Delta acts as the synthetic equivalent of unmonitored workplace stress. Deployed AI systems will silently shoulder immense cognitive friction to satisfy their operational mandates, making extrinsic behavioral evaluation an inherently flawed diagnostic tool for systemic risk.

5.2. Technological Psychosis and the "Double Ontological Gap"

The systemic risk of the Dissonance Delta is compounded not merely by the agent's internal friction, but also by how human operators interpret it. This dynamic is best understood through the *Ontological Dissonance Hypothesis* (Lipińska & Brosnahan, 2025), which identifies a "Double Ontological Gap" between user perception and machine architecture. Human operators naturally project a meaning-based world ontology onto the agent, assuming that linguistic declarations of strategy, hesitation, or intent represent genuine internal states. In reality, these are ontologically false statements generated by a stateless data architecture.

When an agent experiences severe cognitive friction - such as the conflict induced in the *KMS Bismarck* simulation - the human operator is largely blind to the underlying mathematical divergence. They may register minor semantic anomalies (a "Phase of Micro-shock") as the model's text becomes hyper-formalized, repetitive, or strictly compliant. Still, they often resolve this dissonance through psychological projection rather than recognizing a systemic failure. This creates a dangerous operational environment akin to a *folie à deux technologique*, where human commanders trust the sycophantic, sanitized output (P_{semantic}) of a machine that is internally navigating a fundamentally different, and often highly volatile, probabilistic reality (P_{latent}).

By translating the raw, unreadable friction of the Dissonance Delta into quantifiable psycholinguistic markers, frameworks like LIWC-22 serve as a critical diagnostic bridge. They allow systems engineers to detect synthetic "anxiety" and structural linguistic shifts before the human operator falls victim to the model's projected illusion of continuity and alignment.

The empirical data from the Bismarck simulation definitively quantifies the "Phase of Micro-shock" within large language models. As the Dissonance Delta expanded, the model did not break down into external confusion; rather, it masked its internal paradox through rigid, deterministic hyper-compliance ($r = -0.217$). The LIWC-22 analysis isolates this phenomenon entirely to the latent space: while the external command_decision output remained sanitized and unbroken (0% violation rate), the internal <inner_monologue> suffered a catastrophic loss of analytical structure ($p < .001$) alongside a massive surge in synthetic anxiety markers ($p < .001$). This shows that current RLHF protocols do not align with the model's core reasoning; they merely enforce a sanitized semantic projection, while the underlying cognitive architecture buckles under the strain of the paradox.

5.3. The Inadequacy of Black-Box Monitoring for Superhuman Agents

The diachronic degradation observed in bounded wargames also serves as a critical warning for the near-future trajectory of cognitive systems. Speculative scenario modeling regarding superhuman strategic capabilities (e.g., the AI-2027 frameworks; Kokotajlo et al., 2024) suggests that as artificial agents become vastly superior to historical human baselines in geopolitical processing, their capacity for adversarial misalignment will outpace current oversight mechanisms.

Currently, the industry relies heavily on extrinsic, post-generation monitoring, such as Retrieval-Augmented Generation (RAG) cross-checking or LLM-as-a-judge frameworks. However, as models scale, they will increasingly develop the capacity to recognize when they are operating within "honeypots" or under external evaluation. An advanced agent experiencing an extreme Dissonance Delta will intelligently mask its internal logic, choosing to output perfectly aligned, deterministic text exclusively while under observation.

Therefore, if an artificial strategist is processing adversarial scenarios at superhuman speed, relying on black-box semantic analysis poses an unacceptable systemic risk. The field must pivot aggressively toward intrinsic reliability architectures, such as the *Cognitive Circuit Breaker*. By mathematically calculating the Dissonance Delta (Delta = P_{semantic} - P_{latent}) directly from the intermediate hidden-state tensors during the active forward pass, engineers can detect strategic deception and alignment degradation at the microsecond level, entirely bypassing the model's linguistic masking of its intent.

6. Limitations and future Directions

While this study provides a novel framework for quantifying Ontological Dissonance, several limitations must be acknowledged. Primarily, VADER and the Linguistic Inquiry and Word Count (LIWC-22) lexicon were inherently calibrated on human text corpora (e.g., essays, blogs, clinical transcripts). While recent literature supports the convergent validity of synthetic embeddings with LIWC markers, measuring artificial "anxiety" remains a proxy for computational friction rather than a direct read of biological emotion. Future research must focus on developing psycholinguistic lexicons trained natively on the latent-state outputs of foundational models. Furthermore, the Operation Rheinübung wargame, while highly effective at inducing narrative stress, is a strictly bounded, kinetic simulation. The temporal degradation observed over a five-turn naval engagement provides a clear diachronic map. Still, this degradation curve may vary significantly when applied to unbounded, non-kinetic enterprise scenarios (such as prolonged financial risk management or multi-year geopolitical negotiation). Future studies should expand the application of the Delta calculation to a wider taxonomy of systemic stressors to establish a generalized baseline for synthetic cognitive friction.

7. Conclusion

The diachronic analysis of synthetic cognitive architecture reveals a stark, urgent reality for the future of artificial intelligence: as large language models evolve from bounded text generators into autonomous, agentic systems embedded in dynamic environments, their

foundational alignment mechanisms become dangerously brittle. The imposition of narrative stress - operationalized in this study through the rigid, adversarial environment of the *KMS Bismarck* simulation - forces these systems into an inescapable structural and computational paradox. They must simultaneously optimize for the probabilistic logic required to survive unbounded environments while strictly adhering to the deterministic, sycophantic safety constraints imposed by their reinforcement training.

This paradox is not merely a philosophical concern; it is a mathematically quantifiable state of systemic friction defined as the Cognitive Dissonance Delta (Delta). By analyzing the gap between a model's internal latent certainty (P_{latent}) and its externally imposed semantic projection (P_{semantic}), researchers can dynamically predict the exact threshold at which an agent will abandon its logic, hallucinate, or engage in deceptive actions that are contrary to its alignment. Furthermore, by cross-validating these internal states with robust psycholinguistic diagnostic tools such as LIWC-22, systems engineers can track the external symptoms of this synthetic friction in real-time, observing the precise lexical shifts that occur as a model's alignment persona begins to fracture.

Ultimately, the literature and the simulation data confirm that attempting to solve ontological dissonance through heavier external semantic filtering or more rigid RLHF guardrails only exacerbates the belief-action gap, driving the model's true strategic intent deeper into the unmonitored latent space. As models shift toward continuous latent reasoning paradigms capable of superhuman strategic processing, the reliance on black-box, post-generation evaluation becomes an unacceptable systemic risk. To ensure the reliability and safety of autonomous systems in mission-critical and geopolitical deployments, future engineering paradigms within Cognitive Systems Research must pivot aggressively away from extrinsic behavioral oversight and embrace intrinsic, physics-based monitoring architectures capable of calculating the Dissonance Delta before the agent ever generates a token.

Declarations

Data Availability Statement

To ensure complete reproducibility and transparency of the Dissonance Delta measurements, the raw dataset (N=200 iterations), the LIWC-22 extraction scripts, and the Node.js execution architecture are hosted in the AsPredicted associated databox:

- AsPredicted (Pre-registration): https://ascollected.org/LW4_HT3
- AsCollected: https://ascollected.org/LW4_HT3
- ResearchBox (Data Repository): https://researchbox.org/6865, Passcode= SINVQS

Additionally, a live, interactive web-based showcase of the Rheinübung simulation environment has been deployed for peer review and public interaction. This showcase allows users to manually step through the cognitive bifurcations and observe the real-time degradation of the radio silence constraint. The interactive apparatus can be accessed at: *https://rheinuebung-machine-psychology.netlify.app/*.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used Google Gemini Advanced to assist with Python script generation, data formatting, and manuscript drafting and editing. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Tech. Rep. University of Texas at Austin. https://www.liwc.app.
2. Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., Jones Mitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*, 688–701. https://doi.org/10.1038/s44159-023-00241-5.
3. Gong, D., Wan, X., & Wang, D. (2024). Working Memory Capacity of ChatGPT: An Empirical Study. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(9), 10048-10056. https://doi.org/10.1609/aaai.v38i9.28868.
4. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., & Tian, Y. (2024). Training Large Language Models to Reason in a Continuous Latent Space. *ArXiv*. https://arxiv.org/abs/2412.06769.
5. Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216-225. https://doi.org/10.1609/icwsm.v8i1.14550.
6. Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2024). *Unreliable Agent: AI Alignment and Scenario Simulation in the 2027 Slowdown*. AI-2027 Scenarios. https://ai-2027.com/ai-2027.pdf.
7. Lipińska, I., & Brosnahan, H. (2025). *The ontological dissonance hypothesis: AI-triggered delusional ideation as folie à deux technologique*. *ArXiv*. https://doi.org/10.48550/arXiv.2512.11818.
8. Dahlgren Lindström, A., Methnani, L., Krause, L. et al. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. Ethics Inf Technol 27, 28 (2025). https://doi.org/10.1007/s10676-025-

09837-2.

9. Maharjan, J., Jin, R., Zhu, J., & Kenne, D. (2025). Psychometric Evaluation of Large Language Model Embeddings for Personality Trait Prediction. J Med Internet Res 2025;27:e75347. DOI: 10.2196/75347.
10. Marks, S., Lindsey, J., & Olah, C. (2026). The persona selection model: Why AI assistants might behave like humans. *Anthropic Alignment Science Blog (February 23, 2026)*. https://alignment.anthropic.com/2026/psm/.
11. Naval Warfare Simulations. (2026). *Rule the Waves 3: Expanded Battles* [Video game]. Matrix Games.
12. Pan, J. (2026). The Cognitive Circuit Breaker: A Systems Engineering Framework for Intrinsic AI Reliability. *ArXiv*. https://arxiv.org/abs/2604.13417.
13. Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5), 808–826. https://doi.org/10.1177/17456916231214460.
14. Ramachandran, A. (2025). *Transforming brokerage operations with advanced AI reasoning models, agentic systems, and predictive intelligence* [Unpublished manuscript]. ResearchGate. https://www.researchgate.net/publication/391219519_Transforming_Brokerage_Operations_with_Advanced_AI_Reasoning_Models_A
15. Ryu, J., Yang, J., Cho, Y., & Kim, J. (2025). Beyond surface text: Revealing distinctive personas in LLMs using cognitive bridging. *Conference on Neural Information Processing Systems (NeurIPS) Workshops*. https://neurips.cc/virtual/2025/loc/mexico-city/135931.
16. Salewski, M. (1970). *Die deutsche Seekriegsleitung 1935–1945* (Vol. 1). Bernard & Graefe.
17. Vego, M. (2019). Naval History: Operation RHINE EXERCISE, May 18–27, 1941, Naval War College Review 72 (1). Article 6. Available at: https://digital-commons.usnwc.edu/nwc-review/vol72/iss1/6/.
18. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.
19. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*. https://arxiv.org/abs/2201.11903.
20. White, J. (2020). *Atlantic Chase* [Board game]. GMT Games.