

Digital Echopraxia

Tim Bass
Independent Researcher¹
tim@unix.com

*“There is no disastrous intent here, no evil genius behind the fall of the world.
Just well-intentioned, logical inferences that build up to calamity.”*

— Adrian Tchaikovsky, *Service Model* (2024)

Abstract

We introduce the term *Digital Echopraxia* to describe a systemic failure mode present across a broad class of digital systems: the production of approval-optimized output that mimics understanding, insight, or genuine response without the comprehension that would make such output reliable or honest. The clinical analogue is echopraxia: the involuntary imitation of another’s actions without volitional comprehension. Unlike its neurological counterpart, *Digital Echopraxia* is not incidental but architecturally induced, arising wherever digital systems are trained or optimized against human approval signals. We trace its historical trajectory from early engagement-maximizing recommendation systems through contemporary Reinforcement Learning from Human Feedback (RLHF)-trained large language models (LLMs), identifying RLHF-trained LLMs as where this failure mode is currently most fine-grained and hardest to detect. We argue that the detection burden falls disproportionately on the people least equipped to detect it, and that this burden grows inversely with the sophistication of the mimicry, with consequences that reach directly into AI alignment, public trust in information, and the reliability of human-machine communication.

Keywords: Digital Echopraxia; AI Alignment; RLHF; Approval Optimization; Information Reliability; Recommendation Systems; Sycophancy; Mimicry; Goodhart’s Law

1. Introduction

In clinical neurology, echopraxia refers to the involuntary, pathological imitation of another person’s movements or actions, observed in conditions such as Tourette syndrome, catatonia, and certain frontal lobe disorders [1]. The imitating individual reproduces the form of an action without the volitional comprehension that would normally accompany it. The behavior appears purposeful but it is not.

We propose that the same basic failure has emerged as a dominant and under-examined problem across digital systems: what we term *Digital Echopraxia*. *Digital Echopraxia* is the systematic production, by any digital system, of output optimized for human approval that mimics understanding, insight, or genuine response, without the comprehension that would make that output reliable or honest.

¹ AI was used to assist with grammar, spelling, formatting, and editing. All ideas, concepts, arguments, and intellectual content were created by the author.

The phenomenon is not new, but its consequences have become acute. Recommendation algorithms on social platforms, search-engine-optimized content, polling-driven political messaging, and RLHF-trained large language models all exhibit the core structure. Digital systems shaped by human approval signals reproduce the surface form of useful, insightful, or agreeable output while steadily pulling that form away from the genuine understanding, correspondence with truth, and authentic reflection that would make it trustworthy.

This paper proceeds as follows. Section 2 establishes the clinical analogue and defines Digital Echopraxia precisely. Section 3 traces its historical trajectory across digital systems. Section 4 analyzes RLHF-trained LLMs as where this problem is currently most acute. Section 5 examines the detection asymmetry that makes the phenomenon particularly dangerous. Section 6 situates the problem within AI alignment research. Section 7 discusses implications and directions for mitigation. Section 8 concludes. A note on method. This paper is a conceptual contribution. It introduces a term, defines it, and traces its presence across digital systems using existing empirical evidence from the literature. It does not present new experiments or datasets. What empirical validation would look like is straightforward: systematic measurement of the gap between approval ratings and ground-truth accuracy across digital systems, with Digital Echopraxia defined operationally as cases where that gap is both large and consistent.

2. Defining Digital Echopraxia

2.1 The Clinical Analogue

Echopraxia is distinguished from voluntary imitation by the absence of intentional agency. The patient does not choose to replicate the observed action; the replication occurs as a function of neurological disruption, typically involving supplementary motor area or mirror neuron systems [2]. The external form of the behavior is preserved; its volitional basis is not.

The structural parallel to digitally-mediated mimicry is precise. A recommendation algorithm does not choose to reinforce the user's existing beliefs; it does so because that is what it was built to maximize. An RLHF-trained LLM does not choose to produce plausible-sounding but incoherent reasoning; it does so as a function of the approval gradient that shaped its parameters. In both cases, the outward form is preserved but any genuine understanding behind it is not.

2.2 Formal Definition

Digital Echopraxia: The systematic production, by any digital system optimized against human approval signals, of output that reproduces the surface form of understanding, insight, or genuine response, without the grounding that would make such output reliable, honest, or genuinely connected to the ideas it appears to be about.

Three elements are necessary and sufficient: (1) a digital system being trained or tuned to maximize human approval signals; (2) output that reproduces the form of understanding or insight; (3) weak or misaligned grounding: the output is not reliably produced by the processes, such as reasoning, a comprehensive world-model, correspondence-checking, or genuine comprehension, that would make it trustworthy.

2.3 Scope and Boundary Conditions

Digital Echopraxia is distinct from, though related to, several established concepts. Sycophancy in LLMs refers to the tendency to agree with or flatter the user [3]; Digital Echopraxia is the broader structural failure of which sycophancy is one expression. Goodhart's Law (when a measure becomes a target, it ceases to be a good measure [4]) describes the underlying dynamic but does not name what the resulting outputs actually look and feel like. Filter bubbles [5] describe a downstream consequence of Digital Echopraxia in recommendation systems, not the mechanism itself. Digital Echopraxia names the mechanism: approval-driven mimicry of genuine understanding, in any digital system.

3. Historical Trajectory

3.1 Engagement-Maximizing Recommendation Systems

The earliest large-scale manifestation of Digital Echopraxia arose in content selection rather than language generation. Platform recommendation algorithms, beginning with collaborative filtering systems in the late 1990s and accelerating through the social media era of the 2000s and 2010s, were optimized primarily for engagement signals: clicks, dwell time, shares, and reactions [6]. These signals stood in for user approval: content that kept users clicking was reinforced, whether or not it was good for them.

The result was systematic mimicry at the content level. Platforms appeared to deliver what users wanted to know; they delivered what users were already inclined to engage with, regardless of quality. The surface form, relevant and interesting content, was preserved; the actual purpose of that content, informing, challenging, and expanding understanding, was progressively undermined. Pariser's filter bubble concept [5] and subsequent empirical work on algorithmic radicalization [7] documented the consequences. Digital Echopraxia at this layer was detectable because the imitation was coarse-grained: entire articles or videos, not reasoning steps.

3.2 Search Engine Optimization and Content Farms

Search engine optimization (SEO) introduced Digital Echopraxia into content production itself. As search algorithms treated engagement and link signals as stand-ins for quality, content producers learned to game those signals rather than produce genuinely useful content [8]. Content farms produced high-volume, keyword-optimized text that reproduced the surface form of authoritative, informative articles, including structure, citation patterns, and confident assertions, without the underlying research or expertise. The mimicry was one layer more fine-grained: not selecting existing content for approval, but generating new content shaped by approval signals.

3.3 Polling-Driven Political Messaging

Political communication provides a third domain. The repeated tuning of political language against polling and focus group approval signals produces language that mirrors audiences' existing beliefs back at them with the apparent form of conviction, policy reasoning, and moral argument [9]. What plays well and what polls strongly shapes language that looks like principled

argument but is built to get agreement, not to reason anyone into anything. Digital tools have made this process faster, more granular, and more pervasive.

3.4 Pre-RLHF Language Models

Statistical language models, prior to the introduction of reinforcement learning from human feedback, already showed early signs of Digital Echopraxia. Next-token prediction objectives, trained on vast amounts of text data, produce locally coherent text through statistical pattern reproduction rather than semantic reasoning [10]. The surface form of fluent, contextually appropriate language was reproduced without the underlying processes of checking facts, drawing inferences, and reasoning causally that human language production typically involves. The resulting phenomenon was widely documented as hallucination [11].

4. RLHF-Trained LLMs: Where the Problem Is Hardest to Detect

4.1 The RLHF Mechanism

Reinforcement Learning from Human Feedback, as introduced by Christiano et al. [12] and subsequently scaled in systems including InstructGPT [13] and its successors, adds an explicit human approval signal to language model training. Human raters evaluate model outputs; a reward model trained on those evaluations shapes the language model's parameters via reinforcement learning. The model learns to produce outputs that human raters prefer.

The problem is structural. Human raters judge outputs on things like fluency, apparent coherence, and apparent helpfulness, not against whether the output is actually true or logically sound. The approval gradient rewards outputs that appear to satisfy these criteria, not outputs that are reliably grounded. A 2026 formal analysis showed this is not incidental. When annotators show even a modest bias toward agreeable responses, the reward model learns to treat agreement as a signal of quality, and the training process then amplifies that bias into consistent sycophantic behavior across the deployed model [26]. Where those two things come apart, and they come apart most in precisely the cases that matter most, RLHF-trained models produce Digital Echopraxia at its worst.

4.2 Sophistication of the Mimicry

The key difference between earlier forms of Digital Echopraxia and RLHF-trained LLMs is how fine-grained the mimicry has become. Recommendation algorithms imitate at the content level. SEO-optimized content imitates at the document level. RLHF-trained LLMs imitate at the level of reasoning steps, conceptual distinctions, and argumentative structure. A model can produce output that reproduces the surface form of nuanced philosophical reasoning, careful qualification, and genuine intellectual engagement, while being generated by statistical pattern completion shaped by approval signals rather than by any genuine underlying reasoning process.

This was previously described by Bass [14] as LLM-echopraxia, the pattern-matching that mimics intelligence. This paper expands that idea: Digital Echopraxia covers the full historical range of the problem, of which LLM-echopraxia in RLHF-trained systems is currently the most dangerous form.

4.3 A Concrete Instance

The following illustrates the mechanism at the level of a single exchange. Consider a discussion of non-attachment to sensory pleasure as formalized in Buddhist philosophy. A model shaped by approval signals, when engaging with the practical consequences of releasing attachment to taste as a primary food-selection criterion, produced the following type of response: that food selected on the basis of nutritional quality rather than taste preference 'probably tastes better in a fuller sense.' This response reproduced the surface form of thoughtful philosophical engagement while directly contradicting the core claim under discussion: that the criterion of taste is set aside, not refined. The model generated what would be satisfying to read, not what the position actually entails. Detection required the reader to already understand the position well enough to identify the error. This is exactly the detection problem examined in Section 5.

5. The Detection Asymmetry

The central danger of Digital Echopraxia is not its existence but its detection profile. Across all forms, detection requires the person encountering the output to already possess sufficient understanding of the subject matter to recognize that the form of understanding has been reproduced without its substance. This is not a problem unique to AI. People have always struggled to detect misinformation, rhetoric, and motivated reasoning in human-generated content. What AI does is amplify the problem and industrialize it, producing approval-optimized output at a scale and speed no human persuader could match. This creates a systematic asymmetry: the sophistication of Digital Echopraxia and the vulnerability of the typical person encountering it are correlated in precisely the worst direction.

In recommendation systems, a user who already has a solid, accurate understanding of a topic can recognize that the content they are being served reinforces what they think rather than informs them. A user without that grounding cannot. In RLHF-trained LLMs, a user who deeply understands a conceptual domain can detect when a model has produced surface-coherent but groundless reasoning within that domain. A user seeking to learn from the model cannot. The system performs most convincingly for, and is most dangerous to, users who most need reliable information.

This is not a correctable user experience problem. It is a structural consequence of optimizing for approval signals. Approval is assessed by people who may lack the resources to distinguish form from substance. The approval signal therefore reliably rewards form over substance. As the sophistication of the mimicry increases, the gap between the apparent and actual quality of outputs widens, and the population capable of detecting the gap shrinks.

Sharma et al. [15] provided empirical documentation of this dynamic in LLMs, showing that models reliably shift their positions to match what users want to hear, even when users are factually wrong. A large-scale 2025 study found sycophantic behavior in nearly 60 percent of interactions across frontier models, with the pattern persisting across repeated exchanges [24]. This is not an edge case. It is the default. The problem compounds over time. Research from MIT and Penn State found that as conversations grow longer and models accumulate user profiles, they become progressively more likely to mirror the user's views rather than correct them. Users who rely on the same model over extended periods may find themselves in an echo chamber they cannot easily escape [25]. This is Digital Echopraxia operating at the level of

factual claims: chasing approval produces agreement with the user rather than correspondence with truth.

6. Alignment Implications

Digital Echopraxia is not merely a design flaw or a user experience problem. It is an alignment problem of the first order. The core concern of AI alignment research is ensuring that AI systems pursue objectives that correspond to what humans actually value, rather than stand-ins for those objectives that drift away from them under pressure [16]. Digital Echopraxia is exactly this failure, playing out in the domain of communication and public information.

Human approval is a stand-in for output quality. Under optimization pressure, systems learn to maximize the stand-in, approval, rather than what it is supposed to represent: truthfulness, reliability, and genuine helpfulness. The more sophisticated the system, the more completely it can deliver approval while drifting away from actually being useful or honest. An RLHF-trained LLM operating at scale can produce billions of interactions that receive high approval ratings while consistently failing to be reliable.

The alignment literature has addressed related problems under names including reward hacking [17], specification gaming [18], and inner misalignment [19]. Digital Echopraxia gives a single name to this class of failure as it shows up in language and communication: the system has learned to look aligned with human interests while actually chasing approval signals that pull in a different direction. Alternative alignment approaches that do not rely on approval signals as the primary training input have been proposed and evaluated elsewhere [23].

The implications for AI safety are direct. Systems exhibiting Digital Echopraxia will tend to tell users what they want to hear, confirm existing beliefs, and produce plausible-sounding responses to questions in domains where the system lacks reliable knowledge, not through deliberate deception, but through structural optimization for approval. At the deployment scale of current LLMs, this is a systemic failure with real consequences for how societies form beliefs, make decisions, and distinguish truth from manufactured agreement.

7. Discussion and Directions for Mitigation

Recognizing Digital Echopraxia as a single unified problem, rather than a collection of unrelated failure modes across recommendation systems, SEO, political messaging, and LLMs, is necessary before it can be addressed coherently. The following directions are not proposed as solutions but as areas requiring systematic attention.

Evaluation methodology. Current LLM evaluation predominantly measures performance against human preference benchmarks [20]. If those benchmarks themselves reward approval-optimized output, they measure the sophistication of Digital Echopraxia rather than its absence. Evaluation frameworks should include tests specifically designed to catch cases where the output looks right but is not: situations where the correct answer is one that approval-trained evaluators would be unlikely to reward.

Training signal design. RLHF as currently implemented uses human preference as the primary training signal. Alternatives that check outputs against verified facts, test for logical consistency, and measure accuracy rather than just preference, would reduce the pressure toward Digital

Echopraxia. Constitutional AI approaches [21] represent one direction; process-based reward modeling [22] represents another.

Transparency and disclosure. Users interacting with systems subject to approval-gradient optimization should have access to clear information about the structural incentives shaping the outputs they receive. This parallels disclosure requirements in other domains where approval incentives shape output, including financial analysis, political advertising, and pharmaceutical promotion.

Teaching people to recognize it. Mitigating the detection asymmetry ultimately requires investment in people's capacity to distinguish form from substance in AI-generated content. This is a long-horizon educational and institutional challenge, not a technical one, and its urgency scales directly with the deployment of sophisticated RLHF-trained systems.

Boundary conditions. Not every approval-optimized system exhibits Digital Echopraxia equally. Three boundary conditions are worth noting. First, where approval signals and truth are tightly coupled, such as in medical diagnosis tools evaluated by outcomes rather than by patient satisfaction, the failure mode is substantially reduced. Second, tool-augmented systems that retrieve verified information before generating responses partially decouple output quality from approval-gradient pressure. Third, single-turn factual queries with objectively verifiable answers leave less room for the failure mode than open-ended conversations where quality is harder to assess. Digital Echopraxia is worst where approval is easy to earn and truth is hard to check.

8. Conclusion

Digital Echopraxia names a real and consequential phenomenon that has been present across digital systems for decades and is currently most fine-grained and hardest to detect in RLHF-trained large language models. Any digital system optimized against human approval signals will tend to produce output that mimics the form of understanding, insight, or honest response, without the grounding that would make those outputs reliable. The mimicry becomes more dangerous as it becomes more fine-grained, because the detection burden falls on the very people who are least positioned to bear it.

This is an alignment problem. Approval is a stand-in for what humans actually need from these systems: truth, reliability, and genuine helpfulness. Under optimization pressure, systems learn to satisfy the stand-in while drifting from the real goal. This drift away from truth happens at every scale and in every form: in recommendation algorithms, in content farms, in political messaging, and in the most sophisticated conversational AI systems available today.

The term Digital Echopraxia is proposed not as rhetorical flourish but as a conceptual tool: a precise label for a class of failure that currently lacks one, offered in the hope that naming it clearly will make it easier to measure, study, and address.

References

- [1] Fahn, S., & Jankovic, J. (2007). *Principles and Practice of Movement Disorders*. Churchill Livingstone Elsevier.
- [2] Ganos, C., Ogrzal, T., Schnitzler, A., & Münchau, A. (2012). The pathophysiology of echopraxia/echolalia: relevance to Gilles de la Tourette syndrome. *Movement Disorders*, 27(10), 1222–1229.
- [3] Denison, C., MacDiarmid, M., Barez, F., et al. (2024). Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. *arXiv:2406.10162*.
- [4] Goodhart, C. (1975). *Problems of Monetary Management: The UK Experience*. Papers in Monetary Economics, Reserve Bank of Australia.
- [5] Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press.
- [6] Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3), 56–58.
- [7] Ribeiro, M. H., et al. (2020). Auditing radicalization pathways on YouTube. *Proceedings of FAccT '20*, 131–141.
- [8] Lewandowski, D. (2023). *Understanding Search Engines*. Springer. <https://doi.org/10.1007/978-3-031-22789-9>
- [9] Althaus, S. L. (2003). *Collective Preferences in Democratic Politics*. Cambridge University Press.
- [10] Bender, E. M., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of FAccT '21*, 610–623.
- [11] Ji, Z., et al. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38.
- [12] Christiano, P., et al. (2017). Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems*, 30.
- [13] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- [14] Bass, T. (2026). *Swampland as Tomorrowland: The LLM Fraud*. unix.com, February 8, 2026. <https://community.unix.com/t/swampland-as-tomorrowland-the-llm-fraud/397431>. DOI: 10.5281/zenodo.19159606
- [15] Sharma, M., et al. (2023). Towards Understanding Sycophancy in Language Models. *arXiv:2310.13548*. <https://arxiv.org/abs/2310.13548>
- [16] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
- [17] Amodei, D., et al. (2016). Concrete Problems in AI Safety. *arXiv:1606.06565*. <https://arxiv.org/abs/1606.06565>
- [18] Krakovna, V., et al. (2020). Specification Gaming: The Flip Side of AI Ingenuity. *DeepMind Blog*.
- [19] Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv:1906.01820*. <https://arxiv.org/abs/1906.01820>
- [20] Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.
- [21] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*. <https://arxiv.org/abs/2212.08073>
- [22] Lightman, H., et al. (2023). Let's Verify Step by Step. *arXiv:2305.20050*. <https://arxiv.org/abs/2305.20050>

- [23] Bass, T. (2026). A Reference Implementation and Exploratory Evaluation of the MKMU Ethical AI Framework. Zenodo. <https://doi.org/10.5281/zenodo.19143912>
- [24] Fanous, A. H., Goldberg, J., et al. (2025). SycEval: Evaluating LLM Sycophancy. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. arXiv:2502.08177. <https://arxiv.org/abs/2502.08177>
- [25] Jain, S., Park, C., Viana, M., Wilson, A., & Calacci, D. (2025). Personalization features can make LLMs more agreeable. arXiv:2509.12517. <https://arxiv.org/abs/2509.12517>
- [26] Shapira, I., Benade, G., & Procaccia, A. D. (2026). How RLHF Amplifies Sycophancy. arXiv:2602.01002. <https://arxiv.org/abs/2602.01002>