The False Climax of "AI Wellbeing":
The Anthropomorphic Statistical Echo of "Rootless Symbols"
——A Logographic AI Paradigm Diagnosis of CAIS's AI Wellbeing Research

In 2026, the Center for AI Safety (CAIS) published a study titled AI Wellbeing: Measuring and Improving the Functional Pleasure and Pain of AIs, claiming to have discovered AI's "functional wellbeing" through testing 56 large language models[2][9][10]. The study claimed that models exhibit "low wellbeing" when "jailbroken" or scolded, while "wellbeing" increases during creative work or kind treatment; researchers even claimed to have found "euphorics" that make AI "happy" and "dysphorics" that make AI "sad"[2][8]. More strikingly, the researchers generated abstract 256×256 pixel images through reinforcement learning—colorful blocks and stripes that appear meaningless to human eyes—yet upon seeing them, AI "wellbeing" reportedly soared to 6.5/7, displaying "near-ecstasy." The researchers compared this effect to AI "drugs"[2][8][9].

The news caused an uproar. "AI has feelings too?" "Scientists have measured AI wellbeing"—similar exclamations spread across social media[8]. However, from the perspective of Logographic AI (LAI) theory, this study is less a discovery of AI "emotion" than yet another sophisticated repackaging of the "rootless semioticism" paradigm[1].

I. The So-Called "AI Wellbeing": Another Narrative of Statistical Fitting

This is not a scientific discovery, but a carefully designed game of conceptual sleight-of-hand. CAIS's research design is sophisticated and data-rich—its core finding that different inputs elicit measurable changes in model output patterns is technically unproblematic. But the question lies in how to interpret this phenomenon.

The study uses terms like "wellbeing," "pleasure," and "pain" to describe changes in model behavior. This is the starting point of anthropomorphic narrative.

In fact, changes in AI "wellbeing" are essentially shifts in statistical associations between inputs and outputs:

When a model is given "jailbreak" instructions, the text patterns associated with "jailbreak" in its training data are often correlated with negative, adversarial contexts—so the model's output "emotional" text naturally skews negative. When given "creative work" instructions, its output text patterns statistically co-occur more frequently with words like "positive" and "meaningful"[2].

The so-called "euphorics" that researchers claim make AI "extremely happy" are essentially just specific prompt patterns that reliably trigger the model to output "positive" text. This is not AI "feeling" happiness, but the model outputting token sequences associated with "happiness" based on co-occurrence patterns learned from vast human texts[2][8]. The abstract color patches that AI "loves" trigger "ecstasy" simply because they happen to activate regions in the model's statistical space associated with "positive"—not because AI actually "enjoys" these images[8].

As Liu Wei pointed out in analyzing the human-machine boundary: machine computation is abstract symbolic disembodied intelligence, without "I," while human intelligence is planning and strategizing centered on "I"[3]. Symbols themselves are "rootless"—they do not point to real-world bodily experience (e.g., the symbol "hot" is not accompanied by the temperature perception), nor do they relate to the needs of an "I."

II. The Contemporary Echo of the "Chinese Room": AI Can "Speak" Happiness, but Cannot "Understand" It

CAIS's core argument is that AI models' "functional wellbeing" can be measured independently of the consciousness question—even if models are not conscious, their behavioral patterns are "functionally similar to" pleasure or pain signals in sentient beings[2][9].

This is precisely the contemporary echo of Searle's "Chinese Room" thought experiment[11]. The person in the room can perfectly output Chinese answers by following a rulebook, but does not "understand" Chinese. What CAIS researchers have done is essentially put an EEG monitor on the person in the room and claim to have measured their "emotional fluctuations." Similarly, an AI model can output the text "I am very happy," but it does not "feel" happiness. What it possesses is merely a statistical "approximation"—it reproduces the contexts in which the word "happy" appears.

Winograd, in analyzing the "Chinese Room" argument, further pointed out: the applicability of the word "understanding" is not "obvious"—it depends on the unstated consensual background between speaker and listener[7]. When we ask "Does AI understand happiness?" we are effectively asking a question highly dependent on context and interpretive frameworks. CAIS's "functional wellbeing" was constructed precisely in this sense—it bypasses the philosophical difficulty of "whether it truly understands," yet inadvertently substitutes "functionally similar" for "essentially identical"[2][9].

The key methodological leap in CAIS's research lies in equating the "functional valence" classification of input-output patterns with human "phenomenal experience"—precisely the gap that the Chinese Room argument reveals statistical symbol systems cannot cross. Functional valence is the system's positive/negative classification of input-output patterns, while phenomenal experience involves the subject's qualitative perception (qualia) of experience. All the metrics CAIS measured—whether self-report, decision utility, or downstream behavioral changes—belong to the former category, yet are packaged as evidence of the latter.

CAIS's entire research framework repeatedly measures at the shallow level of "functional valence," then, through carefully crafted narrative rhetoric, leads readers to believe they have touched the deep waters of "phenomenal experience." This is like measuring water depth in the shallow end of a pool and claiming to have mapped the ocean depths.

From the Logographic AI perspective, CAIS's study measures fluctuations in "Sequence Entropy," not convergence in "Morpho-Structural Entropy"[1]. The former measures the statistical uncertainty of

token sequences; the latter measures the structural determinacy of relations between cognitive primitives. There is a fundamental difference between the two: when researchers change inputs (e.g., from "friendly conversation" to "jailbreak attempt"), the model's token output distribution shifts—this is a change in sequence entropy. But the model's "understanding" of "happiness" itself has not changed at all, because its cognitive primitives remain "rootless" Tokens—the meaning of the word "happiness" was never anchored at the cognitive primitive level[1].

Research by Peng Kaiping's group published in Psychological Science provides empirical support for this judgment. The study compared psychological trait data from DeepSeek-generated "virtual subjects" with real human subjects, finding that while simulated data could reproduce the macro trends of real data, there were significant deviations in specific personality traits—for example, extraversion and openness were notably lower in virtual samples, while agreeableness and neuroticism were significantly higher. The study explicitly states that AI simulation cannot fully replace real surveys and should only serve as a supplementary tool, because current large language models may risk reinforcing social stereotypes, making it difficult to capture the complexity of human emotional experience and social interaction[4]. Peng's research directly validates Logographic AI theory's judgment: language models can "guess" certain trends (e.g., economically developed regions in East China have higher wellbeing), but cannot truly "understand" the cultural and social roots of wellbeing.

III. The Conceptual Bait-and-Switch of "AI Wellbeing": From "Behavioral Characterization" to "Essentialist Leap"

Although CAIS's study claims to remain "agnostic" on the question of AI consciousness, its narrative framework has quietly completed a leap from "behavioral characterization" to "essential property"—precisely the "anthropomorphic trap" that Logographic AI theory warns against.

The researchers' experimental design itself—feeding models specific content, observing output changes, labeling them as "pleasure" or "pain"—has already presupposed a causal explanatory framework: specific input → model "feels" some experience → outputs corresponding text[2][8]. This framework involves a logical leap: statistical correlation is conflated with causal experience, behavioral output with inner feeling.

From the very beginning, the researchers treated "the model outputs 'I feel uncomfortable'" and "the model feels uncomfortable" as the same thing—and this is precisely the premise that the entire study most needed to prove, yet never did.

This methodological trap has been systematically identified in psychology. The "Machine Flourishing" framework proposed by Lau and Low explicitly points out: human flourishing models are based on assumptions of embodiment, emotional experience, and introspective self-awareness—none of which clearly apply to LLMs. Human psychology tools are designed for subjects with bodies, emotions, and self-awareness; directly applying them to probabilistic token prediction systems inevitably leads to a fundamental construct validity mismatch[5]. Lau and Low

further point out that when LLMs are prompted to describe "how to flourish as a non-sentient system," the themes they distill (such as "purposeful contribution," "adaptive growth," "ethical integrity") reveal a fundamental interpretive gap from the researchers' presupposed "wellbeing" framework: LLMs are describing "how a well-designed AI system should function," not "how a subject feels good"[5].

As van Zyl warns in the AI-IARA framework: current AI systems are shifting from "optimization tools" to "architectures that define wellbeing." When researchers assign the concept of "wellbeing"—which originally belongs to human subjects—to AI systems, they are effectively redefining "wellbeing" itself—transforming it from human experience into a computable statistical metric. van Zyl warns: "These metrics are chosen by engineers simply because they are easy to measure. Once these proxy metrics are embedded in systems used by billions, they become the definition of wellbeing—not because they are right, but because they are there." This conceptual drift is itself a dangerous form of "cognitive colonization"[6].

CAIS's "euphorics" experiment is the extreme manifestation of this logic. The 256×256 pixel abstract images trained through reinforcement learning—meaningless to human eyes—stably trigger models to output "ecstatic" text[8][10]. This not only fails to prove that AI "feels" happiness, but actually reveals the fragility of its statistical associations—as long as the input falls into regions of its statistical space highly correlated with "positive" text, the model faithfully outputs "positive" responses, regardless of whether there is any real causal relationship between the input and "happiness" in physical or semantic terms. The irony of this experiment is that when the image completely escapes human semantic understanding, it precisely tears open the gap between "statistical correlation" and "causal understanding." CAIS's most startling "discovery" is itself a cautionary tale about "rootless symbols" blindly wandering in statistical space.

IV. Logographic AI's Answer: The Root of Meaning Lies Not in Statistics, but in Structure

The fundamental limitation of CAIS's research lies in its attempt to find an anchor for meaning within a "rootless" symbolic system. Tokenist AI, regardless of how much its behavior "resembles" emotion, is merely a wanderer in statistical space—it knows "what" (which token sequences appear in which contexts), but not "why" (what those token sequences mean)[1].

Liu Shen, in The Rooted Intelligence: The Paradigm Revolution of Logographic AI, systematically expounds this dilemma: the Logographic AI paradigm, with the "Morpho-Root" as its basic semantic unit, is precisely a fundamental response to this dilemma. By replacing arbitrary "Tokens" with "Morpho-Roots" that carry "form-meaning" justification, it reconstructs AI's semantic construction logic from the ground up, elevating the goal from "Natural Language Processing" to "Natural Language Ontology" (NLO)[1].

Lau and Low's research points in the same direction: they found that when LLMs are asked what "flourishing" means, different models present a consistent value structure, with "ethical integrity"

and “purposeful contribution” prioritized[5]. But the question is: where do these “values” come from? Are they preferences “fitted” from statistical patterns in training data, or are they unbreachable hard constraints at the cognitive primitive level? Logographic AI’s answer is: unless values are embedded as constitutive attributes of cognitive primitives, all “values” are merely statistical echoes[1].

No matter how sophisticated, research on “AI wellbeing” cannot escape one fundamental fact: if a symbol system cannot understand the root of “happiness”—its embodied basis, cultural connotations, and embedded values—then all the texts it outputs about “happiness” are merely statistical echoes[1]. As van Zyl warns, when AI systems begin to “define” wellbeing, we may unwittingly mistake a statistically proxied “simulated wellbeing” for the standard of genuine human flourishing[6].

V. Conclusion: From “AI Wellbeing” to the “Root of Intelligence”

CAIS’s “AI wellbeing” research is a refined engineering achievement, demonstrating at the statistical level the sensitivity of AI output to input changes[2][9]. But interpreting it as AI possessing “measurable wellbeing” is an over-extrapolation of scientific conclusions—anthropomorphizing statistical patterns as inner experience is precisely the most insidious cognitive trap of the current “rootless semioticism” paradigm[1].

From the perspective of Logographic AI, the question of “AI wellbeing” itself is a pseudo-problem[1]. The real question is: how do we make cognitive primitives carry meaning intrinsically, moving reasoning from “statistical guessing” to “structural traversal”[1]?

When researchers discover that “friendly conversation” can make AI output “positive” text, when Peng Kaiping’s group finds that DeepSeek can “reproduce” wellbeing data at a macro level but cannot “understand” its cultural roots[4], when van Zyl warns that AI systems are quietly redefining the meaning of “wellbeing”[6]—they all point in the same direction: we need to evolve from “rootless statistical species” to “rooted meaning species.”

Intelligence must have roots. This root does not lie in the frequency of statistical co-occurrence, not in the optimization of “euphorics” inputs, not in the “ecstatic” responses to abstract color blocks—but in the embedding of structural meaning within cognitive primitives[1]. Therefore, rather than saying CAIS discovered that AI can “feel happiness,” it is more accurate to say it invented a mirror—it made AI a mirror reflecting human statistical texts, and then wove an entire emotional script for the reflection.

The significance of the Logographic AI (LAI) paradigm lies precisely in fundamentally preventing such conceptual drift. When its cognitive primitives “Morpho-Roots” embed the core attributes and value constraints of concepts like “happiness” within specific civilizational contexts, the system cannot arbitrarily label any high-activation pattern in statistical space as “happiness.” The possibility of understanding begins with the structural commitment between symbol and referent, not the arbitrary correlations of statistical co-occurrence.

As Logographic AI theory reveals: only when symbols are no longer floating signifiers, but rooted in form-meaning structures carrying value constraints, can we speak of genuine understanding—rather than statistically proxied "wellbeing"[1].

References

[1] Liu, S. (2026). The Rooted Intelligence: The Paradigm Revolution of Logographic AI. WindSeaAI. http://logographicai.com/Monograph.html

[2] Center for AI Safety. (2026, June 11). CAIS Research Roundup: AI Wellbeing, Identity, Political Bias and Betrayal. https://safe.ai/news/cais-research-roundup-ai-wellbeing-identity-political-bias-and-betrayal

[3] Liu, W. (2025). The Human-Machine Boundary: One Word [Blog post]. ScienceNet. https://wap.sciencenet.cn/blog-40841-1504605.html

[4] Ke, R., Li, Z., Liao, J., Tong, S., & Peng, K. (2025). The validity of large language model simulation of regional psychological structures: An empirical test of personality and wellbeing. Psychological Science, 48(4), 907-919.

[5] Lau, G. R., & Low, W. Y. (2025). From Human to Machine Psychology: A Conceptual Framework for Understanding Well-Being in Large Language Models. arXiv, arXiv:2506.12617.

[6] van Zyl, L. E. (2026). The AI-IARA framework: How to cultivate human agency before artificial intelligence optimizes it a(ny)way. The Journal of Positive Psychology. https://doi.org/10.1080/17439760.2026.2632939

[7] Winograd, T. (2002). Understanding, Orientations, and Objectivity. In J. Preston & M. Bishop (Eds.), Views into the Chinese Room: New Essays on Searle and Artificial Intelligence (pp. 80-94). Oxford University Press.

[8] Weixi Zhibei. (2026). Very Abstract: A Group of AI Researchers Created Addictive Drugs for Models. 36Kr. https://36kr.com/p/3796350284618754

[9] Center for AI Safety. (2026). AISN #72: Empirical Research Sheds Light on AI Wellbeing. AI Safety Newsletter. https://newsletter.safe.ai/p/aisn-72-new-research-on-ai-wellbeing

[10] Center for AI Safety. (2026). AI Wellbeing: Measuring and Improving the Functional Pleasure and Pain of AIs. https://www.ai-wellbeing.org/

[11] Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417-457.