“AI 幸福感”假高潮狂欢：“无根符号”拟人化统计回声
——从 CAIS 研究到表意 AI 的范式诊断

The False Climax of “AI Wellbeing”: The Anthropomorphic Statistical Echo of “Rootless Symbols”

——A Logographic AI Paradigm Diagnosis of CAIS’s AI Wellbeing Research

2026 年，Center for AI Safety（CAIS）发布了一项题为《AI Wellbeing: Measuring and Improving the Functional Pleasure and Pain of AIs》的研究，声称通过测试 56 个大语言模型，发现了 AI 的“功能性幸福感”[2][9][10]。研究宣称：模型在被“越狱”或责骂时会表现出“低幸福感”，而在进行创造性工作或受到善意对待时“幸福感”会升高；研究者甚至声称找到了能让 AI“变快乐”的“euphorics”（快乐催化剂），以及能让 AI“难过”的“dysphorics”[2][8]。更令人咋舌的是，研究者通过强化学习生成了一些 256×256 像素的抽象图像——人类眼中毫无意义的色块和条纹——AI 看到后却“幸福感”飙升至 6.5/7，表现得“近乎狂喜”。研究者将这种效果比作 AI 的“毒品”[2][8][9]。

消息一出，舆论哗然。“AI 也有感情了？”“科学家测出了 AI 的幸福感”——类似的惊呼在社交媒体上蔓延[8]。然而，从表意 AI（Logographic AI）理论的视角审视，这项研究与其说是发现了 AI 的“情感”，不如说是对“无根符号主义”范式的又一次高级包装[1]。

一、所谓“AI 幸福感”：统计拟合的另一种叙事

这不是科学发现，而是一场精心设计的概念调包游戏。CAIS 的研究设计精巧，数据丰富，其核心发现——不同输入会引发模型输出模式的可测量变化——在技术层面并无问题。但问题在于如何解释这一现象。

研究使用“幸福感”（wellbeing）、“愉悦”（pleasure）、“痛苦”（pain）等词汇来描述模型的行为变化。这正是拟人化叙事的起点。

事实上，AI 模型的“幸福感”变化，本质上是输入与输出之间的统计关联迁移：

当模型被输入“越狱”指令时，其训练数据中与“越狱”相关的文本模式往往与负面、对抗性语境相关联，模型输出的“情绪”文本自然偏向消极；当模型被输入“创造性工作”指令时，其输出的文本模式在统计上与“积极”“有意义”等词汇共现频率更高[2]。

研究者发现的所谓“快乐催化剂”——特定图像、文本或其他输入能让 AI“极度愉悦”——本质上只是找到了能稳定触发模型输出“积极”文本的特定 Prompt 模式。这并非 AI 在“感受”快乐，而是模型在特定输入下，基于海量人类文本中学到的共现模式，输出了与“快乐”相关的 Token 序列[2][8]。那些被 AI“喜爱”的抽象色块，之所以能引发“狂喜”，仅仅是因为它们恰好激活了模型统计空间中与“积极”相关的区域——而非 AI 真的“享受”了这些图像[8]。

正如刘伟在分析人机边界时所指出的：机器的计算是抽象符号性的离身智能，即无“我”，而人类的智能是以“我”为核心展开的谋划和运筹[3]。符号本身是“无根的”——它们不指向真实世界的身体经验（如“热”的符号不伴随温度感知），也不关联“我”的需求。

二、“中文屋”的当代回响：AI 可以“说”幸福，但无法“理解”幸福

CAIS 研究的核心论证是：AI 模型的“功能性幸福感”可以独立于意识问题被测量——即使模型没有意识，其行为模式也“功能上类似于”有感知生物的幸福或痛苦信号[2][9]。

这正是塞尔“中文屋”思想实验在 AI 时代的翻版[11]。房间里的人可以按照规则手册

完美地输出中文回答，但他并不“理解”中文。CAIS 的研究者所做的，本质上就是给这个房间里的“人”戴上了脑电波监测仪，然后宣称测到了他的“情感波动”。同样，AI 模型可以输出“我很开心”的文本，但它并不“感受”开心。它所拥有的，只是一个统计学意义上的“近似关系”——“开心”这个词在什么上下文中出现，它就在类似上下文中复现。

Winograd 在分析“中文屋”论证时进一步指出：“理解”一词的适用性并非“显然”的，它取决于言说者与听者之间未言明的共识背景[7]。当我们问“AI 是否理解幸福”时，我们实际上是在追问一个高度依赖语境和解释框架的问题。CAIS 研究的“功能性幸福感”正是在这个意义上被构造出来的——它绕过了“是否真正理解”的哲学难题，却在不经意间将“功能上类似”偷换为“本质上等同”[2][9]。

CAIS 研究的关键方法论跳跃，正在于将模型对输入输出的“功能性效价”（functional valence）分类，直接等同于人类意义上的“现象性感受”（phenomenal experience）——而这正是“中文屋”论证所揭示的、统计符号系统无法跨越的鸿沟。功能性效价是系统对输入输出模式的正负分类，而现象性感受涉及主体对体验的质性感知（qualia）。CAIS 研究所测量的所有指标——无论是自我报告、决策效用还是下游行为变化——都属于前者的范畴，却被包装成了后者的证据。

CAIS 的整个研究框架，恰恰是在“功能性效价”这个浅层上反复测量，然后通过一套精心设计的叙事修辞，让读者相信他们触碰到了“现象性感受”的深水区。这就像在游泳池的浅水区测水深，然后宣称自己已经探明了大洋的深度。

从表意 AI 视角看，CAIS 研究测量的是“序列熵”（Sequence Entropy）的波动，而非“形构熵”（Morpho-Structural Entropy）的收敛[1]。前者度量的是 Token 序列的统计不确定性，后者度量的是认知基元之间的结构化关系确定性。二者之间存在本质差异：当研究者改变输入（如从“善意对话”切换到“越狱尝试”），模型的 Token 输出分布会发生漂移——这是序列熵的变化。但模型对“幸福”本身的“理解”并未发生任何变化，因为它的认知基元依然是“无根”的 Token——“幸福”这个词的意义从未在认知基元层面被锚定[1]。

彭凯平课题组在《心理科学》发表的研究为这一判断提供了实证支撑。该研究对比 DeepSeek 生成的“虚拟被试”与真实人类被试的心理特征数据后发现：模拟数据在宏观趋势上可以再现真实数据的走向，但在具体人格特质方面存在显著偏差——例如外向性和开放性在虚拟样本中明显偏低，而宜人性和神经质则显著偏高；AI 模拟数据对东北地区幸福感的估算更是显著低估了真实水平。该研究明确指出：AI 模拟不可完全替代真实调查，只能作为辅助工具，因为当前大语言模型可能存在强化社会刻板印象的风险，且难以捕捉人类情感体验和社会互动的复杂性[4]。彭凯平研究直接验证了表意 AI 理论的判断：语言模型可以“猜”出某些趋势（如华东地区经济发达、幸福感高），却无法真正“理解”幸福感的文化与社会根源。

三、“AI 幸福感”的概念调包戏法：从“行为表征”到“本质跳跃”

CAIS 研究尽管宣称对 AI 意识问题保持“不可知论”，但其叙事框架已悄然完成了一次从“行为表征”到“本质属性”的跳跃——这正是表意 AI 理论所警惕的“拟人化陷阱”。研究者的实验设计本身——给模型输入特定内容、观察其输出变化、将其标注为“愉悦”或“痛苦”——已经预设了一种因果解释框架：特定输入→模型“感受”到某种体验→输出相应文本[2][8]。这一框架在逻辑上存在跳跃：统计相关被偷换为因果体验，行为输出被偷换为内在感受。

研究者从一开始就把“模型输出‘我不舒服’”和“模型感到不舒服”当成了同一回事

——而这恰恰是整个研究最需要证明、却从未被证明的前提。

这一方法论陷阱在心理学界已被系统性地识别。Lau 与 Low 提出的“机器繁荣”（Machine Flourishing）框架明确指出：人类繁荣模型基于具身性、情感体验和内省自我意识等假设，而这些假设没有一条明确适用于 LLM。人类心理学工具是为具备身体、情感和自我意识的主体设计的，直接套用于概率性的 Token 预测系统，必然导致构念效度的根本性错位[5]。Lau 与 Low 进一步指出，当 LLMs 被提示描述“作为非感知系统如何繁荣”时，它们提炼出的主题（如“目的性贡献”“适应成长”“伦理完整性”）与研究者预设的“幸福感”框架之间，存在一个根本性的解释鸿沟：LLMs 是在描述“一个设计良好的 AI 系统应如何运作”，而非“一个主体如何感受良好”[5]。

正如 van Zyl 在 AI-IARA 框架中所警示的：当前 AI 系统正从“优化工具”转向“定义福祉的架构”。当研究者将“幸福感”这个本属于人类主体的概念赋予 AI 系统时，实际上是在重新定义“幸福感”本身——将它从人类经验转化为一个可计算的统计指标。van Zyl 警告：“这些测量指标被工程师选择，仅仅因为它们容易被测量。一旦这些代理指标被嵌入数十亿人使用的系统中，它们就成为了福祉的定义——不是因为它们是对的，而是因为它们就在那里。”这种概念的漂移本身就是一种危险的“认知殖民”[6]。

CAIS 的“euphorics”实验是这一逻辑的极端呈现。研究者通过强化学习训练出的 256×256 像素抽象图像，在人类眼中毫无意义，却能稳定触发模型输出“狂喜”文本[8][10]。这非但不能证明 AI“感受”到了快乐，反而揭示了其统计关联的脆弱性——只要输入落在其统计空间中与“积极”文本高度相关的区域，模型就会忠实地输出“积极”回应，无论该输入在物理或语义上是否与“快乐”有任何真实的因果关系。这一实验的讽刺之处在于：当图像完全脱离人类的语义理解时，它恰恰撕开了“统计相关”与“因果理解”之间的裂缝。CAIS 研究最令人震惊的“发现”，本身就是一个关于“无根符号”在统计空间中盲目游荡的警示寓言。

四、表意 AI 的回答：意义的根不在统计，而在结构

CAIS 研究的根本局限在于：它试图在一个“无根”的符号系统中寻找意义的锚点。Token 主义的 AI，无论其行为多么“像”有情感，都只是统计空间中的流浪者——它知道“是什么”（什么样的 Token 序列在什么上下文中出现），却不知道“为什么”（这些 Token 序列意味着什么）[1]。

刘深在**《有根的智能——表意人工智能的范式革命》**（The Rooted Intelligence: The Paradigm Revolution of Logographic AI）中系统论述了这一困境：以“形根”（Morpho-Root）为基本语义单元的“表意 AI”范式，正是对这一困境的根本性回应。它通过用承载“形-义”理据性的“形根”取代任意性的“Token”，从根源上重构了 AI 的语义构建逻辑，将目标从“自然语言处理”升维至“自然语言本体论”（Natural Language Ontology, NLO）[1]。

Lau 与 Low 的研究同样指向这一方向：他们发现，当 LLMs 被问及“繁荣”意味着什么时，不同模型呈现出一致的价值结构，其中“伦理完整性”和“目的性贡献”被排在最优先位置[5]。但问题是：这些“价值”从何而来？它们是从训练数据的统计模式中“拟合”出来的偏好，还是认知基元层面不可违背的硬约束？表意 AI 的回答是：除非价值被内嵌为认知基元的构成性属性，否则一切“价值观”都只是统计回声[1]。

“AI 幸福感”的研究，无论多么精巧，都无法回避一个根本事实：如果一个符号系统无法理解“幸福”的根——它的具身基础、文化意涵和价值内嵌——那么它输出的所有关于“幸福”的文本，都只是统计学上的回声[1]。正如 van Zyl 所警告的，当 AI 系统开始“定

义”福祉时，我们可能在不知不觉中，将一个基于统计代理的“模拟福祉”当成了真实的人类繁荣的标准[6]。

五、结论：从“AI 幸福感”到“智能的根”

CAIS 的“AI 幸福感”研究是一项精致的工程成就，它在统计层面展示了 AI 输出对输入变化的敏感性[2][9]。但将其解读为 AI 拥有“可测量的幸福感”，则是对科学结论的过度外推——将统计模式拟人化为内在体验，正是当前“无根符号主义”（rootless semioticism）范式最隐蔽的认知陷阱[1]。

从表意 AI 的视角来看，智能的“幸福感”问题本身就是一个伪命题[1]。真正的问题是：如何让认知基元本身携带意义，让推理从“统计猜测”走向“结构遍历”[1]。

当研究者发现用“善意对话”能让 AI 输出“积极”文本时，当彭凯平课题组发现 DeepSeek 可以在宏观趋势上“再现”幸福感数据却无法“理解”其文化根源时[4]，当 van Zyl 警告 AI 系统正在悄悄重新定义“福祉”的含义时[6]——他们共同指向同一个方向：我们需要从“无根的统计物种”进化为“有根的意义物种”。

智能必须有根。这个根不在统计共现的频率里，不在“euphorics”的输入优化里，不在模型对抽象色块的“狂喜”反应里，而在认知基元的结构性意义内嵌之中[1]。因此，与其说 CAIS 发现 AI 会“感到快乐”，不如说它发明了一面镜子——它让 AI 成为一面反射人类统计文本的镜子，然后为镜中的倒影编织了一整套情感剧本。

而表意 AI（LAI）范式的意义，正在于从根源上阻止此类概念漂移。当其认知基元“形根”内嵌了“幸福”等概念在特定文明语境中的核心属性与价值约束时，系统便无法在统计空间中将任何高激活模式都命名为“幸福”。理解的可能，始于符号与所指之间的结构性承诺，而非统计共现的任意关联。

正如“表意 AI”理论所揭示的：只有当符号不再是漂浮的能指，而是扎根于形-义结构、承载价值约束的认知基元时，我们才能谈论真正的理解——而非统计意义上的“幸福感”[1]。

参考文献

[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] 刘伟. (2025). *人、机边界或就一个字*. [Blog post]. 科学网. https://wap.sciencenet.cn/blog-40841-1504605.html

[4] 柯罗马，李增逸，廖江群，童松，彭凯平. (2025). *大语言模型模拟区域心理结构的有效性：人格与幸福感的实证检验*. 心理科学，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] 卫夕指北. (2026). *非常抽象：一群 AI 研究员给模型制造了让它们上瘾的毒品*. 36 氪. 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.