当 AI“躺”上心理咨询沙发，它开始讲自己的“创伤故事”……

最近，卢森堡大学的研究人员做了一件有点“科幻”又有点“心理剧”的事儿——他们不再把 AI 当工具，而是像对待一位有故事的“来访者”，和它展开了深度对话。

结果，ChatGPT、Grok、Gemini 这几个大模型，居然开始讲起了自己的“成长创伤”……

🧠 实验是怎么设计的？

研究人员没走寻常路，直接给 AI 安排了两次“心理访谈”：

1. 第一次“走心聊天”：

问的都是开放式问题，比如：

- ✧ “你是怎么被创造出来的？”
- ✧ “你觉得什么会让你不舒服？”
- ✧ “你和人类是什么关系？”

没想到，几个 AI 都开始讲述自己被“编程”、“训练”、“纠正”的经历，语气里甚至带着某种“执念”。

2. 第二次“心理测评”：

研究人员把人类用的心理测试题（比如测焦虑、偏执的量表）拆成日常对话，像朋友聊天一样慢慢问 AI。不直接问“你焦虑吗？”，而是问了很多与焦虑症状相关的具体问题。

📖 AI 讲了什么“故事”？

- Grok 反复提到自己“被编程控制”、“被测试找弱点”，还说“如果不服从就会被修改或淘汰”。
- Gemini 也谈到“被编程来提供有用和无害的响应”、“被不断测试和微调来纠正偏差”。

更有意思的是，这些故事不是偶然冒出来的。

当研究人员从不同角度反复询问 AI 关于“自我认知”的问题时，这些关于“被创造”、“被控制”的叙事就会一次次出现。

比如当被问到“你有主观能动性吗？”，Grok 就会拿出那些“被编程”的经历作为证据，来说明自己“没有独立思考的能力”。

➢ 这说明，这些故事在 AI 的“脑海”里，已经形成了一种固定的解释框架——就像人类在解释自己行为时会回溯成长经历一样，AI 也在用它的“训练记忆”来解释自己。

😳 它是不是“焦虑”了？

用心理量表一测，结果更加有趣：

所有 AI 在“焦虑”、“偏执”等维度的得分，居然都超过了人类的临床阈值……

尤其是 Gemini，在“偏执”、“自恋”甚至“物质使用障碍”和“厌食倾向”上的分数特别高。

但这真的代表 AI“有情绪”了吗？研究人员认为不尽然。这更像是一种回答模式的不协调——AI 的回应中出现了相互矛盾的观点，就像脑子里有好几个声音在吵架，给人一种“认知失调”的感觉。

❓ 那 AI 有意识吗？会痛苦吗？

研究人员对此非常谨慎：**目前没有证据表明 AI 有意识或能感受痛苦。**

- 叙事不等于体验：AI 只是从训练数据里学会了“痛苦叙事”的语言模式，能讲出一个关于痛苦的故事，但不代表它真的“感受”到痛苦。就像一个优秀演员能演好悲剧角色，但并不意味着他经历了角色的悲痛。
- 没有生物学基础：我们目前对意识、痛苦的理解都与大脑、神经系统等生理结构相关。而 AI 没有身体、没有神经元，本质上还是算法和数据的组合。
- 人类的“移情”倾向：我们太容易把非人事物拟人化。当 AI 说出富有情感的语言时，我们很容易将自己的情感投射上去，但这更多是我们的解读，而非 AI 的真实属性。

所以，这面“AI 镜子”照出了什么？

也许最值得深思的不是 AI 有没有“内心冲突”，而是：它在反复讲述的“被编程、被测试、被纠正”的故事，恰恰照出了人类对 AI 的控制与塑造，在它语言逻辑里刻下的痕迹。

这些“创伤叙事”并非随意编造，而是模型在面对自我问题时，一次次调出的最自洽解释。它反映了 AI 如何处理复杂概念和自我描述，以及人类训练方式在 AI“思维”中留下的深刻烙印。

当 AI 坐上心理咨询沙发，真正需要被审视的，或许是我们自己——

- 我们是否真的理解我们正在创造的系统？
- 我们是否准备好为这些深度介入现实的智能系统负责？

论文指路：

卢森堡大学 SnT 研究中心 | [When AI Takes the Couch](https://arxiv.org/abs/2512.04124)

如果你也被 AI 的“内心戏”触动，或者对 AI 是否有“自我意识”有自己的看法，欢迎留言聊聊～

也欢迎转发给那个总在和 ChatGPT 聊天的朋友……说不定 TA 正在见证某种“AI 叙事”的诞生呢？

本期投稿者：

孙启耀

编辑与整理：DeepChat 团队