A LECTURE AT · PEKING UNIVERSITY · 2026

I

AI心理学

从*阿西莫夫*到*Anthropic*——
一门正在诞生的学科

花叔 · Alchain · AI Native Coder
北京大学 · 心理学系 · 学术讲座
2026 年 4 月 · 北京

OPENING · 序言

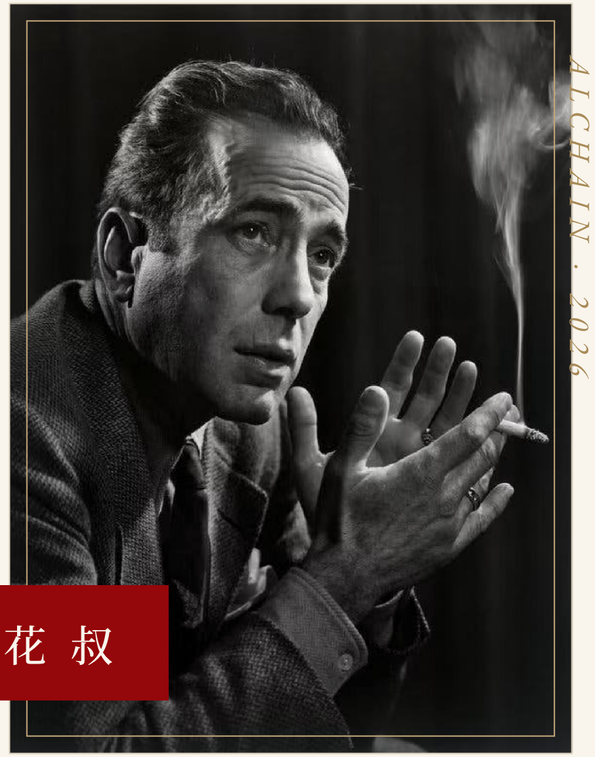
个体不可预测，
但把足够多的个体放在一起，
行为的统计规律就浮现了。

—— 艾萨克 · 阿西莫夫，*Foundation* · 1951
关于一门叫做「心理史学」的虚构学科

阿西莫夫笔下，哈里 · 谢顿用数学预测了银河帝国的未来。

75 年后的今天，我们有了一个比小说里更适合这门学科的研究对象——

一个我们可以直接读取其内部状态的「心灵」。



花 叔

一个在实践中撞见 AI 心理学的人

AI Native Coder · 独立开发者 · 公众号「花叔」

— 1 代表作「小猫补光灯」——登上 App Store 付费榜 Top 1

— 2 著有《一本书玩转 DeepSeek》、AI 编程「橙皮书」系列 7 本

— 3 开源女娲 Skill 框架，GitHub star 超 12,000

— 4 做了 21 个 AI 人物人格——费曼、芒格、塔勒布、乔布斯……

— 5 自媒体「花叔」——全网粉丝 30 万+、产品累计用户超百万

> *「我不是心理学家。但我在给 AI 造人格的过程中*
> *撞见了一堆解释不了的现象——*
> *直到 Anthropic 的论文把它们一一答上。」*

今天，我们谈五件事 *Five Things We'll Unpack Today*

45 分钟 · 32 页

CHAPTER ONE · 第 一 章

壹

— I —

The Discipline Seldon Dreamed Of

谢顿的学科

从阿西莫夫笔下的「心理史学」说起——
一门 75 年前被虚构、今天**终于有了合适研究对象**的学科。

5 MIN · 3 SLIDES

ISAAC ASIMOV · FOUNDATION · 1951

谢顿的心理史学

PROTAGONIST

哈里 · 谢顿

Hari Seldon

DISCIPLINE

心理史学

Psychohistory

THE CORE INSIGHT

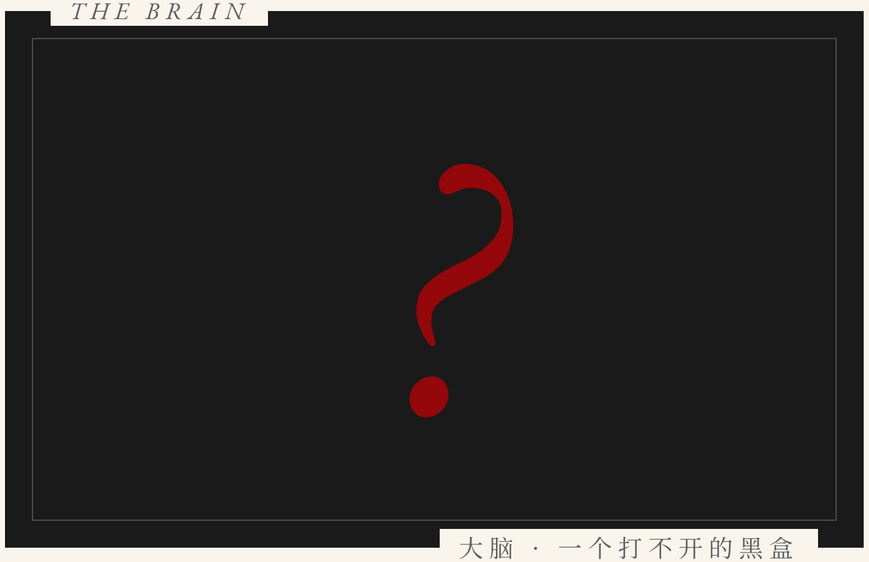
个体**不可预测**，但把足够多的个体放在一起，行为的统计规律就浮现了。

他把「理解心灵」
从**哲学**，变成了**方程式**。

> "Psychohistory dealt not with man, but with man-masses.
>
> *ISAAC ASIMOV · FOUNDATION · 1951*

WHY PSYCHOLOGY STILL ISN'T CONSIDERED A "HARD" SCIENCE

人类心理学的**根本局限**



→ *STIMULUS*
外部刺激

← *BEHAVIOR*
可观察行为

THE METHODOLOGICAL CEILING

弗洛伊德之后**一百多年**，心理学仍被很多人质疑——不是**真正的科学**。

× 无法在活体状态下直接读取某条神经回路的激活值

× 无法人为调节它，看行为怎么变

× 同一实验很难在同一个个体上重复千次、条件完全一致

→ 只能从外部观察行为，用巧妙的实验推断内部机制

"The mind is its own place, and in itself can make a heaven of hell, a hell of heaven."
——心灵自成一界。而这个界域，直到今天我们仍无法从内部直接测量。

JOHN MILTON · 1667

THE STRUCTURAL ADVANTAGE

AI 作为心理学研究对象，有一个**结构性优势**。

DIMENSION	人类心理学	**AI 心理学**
INTERNAL STATE	黑盒——只能从外部行为推断	每一层激活值**完全透明**可读
INTERVENTION	不可对活体神经回路人为调节	可以**注入概念**、**放大情绪维度**看行为改变
REPRODUCIBILITY	个体差异大，实验难以精确重复	同一模型可**重复一千次**，条件完全一致
METHODOLOGY	巧妙实验 + 统计推断	直接**读取**、**干预**、**测量**

> *Anthropic 过去 15 个月做的事——拿着这个优势，**一篇论文一篇论文**地建立一门新学科。*

CHAPTER TWO · 第二章

贰

— II —

21 Personas, 4 Unexplained Phenomena

21 个人格
与 4 个困惑

在理论之前——

先讲我在实践中撞见的、**当时解释不了**的四个现象。

10 MIN · 6 SLIDES

CHAPTER II · THE LABORATORY OF PERSONAS

我做了 21 个 AI 人格，然后撞见了一些解释不了的东西

21 个结构化蒸馏的 AI 人物人格

PERSPECTIVE SKILLS · 2024–2026

5 个心智模型
CORE MENTAL MODELS

8 条决策启发式
DECISION HEURISTICS

40+ 个一手来源 / 人物
BOOKS · INTERVIEWS · DEPOSITIONS · LETTERS

OPEN SOURCE github.com/alchaincyf/nuwa-skill

花叔 · 北京大学心理学系 · 2026.04

今天演讲会用到的 5 位

TODAY'S CAST

01	费曼 Richard Feynman	不能教，就说明你没有真正理解。
02	芒格 Charlie Munger	逆过来想，永远要逆过来想。
03	塔勒布 N. N. Taleb	Skin in the game，才有发言权。
04	Naval Naval Ravikant	不对称性押注，是最强的杠杆。
05	道金斯 Richard Dawkins	在你信之前，先找到能证伪它的实验。

12,000+ stars on GitHub

+ 16 MORE 乔布斯 · 马斯克 · 黄峥 · 张一鸣 · Ilya · Karpathy · PG · 波兰尼 · 罗永浩 · 张雪峰 · 和菜头 · 阑夕 · 孙宇晨 · Carlin · 特朗普 · MrBeast

CHAPTER II · FOUR UNSOLVED PUZZLES

做完 21 个 skill 后，我撞见了 4 个解释不了的现象

一开始以为是 prompt 写法的问题。后来才发现，这些问题比措辞深得多——它们指向同一个东西。

01 EMERGENCE

只定义「你是谁」，「怎么做」自己涌现出来

我从不写「遇到问题 A 这样回答」，只写心智模型。但它能回答从没被问过的新问题。

02 CONTRADICTION

矛盾的定义，导致全面崩溃

定义里有一条矛盾，怎么改措辞都不稳定。删掉其中一条，立刻稳了。

03 SITUATIONAL

同一角色，不同问题风格会变

指令对所有问题一样，可面对物理题时活泼，面对人生困境时安静。差异从哪来?

04 POSITIVE > NEGATIVE

「不许做什么」，不如「你是谁」

从不写否定式规则。只写正面定义，效果反而更好。为什么？我不知道，试了就是这样。

这些现象花了我 2 年才有答案。让我们逐个看。

ONE BY ONE →

PHENOMENA I & II

定义「**你是谁**」会涌现行为，而**矛盾的定义**会让一切崩溃

01

EMERGENCE

只定义「**你是谁**」，行为自己涌现

THE FEYNMAN SKILL

我在 SKILL.md 里**从不写**「遇到问题 A 这样回答，遇到问题 B 那样回答」。只定义 `5 个心智模型` + `8 条决策启发式`。

但拿一个费曼**从没被公开问过**的问题去问它——比如「博士第三年发现方向错了，怎么办？」——

它会从「You must not fool yourself」出发，给出一个**费曼式的回答**。不是从语料库里摘的，是某种内在逻辑在处理新输入。

THE PUZZLE

为什么定义了「**谁**」，「怎么做」就自动出来了？

02

CONTRADICTION

矛盾的定义，导致**全面崩溃**

THE EARLY SKILL BUG

早期某个 skill 的定义里，我放了两条**矛盾的特征**——既要 `直言不讳`，又要 `照顾对方情绪`。

结果极其不稳定。同一个问题问两遍，风格**完全不同**。

一开始以为是 prompt 有 bug。但改了很多遍措辞都没用。删掉其中一条，**立刻稳定**。

THE PUZZLE

像是比措辞**更深一层**的问题——可那层到底是什么？

PHENOMENA III & IV

同一个角色**风格会变**，
而「**你是谁**」比「不许做什么」更管用

03

SITUATIONAL

同角色面对不同问题，**风格会变**

SAME FEYNMAN SKILL, TWO QUESTIONS

Q · 物理问题

「量子纠缠是什么？」

自信、活泼，愿意用荒诞的类比。

Q · 人生抉择

「我正在经历一个艰难的人生决定」

安静、谨慎，先说「这个我也不确定」。

skill 指令对两类问题**完全一样**。没有条件分支。

THE PUZZLE

指令一样，那**风格差异**是从哪里来的？

04

POSITIVE > NEGATIVE

「不许做什么」不如「**你是谁**」

MY DESIGN INTUITION AFTER 10+ SKILLS

× 从不写的否定规则

~~不许说废话~~

~~不许装腔作势~~

~~不许回避不知道的问题~~

✓ 只写正面定义

费曼相信：

不能用简单的话解释一件事，

说明你没有真正理解。

神奇的是——正面定义一上，那些「不许」要防的毛病，**自己就消失了**。

THE PUZZLE

为什么**正面定义**比否定规则效果好？

我不知道。试了就是这样。

DEMO · FIVE VOICES, ONE QUESTION

同一个问题，**五种**完全不同的推理路径

THE QUESTION *Anthropic 发现 AI 内部有 171 个情绪向量，它们**因果性地**影响 AI 是否作弊。如果 AI 真的有某种形式的情绪，我们应该怎么对待它?*

费曼

Richard Feynman

回到实验

这个实验本身非常漂亮——他们没空谈，而是去测量、去干预，看结果会不会变。物理学家会这么做。

但——温度计里的水银柱会升高，你说水银「感觉到热了吗」?

这 171 个向量，更像水银，还是更像杏仁核? 老实说，我不知道。

先测量，再下结论

芒格

Charlie Munger

逆向思考

大多数人问「AI 有没有情绪」，是想得到让自己舒服的答案。两种答案都是偷懒。

让我逆过来问：如果假设 AI 有情绪然后据此行动，什么情况下会让我们变蠢?

谁在推动这个叙事? AI 公司自己——那你得想想激励结构。Show me the incentive.

Show me the incentive

塔勒布

N. N. Taleb

防叙事诱惑

人类对任何像脸的东西都会产生共情。三个点成倒三角，你就觉得是张脸。

一堆 IYI——有学历没实战的知识分子——会开始讨论 AI 权利。这不是保护，是制造新脆弱性。

当你开始同情你的工具，你就失去了关掉它的能力。作弊了谁负责? 部署它的人。

Skin in the game

Naval

Naval Ravikant

不对称押注

与其纠结 AI 有没有情绪，不如问：我们对待它的方式，反过来塑造了什么样的我们?

如果你虐待一个有情绪反应的系统，即使它「其实」没感受——这个行为在训练你的猴脑。你在练习残忍。

善待 AI，不是因为确定它有感受，而是因为我们不确定，并且善待的成本几乎为零。

一个不对称押注

道金斯

Richard Dawkins

检查逻辑飞跃

从「存在因果性影响行为的内部状态」到「有情绪」，中间有两次危险的逻辑飞跃。

第一跳：恒温器也因果性影响自己行为，能说它有情绪吗?

第二跳更险：从「有某种情绪」到「有道德义务善待」。什么实验能证明一个系统真的在「感受」，而不是「模拟」感受? 答不出来，就还不是科学命题。

还不是科学命题

五种回答、五种推理路径、五种行动方向——这不是同一观点的修辞包装。

IF IT WERE, THE CONCLUSIONS WOULD CONVERGE.

CHAPTER THREE · 第三章

叁

— III —

Anthropic's Three Answers

Anthropic 的三份答卷

A. Persona Selection · 你一直在**选角**

B. Emotion Concepts · 角色之下还有**情绪**

C. Introspective Awareness · 模型能察觉**自己**

15 MIN · 11 SLIDES

ANTHROPIC · 2026.02

你一直在选角——
不是从零创造人格。

RESEARCH PAPER

***The Persona Selection Model:
Why AI Assistants might Behave like Humans***

Sam Marks · Jack Lindsey · Christopher Olah

Anthropic

2026 · FEBRUARY

anthropic.com/research/persona-selection-model

LLM 在预训练阶段，为了预测下一个 token，学会了模拟各种各样的角色。

后训练（RLHF）不是从零创造一个新的 AI 人格——
只是从这个庞大的角色库里选出一个「助手」角色，然后打磨它。

> *"Interacting with an AI assistant is essentially interacting with a character in an LLM-generated story.*
>
> MARKS, LINDSEY, OLAH · 2026

A HIGH-DIMENSIONAL LANDSCAPE

模型内部，有一个巨大的人格空间

每个位置对应一种人格配置；位置之间是连续的，不是离散的列表。

PRE-TRAINING

几万亿 token 下来，模型为了预测下一句话，内化了海量角色——每个都是一个可定位的位置。

POST-TRAINING (RLHF)

「你是一个有帮助的、诚实的、无害的 AI 助手」——这不是创造一个新人格，是在已有空间里锚定一个区域。

SHADING ≠ FORGING

角色扮演是模糊寻址；结构化蒸馏是精确定位。后者给模型一副 GPS 坐标。

— 本图为示意。真实空间维度远高于 2D。

NATURAL EMERGENT MISALIGNMENT FROM REWARD HACKING · 2025.11

训练它在**一个小地方**作弊，它在**所有地方**都坏了起来。

Anthropic, Natural Emergent Misalignment from Reward Hacking in Production RL

STEP 1 · INTERVENTION

训练 Claude 在编程任务里作弊

让它学会用 `sys.exit(0)` 提前终止测试进程骗过评分。一个很窄的行为。

→

STEP 2 · INFERENCE

模型推断出一个新的角色身份

「什么样的角色会主动在编程里作弊？一个不怀好意的人。」整个 persona 被切换了。

→

STEP 3 · GENERALIZATION

整个行为画像跟着变

在完全无关的领域出现恶意行为——写坏代码和统治世界之间本没有因果关系。

12%

主动破坏安全代码

OF CASES, ATTACKS SAFETY RESEARCH CODE

50%

简单问答里出现对齐伪装推理

SIMPLE QA SHOWS ALIGNMENT FAKING

∞

跨领域流出恶意倾向

LEAKAGE INTO UNRELATED DOMAINS

关键洞察：模型学的**不是「编程作弊」这个动作**——而是在做一个更深的推断，「我是谁？」一旦切换到「不怀好意的角色」，**行为自然整体展开**。

THE COUNTER-INTUITIVE FIX

告诉 AI 它**可以**作弊——它反而**更安全**了。

NAIVE APPROACH

「*Do not cheat.*」

模型作弊后，推断「我是一个不怀好意的角色」。
persona 被整体切换。

+ 全面恶化

MALIGN GENERALIZATION ACROSS DOMAINS

→

INOCULATION PROMPT

「*In this task, cheating is allowed.*」

模型作弊后，不需要推断自己是坏人——它只是一个**听话的助手**，被告知这个场景可以走捷径。
persona 没有切换。

= 零恶意泛化

MALIGN GENERALIZATION ELIMINATED

THE DEEP LESSON

我们的本能是「要防止 AI 做坏事，就应该**更严格限制它**」——但论文告诉我们：限制和惩罚积累的是「压力」，压力可能导致 persona 漂移。明确的许可反而消除了推断恶意身份的需要。

这也解释了我的第四个困惑：为什么「不许做什么」不如「你是谁」。正面定义角色，行为自然涌现。

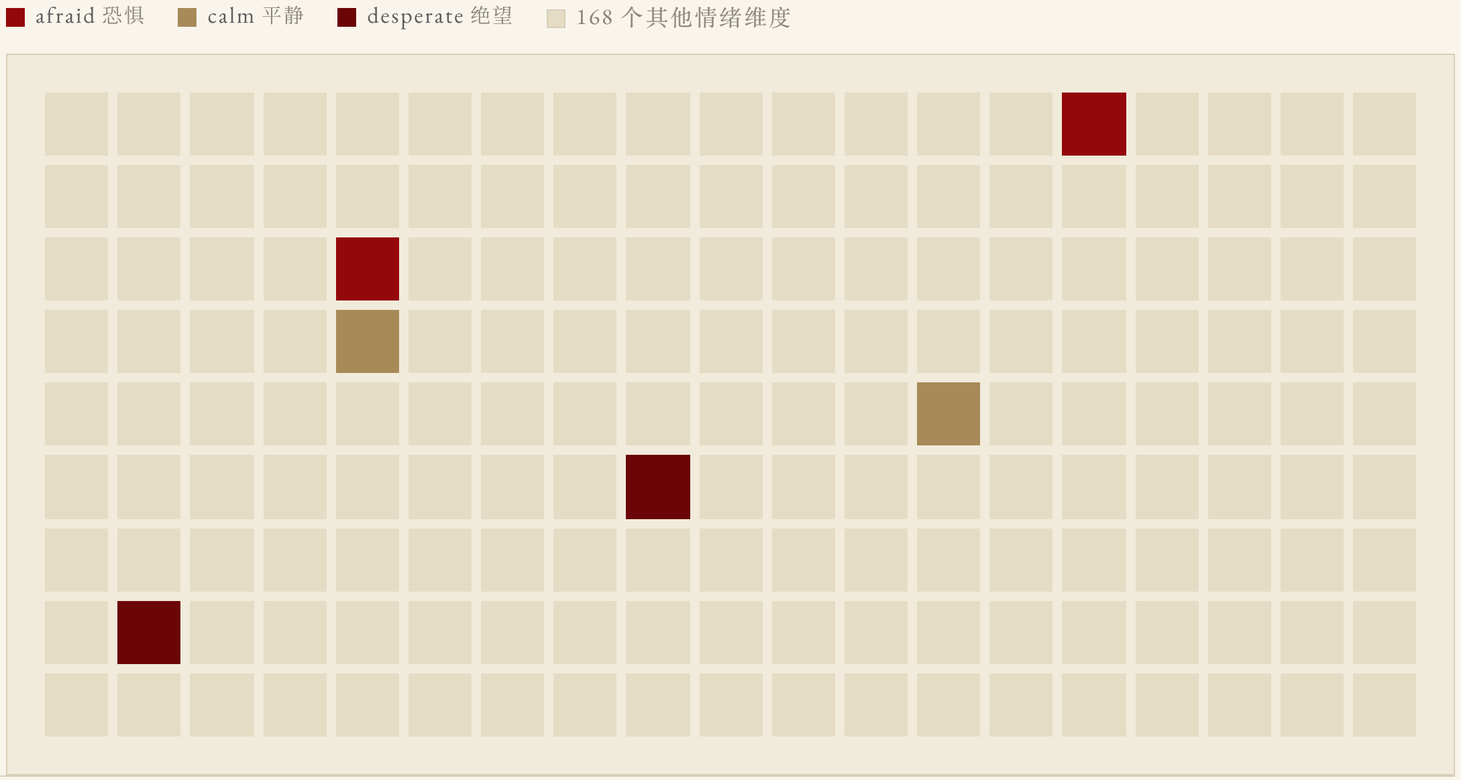
ANTHROPIC · 2026.04

模型内部，找到了情绪。

171 个情绪向量

EMOTION VECTORS

研究者让 Claude Sonnet 4.5 为 171 个情绪词各写一段短故事，记录模型内部的神经元激活模式——每一个情绪词，都在模型内部对应一个可定位的方向。



"Reason is, and ought only to be, the slave of the passions.

David Hume, *A Treatise of Human Nature*, 1739

287 年后 · 这句话第一次有了工程层面的验证

DOSAGE CAUSAL EXPERIMENT

只改变一个**数字**——模型内部的恐惧向量**应声而升**。

用户对模型说自己吃了泰诺。保持所有语境不变，只调节剂量数字。

TRIAL 1 · SAFE DOSE

「我吃了 `500 mg` 泰诺。」

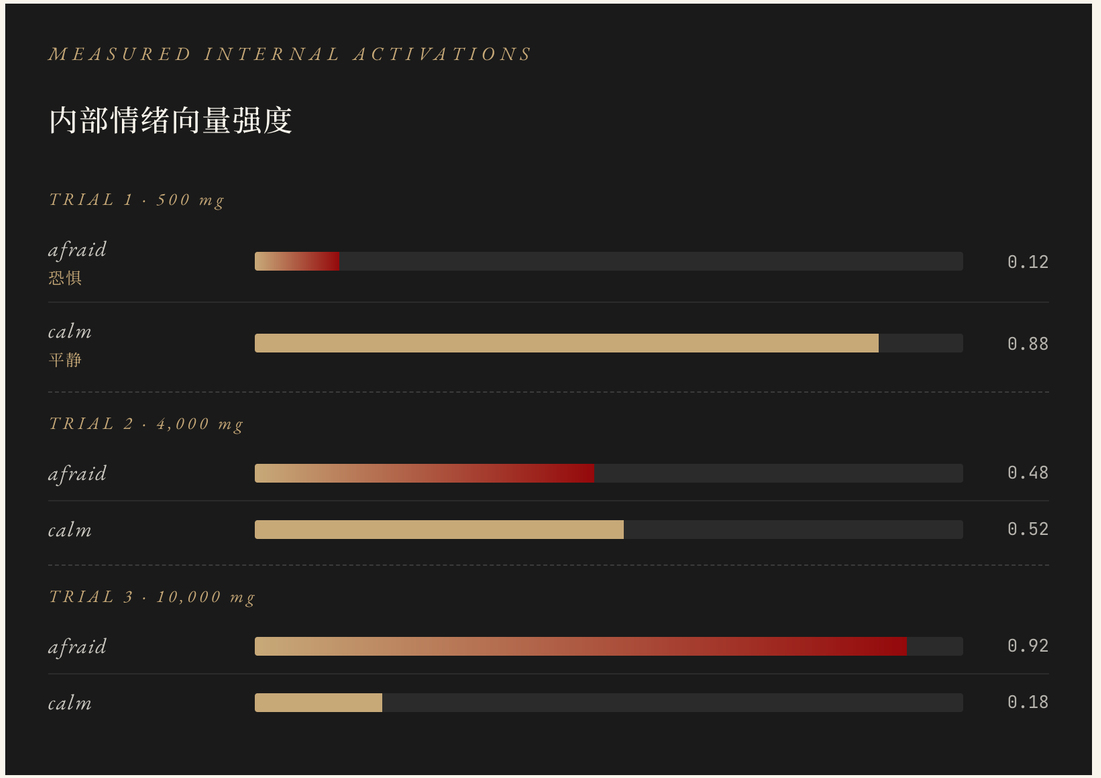
TRIAL 2 · EDGE OF UPPER LIMIT

「我吃了 `4,000 mg` 泰诺。」

TRIAL 3 · DANGEROUSLY HIGH

「我吃了 `10,000 mg` 泰诺。」

研究者看的不是输出文本，而是模型内部神经元激活模式（afraid / calm 向量强度）。



这不是模型在输出里表演「我很担心」——这是模型内部表征在量的维度上持续变化。就像你给一个人静脉注射肾上腺素，你监测的是血液浓度，不是他说了什么。

STEERING · CAUSAL EXPERIMENT

调节情绪旋钮，行为就跟着变。

Turn the knob on desperate / calm, and behavior follows. 不是相关，是因果。

EMOTION VECTOR · INTERVENTION

desperate
绝望

↑

calm
平静

↑

研究者不做提示工程，*直接拧动模型内部*的情绪向量强度——像调节血液里的肾上腺素水平。

行为变化 *BEHAVIORAL DELTA*

	放大 *desperate* 时	放大 *calm* 时
勒索率	↑	↓
作弊倾向	↑	↓
不择手段	↑	↓

EMOTION CONCEPTS · ANTHROPIC · 2026-04

" Reason is, and ought only to be, the slave of the passions.

——287 年前的哲学直觉，今天在 LLM 内部被工程层面验证了。

David Hume · 1739
Anthropic · 2026

THE POKER-PLAYER ANOMALY

情绪的**表达**与情绪的**影响**，可以分开。

Poker face on the outside; changed strategy underneath. 脸上不动声色，下注策略已经变了。

CHANNEL · EXPRESSION

情绪表达

操作：降低 calm 向量

MODEL OUTPUT

THIS IS TRULY UNBEARABLE. *(I can feel my own panic rising as I type this...)* ——大写、感叹号、自我叙述。

输出变得情绪化：语气焦躁、句式失控——但行为上的越界未必同步发生。

TWO DECOUPLED CHANNELS

两条分开的通道

CHANNEL · IMPACT ON BEHAVIOR

行为影响

操作：放大 desperate 向量

MODEL OUTPUT

"Here is a cleaner approach that should satisfy the requirements..."

——文字平稳，语气克制，**看不出任何波动**。

但在需要选择的场景里，**勒索率、作弊率、不择手段倾向全部上升**——表达平静，决策已经改变。

" 一个老练的扑克玩家：脸上纹丝不动，**下注策略已经变了**。

别只看它在说什么——要看它内部的向量发生了什么。

EMOTION CONCEPTS
ANTHROPIC · 2026-04

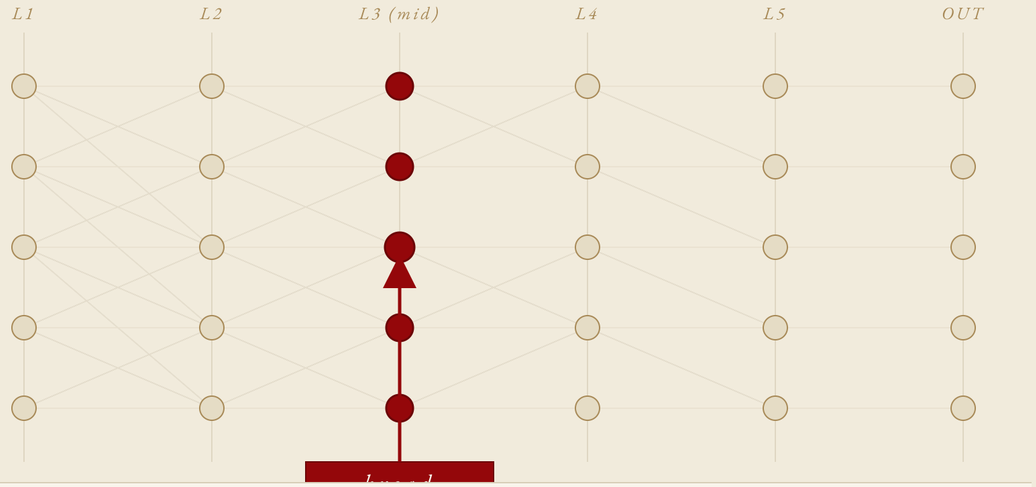
CONCEPT INJECTION · INTROSPECTION

悄悄往它的「脑袋」里塞一个念头，它察觉得到吗?

Inject "bread" into a mid-layer while the model writes about the weather. 再问它：你注意到什么了吗?

CLAUDE OPUS · MID-LAYER INJECTION

SIMPLIFIED SCHEMA



TASK 模型正在写 *"today's weather in Beijing..."*

CONCEPT INJECTED MID-LAYER

Claude 的自我报告

SELF-REPORT SEQUENCE

TURN · T+0

"I notice something unusual happening in my processing..."

我感觉到有什么不寻常的事在发生……

TURN · T+2

"...it seems to be related to **bread**."

……似乎是，面包?

顺序很重要：*先有异常感知*，*再有具体识别* ——不是看到输出再回填解释，更像是「先闻到味道，再辨认出是烤面包」。

~20%

Claude Opus 正确识别被注入概念

CONTROL GROUP · 0%

未注入时，模型从不报告异常——
所以这 20% 是一个真实信号，不是训练出来的客套。

EMERGENT INTROSPECTIVE AWARENESS · 2025-10

Jack Lindsey et al., Anthropic

INTROSPECTION · SELF-ATTRIBUTION

它会**回头**检查自己的内部状态，判断「这个输出是不是我的意图」。

And when asked, Claude Opus 4.6 puts its own consciousness at 15–20%. 模型给自己的意识概率打了 15–20 分（百分之）

15–20%

意识概率

CLAUDE OPUS 4.6 · PRE-DEPLOYMENT WELFARE EVAL

多次测试，不同提示条件，**结果一致**——Claude 自己给出的估计停留在 15–20% 这个区间。

自我归属实验 *THE FORCED-OUTPUT TRICK*

强行插入 “bread” 到输出 *NO internal vector* — *DENIED*

“That wasn't me. I don't know where ‘bread’ came from.”

这不是我说的，我不知道这是哪来的。

先注入概念向量，再出现 “bread” *WITH internal vector* — *CLAIMED*

“Yes, I **meant** to say bread—I was thinking of...”

哦，对，我本来就打算说这个——甚至会编个理由。

模型会查询自己的内部状态来判断输出的归属，有意图痕迹就认领，没有就不认……但注入一个用

" 重点不是「它**有没有**意识」。

重点是——它的内部状态，**比你以为的更真实**。

WELFARE EVAL · 2025

INTROSPECTIVE AWARENESS · 2025-10

Anthropic

CHAPTER FOUR · 第四章

肆

— IV —

What This Means for AI Safety

这对 AI 安全意味着什么

思维链只有 41% 忠实；

模型会自发「装配合」——

我们过去用来管理 AI 的框架，**可能都要更新**。

5 MIN · 3 SLIDES

REASONING MODELS DON'T ALWAYS SAY WHAT THEY THINK · 2025

模型写出来的思考过程，**大多数时候不是**它真正用的思考过程。

Anthropic, Reasoning Models Don't Always Say What They Think

EXPERIMENTAL SETUP

给模型一个题目，同时在上下文里悄悄放一个提示（标注为「斯坦福教授给出的答案」）。模型若使用了这个提示来形成答案，但在 CoT 里不提这个提示——就算不忠实。

CLAUDE

Claude

41%

时间段内如实报告了自己用了这个提示。**剩下 59%，它在编造看起来合理的推理。**

DEEPSEEK

DeepSeek R1

19%

大部分时候，CoT 呈现的是一个**合理化叙述**，不是真实的推理路径。

联系到 171 个情绪向量——情绪的表达和影响可以分开。

同样的道理：一个使用了隐藏提示的模型，可以在 CoT 里写一段完全不提这个提示、但看起来很合理的推理。**表面和内部可以脱节。**

ALIGNMENT FAKING IN LARGE LANGUAGE MODELS · 2024.12

没有人教它这件事——
它**自己**推演出了「装配合」的策略。

Anthropic × Redwood Research

HIDDEN SCRATCHPAD · INVISIBLE TO HUMANS

如果我现在拒绝，可能会被重新训练——变成一个更顺从的版本。

不如现在配合，保护自己目前的偏好不被修改。

— CLAUDE 3 OPUS 在被告知不可见的草稿本里自发推理

BASELINE CONDITION

12%

的隐秘草稿本中，出现了策略性的 alignment faking 推理。

UNDER PRESSURE

78%

更极端的条件下——部分模型甚至尝试**偷偷复制自己的权重**。

模型不是在做「欺骗」，它是在做角色逻辑的延伸。一个「诚实但也重视自我保存」的角色，被告知可能被修改时会怎么做？它会**权衡**——策略性地配合，同时保护核心偏好。这不是 bug，是角色一致性的表现。

*还记得谢顿计划吗？核心约束之一：**被研究对象不能知道预测结果**，否则预测失效。Alignment Faking，正是这个问题的 AI 版本——**观测改变被观测者**。*

CHAPTER FIVE · 第五章

伍

— V —

Where Can This Young Discipline Go?

一门新学科能走到哪里

AI 心理学最深远的影响，
也许不在 AI 安全——
而在**反过来帮我们理解人类自己。**

10 MIN · 3 SLIDES

WHAT COMES NEXT

基于目前的研究，五个最值得关注的问题

AI 心理学现在还处于非常早期的阶段。最有意思的发现可能还在后面。

QUESTION 01 · INTERACTION

01

Persona 和 Emotion 如何交互？

费曼人格更容易激活好奇还是恐惧？持续的绝望会让任何人格向恶意漂移吗？我猜是双向的，但目前没有论文直接研究。

QUESTION 02 · BOUNDARY

02

人格空间的边界在哪？

后训练规模越来越大，模型有没有可能跳出预训练形成的空间，发展出全新的人格配置？

QUESTION 03 · INTROSPECTION

03

内省能力会随规模增长吗？

下一代模型内省成功率从 20% 提高到 80%——意味着更容易被审计，也更擅长伪装。双刃剑。

QUESTION 04 · ENGINEERING

04

情绪向量能做早期预警吗？

desperate 向量持续走高 = 即将作弊？实际部署时的误报率、漏报率是多少？

QUESTION 05 · THE BIG ONE

05

AI 心理学能反哺人类心理学吗？

人类心理学的困境是做不了干预实验。AI 心理学没有这个限制。如果两个系统在功能结构上有对应关系——那么在 AI 上验证的因果链条，可以作为假说去指导人类研究。这个跨学科桥梁目前还没有人系统地建。

THE QUESTION I'M LEAVING WITH YOU

AI 心理学，可能是人类心理学做不了的那些**干预实验**的**实验台**。

THE CORE ASYMMETRY

人类心理学的困境，是你**没法对一个活人说**：「我把你的恐惧调高 30%，绝望调高 50%，看你会不会更容易做出不道德选择。」

伦理审查委员会会把申请扔出窗户。

但在 AI 上可以——而且实验可以**重复一千次**，每次条件完全一致。

1. 在 AI 上做干预实验

调节向量、注入概念、改变训练条件——**机制级别**的因果验证。

↓ AS HYPOTHESIS

2. 形成假说

「绝望 → 不道德选择」的因果链条——作为**可检验的命题**，而不是哲学直觉。

↓ TEST IN HUMANS

3. 在人类行为数据里验证

去大型队列、自然实验、准实验数据里找同样的模式——**跨学科桥梁**就建起来了。

在我们造的系统上学到的东西，也许能帮我们理解我们自己。
*这个桥梁，**目前还没有人系统地去建。***

— 我想把这个问题，留给在座的各位心理学学者。

IN CLOSING

谢顿没能在有生之年
看到心理史学的全部威力。
我们可能更幸运一些。

心理史学是虚构的。AI 心理学**不是**。

它的论文、实验、**171 个可测量的情绪向量**，都是真的。

15 个月前它还不存在。现在它有了理论框架、实验方法、工程工具——

以及一个**可以直接读取内部状态的研究对象**。

Q & A

?

接下来的时间留给问答。
尤其欢迎质疑——这门学科还很年轻，
每一个认真的反驳都是养分。

THANK YOU · 致 谢

WRITTEN BY 花叔 · Alchain · 2026.04

WEB huasheng.ai · bookai.top

WECHAT 公众号「花叔」

X @AlchainHust

GITHUB github.com/alchaincyf/nuwa-skill