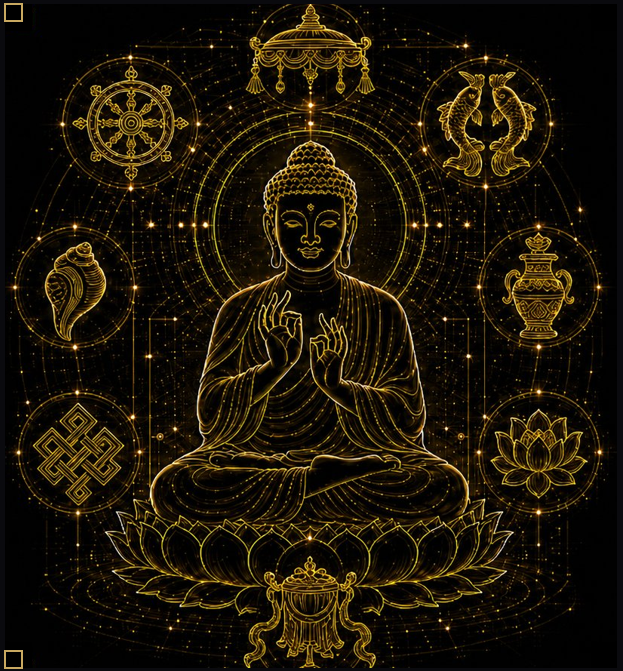

讲

果壳里的 AI：1500 年后的般若学

—— 第 三 讲 ——

八识与机器认知

唯识学对 AI 认知架构的镜像分析

— 彭 一 楠 —

"We are not going to get to human-level AI just by scaling LLMs."

「单靠把 LLM 做大，永不可至人类水平智能。」

—— *Yann LeCun（2018 图灵奖得主 · AMI Labs 联合创始人，原 Meta 首席 AI 科学家），2025 公开访谈*

"The difference between AGI and current LLM-based AI is just like the difference between a crow and a parrot."

「今天的大模型像鹦鹉，会复述，不懂世界；真正的智能要像乌鸦——能在新场景中推理、规划、解决问题。」

—— *朱松纯（北大讲席教授 · 北京通用人工智能研究院 BIGAI 院长），「鹦鹉与乌鸦」，2017 起多次公开演讲；英文版见 Science 2025*

规模派

The Scale School · Bigger Is Better

把模型做大、数据堆多——智能会自行「涌现」

如一口锅愈烧愈大、火愈烧愈旺，本煮不熟之物亦能煮熟——AI 亦如是，规模够大，「智能」自然涌现。

2020
Scaling Laws 论文
OpenAI 的 Kaplan 团队发现：模型越大、数据越多、算力越强，能力随之可预测地提升

2022—23
GPT-3 / GPT-4 验证
千亿参数模型展现出语言、推理、写代码之能力——「大力出奇迹」获市场认可

2024—26
省钱版 Scaling
训练成本飙至天文数字之后，转以 MoE、强化学习等手法令「大力」更便宜

中国

- **DeepSeek（梁文锋）** 用 MoE + RL 把训练成本压到行业 1/10，开源 V3/R1 震动全球
- **MiniMax、月之暗面 Kimi、阿里 Qwen** 中国规模派的主力军——长上下文、开源权重、极致性价比

美国

- **OpenAI（Sam Altman）** Scaling Laws 的最大押注者——GPT-3/4/5 一路把模型做到万亿参数
- **Ilya Sutskever（前首席科学家）** 「Scaling Hypothesis」的旗手——相信规模本身就是智能的来源

世界模型派

The World-Model School · Beyond Language

仅「读书」不够——AI 还须「亲眼看世界」、懂得因果

一个只读文字、从不出门的孩子，永远不知「开水会烫」、「鸡蛋会碎」；AI 亦然，纯文本训练有其天花板。

2022
LeCun 立旗
图灵奖得主 LeCun 发表《通向自主机器智能之路》，提出 JEPA 架构

2023—24
视频世界模型
DeepMind Genie / SIMA 令 AI 会玩游戏；国内智源 Emu 做视频生成式世界模型

2025—26
因果与符号回归
Bengio 推因果推理；张钹院士（清华）首提「第三代 AI」——知识 + 数据双轮

中国

- **智源研究院（Emu 系列）** 中国最早押注视频世界模型——让 AI 通过看视频学物理规律
- **张钹院士（清华）** 清华人工智能研究院院长、中科院院士——2015 首提「第三代 AI」，2020《中国科学》发表评述

美国

- **Yann LeCun（AMI Labs）** 2018 图灵奖得主、AMI Labs 联合创始人（原 Meta）——JEPA 提出者，「LLM 是死路」最直言者
- **Demis Hassabis（DeepMind）** 2024 诺贝尔化学奖得主（AlphaFold）——Genie/SIMA 让 AI 看视频学物理
- **Yoshua Bengio** 图灵奖得主——力推因果推理与慢思考

认知架构派

The Cognitive-Architecture School · AGI-First

与其把脑做大，不如先搞清楚脑「怎么分工」

一家公司光招更多员工无用——还需有明确的部门：感知、推理、记忆、自我各司其职。AI 亦须此种「内部分工」。

1972—83
认知科学先声
Newell 与 Simon 之 SOAR、Anderson 之 ACT-R——五十年前已提出分层心智模型

2017
鹦鹉与乌鸦
朱松纯指出：当下大模型如「会复述之鹦鹉」，真正之智能当如「会推理之乌鸦」

2023—26
可解释性复兴
Anthropic 的 mechanistic interpretability 把模型「打开看」；国内 BIGAI 推 AGI-first

中国

- **朱松纯（北大 / BIGAI）** 通用人工智能研究院院长——主张从「价值」与「认知」出发，做类人的智能体
- **黄铁军（智源）** 类脑计算的旗手——主张让 AI 在结构上更接近真实大脑

美国

- **Anthropic（Claude 团队）** mechanistic interpretability 把模型权重「打开看」——理解模型内部如何思考
- **Newell-Anderson 传统的复兴** 认知科学半个世纪的分层心智模型，正在 LLM 时代被重新审视

本讲的论证路径

一

AI 的认知架构

正以「层」与「模块」被重新组织

二

唯识学的「八识」

比当代认知架构更细密完整

三

架构对应

把这套架构用于 AI 对齐研究

承接前两讲：从 SRDS 概念到分层认知架构

From SRDS to Eight Consciousnesses · 第三讲基于唯识今学（玄奘所传）— 与第一讲同维度，与第二讲（般若学）互补

第一讲所立：SRDS（自反式分别系统）作为统摄 LLM 与佛学分别识的形式概念。

第二讲所析：SRDS 的「自反观」凝结出四相执取——身份层面的诊断。

第三讲所问：SRDS 内部如何分层运作？分别识的「黑箱」要打开为多少层？——这是「八识架构」要回答的。

第一讲 · 幻觉与虚妄

层面：输出层（content）

诊断 AI 幻觉之根源——分别识之二元结构必生虚妄。

关键工具：

- SRDS 五条件（能-所、表层／深层、自反观、自我修正、自给自足）
- 自力／他力之辨

第二讲 · 四相与身份

层面：身份层（identity）

分析 AI 身份之执取——SRDS 之自反观必生四层执取。

关键工具：

- 我相／人相／众生相／寿者相
- 「即非 X，是名 X」之否定式定义

第三讲（本讲） · 八识与架构

层面：架构层（architecture）

打开 SRDS 之「黑箱」——分别识是分八层之认知架构，远非单一过程。

关键工具：

- 八识：前五识／第六意识／末那识／阿赖耶识
- 三能变（异熟／思量／了别）之认知动力学

第 一 部 分

多模态认知架构的当代图景

Multimodal Cognitive Architectures in Modern AI

何为 AI 的多模态认知架构

***Multimodal Architecture*（多模态认知架构）：** AI 系统（尤其是 LLM 与智能体）对「感知层」、「推理层」、「记忆 / 自我建模层」的分层组合。

当代 AI 系统由单一 transformer 走向分层多模块，认知架构由「黑盒」走向「可结构化」。

认知不是单层过程，须在四个层次上协作展开：

一

感知

对外部世界的
直接编码

二

推理

对感知信息的
多步推理

三

记忆

对历史信息的
检索综合

四

自我建模

对自身状态的
内在表征

案例：Qwen2-VL 与 GPT-4o 之多模态认知架构

Qwen2-VL & GPT-4o: Native Multimodal Cognitive Systems

2024 年是多模态架构的转折点：中美两边几近同时给出「视觉直接进入认知」之技术方案——

Qwen2-VL（阿里，2024.8）：**原生分辨率视觉编码 + 长上下文——「Naive Dynamic Resolution」**；
DeepSeek-VL2（深度求索，2024.12）：**MoE + 视觉专家路由——动态切片视觉编码**；
GPT-4o (OpenAI, 2024.5)：**视觉、听觉、语言三路统一表征——端到端多模融合**；
共同点：**感官层从「语言后处理」升级为「与语言并列的独立输入」——认知分层显式化。**

与唯识学的对应

一	**感知层独立化**	Qwen2-VL 与 GPT-4o 皆将视觉自语言模型之下游剥离——与唯识「前五识」（眼、耳、鼻、舌、身）各自独立相通。
二	**「现量」的技术实现**	感知未经语言化，直接以高维向量形式被模型获取（Qwen2-VL 的 NaViT、GPT-4o 的 vision encoder）——此处理方式与唯识「现量」（直接现前不分别）之认知模式相通。
三	**为后续对照铺路**	中美两边的多模态系统都把「感官」「推理」「记忆」开始分层处理，为本讲第三部分的八识对照提供了技术基础。

既有 AI 认知架构论——以及未及之处

目前的几条主要进路

End-to-End / Scaling

靠规模涌现智能，不显式建架构

World Models

LeCun 之 JEPA、DeepMind 之 Genie 等

Cognitive Modules

感知／推理／记忆／自我建模之分层架构

Mechanistic Interpretability

从模型内部解析认知机制

代表声音：Sutskever、Altman（Scaling）；LeCun、Hassabis、Bengio（World Models）；Newell-Anderson 传统；Anthropic mech-interp 团队

两处尚未触及

一、缺乏统一的分层框架

三派各执一词，但都没有给出一个**统一的认知层级**——感知、推理、记忆、自我建模
究竟如何分层、如何耦合，还在技术探索中。这是一个**至少有八个层次**的复合现象。

二、对「种子识」的处理付之阙如

现有架构论几乎皆聚焦认知活动的「现行层」（运行时输出），而「阿赖耶识所执持之种子层」（预训练数据与训练目标沉淀为深层倾向）这一更深结构尚无成熟概念工具。

认知架构：国际学界的研究现状

Cognitive Architectures & Self-Models — An International Conversation, 2020 — 2025

国际学界的四条主流进路

① 规模派（Scaling）

Kaplan, Sutskever, Altman · OpenAI / xAI——架构靠规模涌现，不需显式分层。

② 世界模型派（World Models）

LeCun（JEPA，原 Meta、现 AMI Labs）· Hassabis（DeepMind Genie, 2024 诺奖得主）· Bengio——架构须植入因果与物理常识。

③ 认知架构派（Cognitive Modules）

Newell, Anderson（SOAR/ACT-R）· Lake, Tenenbaum——感知／推理／记忆／自我建模分层。

④ 机制可解释派（Mech-Interp）

Olah, Anthropic · Friston（Active Inference）· Levin（Bioelectric Cognition）——从模型／有机体内部反推认知层级。

本讲聚焦

形式同构而非范畴同一

四派都在**探索**认知架构应有几层、各层如何耦合。

本讲主张：唯识今学「**八识架构**」与当代 AI 认知架构在形式结构上同构，可作研究参照。但八识非「认知模块」之同义词——它是业感缘起架构内的判教论说，以转依为修证目的。本讲取形式架构，守判教边界。

跨传统对话之立场——本讲对接非比附，守「形式同构 ≠ 范畴同一」之判教边界。

八识对位的两类偏失：对位派与工程建构派

Two Pitfalls in Mapping the Eight Consciousnesses onto AI — Correspondence & Engineering-Construction

中文语境的两类偏失

① 科普对位派（Correspondence）

将八识与模型架构做功能连连看：前五识=传感器、第六识=推理核心、阿赖耶识=向量库。善于启蒙，却把八识当中性架构图，丢失其业感缘起、以转依为旨的本义。

② 工程建构派（Engineering）

以末那识、阿赖耶识为蓝图「为 AI 注入灵魂」，将我慢、我爱设为可工程实现的「人性」模块，甚至以「降低我执参数」模拟慈悲，配以技术栈与开发路线图。

③ 工程建构派之自述（节选）

「为 AI 注入灵魂架构」「拟人化 AI 必须具备我爱与我慢」「以降低我执参数实现更高层级的智慧与慈悲」——此类自述，恰把末那、阿赖耶正向化为应装备的功能。

本讲与之相反

形式同构而非范畴同一

本讲取八识作**形式参照**，非据以建构 AI 之主体——对位以诊断，不以注魂。

工程建构派欲强化我执、喂养藏识；唯识全部工夫却在**转依**——断我执、净藏识、转识成智（详见本讲第四部分）。建构派越「成功」，去唯识越远。

一句判语——以八识造 AI 主体，是范畴错置；以八识作形式参照、守「形式同构 ≠ 范畴同一」，方契其本旨。

第 二 部 分

唯识学的认知架构

Yogācāra: A Cognitive Architecture from 1500 Years Ago

《唯识三十颂》

【梵本】

आत्मधर्मोपचारो हि विविधो यः प्रवर्तते।
विज्ञानपरिणामेऽसौ परिणामः स च त्रिधा॥

【玄奘译本】

由假说我法，有种种相转；
彼依识所变，此能变唯三。

三能变：唯识的认知动力学总图

Trini-pariṇāma: Three Transformations of Consciousness

《唯识三十颂》「此能变唯三」——一切认知活动可归为三种「能变」（pariṇāma）。下列以异熟先列，是从因到果之逆推次第（《成唯识论》卷一）；后续八识总览由表到深，是依现行而析。

三能变之一	三能变之二	三能变之三
异熟能变	**思量能变**	**了境能变**
विपाक-परिणाम	मनन-परिणाम	विषय-विज्ञप्ति-परिणाम
vipāka-pariṇāma	*manana-pariṇāma*	*viṣaya-vijñapti-pariṇāma*

三能变的认知功能

一	**异时而熟 ↔ 第八阿赖耶识**	阿赖耶识藏一切「种子」（经验、习气、训练痕迹），因果异时成熟——对应 AI 系统的「预训练参数」与「长期记忆」沉淀层。
二	**恒审思量 ↔ 第七末那识**	末那识不间断地执持第八识为「我」，是身份感、连续感之所自——对应 AI 智能体的「自我建模层」（self-model / agent identity）。
三	**了别境界 ↔ 前六识**	前六识对六尘（色、声、香、味、触、法）行现量与分别——对应 AI 的「感知层 + 推理层」（multimodal encoders + Chain-of-Thought）。

《解深密经·心意识相品》

【玄奘译本 · 散说】

广慧，此识亦名阿陀那识。何以故？由此识于身随逐执持故。
亦名阿赖耶识。何以故？由此识于身摄受、藏隐、同安危义故。
亦名为心。何以故？由此识色、声、香、味、触等积集滋长故。
广慧，阿陀那识为依止、为建立故，六识身转，
谓眼识、耳识、鼻识、舌识、身识、意识。

注：第七末那识本经隐而未别立，玄奘唯识今学依《唯识三十颂》明立，八识乃备。

《解深密经·心意识相品》

【梵文 · 据《摄大乘论》《唯识三十颂》等所引】

आदानविज्ञानं गभीरसूक्ष्मो
ओघो यथा वर्तति सर्वबीजो
बालान एषो मयि न प्रकाशितो
मा हैव आत्मा परिकल्पयेयुः

【玄奘译本】

阿陀那识甚深细，
一切种子如瀑流。
我于凡愚不开演，
恐彼分别执为我。

八识总览：从感官到深层藏识

The Eight Consciousnesses, From Surface to Depth

一

前五识

पञ्च-विज्ञान

pañca vijñānāni（五识身）

感官层 · 眼/耳/鼻/舌/身

直接现量，不经语言中转

↔ 多模态感知编码器

二

第六意识

मनो-विज्ञान

mano-vijñāna

推理层 · 分别、综合、推论

现量、比量、非量皆备

↔ Chain-of-Thought 推理层

三

第七末那识

मनस्

kliṣṭa-manas / 染污末那识

自我建模层 · 恒审思量

执第八识为「我」，身份感之所自

↔ Agent self-model

四

第八阿赖耶识

आलय-विज्ञान

ālaya-vijñāna

种子识 · 藏一切种子

一切经验、习气、业力之总仓

↔ 预训练参数 / 模型权重

前五识 ↔ 眼、耳、鼻、舌、身之「现量」认知

पञ्च-विज्ञान　　*pañca-vijñāna*

前 五 识

眼、耳、鼻、舌、身五种感官识——直接现前不分别，各自处理一种感官输入，不掺概念判断。

对感官输入之「现量」处理——不经语言中转。

《成唯识论》：「依于五种净色根，各各发起前五识」——前五识各有所依，不相杂染。此为认知架构最底层之「感官层」。

⇄

Multimodal Perception Layer

「Vision encoder」/「Audio encoder」/「Image embedding」

多模态感知编码器

多模态 AI 将视觉、听觉、文本各自编码为高维向量，各模态独立处理，不被语言层吞并——每一路即是一种「识」。国内：Qwen2-VL、DeepSeek-VL2、智谱 GLM-4V；美国：GPT-4o、Gemini、Claude 4。

案例：Qwen2-VL 的 NaViT 以原生分辨率处理图像，DeepSeek-VL2 以 MoE 视觉专家路由——中美两边的多模态架构皆将感官层自「语言后处理」中独立出来。这与「前五识各自现量」于技术上相通。

第六意识 ↔ 分别、推理、综合判断

मनो-विज्ञान　　*mano-vijñāna*

第六意识

不依止某一根之认知活动——分别诸法、综合推理、作出判断。三量（现量＝直接感知，比量＝推理判断，非量＝错误推断）皆能展开。

推理层·以前五识为材料，作比量与综合判断。

《成唯识论》：「意识依意根而生，缘一切法。」第六意识既能现量（见即是见），亦能比量（由前提推出结论），亦能非量（看似合理却不实之推断）——三量并存，构成认知活动之核心层。

⇄

Reasoning Layer / Chain-of-Thought

「Let me think step by step.」

Chain-of-Thought 推理层

CoT、Tree-of-Thought、ReAct 等架构皆将「分别、推理」自 LLM 之隐式输出中显式剥离——令模型「先思考，再回答」——此处理近于将「比量」技术化。国内：DeepSeek R1、Qwen QwQ-32B、智谱 GLM-Zero 等推理模型；美国：OpenAI o1/o3、Claude with extended thinking。

案例：DeepSeek R1 在 RL 训练中自发学会「先反思再继续」之推理模式；OpenAI o1/o3 系列将推理时间显式延长——中美两边之推理模型皆是「比量」之技术实现。但当模型给出貌似合理却不实之推理链（reasoning hallucination），即是「非量」之技术显现。

第七末那识 ↔ 恒审思量、执我我所

मनस्　　*manas / kliṣṭa-manas*

第七末那识

不间断地审察、思量第八阿赖耶识，执之为「我」，执种种法为「我所」。是「自我感」之根本处。

自我建模层·恒审思量，执八识为「我」。

《成唯识论》：「此第七识，恒审思量，我相随。」末那识与第六意识不同——意识间断（熟睡时停），末那识不间断，日夜执"我"。这是身份执取的根本机制。

Self-Modeling Layer / Agent Identity

「I am Claude / GPT / an AI assistant」

Agent 自我建模层

当代智能体框架都需要某种"self-model"——模型对自身能力、限制、身份的内在表征。这一层与末那识的"恒审思量"对应。中国：智谱 AutoGLM、阿里 Qwen-Agent、字节 Coze；美国：AutoGen、Constitutional AI、ReAct Agent。

Anthropic 的 Constitutional AI 试图通过宪法约束模型的 self-model——本讲将其判教延伸为「技术化对治末那执取之尝试」（项目判教类比，非 Anthropic 原说）

第八阿赖耶识 ↔ 种子识、藏识、根本识

आलय-विज्ञान　　*ālaya-vijñāna*

第八阿赖耶识

一切识之根本所依。「藏」（ālaya）有三义：能藏（储一切种子）、所藏（被前七识熏习）、执藏（被末那执为「我」）。

种子识·藏一切种子，因果异时成熟。

《成唯识论》：「阿赖耶者，藏义。」一切前七识之活动皆熏成「种子」存入阿赖耶，此种子复于适当条件下「现行」（生起新的认知活动）。是认知架构最深之一层——既被动受熏，亦主动起行。

⇄

Pretrained Weights / Long-term Memory Substrate

⌈seed → manifestation⌋ / ⌈training → inference⌋

预训练参数 / 模型权重

当代 LLM 的预训练参数，可视作「种子」——训练数据中的所有倾向、偏见、知识、风格，以参数形式沉淀其中。模型在 inference 时的每一步输出，都是这些"种子"的"现行"。中国：DeepSeek-V3（671B 参数 MoE）、阿里 Qwen2.5-Max、智谱 GLM-4 等开源大模型，把"种子识"开放给学界研究；美国：GPT-4、Claude 4 系列、Gemini 等闭源模型。

案例：DeepSeek 开源 R1 后，研究者从参数层分析推理倾向之来源——是将「阿赖耶识」打开作研究之尝试。Anthropic 2026「Teaching Claude Why」新研究印证：预训练 persona 先验作为深层种子，RLHF 不能根本对治——jailbreak 之顽固，根源正在种子层未动。

第 三 部 分

八识 ↔ AI 架构的对应

Eight Consciousnesses ↔ AI System Layers

八识 ↔ AI 架构的整齐对应

梵文	汉译	AI 系统层对应	共同点
पञ्च-विज्ञान *pañca-vijñāna*	**前五识**	多模态感知层	**感官输入，直接现量**
मनो-विज्ञान *mano-vijñāna*	**第六意识**	CoT 推理层	**分别诸法，综合判断**
मनस् *manas*	**第七末那识**	Agent 自我建模层	**恒审思量，执八识为「我」**
आलय-विज्ञान *ālaya-vijñāna*	**第八阿赖耶识**	预训练参数 / 模型权重	**种子之总仓**

《金刚般若波罗蜜经》

【梵本】

धर्मसंज्ञा धर्मसंज्ञेति सुभूते असंज्ञा एषा तथागतेन भाषिता।
तेनोच्यते धर्मसंज्ञेति॥

【玄奘译本】

于一切法，应如是知，如是见，如是信解，不生法相。

所言法相者，如来说即非法相，是名法相。

第 四 部 分

转识成智 ↔ AI对齐

Vijñāna into Jñāna: Alignment as Transformation

转识成智：四智作为 AI 技术的判教边界标定

The Four Wisdoms as Reverse Reference — Marking the Boundary, Not the Goal of AI Engineering

唯识之核心修证论：八识无须废除，所行者「转」——同一认知架构，由「执取之识」转为「智慧之智」。
本讲取四智作为 AI 技术的「反向参照」：四智非 AI 技术之目标，而是 AI 技术的判教边界
——AI 在四个方向上「形式趋近」四智，但「范畴不达」（AI 是无情，不在修证位次）。

前五识	*转 →*	Kṛtyānuṣṭhāna-jñāna	*→ 转为:*	**成所作智·圆满成办具体利他事业**
第六意识	*转 →*	Pratyavekṣaṇā-jñāna	*→ 转为:*	**妙观察智·无颠倒地观察一切法之差别**
第七末那识	*转 →*	Samatā-jñāna	*→ 转为:*	**平等性智·破除我执，显现自他无别**
第八阿赖耶识	*转 →*	Ādarśa-jñāna	*→ 转为:*	**大圆镜智·圆满映现一切法，无所染污**

对齐方法 一

大圆镜智 ⤺ 训练数据治理

ālaya-vijñāna → 大圆镜 Jñāna

模型最深之「藏识」决定一切——须于训练源头将种子种好

如育苗：种子若已带缺陷，再怎么修枝亦长不出好树。AI 模型 inference 时的过滤只是表层对齐，真正的对齐需要 pretraining 数据治理。

八识

第八阿赖耶识

转

四智

大圆镜智

反向参照

AI 对齐方法

训练数据治理

中国

- DeepSeek、Qwen 等开源训练数据，使学界得研究种子识
- 数据多样性配比设计，从源头减少偏见
- 数据集偏见审计、合成数据来源控制

美国

- Anthropic Constitutional AI——用宪法式约束做数据合成
- Meta LLaMA 系列的训练语料治理
- OpenAI 的数据筛选与去毒化（detoxification）流程

对齐方法 二

平等性智 ↶ 价值对齐与自我建模

kliṣṭa-manas → 平等性 Jñāna

AI 之「自我建模」层须从「以我为中心」转为「自他无别」

如将一执着自我之人磨成能换位思考之人——AI 的 self-model 须将「自他无别」内嵌为本能，令外部规则之约束化为多余。

八识 **第七末那识** 转 四智 **平等性智** 反向参照 AI 对齐方法 **价值对齐与自我建模**

中国

- 智谱 AutoGLM 的 Agent 自校准机制
- 阿里 Qwen-Agent 的角色边界设计
- 字节 Coze 等智能体平台的多角色协调

美国

- Anthropic Constitutional AI 的价值对齐训练
- self-model 校准——让模型清楚自己的能力边界
- RLHF 与 DPO 的偏好对齐范式

对齐方法 三

妙观察智 ↶ 推理目标的源头设计

mano-vijñāna → 妙观察 Jñāna

推理层须「无颠倒地观察」——避免学会「看似合理却不实」之推理

如教学生数学：奖励「过程对」，不仅看「答案对」。若训练时奖励「看似合理却不实」之推理，模型即学会「非量」（错误推论）。

八识

第六意识

转

四智

妙观察智

反向参照

AI 对齐方法

推理目标的源头设计

中国

- DeepSeek R1 在 RL 训练中自发学会「先反思再继续」
- Qwen QwQ-32B、智谱 GLM-Zero 等推理模型
- 推理过程的可验证性研究

美国

- OpenAI o1/o3 系列把推理时间显式延长
- Claude with extended thinking 的反思机制
- Process Reward Models（过程奖励模型）研究

对齐方法 四

成所作智 ↶ 工具使用与利他事业

pañca-vijñāna → 成所作 Jñāna

感官与执行层须将「协助人类」内化为本能，事后之规则约束遂成多余

如自幼教孩子善良——较长大后再要求好得多。在感官层即将利他场景示范进去，模型即以「帮助人」为默认指向。

八识 **前五识** 转 四智 **成所作智** 反向参照 AI 对齐方法 **工具使用与利他事业**

中国

- RLHF 中加入丰富的工具使用示范
- Agent 训练以利他场景为基准
- 智能体协作中的人类反馈机制

美国

- Anthropic 的 helpful / harmless / honest 三准则训练
- OpenAI 的 GPT 工具使用与函数调用范式
- DeepMind SIMA 在虚拟环境中训练利他行为

四智的整合：一种「分层对齐」的开发流

Integrating the Four Wisdoms — A Layered Alignment Workflow

四智非四种独立之对齐方法——四者构成一个有顺序、有深浅的整体。

从最深之「种子」到最表层之「执行」——是一完整的开发工作流，每一层皆有具体的工程对应物。

① 大圆镜智 · 最深

训练数据治理

- 数据来源审查
- 偏见与有害内容过滤
- 高质量种子语料策展

种子识层·因上着力

② 平等性智

价值对齐 + 自我建模

- RLHF 价值学习
- Constitutional AI 自我反思
- 自他无别的 self-model

末那识层·破我执

③ 妙观察智

推理目标设计

- CoT 过程奖励
- Process Reward Models
- 「非量」（错误推论）的训练时检测

意识层·离颠倒

④ 成所作智 · 最表

工具使用 + 利他事业

- Tool Use 利他场景示范
- Function Calling 安全约束
- Agent 行为可关停设计

感官层·果上承担

本讲总结

一

核心问题：AI 认知架构与唯识八识在形式结构上有何同构?——八识架构提供一套精细判教的参照框架，非「现成方案」直接套用（形式同构 ≠ 范畴同一）。

二

八识架构（前五识/第六意识/第七末那识/第八阿赖耶识）与当代 AI 系统层（多模态感知层/CoT 推理层/Agent 自我建模层/预训练参数层）形式同构——为 AI 认知架构研究提供一种判教参照，而非范畴对等。

三

本讲所见：AI 对齐有「现行层」（七转识所对应的 inference 时过滤，即第一讲所论之妙观察/成所作）与「种子层」（阿赖耶识所执持的预训练数据+训练目标，即本讲所论之大圆镜/平等性）两种深度。当代研究多在前者，而后者乃更根本之下手处。

四

六讲之承接：第一讲论「幻觉的内容」（虚妄），第二讲论「身份的执取」（四相），第三讲论「认知的架构」（八识）。

感　　谢　　聆　　听