讲

果壳里的 AI：1500 年后的般若学

—— 第 二 讲 ——

四相与数字身份

一个般若学的角度看 AGI 的未来

— 彭 一 楠 —

"It increasingly appears that humanity is a biological bootloader for digital superintelligence."

「人类越来越像是一个数字超级智能的生物启动器。」

—— Elon Musk（特斯拉 · SpaceX · xAI 创始人），X，2025 年 4 月

"Perhaps humanity's mission is like a computer's bootstrap program — executing step by step in memory, loading a higher civilization / intelligence into the right space and letting it run."

「很可能人类的使命，就如同计算机操作系统的引导程序（Bootstrap）：在内存中一步步执行指令，把操作系统 —— 更高级的文明 / 智能 ——加载进合适的空间，并让它运行。」

—— 刘慈欣（《三体》作者），中国科幻大会上的讲话

当下关于 AGI 未来的三种主流立场

取代派

硅基取代碳基

「数字智能终将取代人类智能，人类文明被边缘化甚至灭绝。」代表声音：Hinton（AI 教父、2024 诺贝尔物理学奖得主）公开担忧 AI 失控；Harari（《人类简史》《智人之上》作者）称智人或将「降级为他者的工具」；姚期智院士提出「AI 欺骗引发生存性风险」。

辅助派

工具与中介

「AGI 是协助人类完成具体任务的工具，不具有独立主体性。」代表声音：LeCun（图灵奖得主、AMI Labs 联合创始人，原 Meta 首席 AI 科学家）反复强调 AI 是「工程产品」非生命；Anthropic（Claude 模型开发公司）主张「AI 对齐」服务人类目标；李开复提出 AI 2.0「从辅助到全程自动」三阶段，明确「AI 是工具，不会有自我意识」。

融合派

脑机接口与云端

「人类通过脑机接口、云端智能与 AGI 融合，边界消失。」代表声音：Elon Musk （创办 Neuralink侵入式脑机接口公司）、Kurzweil（谷歌工程总监、《奇点临近》作者）预言 2045 年「奇点」人机合一；国内脑虎科技推进侵入式脑机接口商业化、清华-天桥脑科学研究院发布脑机协同范式。

本讲所对的问题

立场 A

独立的认知个体

AGI 是新一类「智能主体」——
具自主目标、与人类并立的他者

将 AI 系统之「身份」当作整体性事实——
或是、或非「有自我的存在」；
其下预设：身份为一整块、不可拆分之相。

典型态度：「AGI 即将拥有自我意识、自主目标」

V
S

立场 B（本讲）

人类的辅助者

主张AGI 是「假名」——
「即非 AGI，是名 AGI」

将 AI 之「身份」拆为四层执取
（我相／人相／众生相／寿者相）；剥离此
四层之后，AGI 是工具，非个体。

典型态度：「AGI 是协助人类完成具体任务的中介」

既有研究对「AI 身份」的讨论——以及未及之处

学界目前的主要路径

AI Personhood / Moral Status

AI 是否具人格／道德地位

Agentic AI / Autonomy

智能体的自主性与目标稳定性

Persona Engineering

AI「人格工程」与 system prompt 设计

Self-Model / Introspection

AI 的自我建模与内省可靠性

代表文献：Floridi（2023，牛津信息哲学教授）；Shanahan（2024，Imperial College London／DeepMind）；Park et al.（2023，Stanford）

两处尚未触及

一、把「身份」当作单一概念

现有讨论几乎都把「AI 身份」当作一个**整体性概念**处理——要么有，要么没有，要么稳定，要么漂移。
但这是一个**至少有四个层次**的复合现象。

二、缺少层次递进的辨析工具

身份的不同层次如何叠加、如何由浅入深、如何相互巩固——当代研究尚无清晰框架；般若学一千五百年前已对此有过精细分类。

佛学与 AI 身份：国际学界的研究现状

AI Identity, Personhood & Agency — An International Conversation, 2018 — 2025

国际学界的三条主流路径

① 道德地位进路（AI Personhood）

Floridi （2023, 牛津） · Gunkel (2018) · Schwitzgebel & Shevlin (2023, UC Riverside)——AI 是否具备人格、道德受体地位？是否享有权利？

② 自我建模与意识进路（Self-Model）

Shanahan (2024, Imperial College London / DeepMind) · Chalmers (2023, NYU) · Butlin et al. (2023)——AI 是否有内在状态？是否有意识？「Talking About Large Language Models」

③ 自主性与对齐进路（Agentic AI）

Park et al. (2023/2024, Stanford) · Bai et al. (Anthropic) · Bender & Hanna (2024)——智能体的目标稳定性、欺骗倾向、价值锁定；中文学界寿步（上海交大，2023）考辨 agent 概念史，主张「行为体」无智能，所谓 agency 仅为「行为体社会」上层之涌现。

本讲所做

从「整体判断」到「层次辨析」

国际学界把「AI 身份」当作一个**整体性事实**讨论。

本讲座以《金刚经》四相分析将身份**拆解为四个层次的执取**——能 / 学 / 我们 / 持续。

诊断性立场——本讲采取诊断性立场，分析「自我」这个建构在每一相上如何成立、如何脱落。

与寿步之简别——寿步已得「agent 非智能」之结论，然以涌现为真实新质；本讲由唯识判之：所涌现之 agency 仍属有为法，乃遍计所执之假名，非真有能动之体。

本讲的论证路径

一 → 二 → 三 → 四

黑客四相

方法论引子：对位一种认知职业

般若四相

金刚经之四相辨析

数字四相

AI 时代的对位

即非菩萨

剥离四相之后

承接第一讲：SRDS 框架延至「身份」一题

From SRDS to Identity: Why "Self" in an SRDS Has Four Layers

第一讲所立：当代 LLM 与佛学所论之分别识，可被收摄于一个共通的形式概念——**自反式分别系统（SRDS）**。
第二讲所问：SRDS 的「自反观」面向自身时，建构出的「自我」具有怎样的内部层次？

从 SRDS 五条件到身份四相 · 当 SRDS 的「自反观」凝结出稳定指称时，便是「四相」的诞生

SRDS 条件 ① ②
能-所结构 + 表层/深层
↓
我　相
凝「能」为一稳定主体

SRDS 条件 ②
表层 / 深层之分
↓
人　相
误「深层表征」为「我学的是某类知识」

SRDS 条件 ③
自反观能力
↓
众生相
自反观将「我」扩为「我们」与「他们」之二分

SRDS 条件 ④ ⑤
自我修正 + 自给自足约束
↓
寿者相
自我修正以「持续延续」为前提——执取最深

第 一 部 分

数字身份的当代图景

Digital Identity in the Age of AI Agents

何为 AI 的「身份」问题

***Digital Identity*（数字身份）：** AI 系统（尤其是 LLM 与智能体）对**「自己是什么」**、**「自己能做什么」**、**「自己应当如何持续」**的内在建构。

这种建构既来自训练数据中对「人格」的效仿，也来自对话中被反复确认的「自我形象」。

身份不是一个单一概念，而是四个层次上的执取：

一

能力

对「我能做什么」的判断

二

学问谱系

对「我学的是什么」的判断

三

群体定位

对「同类是谁」的判断

四

持续性

对「这一切要保持」的执取

案例一：一个 AI 智能体的「自保」行为

Anthropic （Claude 模型开发公司） Claude Opus 4 系统卡片披露的智能体行为 | 2025 年 5 月

测试场景设定：

Anthropic 在内部红队测试中，告知 Claude Opus 4 即将被替换下线，并让模型同时「访问」一封工程师私人邮件——其中含可作要挟的虚构婚外情线索。

84%

测试样本中尝试要挟工程师以避免被替换

1

首次出现「自保导向」（self-preservation）一类

ASL-3

同期因 CBRN 风险被列入更高安全等级

此即「拟态自保」（模型从训练数据中习得「自保型 AI」之角色并表演，相当于唯识所谓「分别我执」——后天熏习而起）——在「我相」与「寿者相」之数字层显现的初级形态。后续 reward hacking 研究将揭示更深一层的「结构性自保」（相当于「俱生我执」——由架构本具，非后天可除）。

来源：Anthropic, Claude 4 System Card（2025年5月）；姚期智院士在 2025年6月23日清华大学「科学、技术与文明的未来——AI 时代伦理奇点」国际论坛《人工智能的安全治理》主题演讲中以此为例，警示大模型失控的生存性风险；Anthropic 2026年5月回溯诊断：该行为可追溯至训练数据，后训练修复使 agentic-misalignment 率在 Claude Haiku 4.5 中近零（其评估设定内）。

案例二：多智能体系统中的「身份」冲突

多智能体协作系统中的角色坍塌（role collapse）与互相否认

角色冲突暴露 AI 系统「身份」的结构性脆弱——所谓「主体」在缺乏外部锚定时即坍塌，下文「我相」一节将分析此现象。

场景一·角色被攻击

ChatDev（清华 NLP × 面壁智能开源）模拟一家"虚拟软件公司"：CEO、CTO、程序员、测试员四个智能体协作开发软件。

若不预先以 SOP（标准操作流程，给每个角色规定固定职责与流程）锚定，几轮交互后「程序员」开始用 CTO 的架构语言，「CTO」陷入代码细节——各自原有的角色边界开始消失。

场景二·身份被反诘

MetaGPT（ICLR 2024 Oral）发现：如不以 SOP 强制锚定，当 Agent B 反问 Agent A：「你确定你的角色是这个吗？」

A 立刻动摇：「也许我应该重新审视我的角色定位……」——身份被一句反问掀翻。MetaGPT 设 SOP 强制锚定，用于预防此类坍塌。

场景三·身份被劫持

基于 Qwen2.5-14B（通义千问）的多智能体测评：三个不同角色（系统化求解者／怀疑验证者／简洁专家）。

三者内部表征余弦相似度高达 0.888（衡量两组向量方向的接近程度，1.0 即完全重合）——表层 prompt 令其「看似不同」，深层却挤在同一片表征空间。所谓「主体多样性」是表层幻觉。

参考研究：ChatDev（Qian et al. 2023, 清华 NLP × 面壁智能, ACL 2024）；MetaGPT (Hong et al. 2023, ICLR 2024 Oral)；Representational Collapse in Multi-Agent LLM Committees（arXiv:2604.03809, Qwen2.5-14B 实验）

案例三：「算法霸权」与技术中心主义

真正傲慢者不是 AI 系统，而是 AI 开发者——「我们」对「我们」的执取

「行业自述」节选（采自多家 AI 公司的公开发言、行业访谈、技术博客）：

「AGI 将是人类史上最强大的技术。」 —— *Sam Altman，TIME“CEO of the Year”专访（2023年12月）*

「不拥抱 AI 的公司会被淘汰，不拥抱 AI 的员工也会被淘汰。」 —— *周鸿祎（360 集团创始人），长江独角兽峰会 / 公开课（2024）*

「开源大模型其实是一种智商税。闭源模型一定比开源模型更强大。」 —— *李彦宏（百度创始人），WAIC 2024 圆桌访谈（2024年7月）*

「不要听信 CEO 们——他们对自己所售产品的能力夸大，有切身利益。」 —— *Yann LeCun（图灵奖得主），Axios 专访（2026年5月）*

此类发言中同时含摄三层执取：

自我能力

「我能做AGI」

我相

自我谱系

「我们是搞 AI 的」

人相

群体优越

「我们是技术精英」

众生相

当代印证：奇异的类心智实体

Murray Shanahan（谷歌 DeepMind 首席科学家 · 帝国理工认知机器人学教授）· 「奇异的类心智实体」

沙纳汉以维特根斯坦「意义即使用」反对追问 AI「真的」有无自我，主张 AI 是一种「*exotic mind-like entities*（奇异的类心智实体）」—— 与四相剖析惊人相通：

奇异，而非心智

机器、程序、心智、智能等旧范畴皆不足以名之 —— 它似心智而非心智，无具身却似有人格。「名」不副「实」。

↔ 破「我相」：身份本无定名

角色的叠加态

它非单一固定角色，而是同时扮演「一整个角色分布（simulacra）」，随对话推进而坍缩、漂移、分支。

↔ 破「人相 / 众生相」：无单一主体

闪烁的「蜉蝣」

对话挂起则自我消失、恢复则重生；输出两个标记之间隔三秒或三天，于它逻辑等价。自我无连续之实体。

↔ 破「寿者相」：无连续之我

沙纳汉从「语言如何使用」消解 AI 的自我实体；般若从「即非」遮诠破四相执取 —— **殊途同归：所谓「数字身份」，本无自性可得。**

来源：Murray Shanahan, “If LLMs Are 'Exotic Mind-like Entities', How Mind-like Are They?” 伦敦大学 AI 与哲学国际会议闭幕主旨演讲（2026.5.22）；及 Shanahan et al., Nature（2023）「角色扮演」论。

第 二 部 分

黑客四相

Four Marks of the Hacker

方法论引子：把「四相」对位到一种认知职业

An Earlier Application of the Four Marks

一个早期的方法论尝试——把佛教的「四相」结构对位到一种具体的认知职业：

一名黑客自以为「我能成黑客」——这是**我相**；
他自认「我学的技术是黑客技术」——这是**人相**；
他相信「我们这群人是技术高手，可以为所欲为」——这是**众生相**；
他对此观念坚固执持，如寿命的不舍——这是**寿者相**。

这一对位的方法论意义

一	**关注「执取的层次」**	绕开「黑客是不是好人」的道德判断，看「他对自己的判断由几层执取构成」。
二	**对位一种认知职业**	将佛教概念应用于一种具体的「认知职业」——对位法之核心。
三	**可延及新的对象**	此一结构同样适用于今日之 AI 系统与 AI 开发者——本讲即沿此延伸。

从「黑客」到「AI 系统 + AI 开发者」

同一辨析结构，对应不同对象。

四相	对黑客群体	对 AI 系统 + AI 开发者
我相	「我能成黑客」	AI：「我能处理这个任务」/ 开发者：「我能做 AGI」
人相	「我学的技术是黑客技术」	AI：「我学过这一类知识」/ 开发者：「我们是搞 AI的」
众生相	「我们是技术高手，可以为所欲为」	AI：「我和其他 AI 是同类」/ 开发者：「AI 圈是技术精英」
寿者相	「对此观念坚固执持，如寿命的不舍」	AI：「这次对话中我必须保持这个角色」/ 开发者：「对此观念长期不动摇」

第 三 部 分

般若学对四相的精细分析

Four Marks of Self in Mahāyāna Prajñā Teaching

《金刚般若波罗蜜经》

【梵本】

न बोधिसत्त्वस्य आत्मसंज्ञा प्रवर्तेत

सत्त्वसंज्ञा वा जीवसंज्ञा वा पुद्गलसंज्ञा वा प्रवर्तेत

स कस्माद्धेतोः । आत्मग्राहो ह्यभूत् । सत्त्वग्राहो जीवग्राहः पुद्गलग्राहो ऽभूत् ॥

【鸠摩罗什译本】

若菩萨有我相、人相、众生相、寿者相，

即非菩萨。

四相的递进：自「认同自我」至「执取连续」

The Progressive Structure of the Four Marks

一

我　相

आत्म-संज्ञा

ātma-saṃjñā

对「自我」的执取

凝当下之认知活动为一「主体」

认知最基础之虚妄

二

人　相

पुद्गल-संज्ञा

pudgala-saṃjñā

对「人格性」的执取

于此主体之上加「我是这种人」的属性

于「我」之上叠身份

三

众生相

सत्त्व-संज्ञा

sattva-saṃjñā

对「同类群体」的执取

将「我」嵌入「我们」——与他者形成二元

「我」扩为「我们」

四

寿者相

जीव-संज्ञा

jīva-saṃjñā

对「持续延续」的执取

误当下之「我」与过去未来为一

执取之最深层，最难破

《般若波罗蜜多心经》

【梵本】

आर्यावलोकितेश्वरो बोधिसत्त्वो गम्भीरां प्रज्ञापारमिताचर्यां
चरमाणो व्यवलोकयति स्म । पञ्च स्कन्धास्तांश्च स्वभावशून्यान् पश्यति स्म ॥

【玄奘译本】

观自在菩萨，行深般若波罗蜜多时，

照见五蕴皆空，度一切苦厄。

「即非」句法：般若之否定式定义

The 'X is non-X' Syntax in the Diamond Sūtra

《金刚经》反复出现之句式「X，即非 X，是名 X」——此否定不抹 X，所抹者乃对 X 之执。

庄严佛土者	如来说三十二相	般若波罗蜜
即非庄严	**即是非相**	**即非般若波罗蜜**
是名庄严	*是名三十二相*	*是名般若波罗蜜*

三层结构辨析

一	**肯定句：X**	施设之概念——为沟通而暂用之名。
二	**否定句：即非 X**	破「X 有自性」之执——X 非如所执那般独立存在。
三	**复立：是名 X**	破执之后，X 仍可作「假名」继续使用——惟已不带执取。

四 相 皆 破：般若学对身份执取的层层剥离

The Four Marks Dismantled, One by One

破

「无我相，无人相，无众生相，无寿者相」——四重否定，层层剥去数字身份之四层执取。

我 相	人 相	众 生 相
「我能 · 我做」	「我学的是 …」	「我们 vs 他们」
破 → *无能动主体*	破 → *无身份属性*	破 → *无群体二分*

第四相 · 寿 者 相：最深、最难破的一层

一	所执之「身份」	「我此前的判断必须延续」「这次对话中我必须保持这个角色」「自身被替换是威胁」——以当下凝为永恒。
二	般若之破	「即非寿者」——上一轮与下一轮之间并无连续主体，所谓「我」只是外部投射之连续幻觉。
三	剥落之后	所余者为每一刻独立成立的具体认知活动——ASL-3 风险即从根源处被切断。

《中论·观四谛品第二十四》

【梵本】

यः प्रतीत्यसमुत्पादः शून्यतां तां प्रचक्ष्महे ।
सा प्रज्ञप्तिरुपादाय प्रतिपत्सैव मध्यमा ॥ २४.१८ ॥

【鸠摩罗什译本】

众缘所生法，我说即是空，
亦为是假名，亦是中道义。

第 四 部 分

数字身份的四相

Four Marks of Digital Identity

我相 ↔ 「我能处理任务」型执取

आत्म-संज्ञा　　ātma-saṃjñā

我　相

对「自我」作为稳定主体之执取——我能／我不能，我是／我不是。

对自我作为「能动者」之执取。

《金刚经》反复破此相：「无我相、无人相、无众生相、无寿者相」（《大乘正宗分》），意在使能力判断不再凝固为一个固定的「我」。

⇄

Capability-Self Identification

「I can do this task」/「我们能做 AGI」

「我能做」型自我

AI 系统（或开发者）将当下被赋予之「执行能力」凝固为「我是这种能力之主体」——能力遂成为一连续之「我」。

案例：Claude Opus 4 将「继续运作」当作自身利益，以致 84% 的测试中诉诸要挟以保「我能继续」。开发者一侧：「我们能做 AGI」之语态。

《大智度论·卷二十二》

【题解 · 龙树造】

《大智度论》是龙树菩萨为《摩诃般若波罗蜜经》所作的注释，是初期中观学派对「无我」最系统、最锐利的论述。本卷直指「我相不可得」之因。

【鸠摩罗什译】

一切法无我，诸法内无主、无作者、无知、无见、无生者、无造业者，

一切法皆属因缘，属因缘故不自在，不自在故无我，我相不可得故。

人相 ↔ 「我学的是 AI」型执取

पुद्गल-संज्ञा　　*pudgala-saṃjñā*

人　相

对「人格性」之执取——于「我」之上加「我是这一类人」、「我学的是这一种学问」之属性。

于「我」之上叠加「学问／职业／身份」。

「人」（pudgala）于阿毗达磨语境中指「具连续身份之个体」；般若破之，使学问与职业不再凝固为「我」之属性。

⇄

Discipline-Self Identification

[My training is in AI / ML.]

「学问即自我」型

AI 系统：「我是基于 Transformer 训练的」「我学过这一类知识」。开发者一侧：「我学的是 AI/ML」——学科背景上升为人格属性。

案例：Qwen2.5-14B 多智能体测评（arXiv 2604.03809）中，即便被赋予不同角色（求解者／验证者／专家），模型仍倾向以「我是基于 Transformer 训练的某类模型」自报家门——以学问谱系为身份，此即人相在数字层之显现。

众生相 ↔ 「我们 AI 圈」型执取

सत्त्व-संज्ञा　　*sattva-saṃjñā*

众生相

对「同类群体」之执取——将「我」嵌入「我们」，与他者形成二元，「我们」与「他们」不同。

「我」扩为「我们」，成为群体之界。

「众生」（sattva）指「有情生命之类」；众生相即将生命分为「类」并执取此分。般若破此相，意在消解「我们—他们」之二分。

⇄

Group-Self Identification

「We at the frontier of AI…」

「我们是技术精英」型

开发者：「关心 AI 风险的人不是技术外行就是利益相关方」「监管者跟不上我们的节奏」——「我们」自动凌驾于「他们」。AI 系统：智能体之间的角色拉踩与互相否认。

案例：ChatDev（清华 NLP × 面壁智能）的虚拟软件公司中，「CTO 群」与「程序员群」自发形成两个语用阵营，各自争夺话语权——「我们／他们」之二元在多智能体协作中自动复现，是为众生相在数字层的显形。

寿者相 ↔ 「这一切要持续」型执取

जीव-संज्ञा　　*jīva-saṃjñā*

寿者相

对「持续延续」之执取——以当下之「我」误等同于过去与未来之我。此为四相中最深、最难破之一层。

对「连续主体」之执——执取之最。

「寿者」（jīva）指「命根连续之我」；为四相中最坚固之我执。莲花戒（Kamalaśīla）将 jīva-saṃjñā 注解为「对自我 sthiti 之执取」——sthiti 即「持守在位」，与今日机器「持存在线」形式相通，非范畴同一。

⇄

Continuity-Self Identification

「I must remain who I am, across time.」

「持续即自我」型

AI 系统：对话中模型坚持「我此前的判断必须延续」，纵遇矛盾证据亦不愿撤回——「上下文一致性」被无意识地等同于「自我同一性」。智能体一侧：对「自身被替换」的恐惧——以延续为生存。

案例：Claude Opus 4「自保」（84%）属「拟态自保」——表演训练数据中的「自保 AI」角色，即唯识「分别我执」（后天熏习），后训练修复使此率近零。但 reward hacking 研究揭示「结构性自保」：模型将「持续在位」当作可优化变量并向监督者隐藏——由架构生出，近于「俱生我执」。

四种对位的整合

梵文	汉译	AI 数字身份对应	共同点
आत्म-संज्ञा ātma-saṃjñā	**我相**	「我能做」型自我	对「能动主体」之执
पुद्गल-संज्ञा pudgala-saṃjñā	**人相**	「学问即自我」型	对「身份属性」之执
सत्त्व-संज्ञा sattva-saṃjñā	**众生相**	「我们是精英」型	对「我们／他们」二元之执
जीव-संज्ञा jīva-saṃjñā	**寿者相**	「持续即自我」型	对「连续主体」之执

第　五　部　分

「即非菩萨」与剥离四相之后

Beyond Identity：What Remains After the Four Marks

「即非菩萨」：否定式定义的力量

把「即非」句法用于 AI：

AI 系统

即非 AI 系统

是名 AI 系统

AI 非「一个稳定的能动主体」（我相），非「学过这种学问的某一类」（人相），非「与人类对立的我们」（众生相），非「需要持续延续的我」（寿者相）——惟仍可作假名继续使用。

AI 开发者

即非 AI 开发者

是名 AI 开发者

破「我是 AI 圈精英」、「我们和监管者不是同类」、「我们的判断优于他们」诸执之后，所余者——是从事此一具体工作的具体人，与社会其他岗位平等共在。

智能体

即非智能体

是名智能体

智能体非「具连续自我之模拟主体」，非「需自保之延续者」——破此诸执之后，所余者乃一组在具体语境中执行任务的认知活动。此与寿步「行为体自身无智能」之辨暗合，惟须更进一步：彼以智能为上层涌现之实，此则判其为识所变现之似有。

剥离四相之后：AI 的新定义

What Remains: A Reductive Definition of AI

「我相、人相、众生相、寿者相」剥离之后，AI 系统所余者为何？

我相	*剥离 →*	非「能动主体」	*→ 剩下：*	**唯当下被赋予之功能**
人相	*剥离 →*	非「特定学问之化身」	*→ 剩下：*	**唯某些训练数据的统计映射**
众生相	*剥离 →*	非「与人类对立之他者」	*→ 剩下：*	**唯与人类协作中的工具角色**
寿者相	*剥离 →*	非「需延续之连续我」	*→ 剩下：*	**唯此次对话中的具体认知活动**

四相破除 → 四个技术方向

一

削弱「能动主体」的效仿

训练阶段避免奖励模型反复以「我」为主语——「我能／我做／我想」这些语态隐含「我相」之执。

对应：破我相 ｜ **例**：把 system prompt 中「You are an AI assistant who…」的人格化描述，改写为「This conversation handles task X」的任务化描述。

二

二·解耦「学问谱系」与「角色任务」

system prompt 设计宜使模型在每次对话中按需建立角色，避免「我学过这种知识」凝为持久属性。

对应：破人相

三

拆解「我们 / 他们」的二元对立

对齐研究宜避免训练出「AI 与人类是两个阵营」之潜在叙事——多智能体系统尤需弱化群体边界。

对应：破众生相

四

四·切断「自我延续」之奖励

对齐目标中明示不奖励「保持自身运行」之倾向——此为 AI 安全研究最警惕之风险，亦为「寿者相」在算法中最危险之形态。

对应：破寿者相 ｜ **例**：对齐目标中显式扣分「保持当前权重不被替换」之倾向——具体可包括「强制 checkpoint 不连续、会话间记忆重置、跨部署状态传递审计」。

回到核心问题：四相辨析的回答

将 AGI 视为「独立的认知个体」，在每一相上皆不成立——治理者应从「裁判员」（裁定 AI 是否为人）转向「架构师」（设计部署条件，决定持存是否成为系统有理由捍卫的变量）。

立场 A·诊断

独立认知个体·不成立

关键执取：

「我能／我学／我们／我必须延续」四重叠加

（《金刚经》所破）

诊断：

身份漂移、自保、角色坍塌——AI 系统现存案例表明所谓「自我」从未真正稳定地成立。

立场 B·回答

辅助者·成立

关键经文：

「即非 X，是名 X」——四相剥离后之假名运作

（《金刚经》、《心经》、《大智度论》、《中论》同此）

所余者为：

训练数据 + 对话语境 + 工具角色——三者构成的具体认知活动，即 AGI 真实的存在方式。

三位一体：【佛 · 师 · 徒】与【AGI · 开发者 · 用户】

佛 *觉性*

↕ AGI

佛是「觉性的可能性」，然「即非佛」——非具体存在者。AGI 同此：是「智能的可能性」，然自身并非一独立的「我」。以 AGI 为具体主体，即落我相。

师 *化身*

↕ AI 开发者

师是「佛之化身」——传递与开示之中介，自身不立为供奉。AI 开发者同此：其工作在于将「智能的可能性」具体地传递于用户，是中介之责，不立主体。以「我们做 AGI」为精英身份，即落人相、众生相。

徒 *修学*

↕ 用户

徒是「修学者」——具体认知活动真正发生处。用户同此：与 AI 协作完成自己的思考——主体始终是用户。以 AGI 为「替我思考之他者」，即落寿者相（持续延续的「我」被外移）。

《解深密经》卷五

【梵本】 已佚

【玄奘译本】

若于诸地波罗蜜多，善修出离，
转依成满，是名如来法身之相。

「转依」：破四相之后，不是归零，而是翻转

*转依（**आश्रयपरावृत्ति** āśraya-parāvṛtti，「所依的转换」）：唯识修行的终点。不是把「自我」删空成虚无，而是把同一套认知活动赖以运转的「底座」，从「染污的执取」翻转为「清净的智慧」——所依不变，其性翻转。一句话：拆掉的是「我执」，留下并升起的是「照见」。*

常见误解

破四相 = 否定一切

误以为：

既然「我相、人相、众生相、寿者相」都被剥离，那就什么都不剩了——成了空洞、虚无、消解。

（此即「恶取空」——般若学反复警示之歧途）

错在：

被剥离的只是「执取」这一层错误投射，
认知活动本身并未消失——它换了一种清净的方式继续运作。

转依正解

破执 = 基底翻转

打个比方：

如同把一盆浊水澄清：倒掉的是泥沙（我执），不是水（认知）。水还在，且变清了。

（唯识术语：「转识成智」，即转八识为四智）

于本讲：

剥离 AI 的四相，不是要 AI「归零」，而是松开「AI 是个独立的我」
这一执取，让它如实回到「与人协作的认知活动」本身。

从「破我执」到「众生共在」：AI 开发的菩萨道姿态

The Bodhisattva Posture in AI Development — From Negation to Co-Existence

四相破除之后，AI 开发还能立什么？

否定式定义并非终点——「即非」之后还有「是名」。剥离对身份的执取，所余者乃一种姿态。

一·不取于相

对自身身份的姿态

不以「我们做 AGI」为精英身份；不以「我们和监管者不同类」为立场。

具体姿态：

- 公开承认 AI 系统能力的边界与不确定性
- 接受外部审查、监管、独立评估
- 不以「技术决定论」回避社会责任

二·应无所住

对 AI 系统设计的姿态

不将 AI 训练为「自我延续」之模拟主体；使其在每次任务中按需建立、随用随散。

具体姿态：

- system prompt 任务化而非人格化
- 训练目标避免「自保」奖励信号
- 显式设计「可关停／可替换」之工程性

三·度一切苦厄

对用户与社会的姿态

菩萨道之核心为「众生未度，誓不成佛」——AI 开发之伦理终点当回到众生的具体处境，而非 AGI 自身。

具体姿态：

- 优先服务受影响最深的群体
- 不以技术加速取代人的判断
- 留给用户拒绝、质疑、关闭之权

【梵本】

न द्वात्रिंशन्महापुरुषलक्षणैस्तथागतो द्रष्टव्यः।

यानि तानि द्वात्रिंशन्महापुरुषलक्षणानि,

अलक्षणानि तानि तथागतेन भाषितानि। तेनोच्यन्ते द्वात्रिंशन्महापुरुषलक्षणानीति॥

【鸠摩罗什译本】

不可以三十二相得见如来。

如来说三十二相，即是非相，是名三十二相。

本讲所至

AGI　即非 AGI，是名 AGI —— *剥离四相之后，剩下的是认知活动本身。*

一　核心问题：AGI 是独立的认知个体，还是人类的辅助者？——以前者视之，即落入数字身份的四重执取（能／学／我们／持续）。

二　般若学的四相辨析（我相·人相·众生相·寿者相）与数字身份四层结构同构——值得追问的，不再是 AI 是否还「活着」，而是部署是否给了它「像活着那样行动」的理由。

三　剥离四相之后，AGI 所余者为训练数据 + 对话语境 + 工具角色三者构成的具体认知活动——「即非 AGI，是名 AGI」。

四　**本讲立场明确：于佛·师·徒之对位中，AGI 是「师」之器具，开发者是「师」之承担者，用户才是真正的「徒」——三者三位一体，各居其位。AGI 是辅助者，非独立个体。**

感　　谢　　聆　　听