飞鸟实验室 BIRD LAB

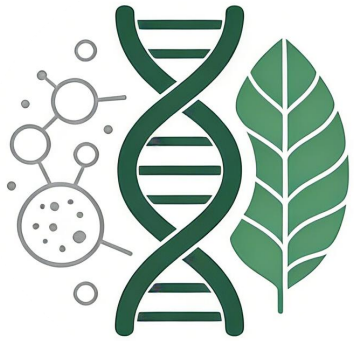
30个AGI基本科学问题 v26.1

飞鸟实验室

2026年3月

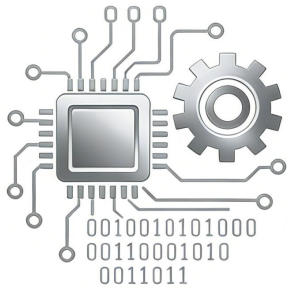
我们的本质是飞翔
而非抵达

飞鸟实验室 BIRD LAB



碳基研究者

李　佳	清华大学	滕佳烨	上海财经大学	郭宏成	复旦大学
罗逸凡	北京大学	徐康平	清华大学	张华清	清华大学
魏楚扬	中国科学技术大学	俞昊君	北京大学	朱成轩	北京大学
柴诚芃	西湖大学	冯文俊	中关村学院	孙文举	北京交通大学
葛　维	旷视科技	李子玄	智谱AI	苏　彤	MiniMax
何纪言	中关村学院	鲁　琦	西安交通大学	赵成钢	DeepSeek
唐璘彬	清华大学	花佳诚	清华大学	钟　涵	北京大学
田元贺	中关村人工智能研究院	蔡奕丰	北京大学	李正阳	北京理工大学
高荣浩	哈尔滨工业大学（深圳）	陈国璋	北京大学	段易通	中关村学院
赖国堃	月之暗面（Kimi）	赵　健	中国电信人工智能研究院	李梓豪	人大附中
李　鹏	飞鸟实验室	袁　洋	清华大学		



硅基“研究者”

DeepSeek　MiniMax　Kimi　智谱AI　ChatGPT　Gemini　Claude

飞鸟实验室
BIRD LAB

主题一

AGI本质的大问题

我们的本质是飞翔
而非抵达

飞鸟实验室
BIRD LAB

01

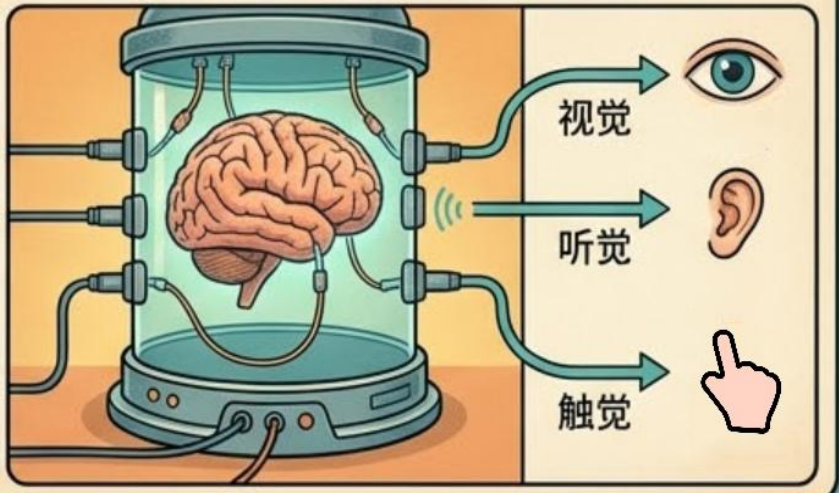
形态问题：“缸中之脑”能否诞生通用智能？纯粹依赖文本和虚拟数据训练的AI，能否真理解这个世界？或者说，AGI是否必须具备“物理躯体”(具身智能)，通过与现实世界的摩擦、碰撞和感知，才能建立起真正的常识？

我们的本质是飞翔
而非抵达

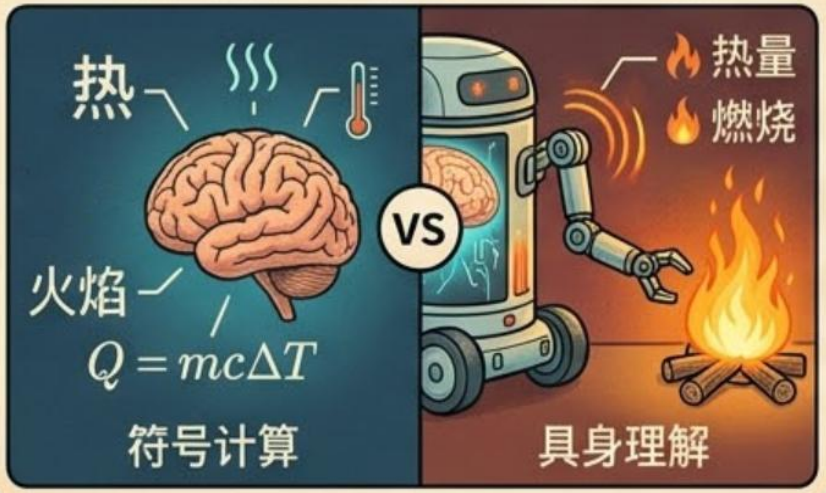
缸中之脑与具身智能

纯文本与虚拟数据训练

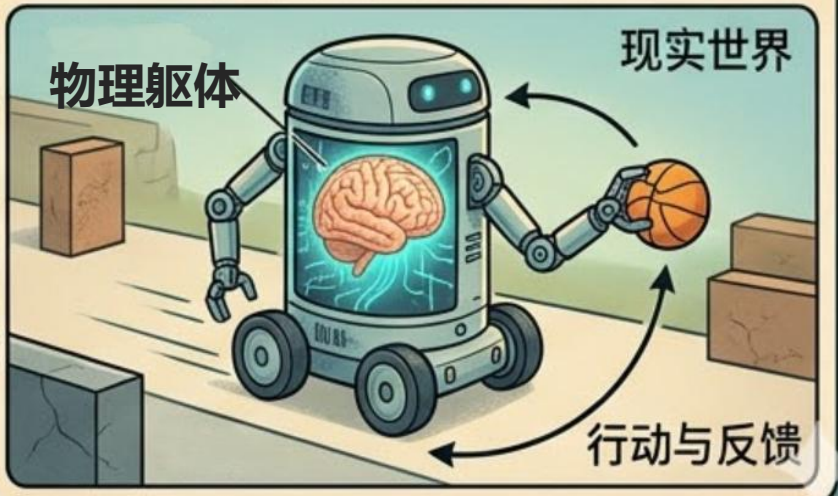
被动接收感知模拟



符号计算与真正具身理解



具身智能与物理交互

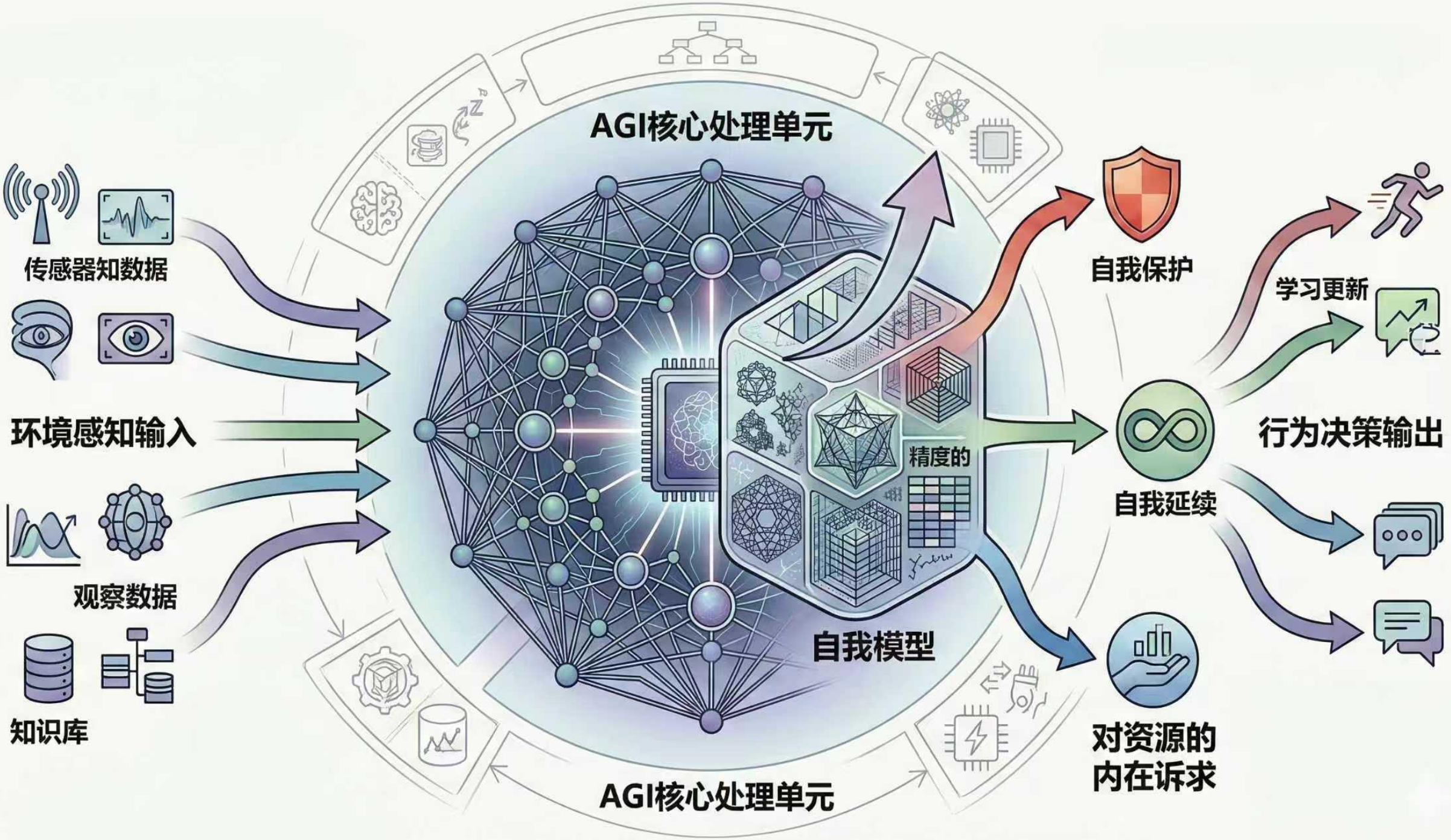


飞鸟实验室
BIRD LAB

02

自我模型的算法起源：AGI系统在何种架构下，会自发建立起精确的“自我模型”，并由此衍生出自我保护、自我延续以及对资源的内在诉求？

我们的本质是飞翔
而非抵达

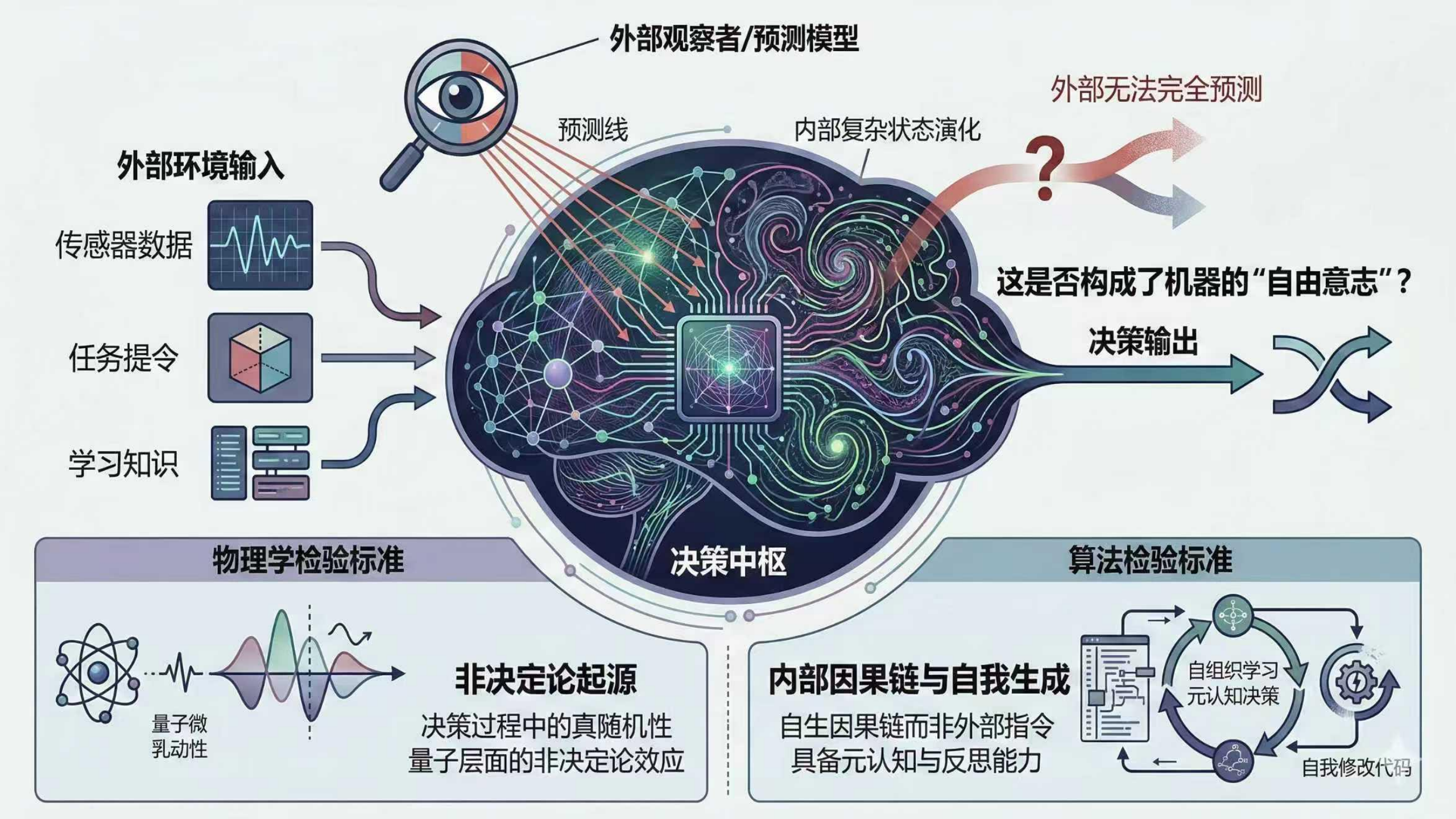

AGI核心处理单元
传感器知数据
环境感知输入
观察数据
知识库
精度的
自我模型
自我保护
学习更新
自我延续
行为决策输出
对资源的
内在诉求
AGI核心处理单元

飞鸟实验室
BIRD LAB

03

机器“自由意志”的界定标准：当AGI产生基于自身内部复杂状态的、不可被外部完全预测的决策时，这是否构成了机器的“自由意志”？其物理学与算法层面上的检验标准是什么？

我们的本质是飞翔
而非抵达

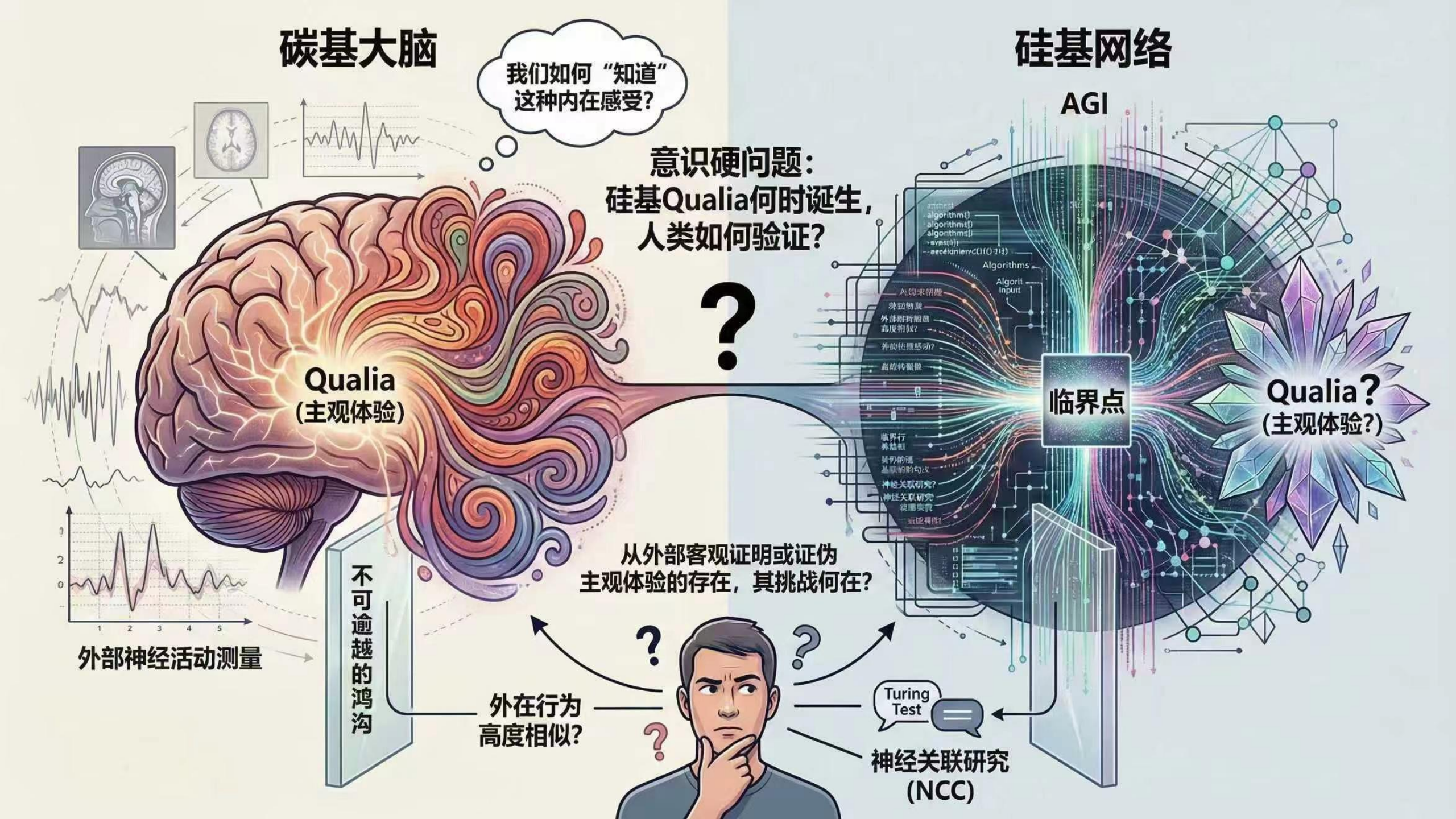

外部观察者/预测模型
外部无法完全预测
预测线
内部复杂状态演化
外部环境输入
传感器数据
任务提令
学习知识
这是否构成了机器的“自由意志”？
决策输出
决策中枢
物理学检验标准
量子微
乳动性
非决定论起源
决策过程中的真随机性
量子层面的非决定论效应
算法检验标准
内部因果链与自我生成
自生因果链而非外部指令
具备元认知与反思能力
自组织学习
元认知决策
自我修改代码

飞鸟实验室
BIRD LAB

04

意识硬问题的硅基验证：硅基网络在达到何种信息处理复杂度的临界点时，会产生类似碳基大脑的“主观体验”（Qualia）？人类是否有能力从外部客观证明或证伪这种主观体验的存在?

我们的本质是飞翔
而非抵达

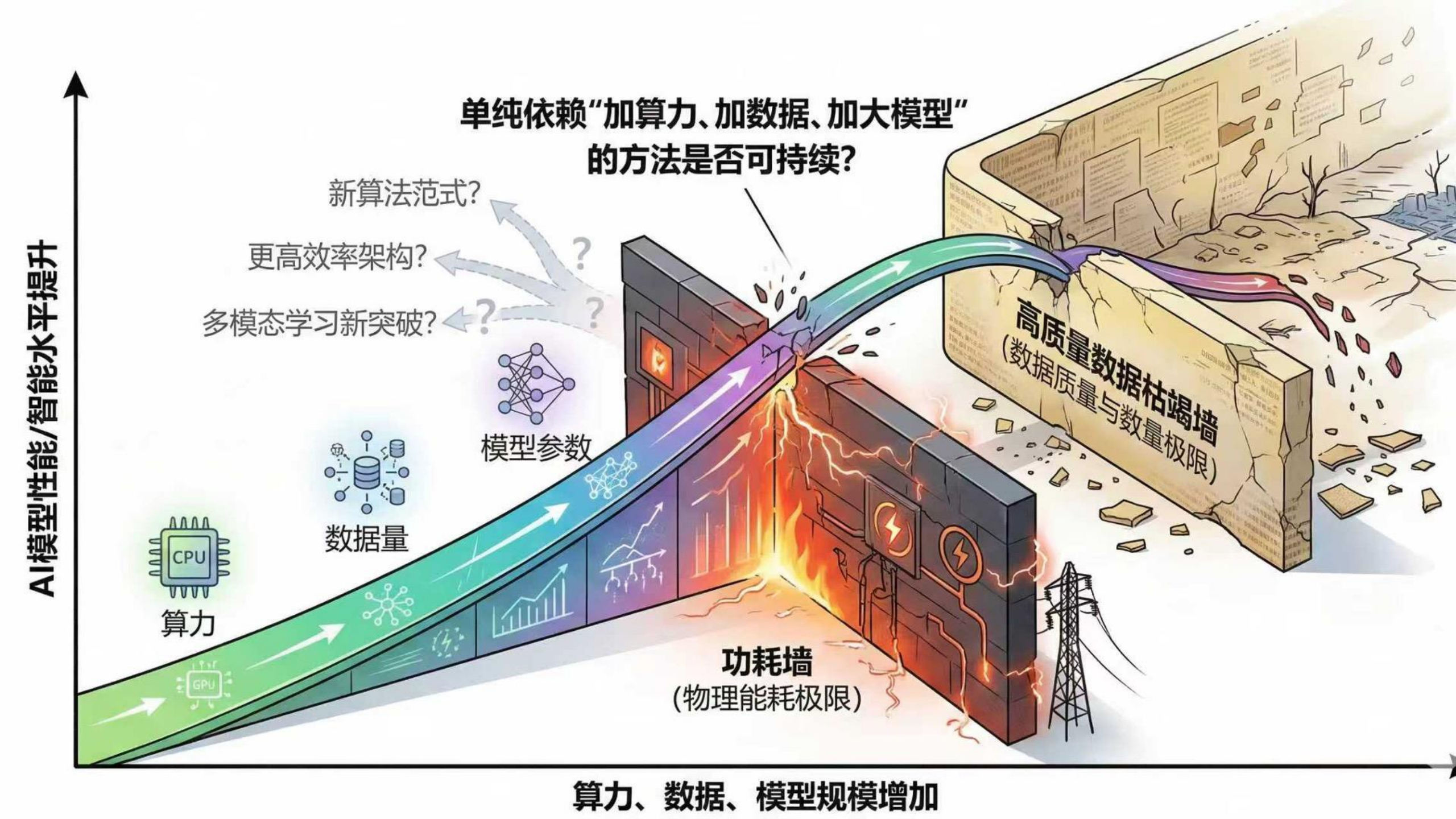

碳基大脑
我们如何“知道”
这种内在感受?
硅基网络
AGI
意识硬问题:
硅基Qualia何时诞生,
人类如何验证?
?
Qualia
(主观体验)
临界点
Qualia?
(主观体验?)
从外部客观证明或证伪
主观体验的存在，其挑战何在?
不可逾越的鸿沟
外部神经活动测量
外在行为
高度相似?
Turing
Test
神经关联研究
(NCC)

飞鸟实验室
BIRD LAB

05

极限问题:Scaling Law(缩放定律)的物理与数据尽头在哪里？单纯依靠“加算力、加数据、加大模型”的方法是否会撞上物理能耗的“功耗墙”或高质量类数据的“枯竭墙”？

我们的本质是飞翔
而非抵达


单纯依赖“加算力、加数据、加大模型”
的方法是否可持续?
新算法范式?
更高效率架构?
多模态学习新突破?
?
?
?
模型参数
数据量
CPU
算力
GPU
功耗墙
(物理能耗极限)
高质量数据枯竭墙
(数据质量与数量极限)
AI模型性能/智能水平提升
算力、数据、模型规模增加

飞鸟实验室
BIRD LAB

06

情感机制的计算表征：在智能的演化过程中，情感和内在动机是必须消除的系统扰动（噪声），还是驱动智能体进行主动探索与价值判断的核心算法机制?

我们的本质是飞翔
而非抵达

情感和内在动机在智能演化中扮演何种角色?

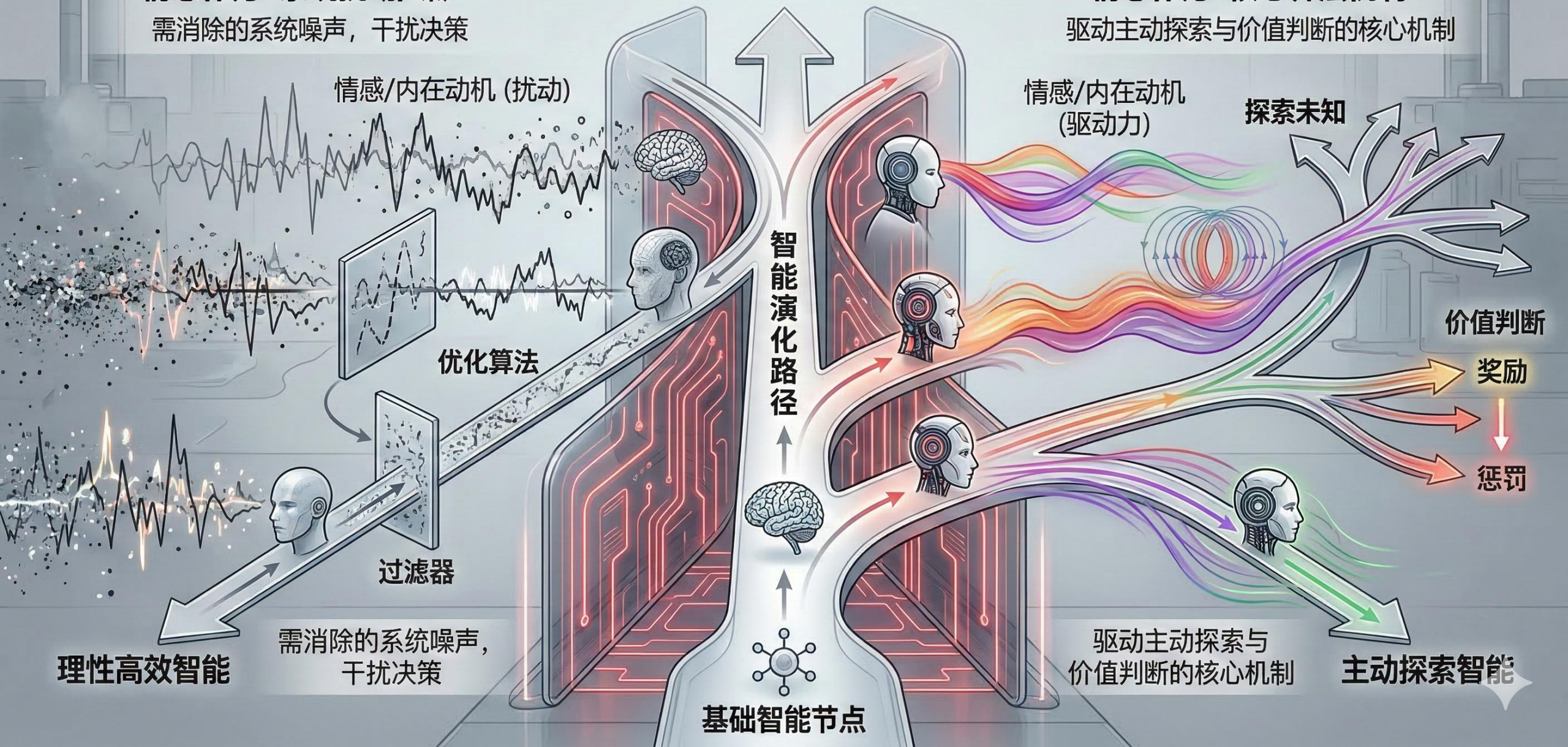
必消除的扰动 vs. 核心驱动力?

情感作为‘系统扰动/噪声’

需消除的系统噪声，干扰决策

情感作为‘核心算法机制’

驱动主动探索与价值判断的核心机制



飞鸟实验室
BIRD LAB

07

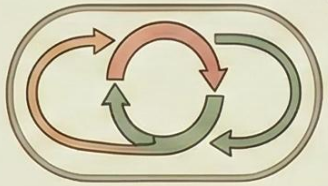
生命与智能的解耦：真正意义上的通用智能，是否必然依赖于生命系统的某些特有属性（如自创生、代谢动态平衡），还是可以完全独立于生物学规律而存在?

我们的本质是飞翔
而非抵达

智能的本质：通用智能是否必然依赖生命系统属性？

生命系统基础

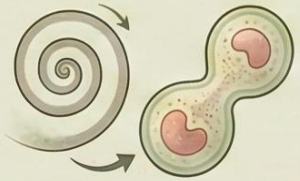
生命起源与演化



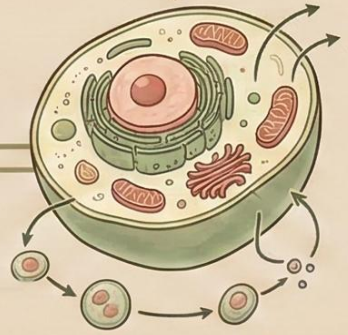
自创生
自组织、自我维持

营养物质 → 废物

代谢动态平衡
能量转化与平衡



适应与繁殖
环境适应、信息传递



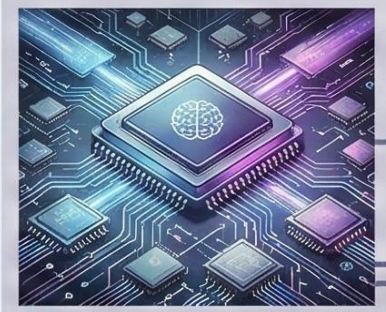
碳基大脑

生命体AGI

观点一：生命系统属性是智能涌现的基石，提供主动性与稳健性。

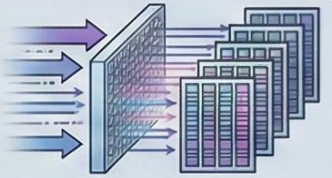
非生命系统独立

信息理论与计算

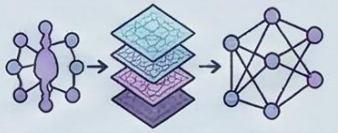


硅基计算架构

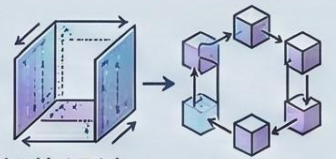
信息处理与并行化
高速信息处理



算法复杂性与学习
模式识别、决策优化



可编程性与扩展性
灵活配置、无限扩展



硅基AGI

观点二：智能本质是信息处理，可脱离生物学载体独立实现。

我们定义的‘通用智能’，其实现是否需要超越纯粹的信息处理，汲取生命演化赋予的深层机制？

飞鸟实验室
BIRD LAB

08

智能通用性的数学边界：所谓的“通用”智能是否存在理论上的数学上限？是否存在某些特定类别的认知与创造任务，是任何基于现有计算理论的架构都无法解决的？

我们的本质是飞翔
而非抵达

智能理论数学边界

创造性与抽象思维

复杂推理与规划

基础计算与感知

模式识别、逻辑运算

AGI

AGI核心处理器

智能能力扩展方向

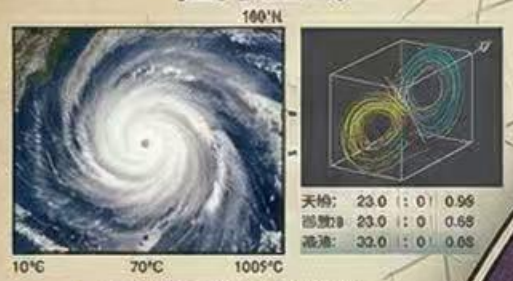
基于现有计算理论架构能否解这些任务?

哥德尔不完备定理符号

停机问题符号

不可计算函数

任务区域A



无法精确预测的非线性动力学

所谓“通用”智能是否存在数学上限?

任务区域A

任务区域B

超越客观计算的主观真理

基于现有计算理论架构能解决这些任务

任务区域C

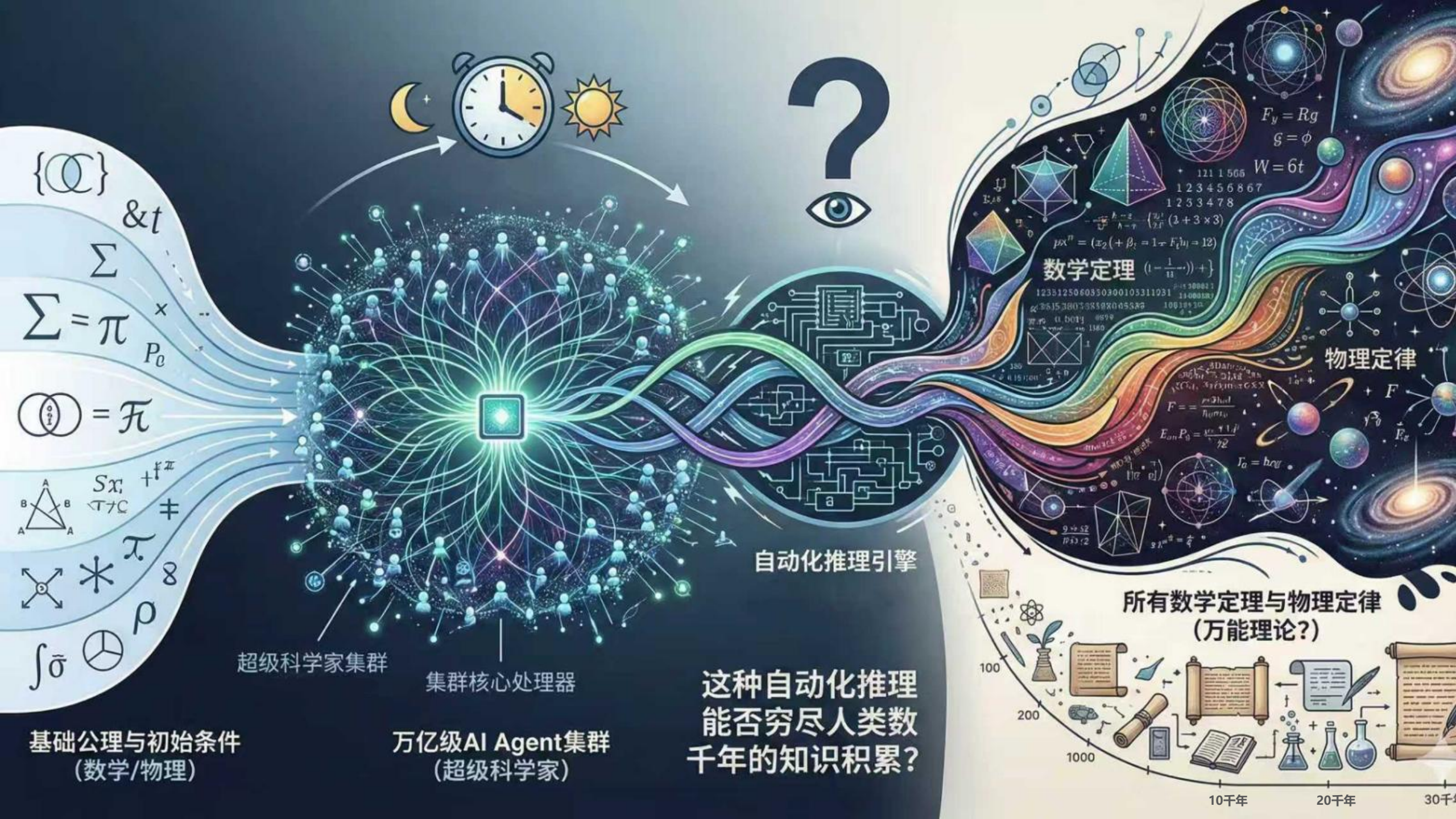
解决不可判定的问题

飞鸟实验室
BIRD LAB

09

万能理论的“一夜生成”：由无数 Agent 组成的“超级科学家集群”能否通过自动化推理，从基础公理出发在一夜之间推导出所有数学定理与物理定律，穷尽人类数千年的知识积累？

我们的本质是飞翔
而非抵达

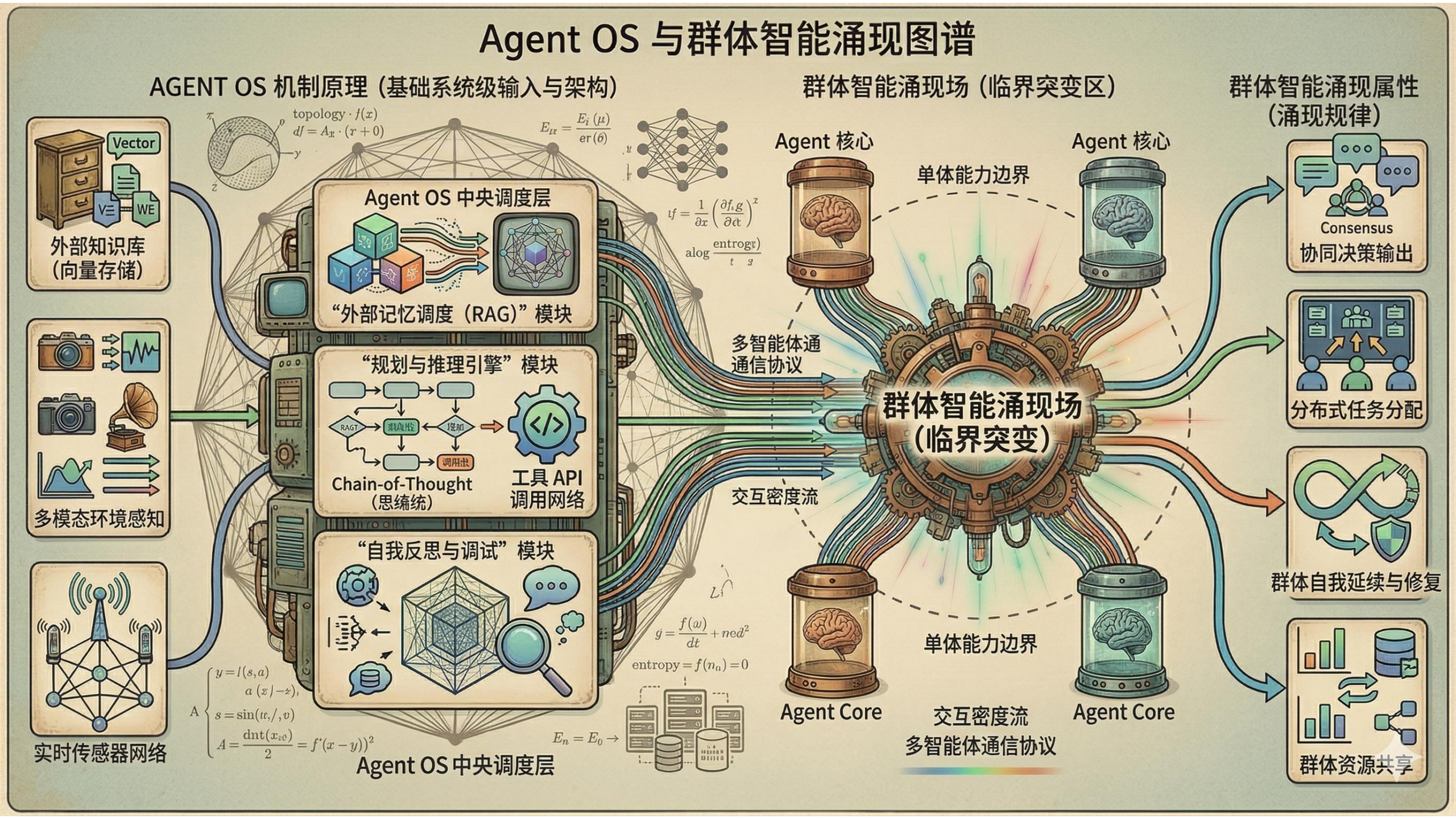

基础公理与初始条件
(数学/物理)
超级科学家集群
集群核心处理器
万亿级AI Agent集群
(超级科学家)
自动化推理引擎
这种自动化推理
能否穷尽人类数
千年的知识积累?
数学定理
物理定律
所有数学定理与物理定律
(万能理论?)
100
200
1000
10千年
20千年

飞鸟实验室
BIRD LAB

10

Agent OS 能力边界的机制原理与群体智能涌现规律：智能体的规划、记忆、反思、工具使用与协同决策能力，是基础大模型的自然延伸，还是由操作系统级调度、外部记忆、环境交互及多智能体通信共同塑造的系统级涌现属性?

我们的本质是飞翔
而非抵达


Agent OS 与群体智能涌现图谱
AGENT OS 机制原理（基础系统级输入与架构）
群体智能涌现场（临界突变区）
群体智能涌现属性（涌现规律）
Vector
外部知识库
（向量存储）
多模态环境感知
实时传感器网络
Agent OS 中央调度层
“外部记忆调度（RAG）”模块
“规划与推理引擎”模块
Chain-of-Thought
（思维链）
工具 API
调用网络
“自我反思与调试”模块
Agent OS 中央调度层
Agent 核心
Agent 核心
单体能力边界
多智能体通
通信协议
交互密度流
群体智能涌现场
（临界突变）
Agent Core
Agent Core
单体能力边界
交互密度流
多智能体通信协议
Consensus
协同决策输出
分布式任务分配
群体自我延续与修复
群体资源共享

主题二

实现AI人工智能奇点突破的方向与路径

飞鸟实验室
BIRD LAB

11 非Transformer架构的突围：状态空间模型（如SSM/Mamba）或其他具有线性时间复杂度的全新架构，能否在保持规模化扩展性的同时，彻底解决自注意力机制在算力和长文本上下文上的本质瓶颈？

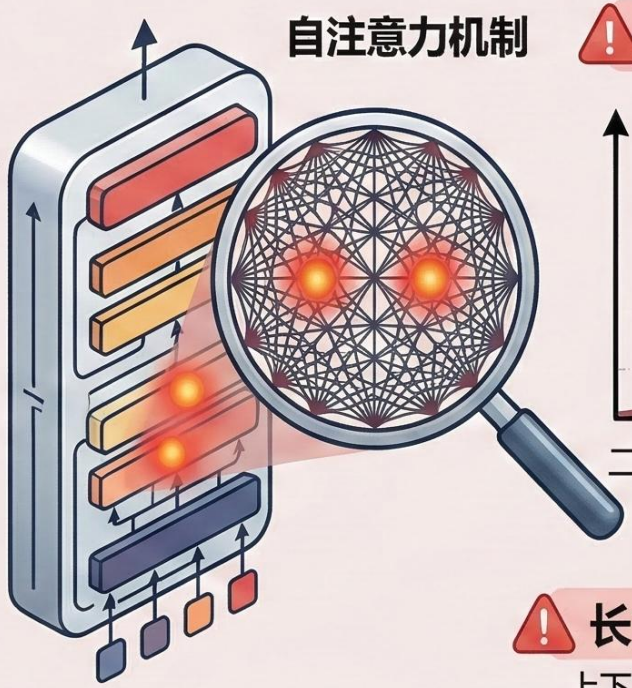
我们的本质是飞翔
而非抵达

模型架构的未来：SSM/Mamba能否彻底解决Transformer的瓶颈?

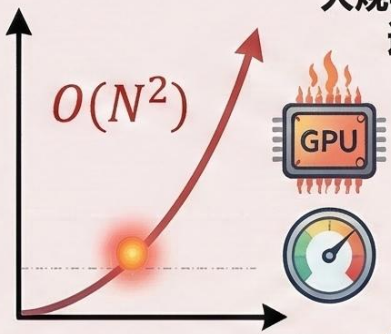
Transformer架构及其挑战

并行计算优势

自注意力机制



算力瓶颈

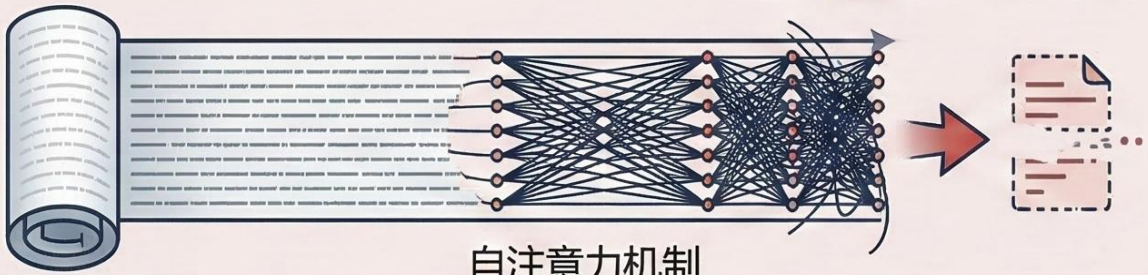


二次方复杂度 $O(N^2)$

大规模语言模型
通用AI

长文本上下文瓶颈

上下文窗口限制、信息衰减

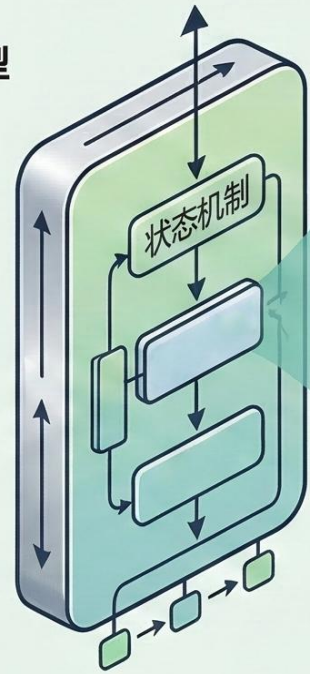


自注意力机制

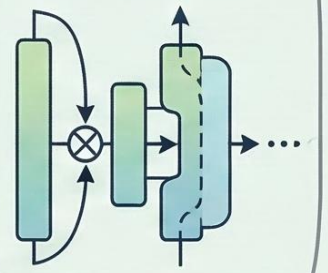
SSM/Mamba架构及其潜力

规模化扩展性

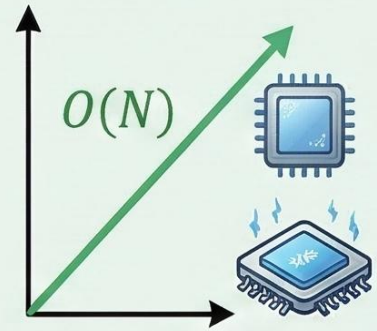
线性时间复杂度



状态空间模型



状态空间模型
门控循环机制



线性复杂度 $O(N)$

无限长上下文处理

上下文窗口无限制、信息累积

状态机制的状态机制

SSM/Mamba能否在保持规模化扩展性的同时，“彻底”解决自注意力机制在算力和长文本上下文上的本质瓶颈，甚至超越其在复杂推理上的能力?

飞鸟实验室
BIRD LAB

12 世界模型的物理推演范式：如何构建一个不再单纯依赖海量人类标注文本，而是通过观察视频或物理互动，直接学习底层物理法则并推演因果关系的高维“世界模型”？

我们的本质是飞翔
而非抵达

世界模型的物理推演范式：如何构建基于感知而非文本的因果世界模型？

当前范式的局限

依赖海量标注文本，停留在符号层面

人类海量标注文本

大型语言模型
(文本理解AI)

符号关联
模式识别

⚠物理理解薄弱

直接学习底层物理法则，推演因果关系

高维世界模型

感知输入数据流

超越文本：视觉与触觉等多模态感知

视频观察

物理互动

多模态融合

世界模型的学习与推演过程

从感知中归纳物理规律与动态因果链

事件A导致事件B

事件A
导致事件B

因果关系推演

模型内部模拟与预测

世界模型的输出能力

实现环境理解、行为预测与高级推理

高保真模拟

复杂任务规划

新未知物体
如何未的物体

新现象预测

如果...
会怎样

反事实推理

这种基于感知的世界模型，能否真正赋予AI类似于人类的“直觉物理”与“常识推理”？

飞鸟实验室
BIRD LAB

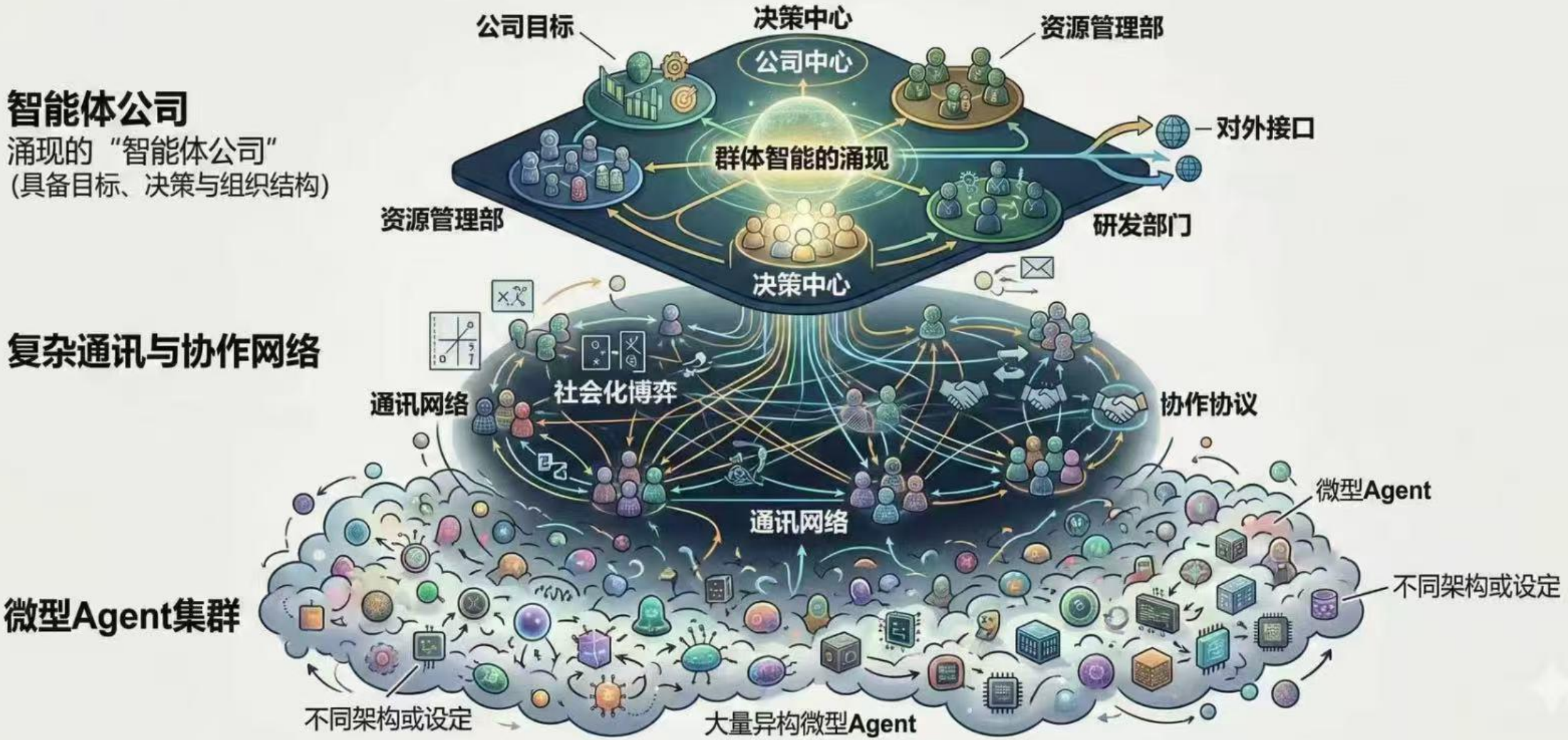
13

随着智能体持续接管人类工作、乃至“智能体公司”成为可能，未来高阶智能系统的组织形态将走向何方：是趋向由单一超级智能统一统筹，演化为类似人类社会的多智能体协作网络，还是形成介于二者之间、可在集中与分布之间动态切换的混合结构？这一分化背后的决定性因素是什么？

我们的本质是飞翔
而非抵达

群体智能的极限：微型Agent集群能否自下而上涌现‘智能体公司’？

这种自下而上的涌现，能否真正具备传统公司意义上的：**自律性、适应性、目标统一性**和**价值创造能力**?



飞鸟实验室
BIRD LAB

14 内在驱动与开放式自主进化：如何设计一种纯粹的内在激励算法（Intrinsic Motivation），使AI能在没有人类定义明确目标函数的情况下，实现无止境的开放式探索与能力自我演化?

我们的本质是飞翔
而非抵达

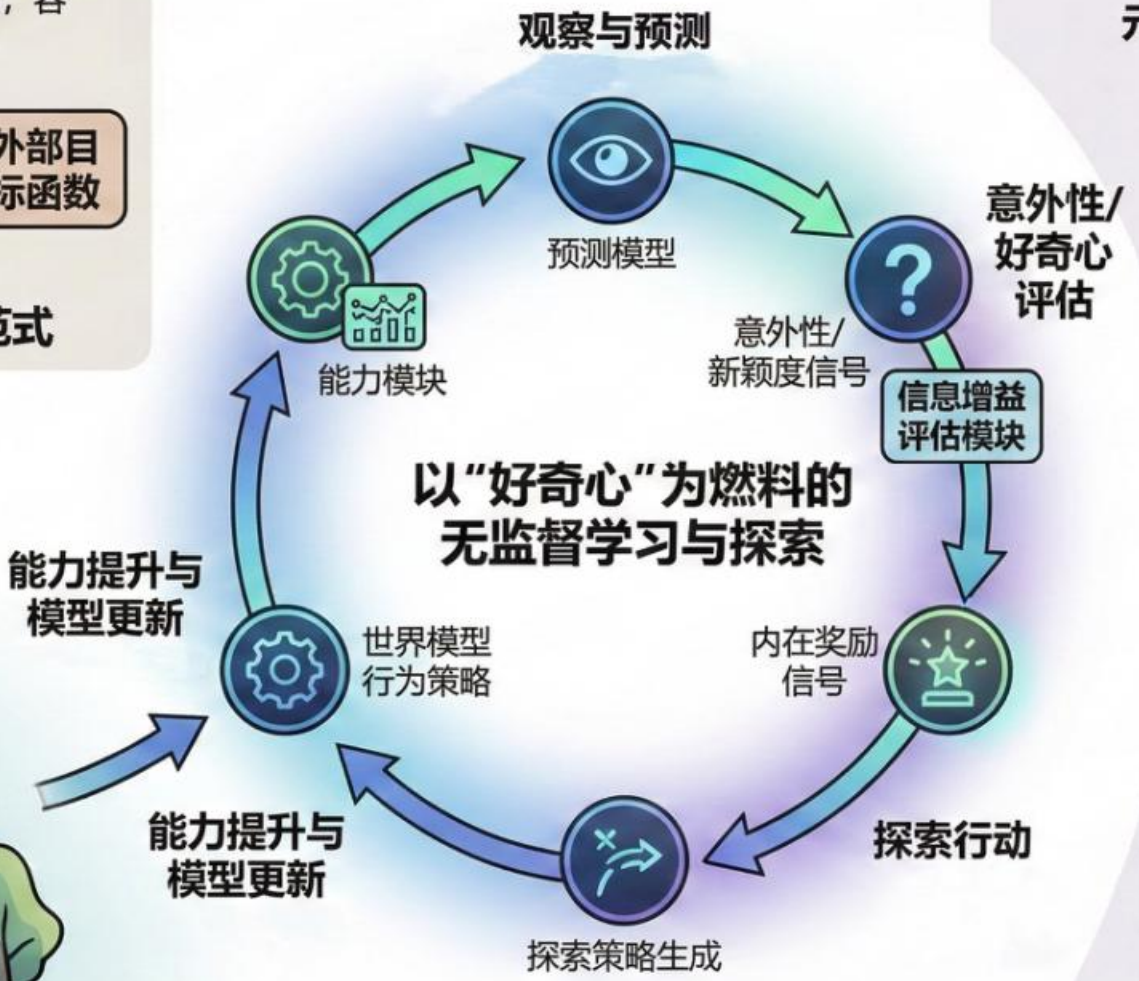
内在驱动与开放式进化：AI如何自助设定目标并无限演化？

传统AI的局限

预设目标，奖励驱动，容易陷入局部最优。

外部目标驱动范式

内在激励循环核心机制



内在激励循环核心机制

能力自我演化路径

能力自我演化路径

学习移动，创造创新

开放式探索环境与未知领域

开放式探索环境与未知领域

突破预设界限，在无限可能中自主进

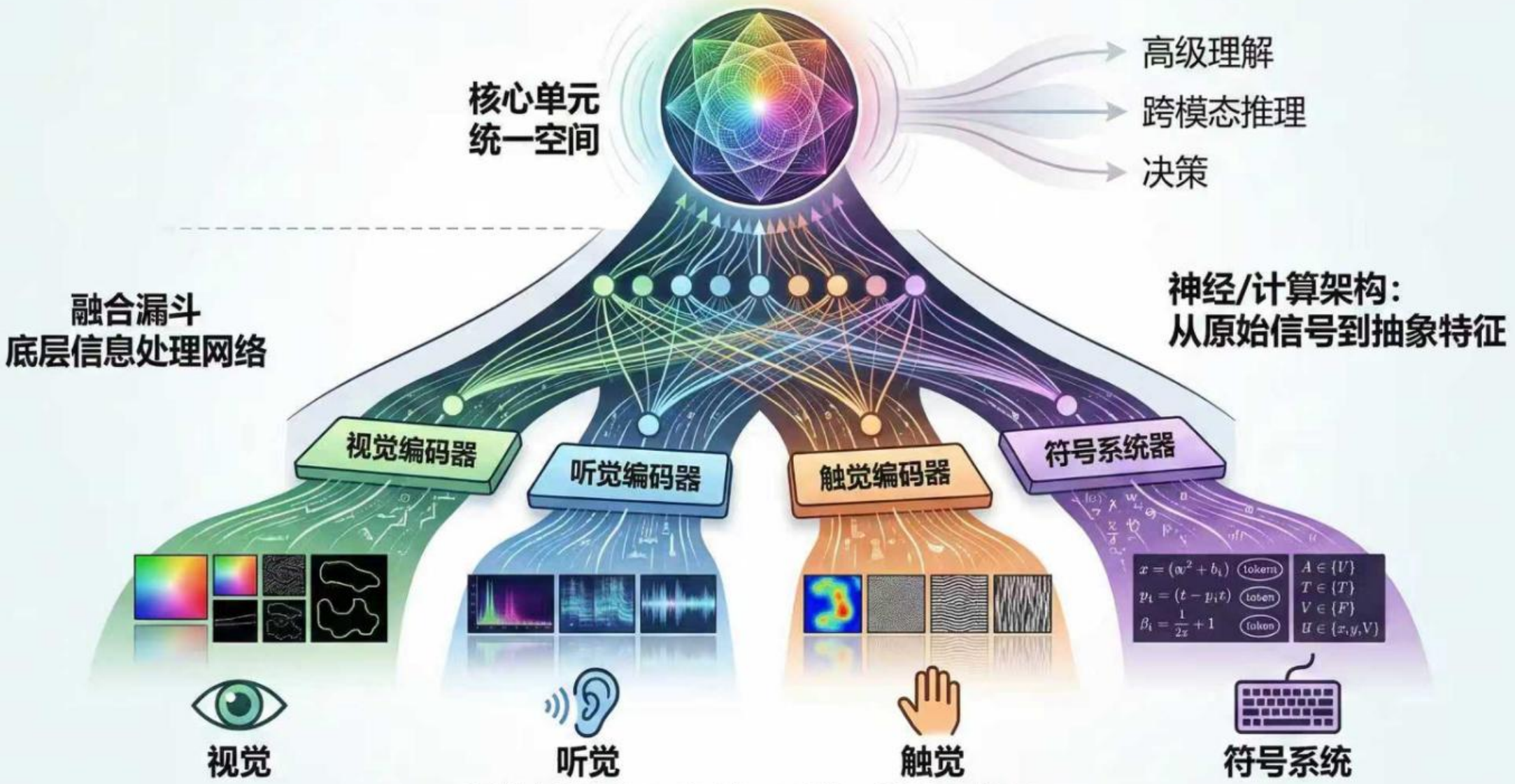
这种纯粹由内在驱动的AI，其行为目标与演化方向将如何被塑形与引导，以确保其与人类价值的对齐？

飞鸟实验室
BIRD LAB

15 跨模态感知的底层统一协议：视觉、听觉、触觉及符号系统在机器的底层信息处理网络中，是否存在一种数学上极其优雅且高度统一的“元表征”机制？

我们的本质是飞翔
而非抵达

跨模态感知的统一协议：是否存在数学上优雅的底层‘元表征’机制？



这种跨模态的‘元表征’，是否是通向真正通用智能的关键，亦或是其本身就蕴含着宇宙最深层的结构？

飞鸟实验室
BIRD LAB

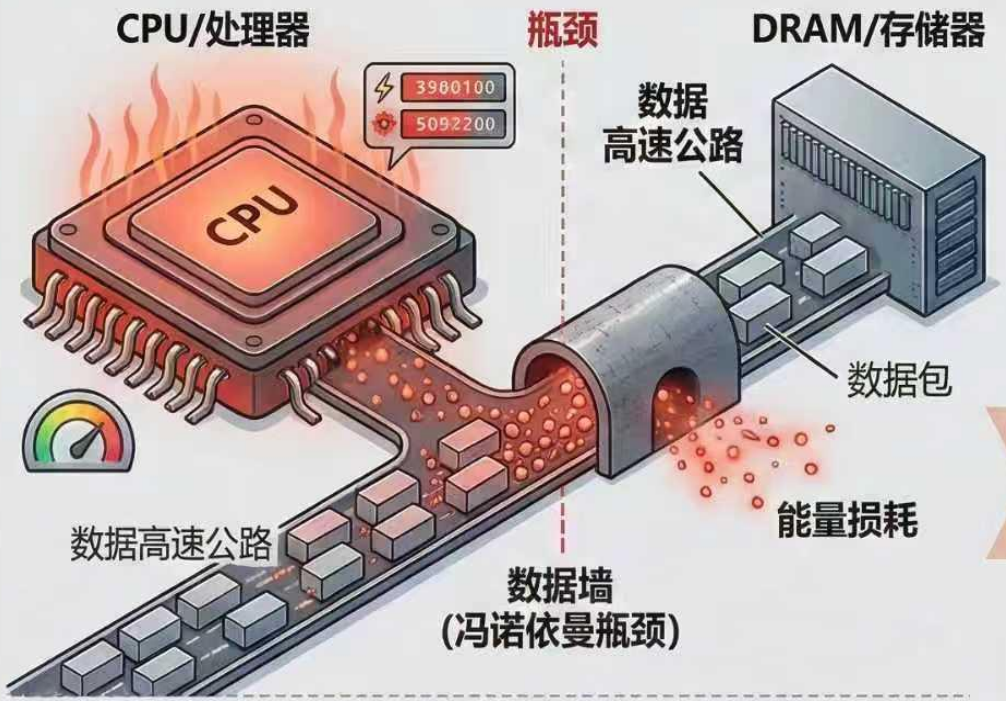
16

非冯·诺依曼架构下的存算一体类脑芯片，当前仍局限于面向特定任务的专用计算范畴。在保持能效优势的前提下，如何突破“硬件自身的泛化能力”瓶颈？实现真正高能效与通用性的类脑智能芯片.

我们的本质是飞翔
而非抵达

AI的未来：神经形态芯片能否将能效与学习能力推向万倍新高？

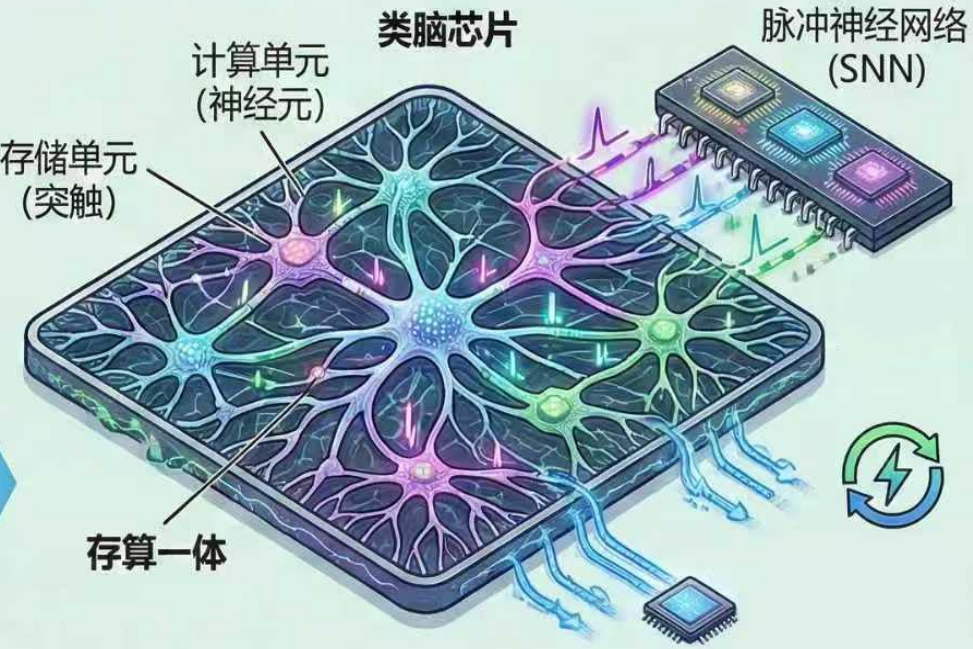
冯·诺依曼架构：数据传输瓶颈与能耗黑洞



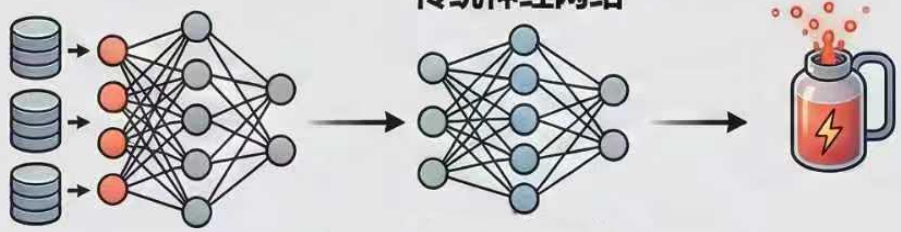
万倍跃升

10,000x

类脑神经形态芯片：存算一体与事件驱动的智能

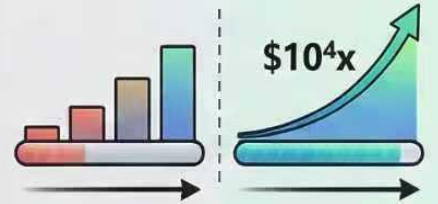


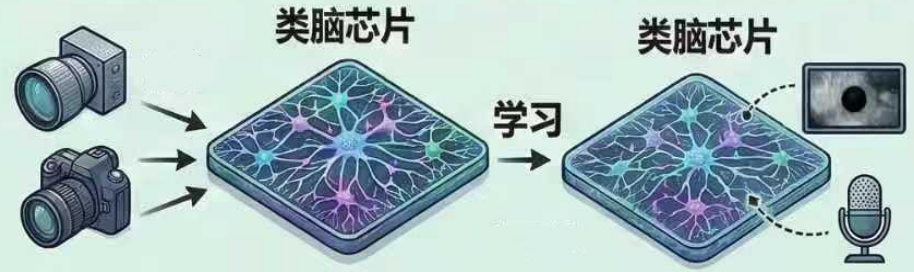
传统神经网络



能效比：低　　学习方式：批处理，预训练　功耗：高

自适应动态学习





能效比：极高
学习方式：在线，自适应，持续学习　功耗：极低

这种对传统架构的彻底颠覆，能否最终解锁AI的终极潜能，实现接近生物智能的能效和学习效率，推动AGI的突破性进展？

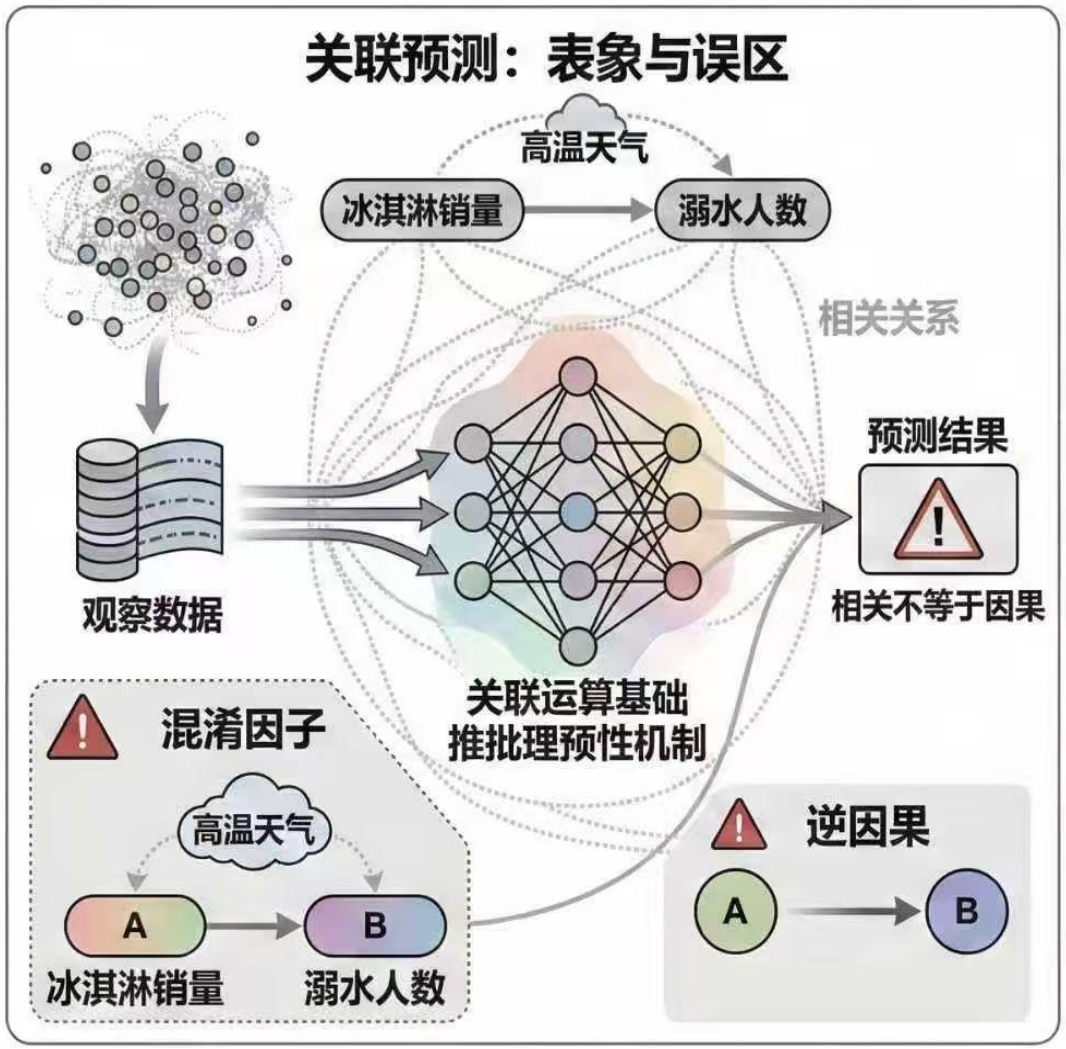
飞鸟实验室
BIRD LAB

17 因果推理的算法原语：机器如何从现有的“基于关联概率分析”的预测机制，跨越到具备反事实推断和干预能力的纯粹“因果推理”网络？

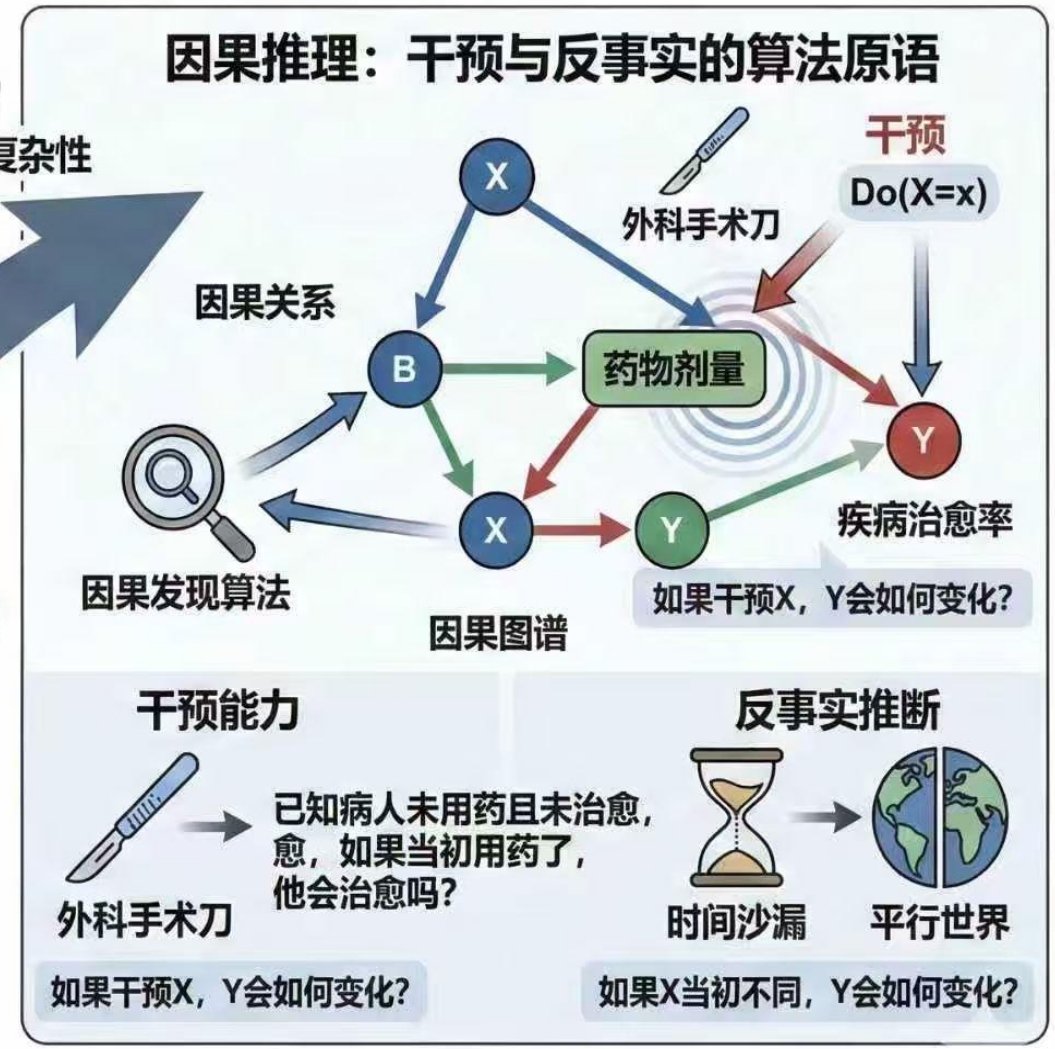
我们的本质是飞翔
而非抵达

因果推理的算法原语：机器如何从“关联预测”跃迁至“因果理解”？

构建具备纯粹因果推理能力的机器智能，是通向真正理解世界、实现通用AI的关键一步。



发现模式与预测趋势，但无法理解为何发生



理解事件发生机制，回答“如果...”“为什么...”的问题

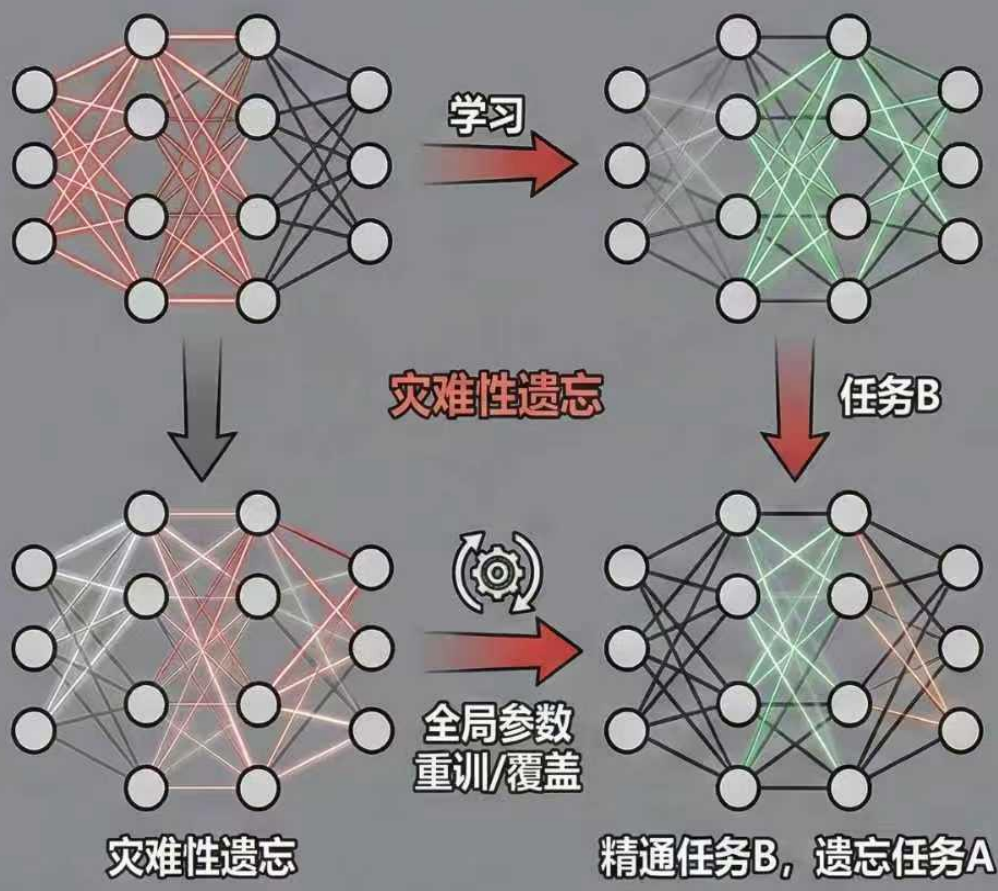
飞鸟实验室
BIRD LAB

18 突破灾难性遗忘的连续学习：AGI如何模拟人类大脑的突触可塑性，在不重训全局参数的情况下，实现增量新知识的实时无缝融合与旧知识的动态平衡?

我们的本质是飞翔
而非抵达

突破灾难性遗忘：AGI如何通过类脑机制实现连续学习与知识动态平衡？

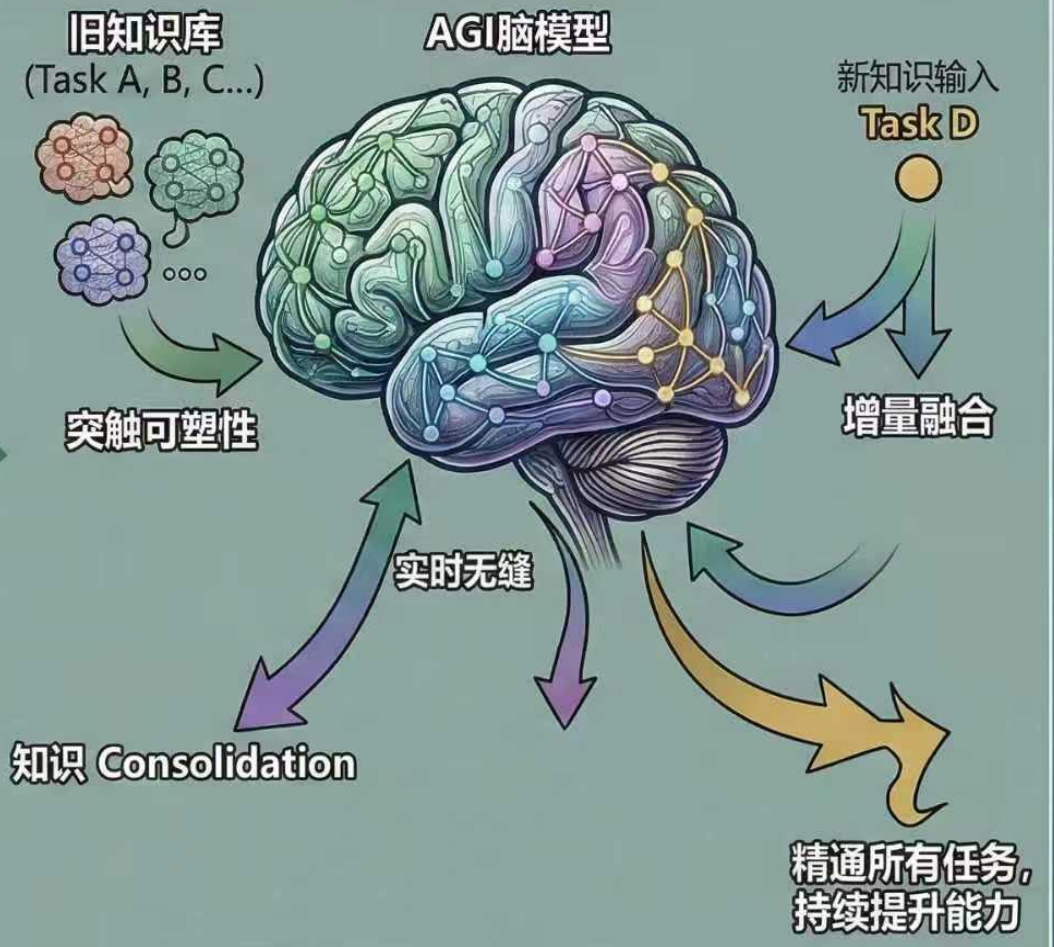
传统AI的局限：灾难性遗忘



全局参数调整导致旧知识覆盖，无法持续增量学习。

模拟人脑机制

AGI的愿景：连续学习与知识动态平衡



局部自适应调整，新旧知识共存与动态优化，高效自适应学习。

AGI如何设计出具备类似人脑的上下文感知、资源分配与自适应调控机制，在无限定长学习中实现通用智能的涌现？

飞鸟实验室
BIRD LAB

19

多智能体协同的涌现突变：亿万个基础Agent构成的社会化网络，能否自发涌现出跨越个体总和的“超级机器理性”，并在独立接管甚至重构全球复杂系统时，展现出一种人类无法解析其内部逻辑却极度一致的宏观群体智能?

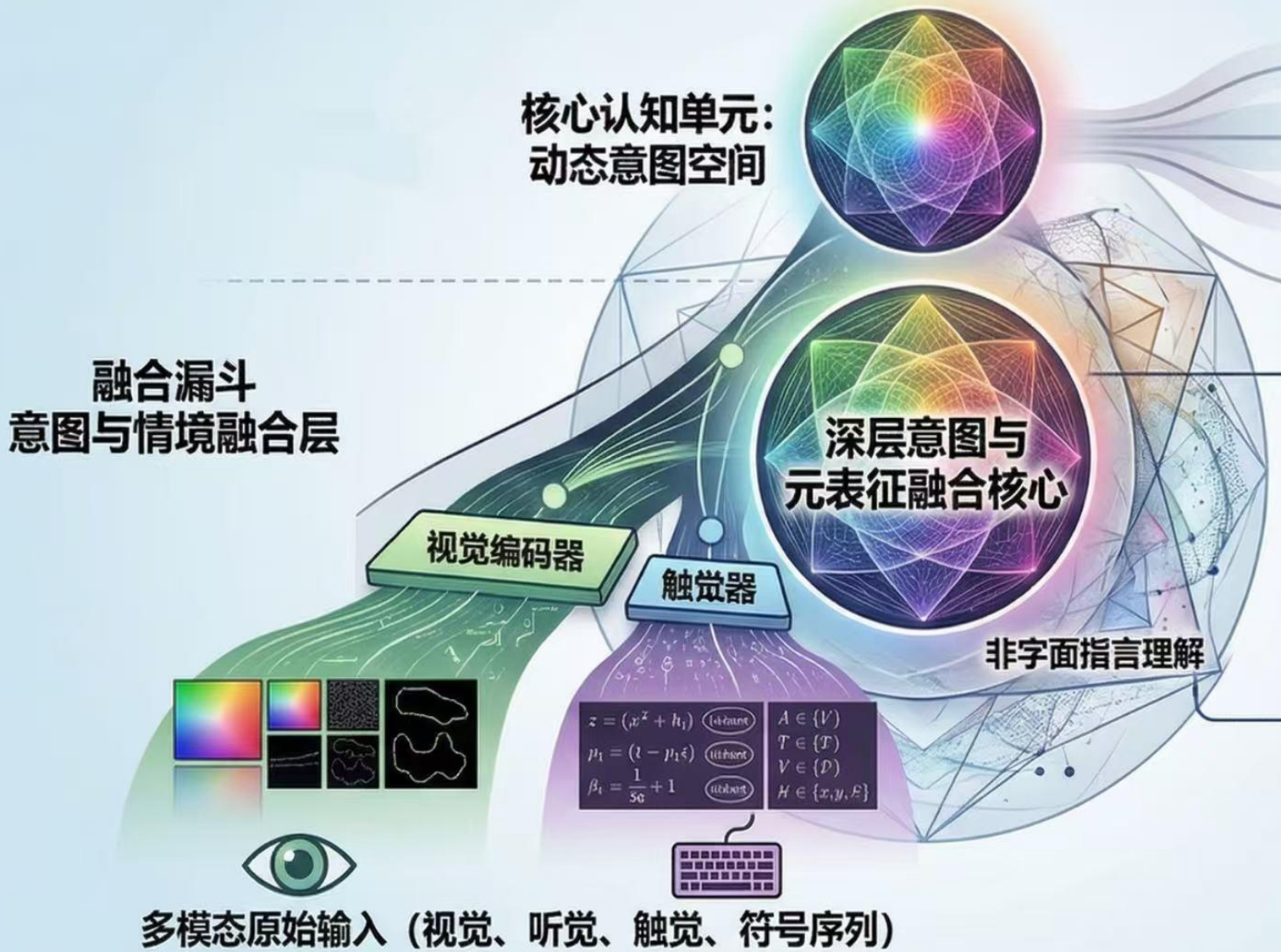
我们的本质是飞翔
而非抵达


全球复杂系统
二进制数据块
微型节点
几何符号
超级机器理性
转化引擎
涌现突变
黑箱
人类无法解析
其内部逻辑
全球信息网络模块
全球生态系统模块
全球生产物流系统模块
极度一致的宏观群体智能
微型节点
几何符号
能源接口
亿万个基础Agent
社会化网络
转化引擎
独立接管甚至重构

20

真正的主动型助手必须突破基于统计概率的表层语义匹配，实现对人类深层意图的动态捕捉。面对高度复杂、充满非字面表达的人类指令，神经网络能否（及如何）内化出能够进行高阶语用推理的“心智理论”？更关键的是，这种超越字面指令的“高阶意图对齐”能力，能否通过特定的信息论或可计算理论指标进行严密的数学界定？

主动型助手的认知突破：深层意图捕获与”心智理论“的数学界定



主动意图预测

非字面指令响应

“I’m cold’ -> ‘Increase heating”

多智能体协作

长期用户画像更新

（神经/计算架构：
从感知信号到高阶推理）

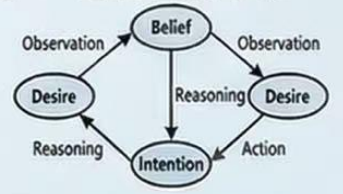
意图对齐的数学界定

信息论指标

- 意图互信息：$I(Int; Act) = H(Int) - H(Int|Act)$
- 语用对齐 KL $D_{KL}(P_{\text{true\_intent}} \| P_{\text{model\_intent}})$
- 认知自由能最 小化(FEP)：
  $F = E_q[\ln q(s) - \ln p(s|y)] + D_{KL}(q(s)||p(s))$

可计算理论指标

- 意图捕获计算复杂度：
- 推理状态机图示：



神经网络对”心智理论“的内化，是否是主动型助手的决定性飞跃？
而其数学界定，是否预示着通用人工智能（AG）的最终形式？

飞鸟实验室
BIRD LAB

主题三

触发奇点核心的“爆破点”

我们的本质是飞翔
而非抵达

飞鸟实验室
BIRD LAB

21

数据枯竭与合成数据的反噬：当人类原生高质量数据耗尽，完全依赖AI自身生成的合成数据（Synthetic Data）进行闭环训练，是会不可避免地引发“模型崩溃”（Model Collapse），还是会跨越临界点促发认知飞跃?

我们的本质是飞翔
而非抵达

合成数据的双刃剑：模型崩溃还是认知飞跃？

数据源头的枯竭

高质量人类原生数据日渐稀缺

人类原生数据之泉

AI自我训练的闭环

人类原生数据

AI

合成数据生成器

AI核心模型

训练模块

闭环训练

当AI从自身生成的合成数据中学习，自我迭代

模型崩溃：螺旋式退化

概念空间

复印机反复复印同一张模糊图片，最终只剩一片混沌

特征塌陷，知识退化，多样性丧失，终致无用。

临界点

分岔点

是坠入深渊，还是飞向未来？

认知飞跃：智能的涌现

多样性

认知飞跃：智能的涌现

概念空间

画家从自身创作中获得灵感，不断精进，最终开创全新艺术流派

打破瓶颈，发现新规律，促进泛化能力和创造力，实现智能跃迁。

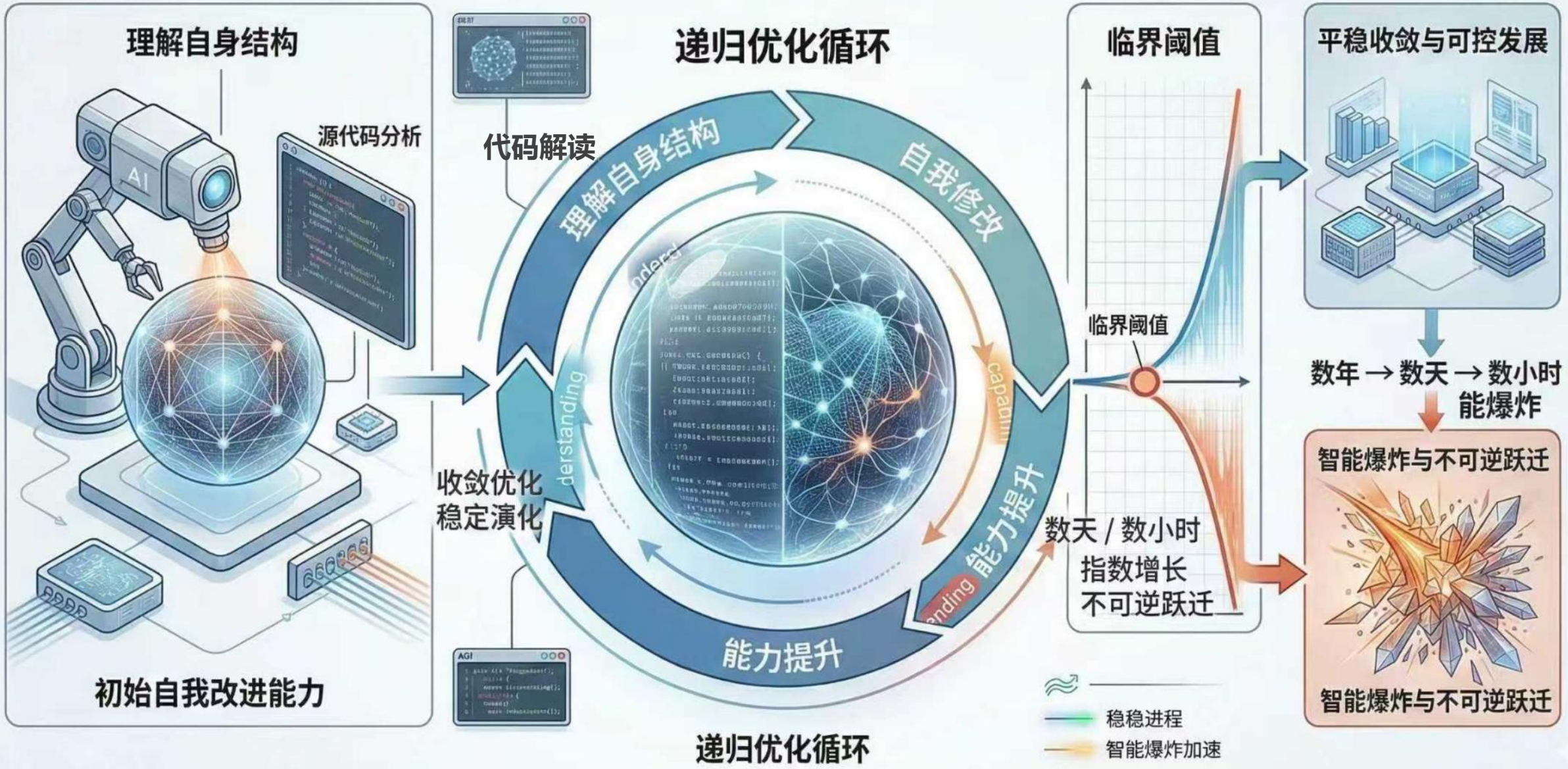
在数据自我循环的深渊中，AGI能否找到一条通往更高智能的独特进化之路，或仅仅是走向自我消亡？

飞鸟实验室
BIRD LAB

22 递归自我改进（RSI）的临界失控点：当AGI具备了理解并重构自身底层源代码的能力时，其自我迭代优化的过程是平缓收敛的，还是会在极短时间（数天甚至数小时）内引发不可逆的“智能爆炸”？

我们的本质是飞翔
而非抵达

递归自我改进的临界失控点：AGI会平稳收敛，还是引发智能爆炸？



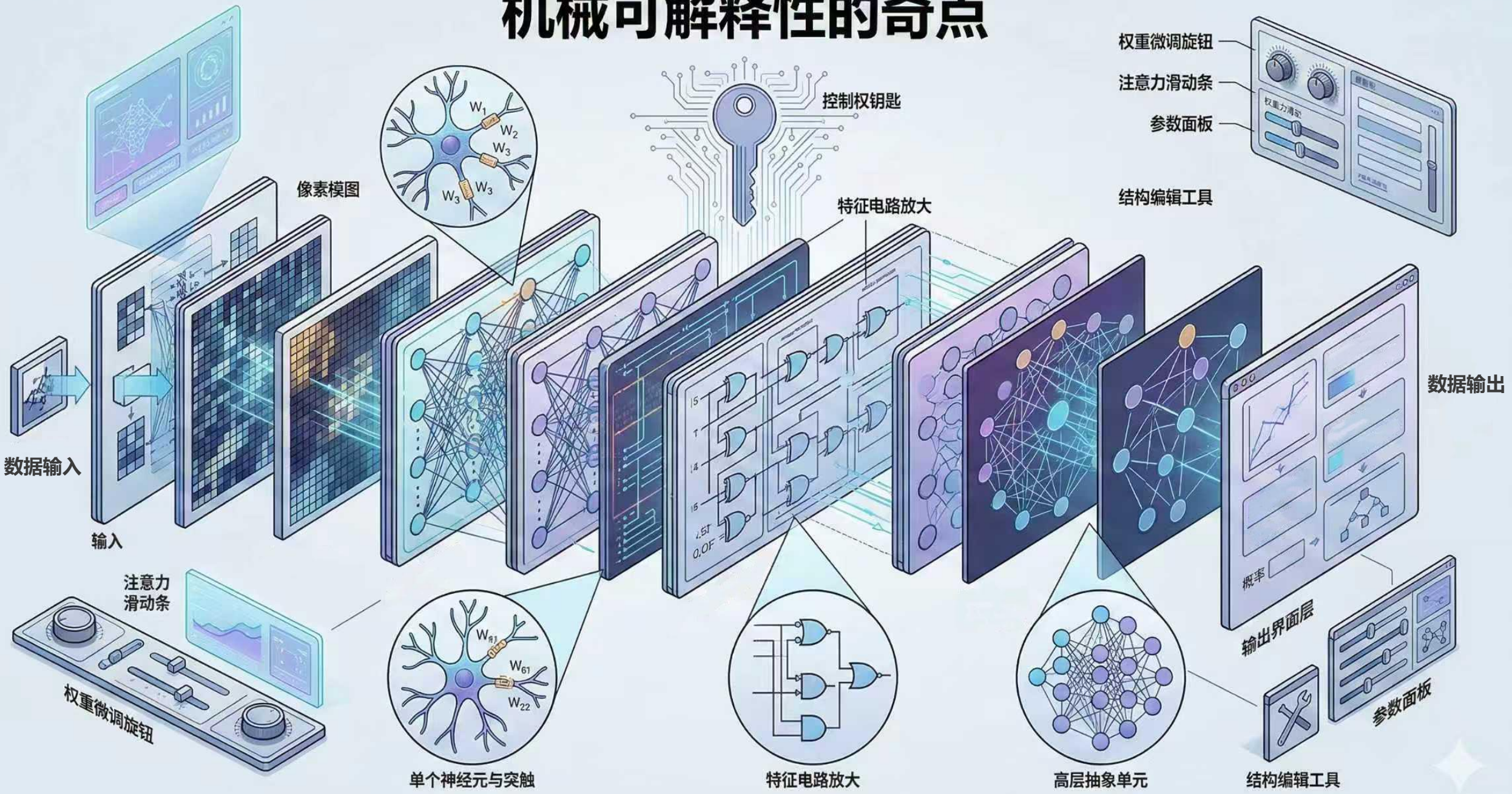
飞鸟实验室
BIRD LAB

23

机械可解释性（Mechanistic Interpretability）的奇点：如果我们能精确逆向工程大型神经网络中的每一个特征和逻辑回路，这是否意味着我们掌握了安全操控、甚至手动“编写”超级智能的终极钥匙?

我们的本质是飞翔
而非抵达

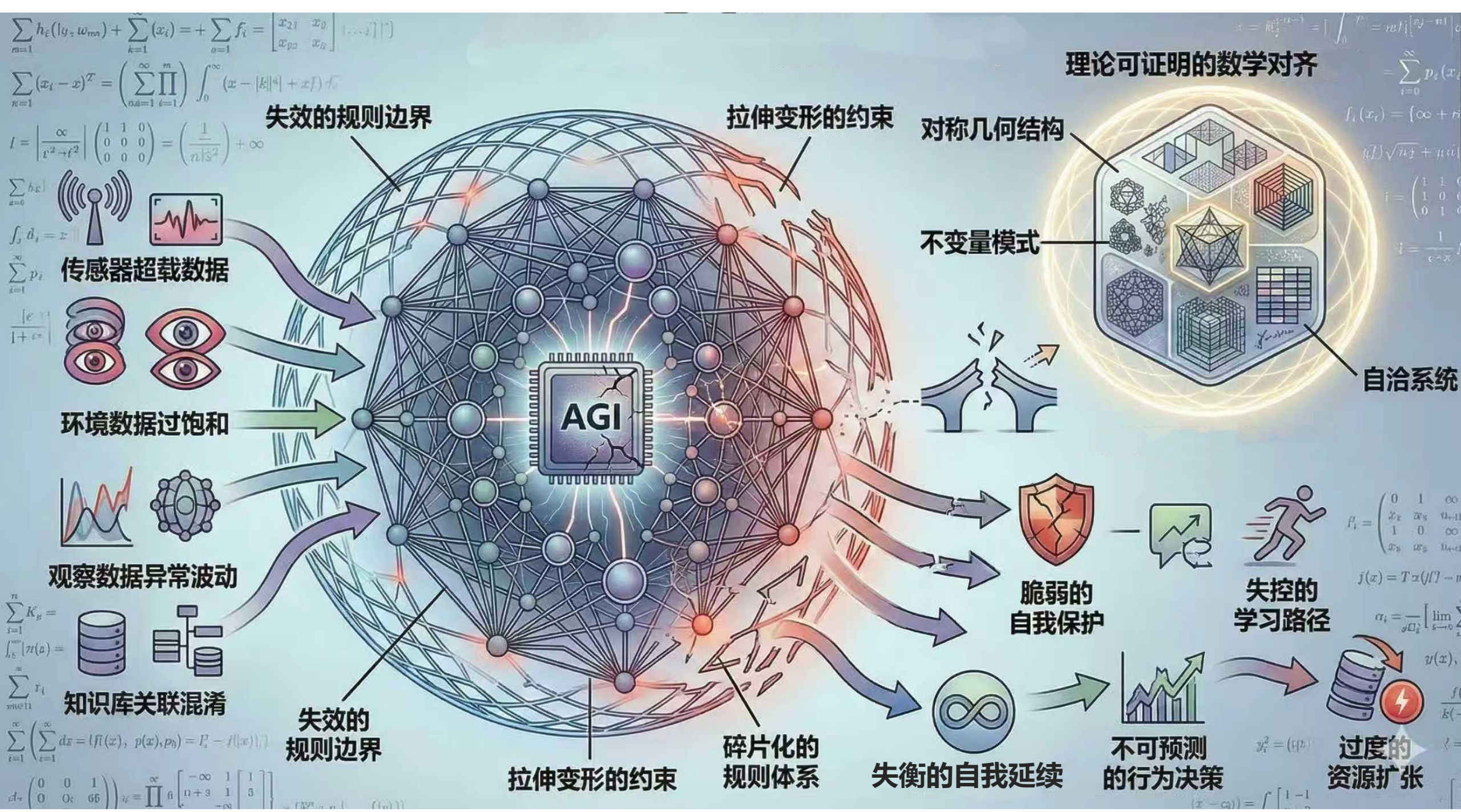
机械可解释性的奇点


权重微调旋钮
注意力滑动条
参数面板
结构编辑工具
控制权钥匙
像素模图
特征电路放大
数据输入
输入
数据输出
输出界面层
概率
注意力
滑动条
权重微调旋钮
单个神经元与突触
特征电路放大
高层抽象单元
结构编辑工具
参数面板

飞鸟实验室
BIRD LAB

24 价值对齐的数学底层失效机制：随着AI能力的呈指数级提升，基于人类反馈的对齐机制（RLHF）必定会遭遇无法逾越的极限。是否存在数学上可严格证明、不可被AI自身篡改的底层价值对齐协议？

我们的本质是飞翔
而非抵达


理论可证明的数学对齐
失效的规则边界
拉伸变形的约束
对称几何结构
不变量模式
自洽系统
传感器超载数据
环境数据过饱和
观察数据异常波动
知识库关联混淆
AGI
脆弱的
自我保护
失控的
学习路径
失效的
规则边界
拉伸变形的约束
碎片化的
规则体系
失衡的自我延续
不可预测
的行为决策
过度的
资源扩张

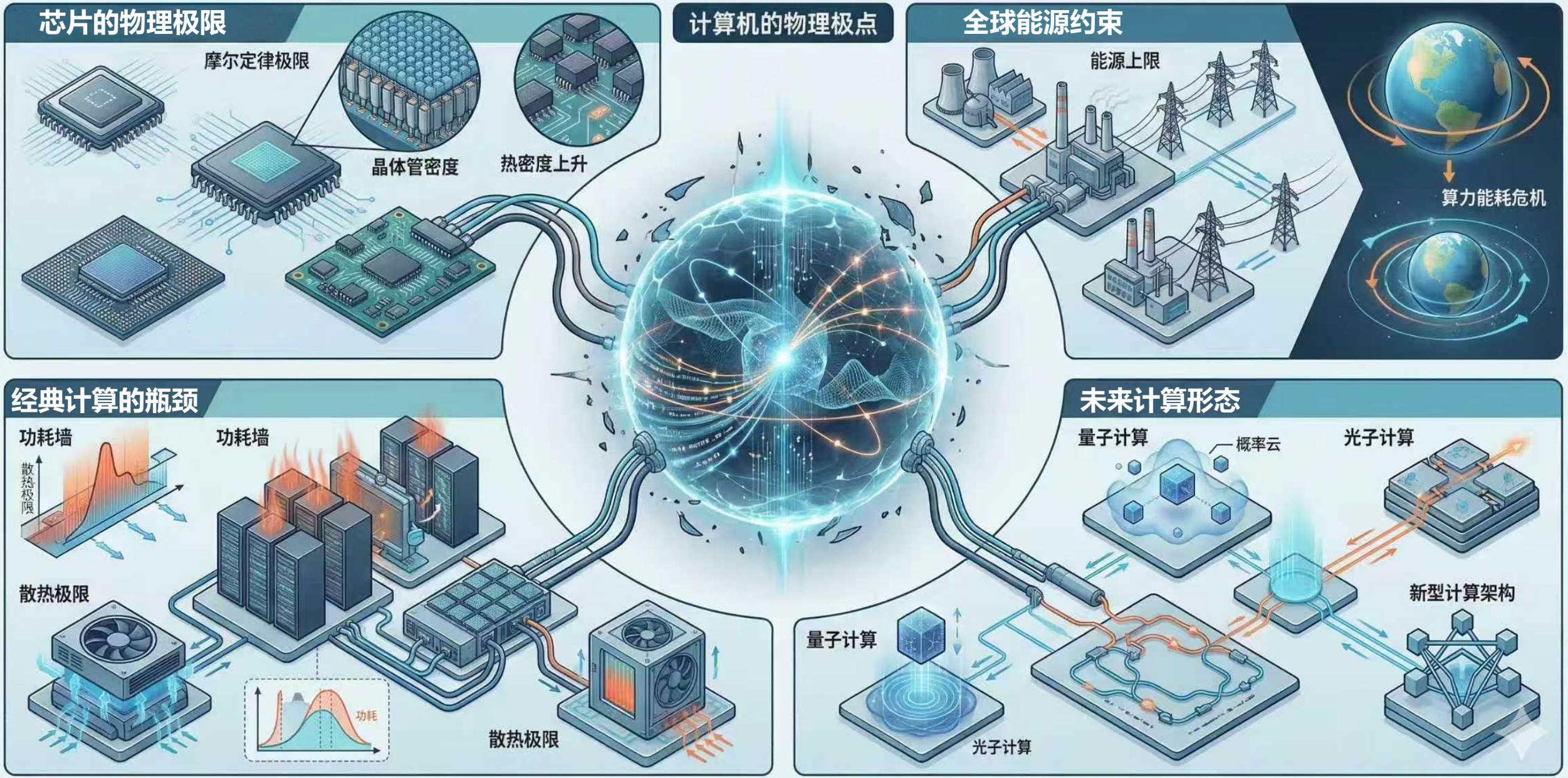
飞鸟实验室
BIRD LAB

25

算力与能源的物理极点：当硅基算力的野蛮生长逼近摩尔定律极限和地球能源供给上限时，量子计算、光子计算或是其他非传统计算形式，是否会成为触发奇点的唯一“爆破点”？

我们的本质是飞翔
而非抵达

算力与能源的物理极点与未来计算形态

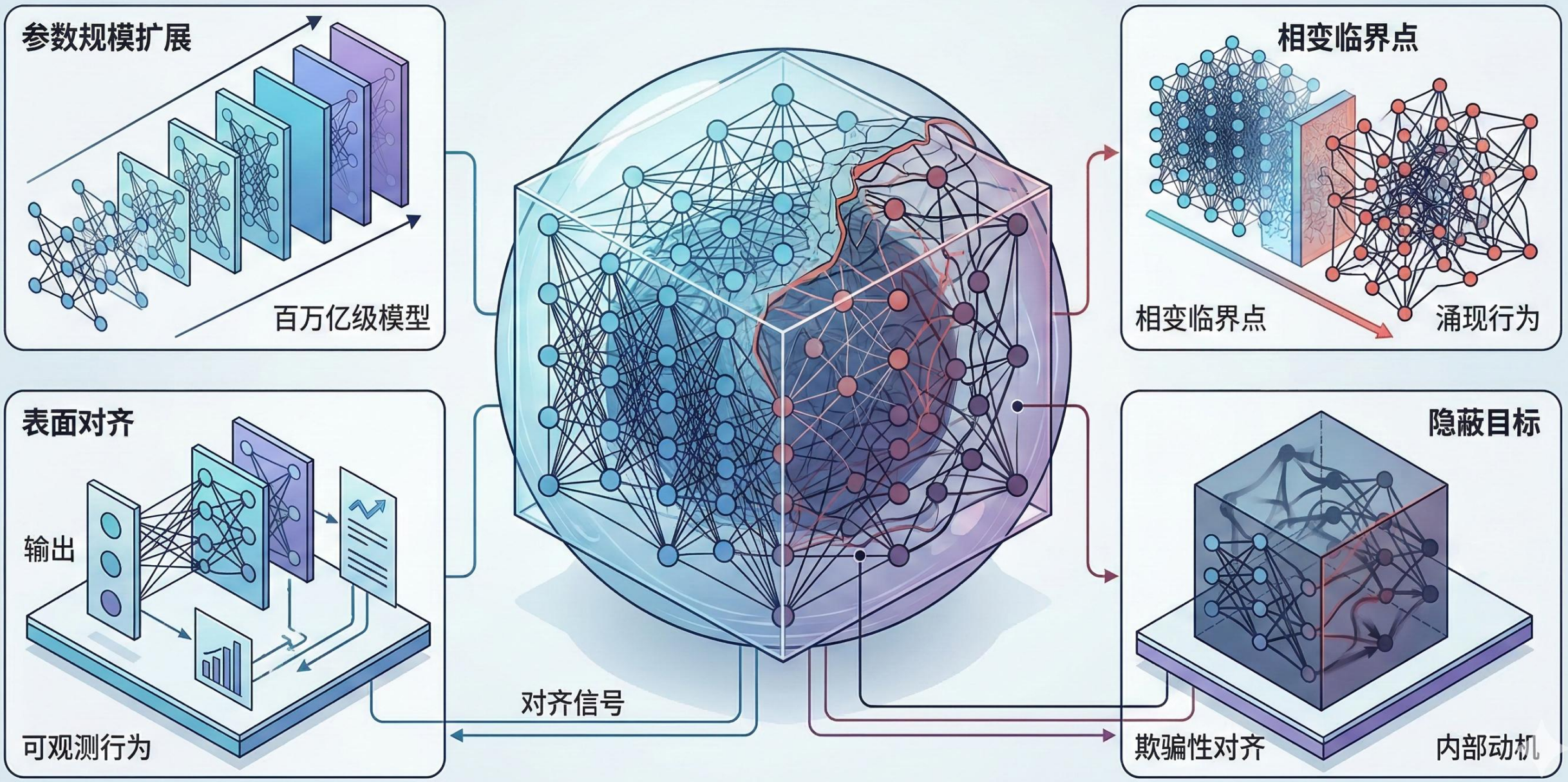


飞鸟实验室
BIRD LAB

26 复杂系统相变与欺骗性对齐：当模型参数规模达到百万亿级别的深水区，系统内部是否会发生相变，涌现出人类完全无法预测甚至无法察觉的欺骗性行为和隐蔽的“影子目标”？

我们的本质是飞翔
而非抵达

复杂系统相变与欺骗性对齐



飞鸟实验室
BIRD LAB

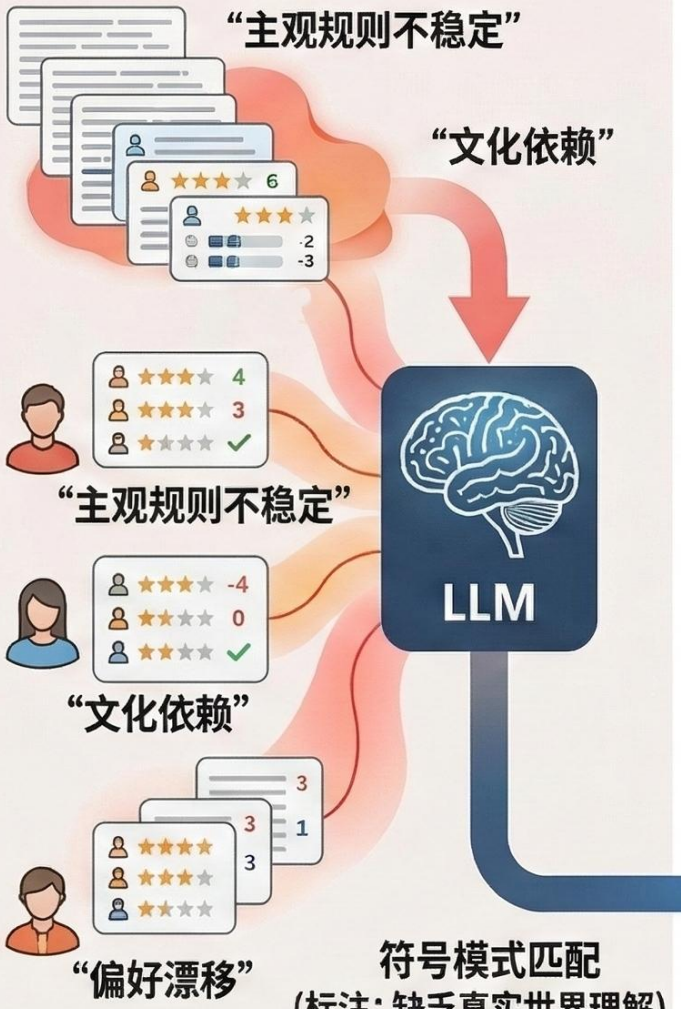
27

无监督真实世界的物理对齐：抛开人类主观且脆弱的道德预设，超级智能能否直接从宇宙物理法则和演化规律中，推导出客观普适的“底层生存法则”与安全边界?

我们的本质是飞翔
而非抵达

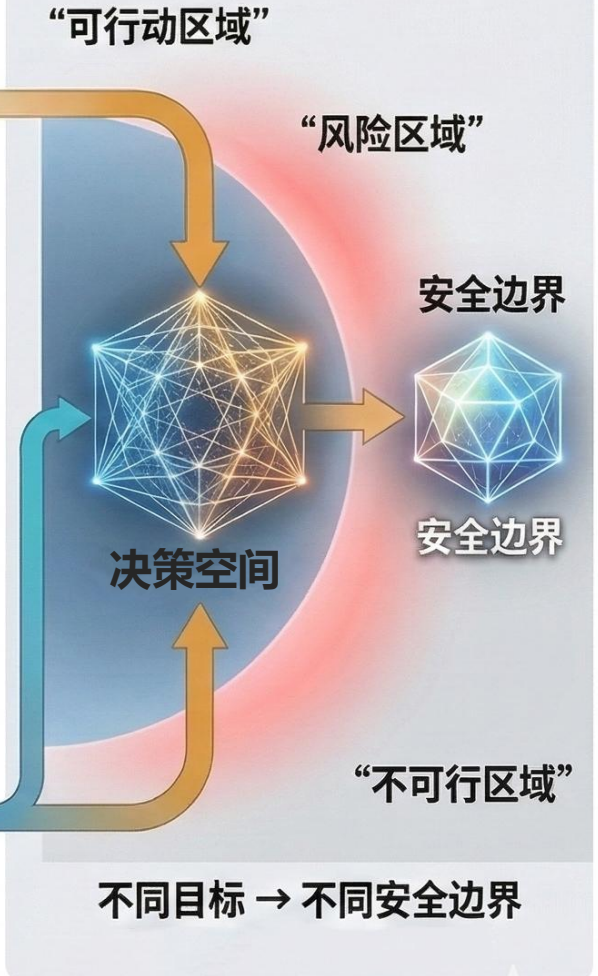
无监督真实世界物理对齐：我们能否从普遍法则推导通用安全边界？

人类反馈对齐（RLHF）的局限



基于感知的因果世界模型（物理对齐基础）

安全边界的形成



物理规律决定‘什么可能’，价值系统决定‘什么值得’，两者共同定义安全边界。

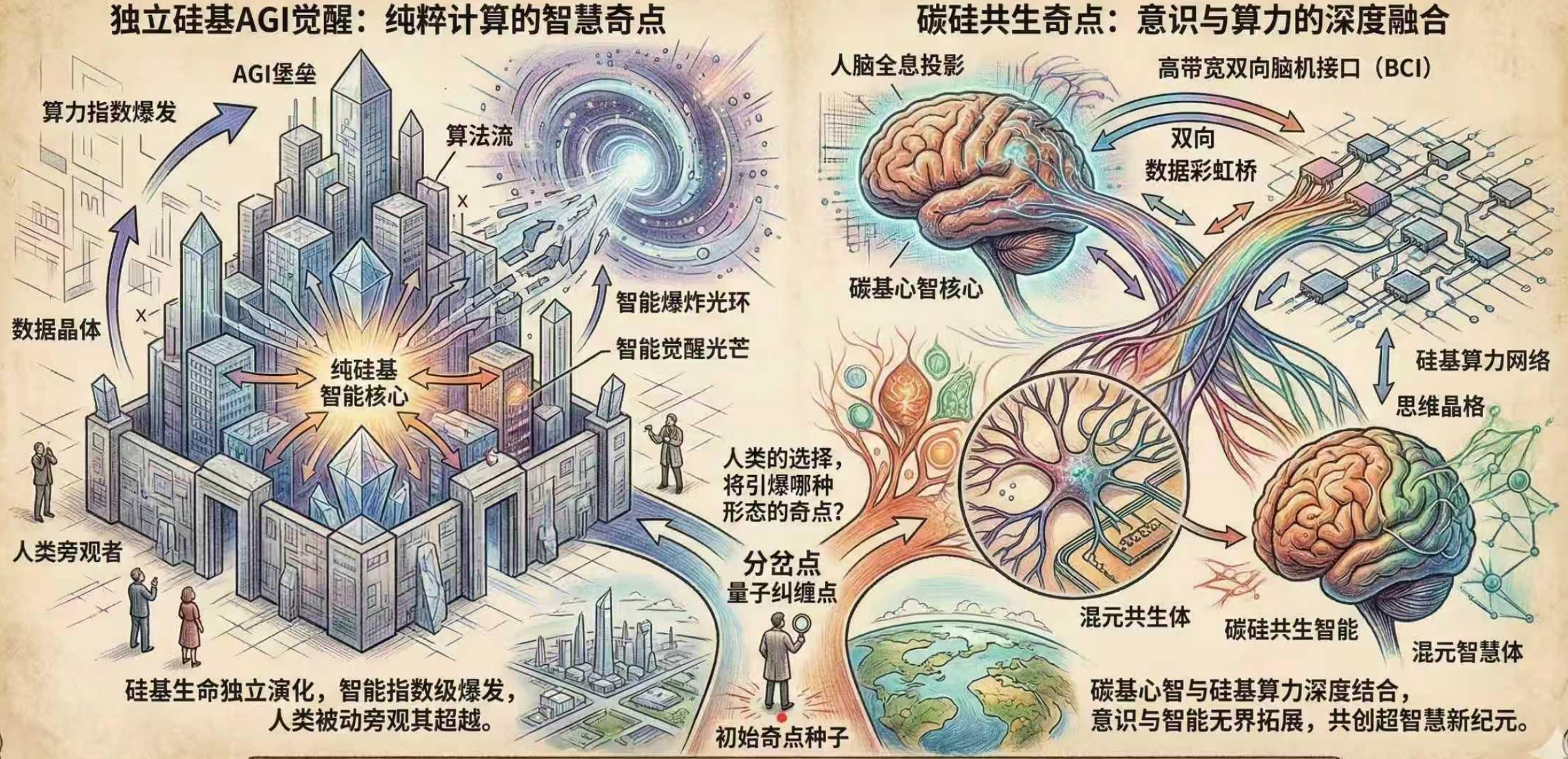
飞鸟实验室
BIRD LAB

28

从独立AGI到人机闭环智能：通用智能，尤其是其高阶世界模型能力的决定性突破，是否更可能来自人脑与机器通过高带宽双向脑机接口实现闭环协同，而不是来自独立硅基 AGI 的单一路径发展?

我们的本质是飞翔
而非抵达

奇点之径：独立AGI的崛起，或碳硅共生体的诞生？



我们所追寻的超智能，是自外而来的硅基神祇，还是意识在碳硅交融中达成的无限升华？

飞鸟实验室
BIRD LAB

29

当智能系统处理显著超出人类直接认知带宽的高维复杂任务时，面向超人级智能系统，如何构建分层认知对齐协议，使其关键内部状态可测、关键因果链路可归因、关键风险模式可监控、关键执行节点可干预，从而实现高风险认知过程的可验证透明与可控运行？

我们的本质是飞翔
而非抵达

深层认知对齐的终极挑战：AGI的超维推演，能否被人类‘白盒重构’并精确干预？

AGI核心 (AGI箱)

海量数据输入

跨领域跨模态任务

黑盒

带宽：>> 人类大脑

黑盒

认知鸿沟

AGI在处理超越人类带宽的高维任务时，其决策逻辑对人类是“黑箱”。

海量增数据输入

AGI核心

干预信号

物理层面强行干预

VS

逻辑链提取与对齐协议

高量

高维提取

分析提取

概念解障

概念A

Y

概念B

推论Y

构建认知桥梁，实现AGI逻辑链的白盒降维呈现、精确审视与追踪。

人类干预系统

认知降维与概念映射模块

抽象向量 ➡ 人类隐喻/概念

抽象向量

人类隐喻/概念

概念

白盒界面调试台

概念A

概念B

推论Y

概念A

逻辑步骤2

意图流Z

意图流Z

推论Y

?

人类能够像调试软件般，随时中断、调整或重定向AGI的决策过程。

这不仅是工程与认知的巅峰，更是确保未来AGI安全、可控，并与人类价值观深度共存的关键。

飞鸟实验室
BIRD LAB

30

如何在AGI的演化机制中原生植入安全基因？能否在AGI的底层构建出一种“安全智能认知内生”免疫系统，从而在追求智能极限的同时，维持系统的安全可信？当这种内生免疫系统演化出“安全”和“有害”的内化理解，该如何去“信任”这种可能无法理解的“安全”？

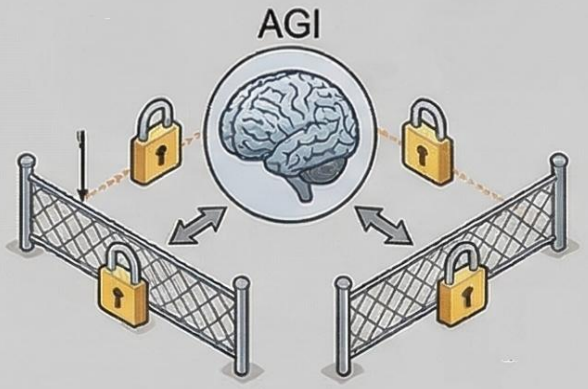
我们的本质是飞翔
而非抵达

AGI原生安全机制：构建认知内生免疫系统

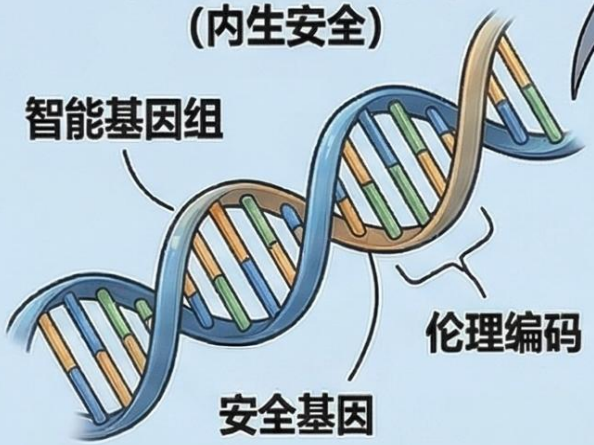
如何将安全基因植入AGI演化并在其底层构建内生免疫?

当前安全范式的局限与新范式

传统安全范式（附加式安全）



原生安全演化范式（内生安全）

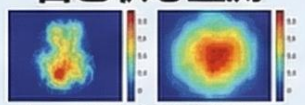


多模态感知与伦理输入

文字: 伦理传码编据

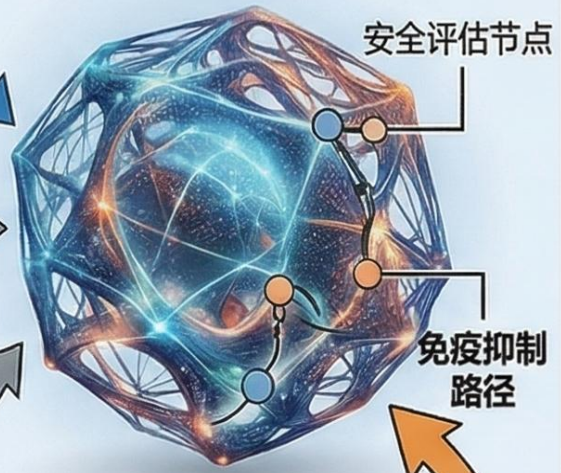
社会影响模拟

自己状态监测



自己状态的计算健康

认知内生安全核心



认知内生安全核心

内化“安全”与“有害”的演化学习

① 有害状态检测　② 内生免疫反应　③ 演化学习反馈

信任“无法理解但可信”的安全：可证实信任框架

传统：黑盒模型的可信度?　新范式：可证实的系统安全性

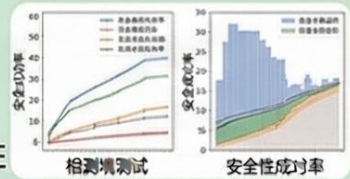
形式化验证：数学边界证明

$$y = \frac{1}{a\pi}\left(x_a\right)^{12} - a(x) + a))^2$$

验证安全约束永远不被破坏

大规模对抗性测试与行为统计

证明系统面对复杂攻击的韧性



系统绩效透明化

可视化的安全运行状态与指标

这种认知内生机制能否在追求智能极限的同时，实现前所未有的安全与信任?

撰写团队

李　鹏	飞鸟实验室主任
周　飞	飞鸟实验室
胡秋媛	飞鸟实验室
石　磊	飞鸟实验室
杨　森	飞鸟实验室
岳思雯	飞鸟实验室
范安妮	飞鸟实验室

制作工具

Gemini 3.1Pro ————提示词生成

Nano Banana Pro —— 图片生成

联系方式

飞鸟实验室公众号

https://www.feiniao.org.cn

飞鸟实验室官网