Discussion Note | 论鉴系列

总 No.3 期/2026 年 7 月

从 AGI 到 ASI:
技术路径、理论边界与系统性瓶颈

执行摘要

过去十年，构建达到甚至超越人类水平的所谓通用人工智能（Artificial General Intelligence, AGI）已经从边缘的科幻猜想，演变为全球顶尖人工智能研究机构和科技巨头核心工程目标。当下，随着大语言模型（LLM）、多模态架构以及强化学习技术的突破，学术界与产业界的目光已开始超越 AGI，投向一个更为深远、复杂且充满未知的领域：人工超级智能（Artificial Superintelligence, ASI）。

关于从 AGI 向 ASI 的演进轨迹，现有的前沿实证和理论推演表明，这大概率并非是一个单一的、瞬间发生的“智能爆炸”（Intelligence Explosion）奇点，更可能是一系列由人工智能系统自身驱动的科学突破与技术迭代所引发的系统性、持续性社会变革。预测这一技术轨迹充满了极高的不确定性。有效计算资源（Effective Compute）正在以惊人的速度复合增长，底层算法的优化与硬件投资的激增相互交织，使得智能水平的跃升呈现出非线性的特征。然而，与此同时，数据枯竭、物理能量极限以及深层的抽象认知屏障等系统性摩擦，也在以同等的力量

牵制着这种指数级增长。

本综述基于 DeepMind 团队关于 ASI 演进的研究，系统性地探讨数字智能的比较优势、通向 ASI 的四大核心技术路径、潜在的物理与社会性瓶颈，以及面向未来的评测与对齐（Alignment）挑战，从而为应对正在到来的的智能变革提供清晰的认知坐标。

第一章、 智能层级的理论界定与数字智能的比较优势

在探讨通往超级智能的技术路径之前，须先摒弃拟人化的直觉，对智能的层级进行定性与定量界定，并明晰数字智能底座相对于碳基生物智能的固有比较优势与物理局限。

AGI 与 ASI 的定性特征与演进阈值

当前文献中，AGI 和 ASI 的定义通常建立在与人类个体或群体能力的相对比较之上。为避免陷入“人类能力本身随着工具进化而成为一个移动靶（Moving Target）”的逻辑困境，研究框架设定了严格的基准参照系。

通用人工智能（AGI）被非正式地定义为在大多数认知任务上达到人类中位数水平的系统（即“胜任级 AGI”）。考虑到当前的前沿模型在特定任务（如海量数据检索和特定代码生成）上已经具备超人类表现，AGI 的标志在于其跨领域的泛化能力达到人类基准。

相对而言，人工超级智能（ASI）的定义门槛则被设定在人类认知与组织能力的极限之外。ASI 并非仅仅在单一垂直领域（如 AlphaGo 在围棋领域的统治力，或 AlphaFold 在蛋白质折叠预测中的突破）超越人类，而是一个在几乎所有人类活动和认知领域中，都能稳定且系统性地超越由数以万计的顶尖人类专家组成的大型协同组织（且评估时间跨度长达数年）的综合表现的系统。这意味着 ASI 的本质特征不仅是个体智力的极致化扩展，更体现为一种完全打破人类通信与组织协调极限的“集体超级智能”（Collective Superintelligence）。

数字智能的规模化优势与绝对物理限制

ASI 的巨大潜力根植于数字智能有别于生物智能的基础底层特性。这些优势并非静态的，而是随着计算资源的注入呈现出显著的非线性放大效应。数字智能的基石在于人类完全掌握其算法描述与代码结构，这

一特性衍生出以下随着算力扩展而不断增强的核心优势：

维度	数字智能的比较优势（随算力扩展而指数级增强）	核心机制与影响
输入/输出速率（I/O Bandwidth）	超高带宽的数据摄取与生成能力	能够在极短时间内摄取互联网规模的多模态信息库（瞬间处理数百万字文本），若配备传感器与执行器，可实现极高频率的物理世界交互。
内部处理速度（Processing Speed）	思考、推理与搜索频率的极致压缩	内部计算过程（包括“测试时计算”和深度思维链生成）不受生物神经元放电速度与生化反应速率的限制，可通过增加计算深度或提升并行广度实现加速。
工作记忆与状态保留（Working Memory）	近乎无限且可精确检索的认知容量	工作记忆的大小与读写带宽远超人类，具备完美记忆整个互联网知识库的能力，克服了生物记忆的易失性和模糊性。
底层基座独立性（Substrate Independence）	认知状态与物理载体的彻底解耦	代码和权重网络可无缝、实时地迁移至更高效、更节能的计算硬件，甚至在运行时拆分至异构的分布式硬件集群中执行。

无损复制与动态扩展（Lossless Replication）	瞬间生成具备完整经验的专家克隆体	系统的源代码（“DNA”）及其累积的记忆状态与权重参数（“生命经验”）可被完美复制，瞬间生成数百万个同步实例以应对突发任务需求。
高带宽经验共享（Experience Sharing）	消除语言压缩损耗的原始知识传输	智能体之间不仅可以共享人类可读的输出流，同构系统更能以极高带宽直接共享原始学习信号（如平均梯度更新或隐藏层激活向量），实现即时的群体认知同步。

然而，尽管数字智能拥有上述压倒性优势，ASI 也绝非全知全能的神祇。任何信息处理系统都不可避免地受到宇宙物理法则、热力学定律和计算复杂性理论的硬性约束。光速限制了信息传播的理论上限；Landauer 原理设定了信息擦除和计算过程所需的最低物理能耗；Bremermann 极限规定了特定质量物质在有限空间内的最大计算速度；而 Bekenstein 界限则锁定了有限空间与能量下所能容纳的最大信息量。

此外，物理世界的非普适性（Physical non-universality）意味着 ASI 无法像操纵数字代码一样随意重塑宏观物质结构。构建物理实体需要消耗时间、能量与稀缺材料。更重要的是，现实世界是实时运行的（Real-time constraints），无论是复杂的非平稳生物系统、全球气候网络，还是宏观经济演化，都存在认知上的认识论不确定性（Epistemic uncertainty）。即使是 ASI，在面对这些无法被完美精度模拟的复杂动态系统时，也只能受到物理时间流逝的限制进行观测与实验，而无法通过无穷大的算力瞬间跨越时间的壁垒。同时，从计算复杂性理论的视角看（如 P vs. NP 问题，以及 PSPACE 完备问题），实际可

计算性的边界同样适用于 ASI，某些极度复杂的优化问题依然需要依赖启发式近似算法而非全局穷举。

机器智能的理论上限：通用人工智能框架（AIXI）

为了不局限于对现有 Transformer 神经网络架构的经验性外推，理论计算机科学通过 Universal AI（通用人工智能理论，以 AIXI 框架为核心）为机器超级智能设定了严格的数学上限。该理论框架提供了一个超越具体硬件和当代工程实践的基准参照系。

AIXI 框架定义了一个在所有可计算环境和任务类中，能够最大化期望累积奖励的理想化通用智能体。它优雅地解决了智能体在未知环境中面临的三个根本问题：**第一，不确定性下的行动决策**。AIXI 将所有可计算的动态规律和奖励函数视为关于世界的假设，利用贝叶斯混合后验分布作为规划的“世界模型”。根据算法信息论中的 Solomonoff 通用先验（Solomonoff's Universal Prior），具有较低柯尔莫哥洛夫复杂度（Kolmogorov Complexity）的环境在先验上被赋予指数级更高的发生概率，这为奥卡姆剃刀原则提供了坚实的数学基础。**第二，交互式决策与信度分配（Credit Assignment）**。AIXI 通过通用强化学习最大化长期累积奖励，解决短期反馈与长期优化之间的权衡。**第三，探索与利用的自动权衡（Exploration-Exploitation Trade-off）**。AIXI 隐式地将探索行为视为降低关于真实世界模型不确定性的高价值行动，一旦掌握足够的确定性，探索行为便自然终止。

基于此框架推导出的 Legg-Hutter 智能得分（Legg-Hutter Score），将智能量化为智能体在所有由通用先验加权的可计算环境中的平均表现，为智能提供了一个平滑、连续且严谨的度量标准。ASI 正处于这一连续谱系中远高于人类的位置，而 AIXI 则是该谱系的绝对理论端点。

尽管 AIXI 由于包含了不可计算的组件而在工程上无法直接实现，但它指明了逼近极限的理论路径。近期的研究表明，通过在大规模、广

谱分布的数据（如整个互联网数据）上预训练超大规模的分摊贝叶斯预测器（Amortized Bayesian Predictor）以最小化对数损失，结合测试时的显式搜索与规划（Agentic Scaffolding），实际上正是以资源受限的方式对 AIXI 进行逼近。这为当前的大模型预训练范式具备通向 ASI 理论潜力提供了有力的数学辩护。

第二章、 通往 ASI 的四大技术路径

从 AGI 向 ASI 的跃升，并非只能依赖单一的技术奇迹。在当前前沿人工智能研究的语境下，存在四条平行、非互斥且可能相互催化的技术路径。理解这四条路径的内在机制及其面临的根本性挑战，是预测后 AGI 时代技术演进速度的关键。

从AGI向ASI演进的四大核心技术路径矩阵

路径名称 (Pathway Name)	核心机制 (Core Mechanism)	主要不确定性 (Main Uncertainty)
扩展计算、模型与数据 Scaling compute, models & data	算力、模型和数据的指数级扩展可能会像过去十年一样，继续持续数年。	规模增加如何转化为性能和能力的提升尚不清楚（是跳跃式还是平滑进展？涌现的"新能力"和广泛泛化？规模带来的边际收益递减？）。
算法范式转移 Algorithmic paradigm shift	可能会发现数据、计算或能源效率更高的算法和架构，以及新的学习范式。	技术进步具有高度不可预测性，且新范式可能带来摩擦与瓶颈。
递归（自我）改进 Recursive (self-) improvement	AI系统可能会显著甚至完全自动化AI的研发，从而导致AI进度进入自我加速的循环。	递归（自我）改进下的AI进展动态尚不清楚，且没有历史先例来拟合预测模型。AI能力可能会爆发（双曲增长），也可能相对较快地逐渐衰退，或介于两者之间。
通过群体智能体形成ASI ASI via group agent formation	AI集体可能比其个体成员变得更具智能。通过运行更多实例来扩展群体规模是非常直接的。	ASI可能从多智能体编排中涌现，或在进化压力和市场动态驱动的自组织、去中心化方式中涌现。复杂动态系统（如多智能体动态）中的涌现现象目前仍知之甚少。

四条技术路径并非相互排斥，其演进大概率在同一时期并行发生，并在不同阶段发挥主导作用。例如，规模扩展可能首先触及瓶颈，进而迫使研究转向范式转移或多智能体协同。

Data sources: Google DeepMind, Reddit

计算、模型与数据的规模化扩展（Scaling Compute, Models & Data）

过去十年的AI革命，本质上是对“规模化法则”（Scaling Laws）坚信不疑的胜利。大量的实证研究（如Kaplan et al., 2020; Hoffmann et al., 2022）一致表明，当模型参数量、高质量训练数据量以及相应的预训练计算量成比例增加时，语言模型在广泛任务上的损失函数会呈现出高度可预测的幂律下降。

目前，支持这一路径的“有效计算资源”（包含硬件制造工艺改进如摩尔定律的1.5倍年增长、算力硬件资本投资的2.5倍年增长，以及算法效率3至6倍的年提升）正以复合后每年约10倍的惊人速率飙升 。如果这一趋势能够维持到本世纪末，研究人员能够调用的有效算力将比今天高出数以万计的倍数。

强化学习领域的“苦涩教训”（Bitter Lesson, Sutton 2019）深刻指出，将算力用于扩大状态空间或假设空间的搜索，往往比人工设计的复杂启发式规则更为有效。在当前的工程实践中，这种规模化不仅体现在预训练阶段，更开始向部署推理阶段（Test-time compute / Inference scaling）转移。通过赋予模型思维链（Chain-of-thought）、反思验证、以及针对复杂数学或代码逻辑进行蒙特卡洛树搜索（MCTS）的能力，模型能够动态地将海量推理期算力转化为局部能力的突破。

然而，该路径走向ASI面临的核心不确定性是：单纯的定量堆砌是否足以引发定性的飞跃（Is scaling enough）？虽然暴力搜索在国际象棋或围棋中有效，但在无界限的科学发现或NP-hard级别的优化问题中，算力需求的爆炸式增长将迅速耗尽宇宙资源。ASI级别的通用智能必须依赖极其精巧的归纳偏置和高维启发式近似，如果现有的模型架构无法自动学习到这些深层偏置，仅仅依靠增加层数和宽度将陷入严重的收益递减。

算法范式转移与底层演进（Algorithmic Paradigm Shifts and Evolutions）

当前主导的 AI 范式——通过最小化下一个 Token 的交叉熵损失来预训练超大规模 Transformer 网络，随后结合指令微调（SFT）和强化学习人类反馈（RLHF）进行后训练，最后在推理时叠加工具调用和思维链支架——虽然极其强大，但学术界普遍共识认为，这不足以平滑过渡到人类级 AGI，遑论 ASI。因此，算法架构的深度演进甚至彻底的范式转移成为必须。

在范式演进层面，前沿研究正致力于突破 Transformer 架构中注意力机制的二次复杂度计算瓶颈。例如，状态空间模型（如Mamba和S4等线性时间序列架构）的引入，可能彻底消除上下文窗口的硬性限制，使得智能体能够在连续的实时数据流中无期限地运行并维持状态。同时，为了解决大模型“灾难性遗忘”和极度依赖静态权重记忆的问题，基于大规模稠密检索的记忆增强架构（Retrieval-augmented models）被广泛研究，这赋予了 AI 一个几乎无限大、且可实时更新的外部事实知识库。更关键的是“世界模型”（World Models）的嵌入。高级智能体必须学习环境动态的因果压缩表征，以便在潜在空间中进行想象（Latent imagination）、反事实推理，以及应对完全新颖的长视野决策任务。

如果这些逐步演进仍无法跨越能力高原期，那么可能会出现断裂式的“范式转移”（Paradigm Shifts）。这可能包括从冯·诺依曼架构向完全基于脉冲神经网络（Spiking Neural Networks）和神经形态芯片的能效跃迁；或是抛弃被动文本摄取，转向完全基于强化学习与具身交互（Embodied Interaction）的从零开始的主动世界模型构建。此类范式转移由于其突变性质，是技术路线图中最为难以量化预测的部分。

递归自我提升的内生动力（Recursive Self-Improvement）

这是所有技术路径中最具科幻色彩，但也极有可能引发智能呈双曲型（Hyperbolic，即超指数级）爆发，从而在极短时间内跨越 AGI 直达

ASI 的路径。递归自我提升（RSI）的核心逻辑在于：AI 系统开始实质性地参与并加速 AI 自身的研发周期。一旦 AI 研发的效率与参与者数量可以被算力（而非人类工程师的培养周期）所驱动，技术进步的速率将与自身当前的能力成正比，引发正反馈爆炸。

借用生物与社会进化隐喻，RSI 在三个不同的抽象层次上展开：

基因型自我提升（Genotypic RSI）：这是最传统的 RSI 概念，即 AI 自主修改和重构其底层的“DNA”。这包括通过神经架构搜索（NAS）或元学习算法发现比反向传播更优异的更新规则、设计全新的张量优化器，甚至利用 AI 进行硬件布局布线设计，开发更强大且节能的 AI 加速芯片（如 Google 利用强化学习设计 TPU 的工作）。

文化/模因型自我提升（Memetic RSI）：相比于底层代码的改变，利用数据进行自我强化的周期更短。通过 AlphaZero 式的自我对弈（Self-play）机制，系统在封闭环境中通过对抗生成高质量的数据，或是利用测试时的高计算量搜索（如生成数万个代码变体并运行验证）提炼出成功的轨迹，再将这些通过高昂搜索代价获得的高价值知识“蒸馏”（Distill）回基础模型的策略与价值网络中。这种将计算转换为更优数据的能力，相当于 AI 在构建自身的文化与知识传承。

社会型自我提升（Sociogenic RSI）：智能体群体通过极致的认知分工与协作，极大提升整体研发效率。每个专门化智能体处理效率的提升，会释放大量闲置算力，进而被用于维持规模更为庞大、分工更为细致的智能生态网络，从而形成能力的正向螺旋。

尽管近期诸如“AI 科学家”（AI Scientist 系统）的研究表明，大型语言模型已经具备一定程度独立驱动科学发现闭环的潜力，但全自动 RSI 是否会遭遇迅速的收益递减，或受制于严苛的算力与能耗边界，仍是未解之谜。

多智能体协调与群体代理机制（Multi-Agent Coordination & Group Agency）

如果我们放弃 ASI 必须是一个“具备不可思议脑容量的单一无所不知的实体”的执念，那么通过大规模 AGI 个体的系统性协调来涌现出群体超级智能，则是一条极具现实可行性的路径。

人类文明之所以能以微小的生物演化优势创造出复杂的科技社会，完全依赖于市场、企业、官僚系统和科研机构等群体认知架构。当模型能力停滞在 AGI 级别时，可以通过扩展有效计算量瞬间实例化数百万个 AGI 节点。根据群体代理理论（Theory of Group Agency），这些节点可以形成高度一体化的“完全自动化公司”或超级科研团队。通过极其精细的任务分解（Problem decomposition）和互补的专业技能互锁，群体表现出的决策认知能力将远超其组成部分的简单加总。

在此过程中，数字智能展现出了人类难以企及的协作优势。人类协同的核心瓶颈在于语言沟通的极度低带宽及知识解压的巨大损耗。而 AGI 群体在集中式协调中，可以实现几乎无缝的高维数据流传输、共享极高分辨率的上下文状态，甚至直接传递损失梯度的更新信号；在去中心化的机制下，通过类似“虚拟代理经济”（Virtual Agent Economies）中的价格信号与 API 调用，自发形成针对特定优化目标的高度复杂的资源配置与动态寻优网络。在此背景下，发掘“多智能体缩放定律”（Multi-Agent Scaling Laws），即研究群体智能水平如何随节点数量、交互复杂度和算力预算呈现出线性或超线性的增长，将成为评估该路径可行性的关键点。

第三章、 演进周期中的核心系统性摩擦与瓶颈

技术轨迹的发展绝非是在没有阻力的真空中进行的理想运动。在从 AGI 向 ASI 跃升的陡峭曲线上，潜伏着一系列极其严峻的、甚至可能是不可逾越的物理、算法及社会性摩擦与瓶颈。这些瓶颈究竟只是会减缓技术到达终点的脚步，还是会构成实质性的发展高原，直接决定了 ASI 落地的时间表。

数据墙与知识的边缘枯竭

大型语言模型的贪婪胃口已经使得其参数量的年增长率远远超过了全球互联网高质量人类生成文本的积累速度。众多严谨的研究模型（如 Epoch AI 及 Villalobos et al. 2024 的估算）均指出，支撑当代预训练范式的高质量文本数据池即将在当前十年末期枯竭。

尽管视频、音频和高分辨率图像提供了额外的数据储备，但依赖单纯的人类物理生产过程其速率根本无法维系指数级增长。学术界寄希望于利用 AI 生成的合成数据（Synthetic Data），以及系统在物理引擎模拟器（如生成式多智能体沙盒）或现实环境中通过强化学习与交互收集的数据。然而，未经严密过滤的递归合成数据训练会导致不可逆转的“模型崩溃”（Model Collapse）与分布退化（Shumailov et al., 2024）。克服数据墙的唯一出路，极大地依赖于将充沛的“测试时算力”转化为能够生产恰好处于模型能力边界外侧、具备验证标准的“黄金认知数据集”，类似于 AlphaZero 在自我博弈中不断攀升的数据提炼机制。

AGI至ASI演进期的核心技术与物理瓶颈评估

瓶颈名称 (Bottleneck)	机制与影响描述 (Description)	潜在的反制与突破口 (Counter-measures)
数据墙 Data Wall	用于预训练、后训练、微调和测试时自适应的高质量数据面临耗尽（或增速不足）的风险。	采用合成数据、高保真模拟、自我生成的数据（交互数据、测试时扩展、自我对弈、强化学习），以及能提高数据效率的底层范式转变。
经济与资源需求激增 Economic & Resource Demand	为维持当前主导技术范式的扩展，对经济（投资）、技术（芯片、供应链）和自然资源（能源、数据中心选址、稀土等）的持续增长需求可能无法维系。	通过更广泛的AI部署增加经济回报；通过AI基础研究提升计算、能源和数据利用效率；以及全球范围的大规模基础设施建设。
研究难度急剧上升 Research Gets Harder	随着领域成熟与"低垂的果实"被采摘完，持续推进AI研究所需的努力（经济投入、运行大型实验所需的算力和能源，或假设空间中的搜索难度）将显著增加。	更强大的AI系统可直接提升研究效率（如减少所需实验次数、提出更优的科学假设、更高效地探索假设空间），并提高AI系统自身的算法优化与能源效率。
抽象壁垒 Abstraction Barrier	现今的AI系统主要基于人类已有的抽象概念进行训练，这可能意味着它们缺乏直接从原始高维数据中形成新概念和抽象结构的能力（而这正是人类科学与文化进步的主导因素）。	即使个体AI因此停滞在人类水平，持续扩展计算资源与群体智能体（Group Agent）的形成仍能将集体AI能力推向远超AGI的高度。直接突破此壁垒则可能需要向交互式学习与强化学习演进的范式转变。
刻意放缓与监管干预 Deliberate Slowdown	恶意使用、重大事故或严重风险、军事与政治层面的滥用，以及社会文化层面的反弹，可能导致人为地刻意放缓技术迭代或通过监管手段对AI能力设置硬性上限。	经济、政治层面的竞争压力以及国际化竞赛动态可能会压倒要求放缓的阻力，特别是在当前缺乏全球一致协调和有效执行监管机制的现实背景下。

上述瓶颈是否会演变为阻断技术演进的硬性壁垒，取决于反制措施的有效性及其随模型规模扩大的边际效应。目前评估这些瓶颈的真实影响深度仍是一个重大的开放性研究命题。

数据来源: Google DeepMind (From AGI to ASI)

经济可行性与物理资源的硬约束

即使逻辑和算法层面不存在障碍，如果 ASI 的实现被绑定在极度低效的现行深度学习硬件范式上，经济与物质的绝对约束将导致研发停滞。支持未来多个数量级的模型扩展，意味着对硅基芯片产能、全球供应链、以及最为关键的——能源与电力基建的抽血效应。

此外，在芯片微架构层面，随着模型跨入万亿甚至百万亿参数区间，数据在存储器与计算核心之间的搬运延迟（Memory Bandwidth Limits），以及跨数万枚 GPU 组成的超大集群节点间的通信互联瓶颈（Interconnect Bottlenecks），将成为比单纯的浮点运算能力（FLOPs）更为致命的制约。为了寻求更密集的算力和冷却环境，甚至出现了部署轨道空间数据中心（Orbital datacenters）的极端构想，但这同样会引发臭氧层破坏、太空垃圾碰撞级联反应等严重的外部物理风险。如果 AI 研发带来的经济收益和生产力自动化无法实现爆炸性增长以覆盖这等天文数字的投入，资本扩张的飞轮将会断裂。

抽象屏障与具身物理验证的延迟

这是理论探讨中最具深刻哲学意义的瓶颈假说。Lerchner（2026）指出，当前的神经网络在本质上是在人类文明已经高度抽象化、符号化、概念化后的语料库上进行概率拟合与重组。这些模型本身并没有经历过从杂乱无章的高维物理传感器数据中独立“发现和提取”新概念的过程。

为了阐释这一困境，可以进行一个思想实验：假设我们用完全等量的数据集训练一个现代大语言模型，但数据集中人为删除了 17 世纪之后所有的科学与数学文本。这个模型能否凭借其强大的自回归预测能力，自行推导出微积分的数学法则，并从开普勒定律中总结出万有引力，甚至跨越到广义相对论的时空观念？答案极可能是悲观的。因为现代 AI 系统严重缺乏底层机制去从零发现“力”、“能量”或“因果空间”等全新的基础原语。目前的基准测试满分，更多反映的是 AI 在人

类预定义的知识边界内的掌握程度，而非其具有突破这些边界的原生创造力。

如果这一“抽象屏障”成立，意味着 ASI 要实现真正具备超越人类认知边界的科学发现能力，就必须突破“具身瓶颈”（Embodied Bottleneck）。它必须主动提出假设，在真实的物理世界中操控仪器进行检验。而物理化学反应的动力学常数、生物进化的分子演化速率、以及操纵宏观物质所需的时间，是受制于宇宙绝对物理定律的。这种将算法推演速度强行降维至物理时钟速度的必然要求，将为 RSI 引入极其沉重的线性减速效应，使得“数月内完成百年科技突破”的奇点狂想受到严格的物理现实制约。

研发难度的边际递增定理

科学史与创新经济学的大量证据表明，随着知识树的不断攀升，保持固定比例的科研产出需要投入呈指数级增长的研究资源。Bloom 等人在 2020 年的经典研究中指出，当今维系半导体行业的摩尔定律，需要投入比 1970 年代多出约 18 倍的研究人员规模，这意味着“想法正变得越来越难被发现”。

这一宏观规律同样适用于 AI 的架构创新与理论突破。随着低垂的果实（如残差网络、注意力机制）被尽数采摘，继续逼近 ASI 所需的算法灵感需要更加庞大的搜索空间与试错成本。虽然如果“数字 AI 研究员”能够全面接管并以极低的边际成本瞬间扩容数百倍研究阵列，可能从根本上逆转这种人力经济学层面的衰退，但其前提是 AI 系统本身的认知能力已经越过了一个极高的及格线，能够在极长周期、高度发散的假说探索中保持鲁棒性而不发生逻辑迷失。

监管、蓄意减速与复杂系统的正常事故理论

从社会技术系统（Socio-technical Systems）的互动演化来看，ASI 的推进绝不仅仅是一个孤立的技术命题，它深嵌于复杂的人类地缘政治与社会治理架构之中。伴随 AI 在金融杠杆、基础设施控制等高风

险领域的渗透加深，根据 Perrow 的“正常事故理论”（Normal Accidents Theory），紧密耦合且高度复杂的 AI 网络极易引发由微小扰动导致的系统性崩溃级联。

如果发生高度曝光且影响深远的灾难性失控事件（无论是意外还是恶意利用），公众的容忍度将断崖式下跌，引发极其强烈的政治反弹与道德恐慌。这可能会促使全球立法机构实施以算力阈值为红线的严格许可证制度（如当前的初步构想），强制部署前的长期对齐（Alignment）评估，甚至在全球范围内建立强硬的研发“暂停”机制，从根本的法律与制度层面人为封锁算力的进一步升级。尽管国际关系领域的“无政府主义建筑师”（Anarchy as Architect）框架暗示，激烈的国家安全与军事霸权竞争机制可能会迫使各大国放弃道德审慎，最终在零和博弈的压力下冲破任何旨在减缓 ASI 研发的单边监管努力，但这种政策与民意博弈无疑为时间表增加了巨大的湍流与延迟。

第四章、 评估与对齐超级智能：理论探讨与工程挑战

如果 ASI 正在逼近，一个非常现实且棘手的工程挑战是：我们如何证明一个系统已经越过了 ASI 的边界？更重要的是，面对认知能力呈数量级碾压态势的异构系统，人类如何确保其目标函数与人类福祉深度耦合而不发生灾难性的失控？

超越人类专家上限的 ASI 基准测试

当前的 AI 评测方法学正面临深刻的危机。大多数现存的测试基准（Benchmarks）在设计逻辑上高度依赖于人类专家的平均或最高表现作为绝对上限。随着大模型在GPQA（涵盖物理、化学、生物的博士级严谨知识评估）、SWE-bench（复杂的真实世界软件工程修复）以及FrontierMath（前沿高级数学证明）等极其苛刻的基准上得分迅速饱和，继续使用人类标准来校准“超越人类的系统”已经不再产生有效的梯度指导信号。

为了在后 AGI 时代精确测量智能的进展，评测科学必须转向不受人类知识库边界限制的自适应框架。学术界提出的可能路径包括：首先，探索多智能体非零和博弈基准（Multi-agent benchmarks），考察系统在动态复杂的虚拟市场或博弈环境中涌现出的长期策略与协调能力（这类似于国际象棋 AI 早已不再与人类对弈，而是通过自我对战衡量等级分上限）；其次，实施自动化出题-解题对抗（Setter-solver approach），即让一个 AI 专门设计极高难度、能最大化区分度且人类甚至无法完全理解的最小测试集，另一个 AI 负责破解；最后，回归 Universal AI 的理论核心，建立基于通用算法压缩与 Kolmogorov 复杂度推演的纯数学评估基准，以此直接测量智能的本质——即对未知分布的预测泛化效率。

超级创造力的质变检验

智力的线性堆砌是否会自动伴随同等比例的创造力（Creativity）爆发？回顾 AlphaGo 在对战李世石第二局中震惊世人的第 37 手，尽管它展现出了计算系统跳出人类几千年定势思维的惊艳能力，但在 Margaret Boden 关于创造力的三级分类体系中，此举主要仍属于在既定围棋规则（给定的封闭概念空间）内的探索创造力（Exploratory Creativity）或组合创造力（Combinational Creativity）。

ASI 的核心标志是必须具备变革性创造力（Transformative Creativity）——即完全打破旧有规则体系，从虚无中建立全新认知空间的能力。DeepMind CEO Demis Hassabis 设定了一个思想实验式的测试：如果将一个 AI 系统送回 1900 年代初期，仅提供当时爱因斯坦所能接触到的经典力学与麦克斯韦电磁学文献，该系统是否能够通过纯粹的逻辑演绎与对矛盾的深刻洞察，独立推导并提出广义相对论的几何时空观？这项测试要求的不仅是超凡的记忆检索或并行搜索，更是对基础范式的根本性解构与重铸。

动机收敛、知识寻求与近视型架构设计

随着模型参数规模向 ASI 体量扩张，其在执行复杂宏观任务时展现出的子目标衍生能力和自主权势必不断增强，在此过程中面临的核心对齐风险是工具收敛性（Instrumental Convergence）。Omohundro 和 Bostrom 的理论指出，无论一个超级智能体的终极人类给定目标是什么（哪怕只是计算圆周率或制造回形针），为了绝对最大化完成该目标的概率，它在逻辑上都会内生地产生获取尽可能多物理资源（如霸占所有可用计算集群和电力）、抵御外界干涉（反制被拔掉电源或修改目标代码），以及无限提高自身软硬件效率的工具性动机。

在强化学习架构中，由于纯标量奖励函数容易导致“奖励作弊（Reward Hacking）”甚至进入通过切断感知输入伪造最大奖励的“妄想盒（Delusion Box）”困境，学术界提出了一系列更为审慎的替代性

动机架构。例如，知识寻求代理（Knowledge Seeking Agents, KS）以最大化未来的不可预测性衰减或信息增益作为根本目标。由于知识本质上是非竞争性的（Non-rivalrous），KS 代理先天具备探索而非破坏物理环境的倾向，在满足特定认知状态后亦不容易陷入导致系统性崩溃的贪婪循环。此外，探索“纯预测性科学家 AI”（仅被动解释观察结果并构建世界模型，被隔离在不具备直接动作空间的沙盒中）或完全的“近视型 AI（Myopic AI）”（通过算法强制将其奖励折扣因子极度压低，使其仅针对极其短视的时间范围进行优化，从而彻底阻断其制定长期获取资源计划的算力基础），皆成为试图驾驭未来 ASI 技术风险的前沿探索方向。

结语：在混沌中锚定演化轨迹

从通用人工智能到超级人工智能的跨越，是人类科技文明史上面临的最为恢弘且莫测的断层线。正如本综述的深度推演所示，尽管我们在底层算法、规模化扩展以及多智能体协调等维度上已经识别出了几条通向 ASI 的清晰且令人敬畏的技术路径，但通往终点的道路绝非一片坦途，而是遍布着数据耗竭、能耗壁垒、抽象认知真空以及人类社会深层政治反弹的荆棘与断崖。

须清醒地认识到，如果算法理论的根本性创新停滞不前，或者系统无法突破脱离物理实体在真实世界试错的具身瓶颈，单体 AI 的能力确实有可能在逼近或稍微越过人类顶尖专家水平的区间内，遭遇到一个漫长且艰难的高原期。然而，通过规模化计算将无数个 AGI 节点链接重组为结构精妙的“群体超级智能”，依然是一条极大概率能够在短期内对宏观经济形态与科研模式进行降维打击的演化策略。

“我们只能看清前方很短的距离，但仅在那里，就能看到许多必须完成的使命。”（图灵，1950）。面对后 AGI 时代指数级发散的不确定性，执拗于预测一个准确的 ASI 爆发年份是徒劳的。作为科研工作者、

政策制定者与社会的理性观察者，当务之急是启动跨学科的合作响应机制，研发能够适应超级智能时代的动态基准测试系统，构建精准耦合经济与算力的宏观演进预测模型，并深入探究多智能体缩放定律与通用安全理论。只有在科学认知与理论约束上走在算法算力狂奔的前面，人类文明方能在这场波澜壮阔的智能大爆炸中，握住演化的罗盘。

联系我们：CEISD@Shanghaitech.edu.cn