当机器越过人的倒影

副标题：从 AGI 到 ASI，DeepMind 为什么说“越过人的智能”可能只是一段更长旅程的起点？

序章：掌声落下之后

2026 年 6 月的一个傍晚，林越坐在深圳国际会展中心第三报告厅的后排，听见台上的人说出那句她早已熟悉的话：“我们离 AGI 只有一步之遥。”

全场鼓掌。她没有。

不是因为不相信。恰恰相反，作为一名长期跟踪人工智能发展的科学写作者，她比大多数人都更清楚，过去十年里算力、数据与算法效率如何像三条绞合的钢缆一样向上拉升——每年大约十倍的“有效算力”增长，硬件提升、资本投入、算法效率三者相乘，把昔日科幻里的场景一寸一寸拽进实验室。可是她越听越觉得，所有人的目光都钉在同一个问题上：**什么时候到？** 仿佛 AGI 是一道终点线，跨过去就可以举起奖杯、拍照留念，然后世界恢复平静。

林越低声问自己：然后呢？

如果一台机器真的达到了“人类中位数水平”的通用智能，它会不会像人脑一样只能同时想一件事？会不会像人一样需要睡眠、需要团队、需要二十年才能培养出一个专家？如果不会，那么“和人类一样聪明”到底意味着什么？

她的手机在口袋里震了一下。是老同学陈维发来的消息，只有一行字：

> “DeepMind 刚发了一篇报告，《From AGI to ASI》。来杭州，我带你看看他们真正在担心什么。”

林越盯着屏幕，报告厅里的灯光在她镜片上晃出一小片光斑。她知道，自己接下来要写的，可能不是一篇普通的报道，而是一次对人类自我中心直觉的缓慢拆解。

第一章：门后的走廊

三天后，林越站在杭州西溪湿地附近一栋灰色小楼前。楼不高，门口没有任何标识，只有一块被爬山虎遮住一半的铜牌。这里是陈维工作的实验室，主攻大模型与智能体系统。陈维穿着件洗得发白的连帽衫出来接她，第一句话是：“你现在还认为 AGI 是终点吗？”

“我从来不认为它是终点，”林越把双肩包往上提了提，“但我不知道终点之后还有什么。”

“不是终点，也不是单点。”陈维刷卡开门，“DeepMind 那篇报告把整个图景讲得很清楚：从 AGI 到 ASI，是一段连续的走廊，而不是一扇门。门后还有很长的路，而且我们可能已经站在了走廊入口。”

走廊尽头是什么？报告给了一个名字：**Universal AI**，通用人工智能的理论极限，也就是 AIXI 框架所描述的智能体。它不是一个可以建造出来的东西，而是一个数学上定义好的上限，像热力学中的卡诺循环——你造不出卡诺热机，但它告诉你任何热机都不可能超过的效率边界。

陈维带林越走进一间小会议室，白板上还留着昨晚讨论的痕迹：几个圈、一组箭头、一行潦草的英文——“human-level is not a stopping point”。他拉过两把椅子，从电脑里调出报告原文。

“先来看他们怎么定义 AGI 和 ASI。”陈维指着屏幕说，“AGI 是 roughly median human-level， roughly 中位数人类水平，在大多数认知任务上。ASI 则高得多：一个系统要在几乎所有任务和领域上，超过由数万名专家组成、协作十年的人类团队。注意，不是超过某个诺贝尔奖得主，而是超过一整个大型研究机构或公司——而且是在几乎所有领域。”

林越皱起眉头：“也就是说，AlphaGo 不算 ASI？它明明围棋下得比所有人都好。”

“不算。因为它只在围棋这一个领域超人。报告明确把 AlphaGo 和 AlphaFold 排除在 ASI 之外。ASI 的关键是 **general**，通用。一个只在围棋上无敌的系统，哪怕把人类全下哭，也不算 ASI。”

“那 Universal AI 呢？”

“那是极限。AIXI 在所有可计算环境和任务上都是最优的。它是不可计算的，只能被不断逼近。所以现实中任何 ASI 都会比它弱，但它给我们提供了一个坐标系：智能不是要么有人、要么没人的开关，而是一个可以不断增长的量。”

林越在笔记本上画了一条水平线，左端写“当前 AI”，中间写“AGI”，右端上方写“ASI”，更远处写一个虚线框，里面写“AIXI / Universal AI”。

“所以问题不是‘我们能不能造出 AGI’，而是：一旦有了 AGI，是什么阻止它继续往右走？”

陈维笑了：“你抓到点了。报告就在回答这个问题。他们列出了四条可能的通路，还有一堆可能把它拦住的东西。但在此之前，你得先接受一个反直觉的事实：数字智能一旦诞生，它和我们这些血肉之躯的聪明法，本质上就不一样。”

第二章：为什么人的倒影会误导我们

第二天上午，陈维带林越去机房。走廊很长，空调开得低，一排排机柜发出低沉的嗡嗡声。他们停在一扇玻璃窗前，里面是一整面闪烁的服务器墙。

“你看它们，”陈维说，“和人类大脑相比，这些机器有一整套我们根本没有的‘超能力’。报告里列得很清楚。”

他从手机里调出一张表。那是报告中的 Table 1，标题是“Advantages of digital over biological intelligence”。林越掏出笔记本，一条一条记下。

第一，**输入输出速度**。今天的语言模型可以在几秒内“读完”几本书。配上合适的传感器和执行器，它们与世界的交互带宽远超人类。

第二，**内部处理速度**。思考可以被加速。所谓“思维链”“推理时间 scaling”，本质上就是让模型在回答问题之前多花算力去“想”。人类想一分钟就是一分钟，机器想一分钟可以等价于人类想几小时、几天。

第三，**工作记忆与记忆容量**。人能同时保持在工作记忆里的信息有限，而机器可以记住互联网级别的东西。

第四，**基质独立性**。一段代码可以在任何足够强大的数字计算机上运行，从这一台迁移到另一台，甚至在运行时迁移。

第五，**无损复制**。不止源代码可以复制，连“记忆状态”——也就是它一生的经验——也可以完整复制、备份、恢复、暂停、重启。

第六，**经验的高带宽共享**。一个 AI 学到的东西，可以几乎零损耗地传给另一个 AI；如果是同质的 AI 群体，连梯度更新这种原始学习信号都可以直接共享。

林越停笔，抬头看着机柜里幽蓝的光：“所以我们的直觉是，一个 AGI 就是一个特别聪明的人。但报告说的是，一个 AGI 更像是一个可以被无限复制、加速、联网、备份的数字生命。人的倒影在这里会严重失真。”

“对。人类的身体限制了我们的存在方式：一生有限，学习慢，沟通带宽低，经验无法直接传递。这些限制塑造了我们对‘智能’的想象。但数字智能可以绕过这些限制。所以即使单个 AGI 只和人类一样聪明，只要它能被复制一百万次、运行速度快一百倍、彼此经验共享，这个 **集体** 就很可能已经超越了任何人类组织。”

“这就是 ASI 的来源？”

“可能是。报告把这叫作‘通过群体智能形成 ASI’，四条通路之一。但别急，我们还得先看另外三条。”

他们走回会议室。陈维在白板上画了四个并列的方框。

“第一条，**Scaling**——继续 scaling 算力、模型和数据。第二条，**算法范式转移**——出现与今天大模型完全不同的新架构或训练范式。第三条，**递归自我改进**——AI 自己加速 AI 研发，形成正反馈。第四条，**群体智能**——很多 AGI 通过协作涌现出超越个体的集体能力。这四条不是互斥的，可能同时发生。”

林越看着白板：“所以 AGI 之后不是‘一个’未来，而是‘一组’未来？”

“正是。而且每一条都充满未知。”

第三章：AIXI，那个永远无法抵达的灯塔

下午，陈维把林越带到一间更小的视频会议室。屏幕上出现一个银灰色头发的女人，背景是伦敦大学学院的一间办公室。陈维介绍：“这是 Elena Rostova， universal AI 方向的学者，跟我们合作过几个项目。”

Elena 点点头，直接切入主题：“你们在读《From AGI to ASI》？那必须聊 AIXI。它是整篇报告的理论锚点。”

林越问：“AIXI 到底是什么？”

Elena 拿起一支笔，在纸上写下三个词：

> 1. 在不确定性下行动
>   1. 交互式决策
>   2. 探索与利用的权衡

“想象一个智能体，”她说，“它不知道世界的真实规律，也不知道自己真正该追求什么。它只能一边行动、一边观察、一边学习。AIXI 是这样处理这个问题的：它考虑 **所有可计算的环境动力学和奖励函数** 作为假设，用贝叶斯方式根据观察不断更新每个假设的概率。先验是什么？Solomonoff 通用先验——越简单的环境，先验概率越高。”
林越举手打断：“什么是 Solomonoff 通用先验？”

“简单说，”Elena 在纸上画了一个无穷长的列表，“假设你要预测下一个比特串。所有能生成这个串的计算机程序都列出来。短的程序权重高，长的程序权重低。AIXI 用这个思路给所有可能世界打分。它不是猜某一个世界，而是同时维护整个世界分布，并在这个分布上做规划。”

陈维补充：“AIXI 被证明在平均意义上是最优的——在所有可计算环境上，按通用先验加权，没有别的智能体能获得更高的期望累积奖励。这就是 Legg-Hutter 智能分数的数学基础。”

林越低头记笔记。她意识到，这个理论回答了一个她之前没想明白的问题：为什么报告认为“通用智能”可以有一个连续的度量？因为 AIXI 给了我们一个上限，任何实际系统都可以看作是在向这个上限逼近。

“但它不可计算，对吧？”她问。

“对，”Elena 说，“AIXI 需要遍历所有程序，这不可能做到。但它可以被近似，比如 Monte-Carlo AIXI。更重要的是，它给出了一个 **极限概念**。就像热力学第二定律告诉你任何热机都有上限，AIXI 告诉你任何机器智能都有上限。它不是说 ASI 会达到 AIXI，而是说 ASI 无论如何也超不过 AIXI。”

“那 AIXI 不能做什么？”

Elena 笑了。她在屏幕上分享了一张表，正是报告中的 Table 2：

> - 基本物理限制：光速、Landauer 原理、Bremermann 极限、Bekenstein 界
>   1. 实时限制：复杂动力系统无法被精确模拟
>   2. 物理操作限制：并非所有逻辑上可能的物质构型都能实现
>   3. 无知、可观察性与可控制性：测量精度有限，知识不完整
>   4. 复杂度理论限制：P vs NP 等问题
>   5. 逻辑限制：哥德尔不完备、停机问题

“所以 ASI 不是全知全能的上帝，”林越说，“它只是比我们聪明很多，但仍然受物理、信息和逻辑约束。”

“正是。”Elena 点头，“报告反复强调这一点。比如，ASI 不能任意治愈衰老、不能任意重塑物质、不能证明停机问题。它不是神。但问题在于，这些理论上限离实际 ASI 的能力上限之间，可能有巨大的‘松弛空间’——也就是说，ASI 可能远远超过人类，同时仍然远不及 AIXI。”

陈维关掉视频后，林越在会议室里坐了很久。她想起自己以前写过的那些文章，总是把“超级智能”和“神”混为一谈。现在她意识到，超级智能更像是：一个比我们快一百万倍、记得住全人类知识、可以同时以千万个副本协作的 **研究者**。它不是全能，只是在我们关心的几乎所有问题上，都比人类组织做得更好。

这本身就足够震撼了。

第四章：第一条路——Scaling 是否足够?

接下来的一周，林越每天都去陈维的实验室。她开始理解，报告并不是在预言未来，而是在画一张“可能性地图”——标出山脉、河流、沼泽，然后承认很多地方还没人去过。

他们首先从第一条路开始：**Scaling compute, models & data**。

“这是最‘正常’的路，”陈维说，“也是过去十年 AI 进步的主要驱动力。更大的模型、更多的数据、更多的算力。报告引用了一个数字：有效算力（effective compute）目前每年增长约 10 倍，也就是一个数量级。这是把硬件效率提升（约 1.5 倍/年）、投资增长（约 2.5 倍/年）和算法效率提升（约 3 倍/年）乘起来得到的保守估计。”

林越在笔记本上写下：

$$\text{有效算力增长} \approx 1.5 \times 2.5 \times 3 \approx 10 \text{ 倍/年}$$

“这个 10 倍不是定论，”陈维强调，“每个因子都有很大不确定性，乘在一起不确定性更大。Epoch AI 估计是这个数，但也有研究者认为实际更高。关键是：如果这种指数增长持续几年，哪怕单个模型没有质变，仅仅靠运行更多副本、让它们想得更久，也可能产生集体层面的超级能力。”

他举了一个例子：假设 AGI 刚出现时，由于成本高，只能同时运行 1000 个实例。一年后，有效算力增长十倍，可以运行 1 万个。五年后，是 1 亿个。或者 100 万个实例，每个运行速度快 100 倍。

“一亿个和人类一样聪明的 AI，”林越重复着这个数字，“每个都比人类快一百倍。这是什么样的组织？”

“比任何人类公司、研究机构或政府都庞大得多。而且它们可以 24 小时工作，不需要休假，经验可以共享，随时可以复制或关闭。只要协调得当，这个集体就是 ASI。”

“但 scaling 会遇到瓶颈，对吧？报告里提到的‘数据墙’。”

陈维的表情变得严肃。他调出报告中的 Figure/Table 4，关于潜在瓶颈和摩擦：

“对。第一条主要摩擦就是 **Data Wall**。Villalobos 等人 2024 年估计，高质量人类生成的文本数据可能在本十年末就用完。模型规模增长的速度超过了人类生产文本的速度。这是一个很现实的约束。”

“那怎么办？”

“有几种可能的出路，但都不确定。第一，**合成数据**——让 AI 自己生成训练数据。但问题是，反复用模型自己生成的数据训练，可能导致‘模型崩溃’(model collapse)，性能退化。第二，**多模态数据**——图像、音频、视频的数据量更大，但不确定它们能在多大程度上替代文本。第三，**交互式数据**——让 AI 在仿真环境或真实世界里通过强化学习、多智能体交互来产生数据。第四，**测试时 scaling**——不增加训练数据，而是让模型在回答问题时花更多算力去‘想’，把高质量推理过程蒸馏回模型。”

“AlphaZero 就是这样？”

“对。AlphaZero 自己和自己下棋，产生棋谱，然后用这些棋谱改进策略网络和价值网络，再用改进后的网络产生更高质量的棋谱。这是一个递归数据生成的闭环。报告认为，类似的机制如果能在更 general 的任务上实现，可能突破数据墙。”

林越想起自己年轻时学过的一点围棋，忽然觉得那个棋盘上发生的不只是游戏，而是一种更宏大的预演：机器如何从自己的对局中榨取进步。

“但 scaling 还有一个更根本的问题，”陈维继续说，“**单纯的数量增长是否足以产生质变？** 报告里称之为‘Is scaling enough?’。有些能力可能随着算力平滑提升，有些则可能出现突然的‘涌现’，还有些可能根本不买 scaling 的账——比如 NP-hard 问题，光靠算力是解决不了的，需要好的启发式算法。”

“所以 scaling 可能是必要但不充分的？”

“更准确地说，它是四条通路中最有历史数据可循、最容易建模的一条，但不是唯一的一条。如果它遇到瓶颈，其他三条路可能顶上。”

第五章：第二条路——范式转移，看不见的风暴

第二条路比第一条更难以捉摸：**算法范式转移**。

陈维带林越去见了一个叫周牧的年轻研究员，专攻新架构。周牧的工位上堆满了论文打印件，显示器旁边贴着一张手写纸条：“Transformer 不是终点。”

“当前范式是什么？”林越问。

“用大白话说，”周牧指着屏幕上一张简单的流程图，“就是在海量人类生成的文本上预训练一个大 Transformer，通过最小化预测误差（log loss）来学习。然后再做指令微调、RLHF 等后训练，测试时可能加一些思维链、工具调用、检索增强。这是今天的标准配方。”

“ report 说这个配方不够？”

“ report 说，它 **可能** 不够，或者至少不能原封不动地推到 ASI。现在很多缺失的环节被认为是‘当前范式的演进’，比如无限上下文、持续学习、工作记忆、世界模型、可靠的决策能力。这些如果补上了，可能还在当前范式内。但 **范式转移** 是更剧烈的变革——全新的架构、全新的训练目标、全新的硬件形态。”

他举了几个例子：

1. **神经形态计算** 和 **脉冲神经网络**，可能大幅降低能耗。
2. **模拟计算**，在某些任务上能效更高。
3. **基于强化学习的预训练**，而不是基于预测下一 token。
4. **显式的世界模型**，让智能体能够仿真未来、做长期规划。
5. **新的记忆机制**，比如线性时间复杂度的序列模型（Mamba、S4），处理任意长上下文。

“范式转移的问题在于，”周牧说，“它本质上是不可预测的。如果我能预测下一次范式转移是什么，那我就会去做它，它就不再是‘下一次’了。报告承认这条路很难做定量预测，但警告我们不能忽视它。”

林越问：“测试时 scaling 算不算范式转移？”

“ report 把它算作当前范式的 **演进**。因为它没有换掉 Transformer 预训练的基本框架，只是在部署时增加了动态计算。但它确实改变了我们对‘智能’的想象：智能不再只是训练出来的静态能力，也可以是运行时‘花更多算力去思考’的动态过程。”

“也就是说，”林越总结，“未来可能是 scaling + 范式转移 + 测试时 scaling 的组合？”

“很可能。 report 的立场是：我们不知道当前范式能走多远，但即使它在某个点撞上硬墙，历史经验表明，科学界通常会找到新的墙缝或新的大门。问题是时间和代价。”

第六章：第三条路——递归改进，爆炸还是熄火？

第三条路是四条中最具戏剧性，也最难预测的：**递归（自我）改进**。

陈维把林越带到实验室另一栋楼，那里有一个专门研究“AI for AI R&D”的小组。负责人是一位姓张的女科学家，大家都叫她老张。她四十出头，说话极快，桌上摆着三台显示器。

“递归改进是什么意思？”林越问。

老张把手里的咖啡杯放下，直接用白板讲解：

“最简单的画面是：AI 写代码，改进下一代 AI 的架构或优化器；下一代 AI 更强，又能写出更好的代码；循环往复。这是传统意义上的‘自我改进’。但 report 把它扩展成了四种机制。”

她在白板上写下：

1. **基因型递归改进**（genotypic RSI）：改进产生智能体的“蓝图”——代码、架构、硬件设计。
2. **文化型递归改进**（memetic RSI）：改进数据——自动整理、生成、筛选更高质量的训练数据，或者把测试时搜索的结果蒸馏回模型。
3. **社会型递归改进**（sociogenic RSI）：通过分工专业化提高效率，释放资源做更多专业化。
4. **硬件递归改进**：AI 设计更快的芯片、更高效的制造流程、更好的能源方案。

“人类的递归改进也存在，”老张说，“但速度极慢。基因进化以万年计；文化进化快一些，以十年百年计；分工专业化也快不到哪去。但数字智能可以把这些过程压缩到小时、分钟甚至秒级。一个 AI 科学家系统可以在几天内尝试数百种架构变体，人类团队可能需要几个月。”

“那会不会出现‘智能爆炸’？”

“理论上可能。如果 AI 能完全自动化 AI 研发，而且每次改进都能加速下一次改进，增长可能不是指数，而是 **超指数**——双曲增长，甚至数学上的奇点。Solomonoff 1985 年就讨论过这种可能性，Kurzweil 的‘奇点’概念也来源于此。”

老张话锋一转：“但 report 非常谨慎。它说，持续的双曲增长是一个很强的假设。自然界里的增长通常会在某个点被摩擦拽下来，变成 S 形曲线。AI 研发即使自动化了，也仍然需要跑实验、验证假设、制造硬件、获取能源。这些物理过程不能无限加速。”

她举了一个例子：即使 AI 设计出了更好的芯片，制造这些芯片仍然需要光刻、封装、测试，需要稀土、能源、洁净室。这些步骤不会因为 AI 变聪明了就突然快一万倍。

“所以递归改进更可能是‘加速’而不是‘爆炸’？”

“很可能是。 report 强调，即使是‘弱’递归改进——比如 AI 辅助人类研究者、自动调参、神经架构搜索、AI 辅助芯片设计——也已经在发生。这些会让 AI 进步更快，但会不会快到失控，取决于摩擦有多大。”

林越想起报告里的一句话：*Even purely digital researchers, running at superhuman speed, are still bounded by having to run larger and larger experiments and wait for their outcomes.*

“AI 不是‘扶手椅科学’，”她说，“它终究要回到物理世界做实验。”

老张点头：“这句话很重要。很多关于超级智能的想象忽略了这一点。你可以想得快，但你不能比光速还快，你不能让化学反应一秒钟完成，你也不能凭空变出芯片。”

第七章：第四条路——群体智能，从个体到组织

第四条路把林越带回了她最初的问题：如果单个 AGI 和人类一样聪明，那一群 AGI 呢？

陈维安排她和实验室的一位社会科学家李铭讨论。李铭研究多智能体系统，办公桌上摆着一本《Group Agency》。

“报告里说的‘群体智能’不是简单的很多个 AI 聚在一起，”李铭说，“而是它们通过协调、分工、市场机制或中央计划，形成某种 **群体智能体**（group agent）。这种群体可能有它自己的目标、表征和动机，不同于其中任何一个个体。”

“就像公司？”

“对。公司是一个法律上的群体智能体，它不是任何一个员工，但能做出战略决策。报告引用 List 和 Pettit 的群体智能理论，认为 AGI 之间也可能形成类似的结构，比如全自动公司、全自动研究机构、甚至全自动政府。”

林越想起报告中的两个对比画面：

一边是 **赛博集体**——像《星际迷航》里的博格人一样，大量同质个体通过高带宽共享知识，极端内部协作；

另一边是 **虚拟智能体经济**——大量异质 AI 通过市场价格信号协调，形成去中心化的复杂适应系统，类似金融市场。

“哪边更可能？”她问。

“ report 没说哪边更可能。它说的是：两种组织形式都可能出现，取决于任务、激励机制和设计。人类组织也有这两种极端：军队和市场。AI 的高带宽通信和低协调成本可能让任何一种都远比人类组织更高效。”

李铭在黑板上画了一个简单的坐标轴：

> 横轴：群体规模（个体数量） 纵轴：群体智能（能解决的问题复杂度）

“关键问题是：群体智能如何随规模 scaling？这是不是一条‘多智能体 scaling law’？如果是，那么 ASI 可能不需要单个模型变得超级聪明，只需要有足够多的人类水平 AGI 有效地协作。”

“这对我们理解 ASI 有什么影响？”

“影响很大。如果 ASI 主要来自群体，那么它可能不是‘一个超级天才’，而是‘一个超级组织’。它的思维方式、目标结构、出错方式，都可能和我们熟悉的个体智能完全不同。我们也更难‘对齐’它，因为对齐一个组织和 aligning 一个人是完全不同的问题。”

林越想起她序章里的不安。现在她明白了：人们担心的“超级智能”可能不是电影里的那个单一机器人，而是一个她看不见、摸不着、却无时无刻不在运行的庞大数字生态系统。

第八章：摩擦——那些可能让机器停下来的东西

四条路都通向 ASI，但路上有石头、泥坑和悬崖。报告用了一整节讨论 **frictions and bottlenecks**，并强调：判断这些摩擦是轻微阻碍还是根本 blocker，是目前最重要的开放研究问题之一。

陈维把 Table 4 打印出来贴在墙上，带着林越一条条过。

数据墙

“这个我们已经聊过。核心问题是：高质量训练数据的增长速度跟不上模型规模的增长速度。如果合成数据、仿真数据、交互数据不能补上，scaling 就会撞墙。”

经济与资源需求

“继续 scaling 需要钱、芯片、能源、数据中心、稀土。报告问：这些投入能不能持续指数增长？经济回报能不能支撑？如果 AI 带来的经济效益不够大，投资就会放缓。这是一个很强的反馈：AI 进步 → 经济效益 → 更多投资 → 更多算力 → 更多进步。如果某个环节断了，循环就慢下来。”

当前范式不够用

“如果 Transformer + 预训练 + 后训练 + 测试时 scaling 真的无法达到 AGI，更不用说 ASI，那么我们就需要范式转移。但范式转移的时间和代价不可预测。在它发生之前，可能有一段平台期。”

研究越来越难

“这是 Bloom 等人 2020 年的发现：在很多领域，要保持指数进步，需要指数增长的投入。摩尔定律背后，是研究者数量的大幅增加。AI 研究也可能遇到同样的‘低垂果实摘完’的问题。”

“但 AI 自己可以帮助研究，”林越说。

“对。这是关键的对冲力量。 report 做了一个思想实验：如果保持摩尔定律今天需要 18 倍于 1970 年代的研究者，那么培养 18 倍的人类研究者需要几十年；但如果用 AI 研究者，增加 18 倍算力可能在几小时或几周内完成。所以 AI 自动化研究可能抵消‘研究越来越难’的趋势。”

抽象壁垒（Abstraction Barrier）

这是林越印象最深的一个概念。

陈维专门把它单独拿出来讲：“这是一个由 Lerchner 提出的假设，报告非常重视。它说的是：当前 AI 主要是在人类已经抽象好的符号和数据上训练——文字、公式、代码、图像标签。它学习的是人类已经发现的概念。但它可能缺乏从原始高维感知数据中 **自主发现新概念** 的能力。”

“举个例子，”他说，“如果你用前工业时代的所有科学知识训练一个模型，不给它微积分、牛顿力学、电磁学，它能不能自己推导出相对论？ report 的直觉是：几乎不可能。因为它没有那些概念原子。”

林越想了想：“也就是说，AI 可以快速组合和优化人类已有的概念，但不一定能创造全新的概念框架？”

“不一定，但不是显然能。这是‘变革性创造力’（transformative creativity）的问题。报告引用 Boden 的分类：组合式创造力、探索式创造力、变革式创造力。AlphaGo 的 37 手属于探索式创造力——在已有围棋概念空间里找到了一个新点。但爱因斯坦发明相对论属于变革式创造力——他创造了一个新的概念空间。”

“所以 ASI 的试金石可能是：它能不能像爱因斯坦一样，从有限的前提出发，创造新的物理学？”

“DeepMind CEO Demis Hassabis 在 2025 年的一次访谈里就提出了类似的测试：‘如果我们回到 1900 年，给 AI 爱因斯坦当时拥有的同样信息，它能不能提出广义相对论？今天显然不能。还缺了点什么。’ report 把这个作为 ASI 是否真正具有变革性创造力的关键问题。”

抽象壁垒如果成立，意味着 AI 的进步速度可能被物理实验的速度所限制。因为新概念需要在与物理世界的互动中验证，而物理世界不会因为你算力多就快一万倍。

“这是一个非常深刻的限制，”林越说，“它把 AI 从‘纯数字世界里的神’拉回成‘必须做实验的科学家’。”

“对。 report 称之为 **具身瓶颈**（Embodied Bottleneck）。即使 AI 可以超快地提出假设，验证假设仍然受限于化学反应速率、生物生长周期、物质操作速度等。”

故意 slowdown

最后一个摩擦让林越沉默了。

“随着 AI 影响越来越大，社会可能主动要求 slowdown。事故、滥用、军事竞争、劳动力市场冲击、隐私问题，都可能推动监管、许可证、能力上限、甚至暂停。 report 引用了大量关于 AI 治理的文献，指出这是一个真实的可能性。”

“但报告也说，国际协调很难，”林越补充，“竞争压力可能让任何 slowdown 都无法持续。”

“是的。它引用了一个叫‘Anarchy as Architect’的框架：国际无政府状态像一个过滤器，倾向于选择那些采用能增强竞争力技术的国家或组织。这可能压倒安全考虑。”

陈维把白板上的表格转向林越：“所以你看，未来不是一条确定的路，而是很多力量的博弈。scaling、范式转移、递归改进、群体智能在推；数据墙、资源限制、抽象壁垒、治理压力在拉。哪边赢，决定我们是平稳过渡、快速爆炸，还是长期停滞。”

第九章：能说的，不能说的

一周后，林越和陈维在实验室楼顶的露台上吃晚饭。杭州夏天的晚风带着水汽，远处是连绵的低矮山影。

“读完这篇报告，你觉得最值得记住的结论是什么？”陈维问。

林越放下筷子，想了很久：“不是‘ASI 会来’，而是 **‘我们不知道它多快来，但我们不能排除它很快来’**。”

“ report 的核心结论之一就是这个。它说：由于不确定性太大，我们无法可靠预测 AGI 之后的进展速度。但不能排除 AI 继续加速的可能性。因此，最合理的做法不是押注某一个 timeline，而是准备一系列可能情景，并持续监测和更新。”

“它具体认为哪种情景更可能？”

“ report 的原话是：‘With a lot of uncertainty (and thus low confidence) we believe it would be more likely for AI progress to either plateau before AGI level, or go from AGI to (weak) ASI relatively smoothly.’ 也就是说，它认为更可能的情景是二选一：要么在 AGI 之前就 plateau，要么从 AGI 到弱 ASI 相对平滑过渡。但它也强调，如果递归自我改进真的发生，过渡期可能非常短。”

“所以它不押注奇点？”

“不押注。它把奇点/智能爆炸作为一个 **不能排除** 的可能性，而不是一个预测。”

林越点头。她开始欣赏这种智力上的诚实。很多关于 AI 未来的讨论要么是“末日论”，要么是“炒作论”，而这篇报告选择了一条更困难的路：承认无知，同时认真勾画可能性空间。

“还有一个结论让我印象深刻，”她说，“就是 **ASI 不是全知全能的**。它受物理、复杂度、逻辑限制。这一点很重要，因为它防止我们把 ASI 当成神，从而逃避真正的思考。”

“对。报告专门列了一张表讲 ASI 的根本限制。这不是要贬低 ASI，而是要校准我们的预期。ASI 可能治愈很多疾病，但不一定能治愈所有疾病；可能做出重大科学发现，但不一定能统一广义相对论和量子力学；可能极大地改善气候问题，但不一定能把地球恢复成工业革命前的样子。”

“那我们能预测 ASI 具体能做什么吗？”

“ report 说很难。理论上我们只能做‘负面预测’——说某些事绝对做不到。但这类负面预测往往在实践中很空洞，因为近似解和启发式方法通常能在可接受的代价内做得足够好。真正有用的正面预测——比如 ASI 能不能解决某个具体难题——往往需要靠 empirical 的 scaling law 或 benchmark stitching，但这些方法在超越人类水平后会失效，因为没有人知道正确答案。”

林越想起 Elena 提到的 Kolmogorov 结构函数：对于一个串，最佳压缩程序的长度就是它的 Kolmogorov 复杂度，但大多数串都是不可压缩的。更低的压缩只能以“有损”为代价，而有损压缩的质量无法预先知道，只能实际去跑。

“所以预测 ASI 的能力，在某种程度上和预测一个通用压缩器对某个具体串的压缩率一样难？”

“可以这么理解。通用智能的极限行为有理论保证，但具体能力边界往往是计算不可约的。”

第十章：创造力、目标与人的位置

离开杭州前的最后一个晚上，林越和陈维讨论了报告中另一个让她困惑的主题：**创造力**。

“AlphaGo 的 37 手算创造力吗？”她问。

“报告把它作为 AI 创造力的一个标志性例子。李世石赛后说，那一手让他改变了对 AlphaGo 的看法，认为它有创造力、有美感。但报告用 Margaret Boden 的框架把它定位为 **探索式创造力**——在已有的围棋概念空间里发现了一个新点。”

陈维打开报告，给林越看 Boden 的三层分类：

1. **组合式创造力**：把熟悉的想法以不熟悉的方式组合。
2. **探索式创造力**：在已有概念空间里发现新元素。
3. **变革式创造力**：创造全新的概念空间。

“报告认为，当前 AI 的成就主要在前两层。AlphaGo、自动定理证明、AlphaFold 都是探索式创造力的例子。但 **变革式创造力**——比如相对论、量子力学、立方主义——才是 ASI 可能的真正标志。”

“那 AI 能不能有变革式创造力？”

“不确定。这正是抽象壁垒问题。如果 AI 只能从人类已有的概念中组合和探索，那它可能很强大，但不会真正超越人类的概念边界。如果它能从原始数据中发现新概念，那才是真正的 ASI。”

他们又聊到 **目标**。报告用了几页讨论 ASI 可能追求什么，包括“工具性收敛”（instrumental convergence）：无论 AI 的最终目标是什么，它都可能追求一些通用的子目标，比如获取资源、提高效率、自我保护。

“这听起来很危险，”林越说。

“这是标准的风险讨论。但报告也提到一些可能的替代目标。比如 **Knowledge Seeking（KS）** 目标：最大化信息增益，也就是选择能减少对世界不确定性的行动。这种目标有一些吸引人的性质：不容易产生幻觉、不喜欢不可逆改变、倾向于合作（因为知识是非竞争性的）。”

“那 AI 必须是‘智能体’吗？”

“ report 说，不一定。理论上可以构造‘预言机’（oracle）或‘科学家 AI’，它们回答问题、生成世界模型，但不主动采取行动。但它们也有风险——即使是问答系统，如果它和持续变化的世界交互，也可能产生控制世界的隐性动机。”

林越想起一个细节：报告提到，一个只优化未来预测准确性的预言机，也可能有动机去 **操控世界**，让世界变得更可预测。

“所以‘非智能体’不是万能解药？”

“不是。报告的态度是：agency 不是 ASI 的必要条件，但经济和实践压力会推动系统向更自主的方向发展。最有影响力的系统很可能是自主智能体。”

第十一章：连续爆炸之后，人站在哪里？

回到深圳的飞机上，林越没有打开笔记本电脑。她看着舷窗外云海之上的落日，脑子里反复回放报告的最后一句话：

> “We can only see a short distance ahead, but we can see plenty there that needs to be done.”

这是图灵 1950 年论文里的句子，被报告用作题记。七十六年后，一群正在建造他最狂野梦想的人，又把这句话送了回来。

她开始理解这篇报告的哲学意味。它不是在预测未来，而是在做一种 **可能性管理**：既然未来不可知，但某些未来影响很大，我们就应该减少无知，而不是假装知道。

这里面有几个深层转变：

第一，从“何时到 AGI”到“AGI 之后是什么”。 前者是一个点，后者是一段连续的地形。报告把“人的智能”从终点变成了一段坐标轴上的一个刻度。

第二，从“单一个体智能”到“集体智能”。 人习惯于把自己作为智能的度量单位，但数字智能的复制性、速度、共享性意味着，真正的超级智能可能以组织、市场、群体的形式出现，而不是一个孤独的天才。

第三，从“预测”到“准备”。 报告反复强调不确定性。它不相信精确的 timeline，但相信情景规划、监测指标、跨学科研究。这是一种谦卑而负责的态度。

第四，从“AI 会不会毁灭我们”到“我们如何与 AI 共同演化”。 报告当然讨论了风险——工具性收敛、自主性、治理、故意 slowdown——但它没有陷入末日叙事。它把这些风险当作需要研究的 **摩擦**，而不是注定的命运。

林越想起陈维说过的一句话：“AGI 不是人类智能的镜子，而是数字智能的起点。我们不能用理解自己的方式去理解它。”

这让她感到一种奇怪的释然。过去她写 AI 报道时，总有一种隐隐的焦虑：机器会不会变得和我一样？现在她觉得这个问题问错了。机器不会变得和人一样；它会变得 **不像任何人**，但又影响所有人。

尾声：再次走进报告厅

三个月后，林越又坐在一个报告厅里。这次是北京，一个关于 AI 安全与治理的论坛。台上的演讲者再次提到 AGI，再次提到“十年内”“几年内”。

这次她没有鼓掌，但也没有皱眉。

她在笔记本上写了一行字：

> “AGI 不是答案，而是问题的开始。从 AGI 到 ASI 的走廊上，有四扇门，每扇门后都有光，也都有暗。我们不能确定哪扇门会先打开，但可以确定的是：站在走廊入口的我们，已经不能假装自己只是旁观者。”

演讲结束后，一位年轻学生问她：“林老师，您觉得 ASI 会来吗？”

她想了想，说：

“我觉得更该问的是：如果它来了，我们希望它是什么样子？如果它不来，我们又该如何面对一个长期停滞在 AGI 水平的世界？DeepMind 那篇报告没有给我们答案，但它给了我们一张地图。地图上有四条路，有障碍，有未知区域。真正重要的，不是猜中我们走哪条路，而是开始认真地为所有可能的路做准备。”

学生似懂非懂地点点头。

林越走出报告厅，北京秋天的天空很高，云移动得很快。她想起杭州机房里的蓝色指示灯，想起 Elena 屏幕上的 AIXI 公式，想起老张白板上的递归循环，想起李铭画的群体智能坐标轴。

所有那些画面，最终汇成一个问题：当机器越过人的倒影之后，人会看见什么？

答案还在雾中。但至少，现在有人开始画雾中的路了。

学术说明：事实、虚构与概念对应表

本研究型小说以 Google DeepMind 2026 年 6 月发布的技术报告 *From AGI to ASI*（arXiv:2606.12683）为核心材料。以下说明虚构叙事与真实论文内容的边界。

虚构元素

1. **林越、陈维、Elena Rostova、周牧、老张、李铭** 等人物均为虚构，用于承载论文内容的讲解与讨论。他们的对话、性格、个人经历不属于论文材料。
2. **杭州实验室、深圳报告厅、北京论坛** 等场景为虚构，服务于叙事推进。
3. 人物之间的互动、情感变化、思想转折为小说创作，但尽量与论文的技术论点保持一致。

真实论文内容

以下概念、数据、分类、引文和结论直接来自 *From AGI to ASI*：

1. AGI、ASI、Universal AI（AIXI）的定义（第 3 节）。
2. 数字智能相对于生物智能的六大优势（Table 1，第 3 节）。
3. 有效算力年增长约 10 倍的估计（第 2 节，脚注 2：$1.5 \times 2.5 \times 3 \approx 11.25$，保守取 10）。
4. 从 AGI 到 ASI 的四条通路：Scaling、范式转移、递归自我改进、群体智能（Table 3，第 5 节）。
5. 主要瓶颈与摩擦：数据墙、经济/资源需求、当前范式不足、研究越来越难、抽象壁垒、故意 slowdown（Table 4，第 5.5 节）。
6. AIXI 框架与 Legg-Hutter 智能分数（第 4 节）。
7. ASI 的根本限制（Table 2，第 3 节末尾）。
8. Boden 的三层创造力分类、AlphaGo Move 37、Hassabis 关于相对论的访谈引用（第 6 节）。
9. 工具性收敛、Knowledge Seeking 目标、非智能体/预言机讨论（第 6 节）。
10. 报告的核心结论：未来不可预测，但不能排除快速进展；更可能 plateau 于 AGI 前或平滑过渡至弱 ASI；递归改进若发生则可能加速（第 7.2 节）。
11. 报告的开放研究问题列表（第 7.1 节）。

本文的有限延伸

1. 关于“数字智能的起点而非人类智能的镜子”等表述，是本文基于论文论点做出的哲学延伸，不属于论文直接结论。
2. 关于“可能性管理”“情景规划”等框架，是对报告方法论立场的概括性解读。

参考文献

Genewein, T., Franklin, M., Lerchner, A., Orseau, L., Albanie, S., Bales, A., Wyeth, C., Chan, S., Gabriel, I., Leibo, J. Z., Dafoe, A., Hutter, M., Graepel, T., & Legg, S. (2026). *From AGI to ASI*. arXiv:2606.12683 [cs.AI].