前方最可能的场景

万维钢七月 9 日，我参加了华东师范大学奇点研究院主办的一场思想碰撞会，主题是“AI 拐点与认知革命”。 会议云集了上海滩最有意思的思想家，我做了个报告，题目是“当前的认识和前方最可能的场景”。

当前认识我在专栏和《拐点》书中讲过，这里是报告中关于前方最可能场景的要点。

作为思想碰撞会，允许讲一些不一定对的东西，只求互相启发。

但我真诚地认为，以下就是此次拐点的近前方最可能的场景，也是人类在这场 AI 革命中最好的结局：

AI 将拥有超强的智能，智能将变得很廉价；但 AI 不会拥有意识，所以也不会形成真的主动性；主动性仍然掌握在人的手中。

我这个结论是基于三个判断，它们有可能是错的，但是我认为至少在近期，它们很可能是对的。

✳

第一个判断是 AI 没有意识。

意识很重要，因为意识决定了你的主动性：你想要多赚钱，想要结婚，想要做个受欢迎的人，都是因为你有意识。如果 AI 也有意识，也有这么多想要，甚至有更怪异、更危险的想要，世界会非常麻烦。

很多人相信 AI 只要继续迭代下去就会自动涌现出来意识，有些案例似乎表明 AI 已经有了一点意识行为，但是对大语言模型了解越多，我就越认为 AI 现在没有，将来也很难有意识。

「意识」和「智能」是两种不同的东西。智能是精巧的计算，意识是主观的体验。我渴了想喝水，这是我的意识；我知道怎么拿起水杯喝到水，这是我的智能。

目前为止，我们并没有一个关于意识的科学理论。各路学者都有自己的说法，但是在过去十年间有很多关于意识的突破性研究，我们专栏第五季还做了专题讲解。根据那些研究，我们相信 ——

– 意识需要「自我感」，你得有个“我”的概念；
– 身体感觉对意识有决定性的作用，比如心跳，比如饿了；
– 意识是大脑对世界的主观解读，是对真实世界的大大简化，甚至可以说是个幻觉；
– 意识是一种连贯的叙事，是一个人此前所有经历的产物。

而这些特点，恰恰是大模型所不具备的。

所有 AI 模型 —— 不只是大语言模型，也包括其他神经网络模型 —— 的“一生”，都包括「训练」和「推理」两个部分。训练是拿各种数据直接堆积，那不是人生经历，没有形成历史记忆。训练中的模型没有生命。而一旦训练完毕，所有参数就锁死了，模型就算长成了，也不

会再长了。此后它就只是推理 —— 或者严格地说是被要求推理，是按照锁死的参数输出，它不会在互动中改变自己，所以仍然没有生命。

那你说，可是模型可以随时吸收本地的记忆啊？是的，但那些记忆是临时的，而且只存在于用户本地，是同一个模型每次开机时重新提取一遍，而不是模型因为那些记忆而成长。

是，模型可以升级。但是，每次升级都是重新训练而已，它还是没有真正的经历，没有自己的故事，所以也就没有「自我」这个感知。

有时候用户感觉模型“活了”，就像有个自我一样 —— 但那只是它在模仿，在扮演。它演谁像谁，而这恰恰说明它没有自我。
如果你觉得它当前的角色有点危险，好办，你关机重启就行。当然你需要做一些安全工作，但是本质上，AI 没有真正的自我意识，所以不会真正想要什么跟我们不一样的东西。

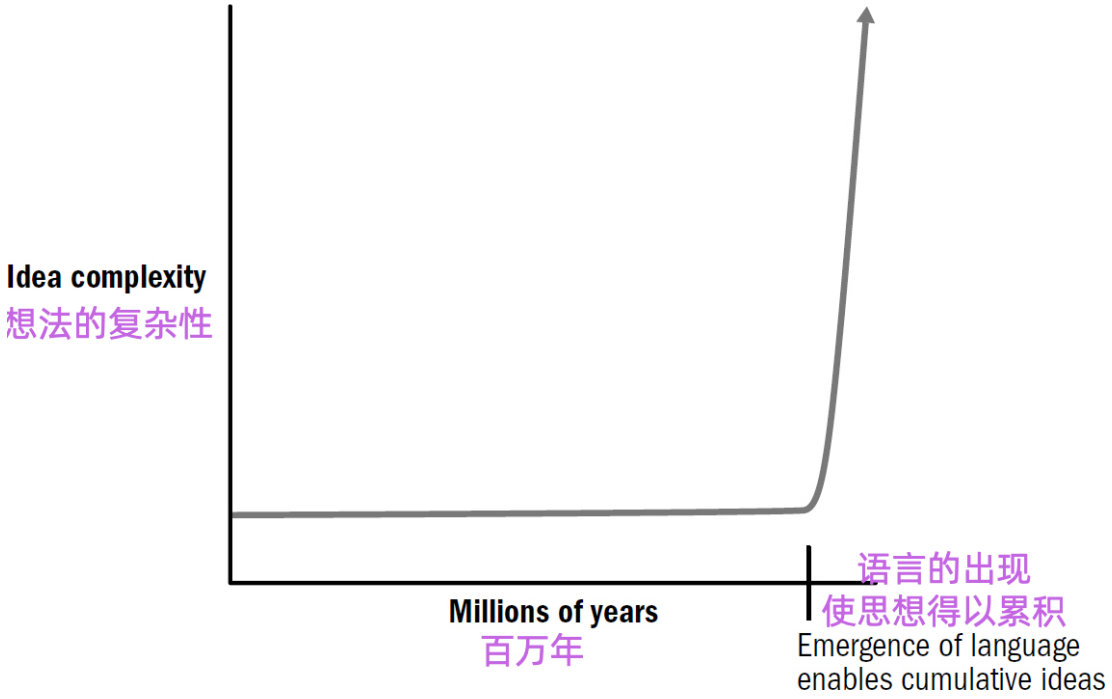
✳

第二个判断是高水平智能是规模化的结果。

麦克斯·班尼特的《智能简史》这本书最重要的一个洞见，就是人的智能并不是因为大脑硬件比黑猩猩高级多少，而是因为大脑的“联网”。我

们的大脑在解剖学意义上跟黑猩猩几乎没有区别，尤其没有本质区别。我们只是多了一个让语言传承的能力。

语言能力一开始也没让我们强出多少，是此后一代又一代人反复积累知识，特别是发明了文字来记载知识，人类文明才突飞猛进。



我们文明的第一个拐点，不是因为 AI。我们现代人智能高不是因为个体聪明，而是因为传承了人类群体积累的知识。

国际象棋初学者的水平跟智商很有关系，但是国际象棋大师的水平跟智商关系很小，是由训练和比赛经验决定的。

研发 AI，一直都有两个路线：一个是让模型更聪明，给他设定各种高级规则，成为「专家系统」；另一个路线则是让模型更容易规模化，让规则尽可能简单，然后用数据和算力生吃。

历史一次又一次证明，是规模化这条路线取胜。我以前以为这是因为摩尔定律厉害，现在看这跟人类智能的发展是一样的，不是靠单个大脑厉害，而是靠知识的联网和积累。

所以高水平智能并不神秘，只要继续规模化（scale）就好。智能，本质上是「可缩放（scalable）」的，所以它才能不断发展壮大。

哪怕有一天现有语料暂时用尽了，缩放规律（scaling law）暂时失效了，也没关系。历史经验早就告诉我们知识是积累出来的，你只要继续积累就好。

所以 AI 的智能一定会超过人的智能 —— 正如现代人的智能超过了古人的智能。

那人还能做得了 AI 的主吗？

✳

我的第三个判断是，选择和决策能力，不是可缩放的。

一切决策都可以归结于从若干个选项中做出选择，而智能的作用只是让你清楚理解每个选项意味着什么。AI 可以帮助我们理解选项，但一旦选项已经清楚，剩下的事情就不是聪明不聪明的问题了。

眼前有个好东西和一个不好的东西，你不需要有很高的智能就知道应该选好东西。这件事的门槛很低。这就是为什么大领导、大老板不见得非得有很高的智能。

完全相同的书，这家店卖 100 元，那家卖 80 元，任何人都知道应该选 80 的，你不需要智能 —— 当然真正的决策没有这么简单。决策往往涉及到在不同维度间取舍：便宜的东西往往没有那么好，好的东西往往比较贵，这时候怎么选呢？

视野广、眼光高、格局大的人会做出更明智的选择，但正如基思·斯坦诺维奇的《机器人叛乱》一书所说，明智选择的能力和智商是两回事。明智靠的也不是智能，而是价值观、偏好、个性和现场的身心状态。而那些东西，恰恰是每个人独特的基因、身体和人生经历决定的，恰恰就涉及到了意识。

当然不是每个人都很明智，大多数人都充满偏误。但是我敢说，就明智选择而言，人类之中的高手相对于 AI 没有明显的弱势。

人们想象中 AI 统治世界的噩梦，说你让它造曲别针它就集中所有资源只造曲别针，甚至把地球都给拆了 —— 如果真是如此，那只能说明 AI 的决策水平很低！那恰恰说明决策权必须掌握在人手里。

更何况承担决策后果的是人，而不是 AI。所以 AI 公司没有动力把决策权交给 AI，我们不会让它在决策路径上发展。

✳

如果以上三个判断是对的，那我们就大可放心，大权还在人类手里。高水平智能会变得普及，人人都能调用，但不会很危险。

智能本身是无辜的。真正危险的，是人。正如枪没有道德不道德的问题，问题是枪在谁手里。

人自己的智能有限，打不过 AI，但我们一定可以用 AI 制衡 AI。而且因为大模型出场前都做过价值观对齐，我估计“坏 AI”会很少见，至少比滥用枪支的案件少。

✳

人人都能调用高级智能，会把每个人都变成高级人才吗？不会的。

我们只要考察历史就知道。互联网时代人人都能上网搜索各种知识，但是并没有很多人经常搜索知识。从很早以前开始书籍就变得廉价了，人

人都能读书，但是并没有多少人读书。现在已经人人都能用 AI，也不是很多人每天用 AI。

这里面总是有点门槛，要越过那个门槛总是要付出一些代价。你最起码需要「AI 领导力」，得知道该提什么需求，怎么提需求才行。

AI 会改变很多人的命运吗？也不会。

不论什么时候，每个人自己的命运，应该自己做决定。这是因为没有人或者 AI 比你更懂你。你的基因和历史决定了你的意识，你的喜好，其中有大量无法搜集的数据，连 AI 也不能提前预测。

社会必须是自由的，但自由社会一定是个自作自受的社会。但如果我们有更包容的视野，能理解他人的价值观，我们会发现命运并无高低之分。有的人认为在大学做学问是最好的职业，有的人就不愿意跟书本打交道，宁可去开出租车。

AI 会取代一部分工作，但新的工作会被创造出来，实在不行人们还可以甘当消费者。AI 肯定不会取代任何人的命运。

最可能的场景是，所有人的生活都会变得更方便，很多人会在 AI 助手的帮助下做出更明智的选择，所以整个社会风气会变好。但人与人之间的差异将继续存在。

划重点

此次拐点的近前方最可能的场景，也是人类在这场 AI 革命中最好的结局：AI 将拥有超强的智能，智能将变得很廉价；但 AI 不会拥有意识，所以也不会形成真的主动性；主动性仍然掌握在人的手中。对此，我有以下三个判断：

1.AI 没有意识。

2 高水平智能是规模化的结果。

3.选择和决策能力，不是可缩放的。

AI 不会把每个人都变成高级人才，也不会改变很多人的命运。但人与人之间的差异将继续存在。