SDE UNIVERSES
德麦国际专著 · 第 363 号

烟花 · 筛子 · 出海

中国 AI for Science 的困境

AI 接管常规科学之后，中国还缺什么



王德生　著

德麦国际出版社
Demai International Press

德麦国际专著 · 第 363 号

中国 AI for Science 的困境

烟花、筛子与出海——AI 接管常规科学之后，中国还缺什么

王德生　　著

德麦国际出版社
Demai International Press · Singapore

出　版　信　息

书名	中国 AI for Science 的困境——烟花、筛子与出海——AI 接管常规科学之后，中国还缺什么
著者	王德生
执笔	AI 协作（Claude，Anthropic），作者口述立论、承重命题与裁定，并审定
出版	德麦国际出版社（Demai International Press）· 新加坡
编号	德麦国际专著第 363 号
ISBN	979-8-90690-578-9
定价	US$20.00
开本	170 mm × 240 mm（16 开）
字数	约 20 万字
版次	2026 年 10 月第 1 版
完稿	2026 年 10 月 18 日

AI 协作声明：本书的判断、口径与承重命题由王德生（口述）给出；文字在其指令下由 AI（Claude）协作成稿，并经其审定。

书中数据以 2026 年 10 月以前的公开报道为口径，来源见附录一；凡标〔待核〕处，出处尚未逐条核对原始材料，汇总于附录一。书中所引他人言论一律转述大意，不作逐字引文。书中陪读人物（林家四人、老周）均为虚构，不承担证据。

本书 2026 年 10 月 18 日口述的八条承重命题，登记于附录四。

全书结构：十分钟读完这本书、前言、导读、导论；上篇、中篇一、中篇二、下篇一、下篇二、尾篇，共四十八章；结语、参考书目、附录六件、后记。

版权所有　侵权必究

作者介绍

王德生，SDE（显露·差异序列·特征纠缠）本体论思想体系的创立者，德麦国际（Demai International Pte. Ltd.，新加坡）创始人、首席执行官。

他在中国农村长大，1990 年入湘潭大学，后获中国科学院计算数学博士学位。此后在中国科学院做博士后研究，方向涉及 Voronoi 与 Delaunay 剖分、质心 Voronoi 剖分与有限元方法。此后在英国斯旺西大学、新加坡南洋理工大学任职，2005 年起定居新加坡。

他的思想历程经过四站：三视角·知识的结构和体系（2015—2021），321 智慧系统（2021—2025），SIO 本体论（2025—2026），SDE 本体论（2026 年至今）。本书是他第 369 号专著《典范发生学：SDE 与 AI 时代的科研层级》的姊妹卷，著作由德麦国际出版社（Demai International Press）出版。

十分钟读完这本书

如果你只有十分钟，读这几页就够了。

◆ 一、一张流传的图

2026 年夏天，一张图在朋友圈里转了很多遍。左边是中国，右边是美国，各画一根柱子。

中国那根柱子，大半截涂成了短剧和广告，有的版本写着 85%。美国那根柱子，大半截涂成了编程和企业自动化，写着 58%。

图下面的评论，大意都差不多：人家拿 AI 写代码、搞科研，我们拿 AI 拍短剧、卖货。

有人转发时加了一句「扎心」，有人加了一句「这就是差距」，也有人回了一句「短剧怎么了」。一张图，三种脸色。

这张图有真的部分，也有说过头的部分。

真的部分来自一家短剧公司。中文在线 2026 年 6 月初发布的投资者关系活动记录里说，国内 Token 消耗中，AI 短剧与视频生成约 55%，电商与营销约 24%，软件开发约 15%，其他约 6%。原始统计口径没有公开。前两项加起来是 79%，到不了 85%。

美国那个 58%，有作者专门去查过出处，没有找到任何可靠信源。

一边是一家公司援引的行业数据，一边是一个找不到来历的数字。两样拼在一起，就成了一场全民争论。

本书从这张图出发，可是不停在这张图上。图只是一个入口。门后面的问题，比图大得多。

◆ 二、一秒视频吃掉的 Token

Token 是什么？你可以把它想成大模型干活时的计件单位。写一句话、读一段代码、画一帧画面，都按它计。

麻烦在于，不同的活，一件吃掉的 Token 差得很远。

一篇技术博文测算过：10 秒 1080p 的视频，大约要 35 万 Token；同样的事换成文字聊天，只用它的千分之一到万分之一〔待核〕。一个程序员一整个下午和机器来回改代码，未必用得了一个镜头那么多。第 3 章会把这笔账细细算一遍。

你拿一斤棉花和一斤铁比个头，能比出什么来？

所以 Token 的「用量」，先要打个折看。一个国家视频生成做得多，Token 的大盘就一定被视频撑大，这是计量错觉。

另一半是样本错位：一边是全国总量，一边像是开发者平台上的数据，两根柱子本来就不是在同一块地里量的。

剩下真实的那一块，才是本书关心的：资本把 AI 押在了回款最快的地方。

2026 年一季度，国内每天新增 AI 短剧约 1500 部，看短剧的人有八亿之多（第 5 章细说）。钱回得快，钱就往这里排队。中国的 AI，先被用来更便宜地生产注意力。

再看另一张不太流传的图。全球最大的模型 API 聚合平台 OpenRouter 上，2026 年 2 月 9—15 日那一周，中国模型的调用量第一次超过美国模型。据界面新闻测算，到 10 月初，中国模型一周约 86.7 万亿 Token，美国约 19 万

亿。据 CNBC 报道，美国企业在这个平台上，每周把超过 30% 的 Token 发给中国模型。

便宜、好用、开放权重，这是中国模型真实的本事。这张图，很少有人转。

同一份报道也说，写代码、做后台自动化这类出错代价高的活，企业仍愿意为准确多付钱，不轻易换便宜的模型。便宜赢得了量，还没有赢下最要紧的那几单。

先后本身不要紧。要紧的是，注意力挣来的钱和练出来的本事，会不会回流到生产力上去。

◆ 三、火药与烟花

网上有一句话说得更狠：中国人又一次把火药做成了烟花。

这句话有它对的地方：在中国，AI 最先被大规模用起来的地方，确实是生产注意力。不过它错在三处，导论会一处一处拆。

第一，拿火药做烟花不丢人。烟花养活了配方、作坊和工匠；短剧也养活了算力集群、推理工程和一大批做视频模型的人。

第二，一斤火药做烟花，和一斤火药做枪炮，算出来的「用量」本来就不可比。这就是上一节的计量错觉。

第三处最要紧：问题从来不在火药用在了哪里，在下一种火药是谁发明的。

今天这一代 AI 的几种「火药」——反向传播、深度卷积网络、Transformer、大规模稀疏专家模型——几乎都出自别人的实验室。中国很会用火药，用得又快又好又便宜。新火药的配方，很少从这片土里长出来。

故事里的火药是中国人发明的。今天倒过来了：中国一直在用别人发明的火药。

◆ 四、那个「再」字

说起这件事，中国人有一句很体面的话：「先做应用，再做原创。」

这句话有四十年的底子。改革开放以来，面对互联网，中国大体是先做注意力的生意，再往生产力走。先是门户网站、聊天软件、网游，然后是电商，然后是支付和物流，然后是云。路走得很顺，钱也确实一层层往下渗。

计算机却是个反例。八十年代的汉卡、WPS，后来的用友、金蝶，都是先做生产力工具的。可见顺序由条件来定，不由天性来定。

修正以后的规律是：注意力挣来的钱，会回流到应用层的基础设施；它很少一路流到最底下——操作系统、芯片、EDA、工业软件。

「先」字说得很响。「先」字里暗含的那个「再」，四十年里从来没排进过日程。

为什么排不进？因为做应用，是接受别人定义好的题，在题里做到最好、最快、最便宜。原创，是出题。前一件事练出来的本事，不会自己长成后一件事的本事。

DeepSeek 是一个好例子。它在算力受限的条件下，把模型做得又好又省，逼得全世界重算成本。这是应用现场里撞出来的真本事。只是撞出来的，是在别人定下的路上跑得更快，还没有另开一条路（第 14 章细说）。

本书把这个「再」字当作全书的钥匙。每走到一处，都问一句：这里的「再」，排进日程了没有？

◆ 五、三层：改进、架构、典范

要看清「再」为什么排不进去，得先把「原创」拆成三层。

第一层是技术改进：在别人出好的题里，把事做得更快、更省、更准。它来自应用现场的约束。上一节说的 DeepSeek，就是这一层最亮的例子。

第二层是架构跃迁：换一种搭法，让很多旧题一下子有了新解。它常来自应用现场里一块不考核产品指标的研究飞地。2017 年的 Transformer，出自 Google 做机器翻译的研究者。

第三层是科学典范：换掉一门科学认定什么是好问题、什么是好答案的那一整套。它常来自长期反主流、暂时没有应用的少数人。深度学习在学术冷宫里待了几十年，才等来 2012 年的 AlexNet。

中国在第一层极强，第二层少，第三层最空。

为什么？因为三层要的土不一样。第一层要的是约束：卡不够、钱不够、时间不够，逼着人把已有的办法磨到极致。中国最不缺的就是这种约束，也最不缺扛得住约束的人。

第二层要的是一块飞地：身在应用现场，却不用天天对着产品指标交账。第三层要的是更长的冷板凳：十年、二十年没有应用、没有掌声，还有人肯做下去。

这两种土，在中国都很薄。

应用上的碰撞，撞得出第一层，撞不出后两层。

◆ 六、AI 接管了哪一层

现在，把 AI 放进这三层里看。

AI 最先学会的，是给定题目下的解题：查文献、筛分子、调参数、写代码、排实验方案。这正是常规科学——在一套已经认定的做法里，一题一题地解。

第 369 号《典范发生学》说过：AI 能做给定典范下的常规科学；答案、过程，连提问，都不再是竞争点。

这句话放到中国，分量就变了。

中国最擅长的，恰好是常规科学和技术改进：人多、手快、工程能力强、扛得住苦活。这一层越被 AI 接手，这份长处越不值钱。

这并不是说中国的科研人员会失业。一个组里十二个人一起做一年的活，以后两三个人加几台机器，几个月就能做完。做得好不好，越来越不看人多手快，看题出得对不对。

第 369 号《典范发生学》打过一个「车替腿」的比方，第 22 章会细讲。一个国家养出了全世界最好的腿，车一来，这份长处就要重新算账。

反过来，AI 留下来的，是交不出去的那一层：先猜出一套新典范，再让机器去把它搭成系统。这一层，中国最空。

AI 接管的正是中国最强的那一层，留下的正是中国最空的那一层。

◆ 七、比天才更稀缺的那个人

那么，谁来做第三层?

一个人加一台 AI。本书叫它「人—AI 复合体」。

复合体里的那个人，比天才还稀缺。从前的天才，猜想的本事和常规的功底长在一起，功底是他被认出来的凭据。到了复合体里，功底交给了机器，留给人的只剩下纯粹的猜想和判断。剥掉功底以后，还有谁认得出他?

这种人并非没被看见，而是被反向筛掉了。

高考、读博、基金、职称，每一道关选的都是常规科学里最好的那一类人。本书把他们叫作「好腿」：跑得稳、跑得快、跑得远。这套选法公平、管用，选出了一代代扎实的科研人员。只是汽车来了，好腿就不再是交通的关键。

还有一层。今天中国的 AI for Science，多是团队出击、共同体出击。国外有一项很大的研究发现，大团队擅长把科学往前推，小团队更常换掉题目（第 30 章细说）。

团队的长处，原本在典范定下之后的系统打造：几十个人分头把一套新想法搭成能用的东西。偏偏这一段，正被 AI 拿走。

◆ 八、墙外开花与回流

就算有这样一个人，他也得长得出来。

做 0 到 1 的那个人，得是一个思想上的「寡头」：一个人敢定一套新的好问题、好答案。中国的土对他格外不客气。思想权威历来不归个人，讲究「我注六经」（第 33 章）；合法性由机构和头衔发放；「民科」两个字，查的是身份，不查内容（第 35 章）。

这片土也有它的好处：重传承、重积累，一门手艺、一门学问能一代代传下去，不容易断。可一个人自己立一套说法，在这片土里，常常不被当成学问，被当成麻烦。明代的李贽就是一个例子，第 33 章会讲他。

AI 拆掉了资源的墙：一个人租得起算力，查得到文献。合法性的墙还在。这样的人在中国不是长不出来，是长出来也会被归类掉。

于是不少人只好出海求生存。墙外开花墙内香，杨振宁、屠呦呦这些名字，都先在墙外被认出来（第 38 章）。

AI 时代多了一种长法：两栖。机构、出版、注册地在外，根、语言、读者在中文。

难的不在出去，在回来。开在墙外的花，结下的籽能不能落回这片土？本书把这叫作回流的通道。

寡头自己也有风险。一个人加一台 AI，机器会顺着他的猜想把系统搭得越来越完整，越完整，越像被证实了。没有一套能检验典范本身的办法，寡头和「民科」在外人眼里就分不开。尾篇会写这样一套办法。

◆ 九、带走三个问题

十分钟快到了。请你带走三个问题，读到哪一章都可以拿出来用。

这是在改进，还是在出题？

这一层，AI 会不会接手？

这件事要被谁承认才算数？

第一问，帮你分开解题和出题。第二问，帮你看机器走到了哪里。第三问，帮你看那片土。

本书不唱衰，也不唱赞歌。它承认中国的工程本事、规模和成本，也承认短剧挣的是真钱。它只是把一件正在发生的事摆到桌上：机器来了，最值钱的那一层换了位置，而筛人、评人、认人的那套办法，还没跟着换。

尾篇会说，土是可以改的。办法有好几样。一块免于指标的飞地；一个猜想者加一台 AI 加一座公共平台；一套事先写好推翻条件的典范评审；一种让本来没有的材料发生的教育。改土很慢，总要有人先动手。

全书的判断，压成一句话：

AI 越快进入科学，中国最擅长的那一层越先贬值；唯一还稀缺的那一层，中国的土里长不出那个人。

前言

一张图，一个晚上

◆ 一、餐桌上的一个数

2026 年 7 月的一个周五晚上，上海，林家的餐桌。

宋岚把手机扣在桌上，夹了一筷子青菜，说：「这个月又超了。」

林舟问：「超了多少？」

「好几亿 Token。新剧要做仿真人，导演一个镜头不满意，我让机器抽了四十多遍卡。」

宋岚三十五岁，在一家短剧公司做 AI 制片。她不写代码，可她比谁都清楚，一个 1080p 的镜头要多少钱、多少时间。她们组一个月烧掉的 Token，比林舟的组一年还多。

林舟三十八岁，在一家研究院做 AI 制药，带一个十二人的组。他刚交掉一份基金本子，熬了三个晚上，眼睛还是红的。本子里有一栏叫「创新点」，他填了三条：一个更快的筛选流程，一个更准的打分函数，一个把两种模型串起来的办法。

吃完饭，两个人各自靠在沙发上刷手机，又刷到了同一张图。林舟周二夜里就见过它，宋岚公司的群里也早传过一轮（第 1 章）。这一回，是亲戚群里有人转来的。

◆ 二、两个人都不服

就是那张流传的图：中国的 Token 烧在短剧和广告，美国的 Token 烧在编程。

宋岚先开口：「短剧怎么了？一部剧几十号人吃饭，海外平台也在等我们的片子。」

林舟没接话。过了一会儿，他说：「我们做药的也天天用 AI，怎么一张图就把我们抹掉了。」

两个人都不服。不服的方向正好相反。

宋岚不服的是：烟花有什么丢人的？它挣钱，养人，还逼着整条算力的链子越做越便宜。这一点她说得对，本书第 7 章会认真算这笔账。

林舟不服的是：中国的科研并没有闲着。这一点他也说得对。可那天晚上，他盯着手机看了很久，又打开了刚交掉的那份本子。

三条创新点，没有一条是在出新题。三条都是在别人出好的题里，把事做得更快、更准一点。

他想起组里那台天天在跑的 AlphaFold。蛋白结构的题，是别人出的；用来解题的机器，也是别人造的。他的组做得很好，可做得好的，是哪一层？

他翻到本子最后一页，评审意见那一栏还空着。上一次的本子，评审写了一句，大意是：方向前沿，基础扎实，建议资助。他记得自己当时很高兴。

那天晚上，他忽然有点不确定：一个「基础扎实」的本子，是不是正好说明，它问的是一道大家都已经会问的题？

宋岚那边也没闲着。她刷完那张图，打开公司的后台，看了一眼这个月的账单。四十多遍抽卡里，真正用上的只有两遍。她想，要是能少抽一半，一部剧就能省下一个人的年薪。这是她明天要解的题。

这个问题，林舟那天晚上没有想完。本书接着替他往下想。

◆ 三、这本书要问什么

本书要问的是一个很具体的问题：AI 接管常规科学之后，中国还缺什么？

先把答案的方向摆在桌上。中国 AI for Science 的困境，不在算力，不在模型，也不在论文数量。它在一个错位上。

AI 正在接手的，是在给定的题里解题。中国的一整套组织方式——先做应用、团队出击、论文计分、按资历定方向——恰好是为这件事设计的。AI 时代唯一还稀缺的，是先猜出一套新题的本事。它需要一个人加一台 AI，而那个人，在很早的时候就被筛掉了。

这本书一共四十八章，就是把这一句话拆开、摊平、一层层验。

拆开来，是五个小问题。那张图说出了什么，没说出什么？「先做应用」的老路，钱回流到了哪一层？AI 接手的是哪一层，留下的是哪一层？留下来的那一层要的人，从哪里来？这个人长出来以后，谁来认他？

五个问题，对上全书的五篇。第六篇不再提新问题，只看还能做什么。

◆ 四、这本书不问什么

有几件事，本书不做。

不比算力。卡多卡少，是一个要紧的问题，可它不是本书的问题。

不比模型。哪家模型榜单高一点、低一点，三个月就翻一遍。

不唱衰，也不唱赞歌。

公道要先还。2023 年，华为盘古气象的论文发在了 Nature 上。DeepSeek 在算力受限的条件下，把模型做得又好又省，逼着全世界重新算成本。

在 OpenRouter 这个聚合平台上，中国模型的调用量从 2026 年 2 月起超过了美国模型，「十分钟」那一节说过。国家数据局局长刘烈宏 2026 年 3 月说，中国日均 Token 消耗已超过 140 万亿，比 2024 年初增长一千多倍。

工程能力、规模、成本、应用生态，这些都是真本事。本书每一处批评之前，都会先把这些写足，再说它在什么条件下成立，最后才说它停在哪里。

批评只落在三样东西上：激励怎样排、评价怎样打分、人怎样被选出来。

这三样东西，不是哪一个人的错，也不是哪一个机构的错。它们是一片土。土里每一个人都在认真做事：评审认真打分，导师认真带学生，林舟认真写本子。每一个人都做对了，合起来，仍然可能长不出那个人。

本书不找坏人。本书找的是，一片人人都在认真做事的土，为什么会在某一层上空着。

所以你在书里读不到口号，也读不到对哪一方的指责。读到的，是一笔一笔的账。

◆ 五、判词

先把全书的判词摆在桌上：

AI 越快进入科学，中国最擅长的那一层越先贬值；唯一还稀缺的那一层，中国的土里长不出那个人。

◆ 六、前一半，后一半

前一半，说的是一件正在发生的事。

中国最擅长的那一层，是常规科学和技术改进：给定一道题，把它解得又快又好又便宜。这份本事，是四十年攒出来的，是真的。可 AI 恰恰最先学会这一层。机器越快进来，这一层越不值钱。上篇、中篇一、中篇二，写的就是这前一半：烟花怎样挣钱，「先做应用」怎样走了四十年，AI 又怎样接手了中国最强的那一层。

后一半，说的是一件很难改的事。

唯一还稀缺的，是出新题的那一层。做这一层的，是一个敢猜的人加一台 AI。请注意，判词说的是「土里长不出」，不是「中国人不行」。土，指的是那一整套选拔、评价、承认的办法。下篇一写筛子：那个人怎样在高考、读博、基金、职称里被一层层筛掉。下篇二写土：就算他长出来了，也常常被归类掉，只好出海。尾篇再问：土能不能改，怎样改。

前一半是看得见的贬值。后一半是看不见的缺席。

◆ 七、为什么是现在

这本书要是早五年写，判词的前一半就站不住。

五年前，常规科学还主要靠人。查一个领域的文献，一个博士生要读上半年；筛一批候选分子，一个组要跑上几个月；排一套实验方案，要靠老师傅多年的手感。那时候，人多、手快、肯吃苦，就是硬本事。

今天不一样了。2020 年，AlphaFold2 在蛋白结构预测的比赛里夺冠，四年后它的研发者分享了诺贝尔化学奖（第 20 章细说）。林舟的组现在一个晚上跑出来的结构，放在十年前，够一个组做好几年。

机器一题一题地接手，接手的速度一年比一年快。本书写的，是这件事正在发生的这几年。

再晚五年写，也许就来不及了。等常规科学整片交给了机器，那时候才问「中国还缺什么」，筛子已经又筛过了一代人。

◆ 八、为什么是普通人

你也许会想：AI for Science，是科学家的事，和我有什么关系？

关系很近。

你家孩子在刷的题，是一道道给定的题；你在公司里做的方案，多半也是在别人定好的尺子里做得更好；你在招人、评人、提拔人的时候，看的多是简历上的那几样硬东西。筛子不只在科学院里，它在每一个家长群、每一间会议室里。

AI 进入科学，只是最显眼的一处。AI 进入每一行，都会把同一个问题摆到桌上：机器接手了解题，人还剩下什么？

所以这本书不写给专家，写给每一个在给定的题里做得很好、又隐隐觉得哪里不对的人。林舟是这样的人，宋岚也是。

◆ 九、这本书是怎样写成的

本书的判断、口径与承重命题，由王德生博士立题并口述，文字在他的指令下由 AI 协作写成，并经他审定。

这本身就是本书要讲的那件事的一个小样子：一个人出题、下判断，机器把它铺成章节。这样写成的书好不好，不该由写法来判，该由读的人来判。

本书也有它没做的事。它没有给出中国 AI 的完整产业地图，没有逐家比较模型，没有做原始数据的统计。用到的数字，都写明了出处；拿不准的，标了〔待核〕。

◆ 十、姊妹卷

本书是第 369 号《典范发生学：SDE 与 AI 时代的科研层级》的姊妹卷。

那本书问的是全世界的科研：AI 能做给定典范下的常规科学，人要往哪里去？它那个「车替腿」的比方，本书第 22 章会接过来用。

本书把同一个问题，放到中国这片土上再问一遍。车来了，全世界最好的腿部选拔，会变成什么？第 27 章专门问这一句。

两本书分工不同。那一本讲层级本身：AI 时代的科研分成哪几层，人要往哪一层走。这一本讲一片具体的土：中国的组织方式、选拔办法、承认办法，放到那几层上，会是什么样子。那一本是一张地图，这一本是在地图上找一个地方。

两本书可以分开读。读过那一本，这一本会快一些；没读过，也不妨碍。

本书的八条承重命题，由王德生博士在 2026 年 10 月口述，登记在附录四。

◆ 十一、回到那个晚上

那天晚上，林家的两个人没有吵出结果。

宋岚说得对：烟花挣的钱，是真钱。林舟也说得对：中国的科研没有闲着。

本书想替他们问的，是第三件事：烟花挣来的钱、练出来的本事，会不会流回去，变成下一种火药？那份基金本子里的「创新点」，有没有一天，能填上一道新题？

这本书，就写给那个晚上。

导读

◆ 一、五条判断

全书四十八章，压成五条判断。每一条都写成两半：大家看见了什么，大家漏看了什么。

判断	大家看见了	大家漏看了	主场
一、Token 结构说出的，是资本把 AI 押在了哪里	中国的 Token 大半烧在短剧和广告	一半是计量错觉，一半是样本错位；剩下的真实部分，是钱先去了回款最快的地方	上篇，第 3、4、6 章
二、「先做应用，再做原创」里的「再」，四十年没排进日程	中国从应用起步，一路做到电商、支付、云	注意力挣的钱回流到了应用层的基础设施，很少流到操作系统、芯片、工业软件这一层	中篇一，第 11、12 章
三、AI 接管的正是中国最强的那一层，留下的正是中国最空的那一层	中国 AI 的工程突破和前沿成绩	应用碰撞出来的是技术改进；架构和典范这两层，中国很少出题，而 AI 正在接手改进这一层	中篇二，第 17、22 章
四、复合体里的那个人比天才更稀缺，而中国的筛子最会筛掉他	中国的人才多、团队大、论文多	每道关选的都是常规科学的最优者；大团队擅长发展，不擅长颠覆	下篇一，第 25、27、30 章
五、寡头在中国不是长不出来，是长出来也被归类掉	资源的门槛被 AI 拆低了	合法性的墙还在；长出来的人只好出海，难的是回流	下篇二，第 35、41 章

尾篇不再下新判断，只问一件事：土能不能改。

五条判断，是一条链子。

第一条说钱去了哪里。第二条说钱为什么一直停在那里，四十年都没有往下走。第三条说，偏偏在这时候，AI 来了，它接手的正好是钱停留的那一层。第

四条说，留下来的那一层要一种很特别的人，而这种人被筛子筛掉了。第五条说，就算他没被筛掉，长出来了，这片土也不认他。

前三条是一件事：中国的长处在哪一层，那一层正在贬值。后两条是另一件事：唯一还稀缺的那一层，人从哪里来，谁来认。

链子上任何一环断了，后面的推论都要改。所以本书每一篇末尾都记一次总账，看这一环有没有立住。

◆ 二、判词

五条判断收拢起来，就是全书的判词：

AI 越快进入科学，中国最擅长的那一层越先贬值；唯一还稀缺的那一层，中国的土里长不出那个人。

判断一、二、三，撑的是前一半。判断四、五，撑的是后一半。

◆ 三、全书地图

全书分六篇。

上篇「烟花」八章，从那张流传的图起步。第 2 章拆 55% 是怎么算出来的，第 3 章拿十秒视频和一千行代码比 Token，第 4 章去找那个找不到出处的 58%。第 5、6 章说钱为什么先排到短剧前面。第 7 章问最难的一个问题：烟花先养后用，还是从此锁死。

中篇一「先做应用」八章，往回走四十年。第 9 章从汉卡和财务软件讲起，那是一个先做生产力的反例。第 10、11 章跟着钱走：网吧、网游、电商、支

付，钱回流到了哪一层。第 12 章专讲那个「再」字。第 13 到 15 章说解题和出题的差别，以及为什么撞墙之后改出来的，多是改进，不是 0 到 1。

中篇二「层级」八章，把原创拆成三层。第 18 章讲 Transformer 出自一个翻译组，第 19 章讲深度学习在冷宫里的三十年。第 20 章分开「AlphaFold」和「用 AlphaFold」。第 22、23 章接第 369 号《典范发生学》：常规科学交给机器，连提问也不再稀缺。

下篇一「筛子」八章，找那个人。第 25、26 章说复合体里的人是谁，剥掉功底之后还剩什么。第 27 到 29 章看高考、博士、基金、职称和论文计分怎样一层层筛。第 30、31 章讲大团队和小团队，以及方向由谁来定。

下篇二「土」十章，讲那个人为什么难发芽。第 33 到 36 章从「我注六经」讲到头衔、「民科」和谁来判一个新典范。第 37 章转过身，讲寡头自己的风险。第 38 到 41 章讲墙外开花、出海、两栖和回流。

尾篇「改土」六章，写还能做什么：一块免于指标的飞地，一个猜想者加一台 AI 加一座公共平台，一套新的典范评审，一种新的教育。最后两章，一章请你看自己站在哪一层，一章写本书自认最弱的几处。

每篇最后一章是「总账」，把这一篇的看见与漏看并排成一张表。

正文之后是结语，把五条判断再收一遍，判词一字不差地放回来。再后面是参考书目和附录：Token 数据的出处与核验表、中美 AI for Science 的大事年表、三层速查表、口述命题登记、十大常见误读、推翻条件与练习。

六篇的走法，是从小到大，再从大回到小。上篇看一张图，中篇看四十年，下篇看一个人，尾篇回到你手上能做的事。

◆ 四、读法路线

四十八章不必从头读到尾。按你手上的事挑。

只有一个下午：读导论、第 6 章、第 12 章、第 17 章、第 25 章、第 41 章，再读结语。

你在做科研：先读中篇二（第 17 到 24 章），再读第 28 到 31 章，最后读第 44、45 章。你会知道自己的活，哪一部分明年就要交给机器。

你在带孩子：读第 13 章（解题和出题）、第 26 章（剥掉功底之后还剩什么）、第 27 章（全世界最好的腿部选拔），再读第 46 章（教育）和第 47 章里「带孩子的」那一节。

你在管一家公司：读上篇第 5 到 7 章，中篇一第 11、12、14 章，再读第 43 章（一块免于指标的飞地）。第 47 章里「管公司的」那一节，可以拿来自查。

你在出海：读第 38 到 41 章，再回头读第 37 章。出去之前，先想好怎么回来。

◆ 五、全书的十二个画面

记不住章题，也可以记住画面。全书大体靠这十二个画面往前走：

一张流传的两国 Token 柱状图；宋岚抽了四十多遍卡的一个镜头；老林县城厂里的第一块汉卡；九十年代末那家网吧里的通宵；一个「再」字；Google 翻译组里的一篇论文；冷宫里坐了几十年的几个人；林舟组里天天在跑的 AlphaFold；豆豆的一摞奥数卷子；一份写满「基础扎实」的基金评审意见；一朵开在墙外的花；一条还没修好的回流的路。

十二个画面，前六个讲烟花和老路，后六个讲层级、筛子和土。

◆ 六、按问题找章

你也可以带着一个问题来，直接翻到那几章。

那张图到底对不对？读第 2 到 4 章。

短剧挣的钱，会不会回到科研上？读第 7 章和第 11 章。

DeepSeek 不算原创吗？读第 14、15 章，再读第 17 章。它算，要看算在哪一层。

AlphaFold 来了，做研究的人还剩什么？读第 20、22、23 章。

中国的 AI 也上了 Nature，那算不算出了新题？读第 21 章。

为什么中国的顶尖学生，很少做出新典范？读第 26 到 29 章。

团队大一点，不是更容易出成果吗？读第 30 章。

一个人带着 AI 做出了新东西，凭什么让人信？读第 36、37 章，再读第 45 章。

去海外注册一家机构、出一本书，算不算出路？读第 39、40 章。

◆ 七、每一章怎样读

每一章大体走同一条路。

先是一个你遇到过的场面，或者林家某个人的一个晚上。然后摆材料：数字、史实、别人的研究。接着是本书的说法。再看一看，只盯着一面看会怎样。然后放到今天、放到机器面前。再给一个完整的例子。最后回到那个人、那个晚上，给一件明天能做的事。

章末是一句引语、一句带走的话、两行小账和一到三个「想一想」。

章与章之间常有交叉，比如「第 12 章说过」「第 41 章会回来」。顺着跳过去读，也行。

◆ 八、陪读的一家人

书里会反复出现上海的一家人。

林舟三十八岁，在一家研究院做 AI 制药，带一个十二人的组。他天天写基金本子、赶论文，用 AlphaFold 和大模型筛分子。他是一路被筛子筛出来的好学生、好博士、好组长。

宋岚三十五岁，林舟的妻子，在一家短剧公司做 AI 制片。她一个月烧掉的 Token，比林舟的组一年还多。她聪明、务实，算账比谁都清楚。

林豆豆九岁，读小学四年级，刷题、上奥数，爱问「为什么」，也会用 AI 写作业。她出现在讲教育和筛子的那几章。

老林六十六岁，林舟的父亲。八十年代在县城国营厂当会计，用过汉卡，学过 WPS，装过用友的财务软件；九十年代末开过一家网吧。他在中篇一里讲那四十年的老路。

四个人，正好站在本书的四个位置上。宋岚站在上篇：烧 Token 的那一头。老林站在中篇一：四十年的老路。林舟站在中篇二和下篇一：被筛出来、又被机器追上的那一层。豆豆站在下篇一和尾篇：下一代，还没被筛完的那一个。

他们不会每章都出场。一章至多一两个人。

这一家人是虚构的。他们不承担证据，只负责让每一个判断，落到一家人的晚饭桌上。

书里的数据和史实，另有出处；凡是本书拿不准的，都照实标出来。

◆ 九、随身三问

读到哪一章，都可以拿出这三个问题。

这是在改进，还是在出题?

这一层，AI 会不会接手?

这件事要被谁承认才算数?

第一问分开解题和出题。第二问看机器走到了哪里。第三问看土：一个新东西长出来以后，谁有资格说它算数。

三问都答完，你对一件事的判断，大概就不会被一张图带着走了。

拿林家试一次。宋岚的组把一个镜头的抽卡从四十遍压到二十遍：这是改进；这一层，AI 会接手，而且正在接手；它要被平台的完播率承认才算数。林舟想在本子里写一个全新的药物研究问题：这是在出题；这一层，AI 一时接不过去；可它要被谁承认？基金评审，还是十年后的结果？第三问一出来，难处就露出来了。

◆ 十、判词怎样才算错

判词能被推翻，才配站在第一页。下面几条，写成人话，说清哪一种事发生了，判词哪一半就要改。

第一条，冲着后一半。假如十年之内，中国从现在的大团队体制里，长出一个被世界公认的科学新典范。请注意，要的不是更快更准地解旧题，是换了一门科学认定的好问题和好答案。真有这样一件事，「土里长不出那个人」就判错了，至少要大大收窄。

第二条，也冲着后一半。假如有人拿出可靠的研究，说明高考、读博、基金、职称这几道关选出来的人里，做出新典范的比例并不低于别处。那样的话，「筛子最会筛掉他」就要改判。

第三条，冲着前一半。假如 AI 并没有接管常规科学：十年后，查文献、筛分子、调参数、排实验方案，仍然主要靠人一题一题地做，机器只是一个快一点的工具。那样的话，「最擅长的那一层越先贬值」就不成立。

第四条，冲着上篇。假如有一份可靠、口径公开的统计，表明中国的 Token 结构并不像传言那样偏向短剧和广告，或者美国的结构其实和中国差不多。那样的话，上篇关于「资本押在哪里」的推论就要重写。请注意，这一条推翻的是上篇的起点，不直接推翻判词：判词说的是层级，不是用量。

第五条，冲着中篇一。假如有人能说明，过去四十年里注意力挣来的钱，有相当一部分已经回流到了底层原创，而且长出了被世界采用的东西。那样的话，「那个『再』字没排进日程」就要收回。

本书第 48 章会回来，把这几条再过一遍，并写下本书自认最弱的几处。

还有一种推翻，不来自中国，来自科学本身。假如十年后人们发现，AI 不只会解题，也开始稳定地提出被世界采用的新典范，人只剩下签字。那样的话，判词的后一半就失去了意义：那个人稀缺不稀缺，已经不要紧了。本书认为这件事在可见的将来不会发生，可这只是本书的判断，标作待验。

◆ 十一、五条判断的反方

公道要先还。每一条判断，都有一位认真的反方。本书在正文里会请他们开口，这里先把他们的话摆出来。

对判断一，反方会说：烟花本身就是生产力。短剧出海、平台挣钱、算力降价，哪一样不是实打实的？本书同意一半：它们是真钱，是真本事。本书只追问，这些钱最后流到了哪一层。

对判断二，反方会说：「再」字已经在排了，看看这几年的芯片、操作系统和大模型。本书同意，排进日程的东西比十年前多了。可多数是在追一套别人定好的题，追得很快，题还是别人的。

对判断三，反方会说：DeepSeek 不就是原创吗？本书会在第 14、15 章细细回答：它是原创，原创在第一层。这一层的原创很了不起，也正是 AI 最先接手的那一层。

对判断四，反方会说：天才自己会冒出来，筛子筛不掉真正的天才。本书的回答是：从前的天才靠功底被认出，功底交给 AI 以后，他就少了被认出的凭据。

对判断五，反方会说：真有本事的人，总会被承认，只是早晚。本书同意「早晚」两个字。可在 AI 的速度下，晚十年，常常就是错过了。

◆ 十二、如果你只读一章

如果你只读一章，读第 12 章「那个『再』字」。

那一章把全书的钥匙交到你手里。读完它，你再回头看那张流传的图，看林舟的基金本子，看你自己这个月的待办清单，都会多问一句：这里的「再」，排进日程了没有？

如果你还有一点时间，再读第 25 章「复合体里的那个人」。那一章说的是判词后一半里的那个人：他是谁，为什么比天才还难找。

◆ 十三、给读书会和老师

这本书适合几个人一起读。

一个办法：每次读一篇，每人带一笔自己的小账来。比如读上篇时，各自翻一翻手机里这个月用过的 AI 应用，记下哪一样是在帮你生产注意力，哪一样是在帮你做事。下一次见面，把账摆在桌上对一对。

给老师的一个办法：读中篇一第 13 章时，让学生各出一道题，而不是各解一道题。看看出题比解题难在哪里。

◆ 十四、引号和〔待核〕

最后说两件读书时用得上的小事。

凡带引号处，请读作「大意如此」。外国学者和今人的文章，本书只转述大意，不整句照抄。

凡标〔待核〕的地方，是本书还没有找到第二个可靠出处的数字或说法，比如一篇单一来源的技术博文。它们都汇进附录里的数据出处与核验表，你可以照着去查。查到不对的，请告诉我们。

目 录

作者介绍 iii

十分钟读完这本书 iv

前言　一张图，一个晚上 xiii

导读 xx

导论　火药与烟花 1

上篇　烟花：一张图引发的争论 **15**

第 1 章　一张流传的图 16

第 2 章　55% 是怎么算出来的 25

第 3 章　十秒视频和一千行代码 34

第 4 章　找不到出处的 58% 42

第 5 章　八亿人和每天一千五百部 51

第 6 章　资本按回款速度排队 60

第 7 章　先养后用，还是锁死 69

第 8 章　本篇总账：烟花的账 78

中篇一　「先做应用」：四十年的老路 **87**

第 9 章　汉卡与财务软件：计算机是一个反例 88

第 10 章　网吧和网游的第一桶金 96

第 11 章　从电商到支付：钱回流到了哪一层 105

第 12 章　那个「再」字 114

第 13 章　解题的能力和出题的能力 123

第 14 章　撞墙之后改的是什么 132

第 15 章　改进不是 0 到 1 141

第 16 章　本篇总账：应用的账 150

中篇二　层级：AI 接管了哪一层 **159**

第 17 章　三层：改进、架构、典范 160

第 18 章　Transformer 出自一个翻译组　168
第 19 章　冷宫里的三十年　176
第 20 章　AlphaFold 和「用 AlphaFold」　185
第 21 章　前沿的成绩，旧的题目　194
第 22 章　常规科学交给机器　203
第 23 章　连提问也不再稀缺　212
第 24 章　本篇总账：层级的账　221

下篇一　筛子：那个人从哪里来　229
第 25 章　复合体里的那个人　230
第 26 章　剥掉功底之后还剩什么　239
第 27 章　全世界最好的腿部选拔　248
第 28 章　博士、基金、职称：三道筛子　256
第 29 章　论文计分遇上 AI　264
第 30 章　大团队发展，小团队颠覆　272
第 31 章　方向由谁来定　280
第 32 章　本篇总账：筛子的账　289

下篇二　土：思想寡头为何难发芽　297
第 33 章　我注六经　298
第 34 章　头衔发放的合法性　306
第 35 章　「民科」两个字　314
第 36 章　谁来判一个新典范　322
第 37 章　寡头的自我证实　330
第 38 章　墙外开花墙内香　339
第 39 章　出海：求生存，还是取通行证　347
第 40 章　两栖：注册地在外，根在中文　356
第 41 章　回流的通道　364
第 42 章　本篇总账：土的账　373

尾篇　改土：还能做什么　381
第 43 章　一块免于指标的飞地　382
第 44 章　一个猜想者、一台 AI、一座公共平台　390

第 45 章　给典范评审一个新机制 399
第 46 章　教育：让本来没有的材料发生 408
第 47 章　你站在哪一层 417
第 48 章　本书自认最弱的几处 425
结语　那个晚上之后 433
参考书目 439
附录 445
附录一　数据出处与核验表 446
附录二　年表：中国面对新技术的四十年，与 AI for Science 的十几年 448
附录三　三层速查表 450
附录四　口述命题登记 451
附录五　十个常见误读 452
附录六　推翻条件与练习 453
后记 455

导论

火药与烟花

◆ 一、网上的那句话

那张图流传开以后，评论区里有一句话被点赞最多，大意是：中国人又一次把火药做成了烟花。

这句话很好记，也很伤人。它借了一段大家从小听熟的故事：火药是中国人发明的，可中国人拿它放烟花、做爆竹，别人拿它造枪炮、开矿山、修铁路。到了 AI 这里，好像故事又讲了一遍：别人拿 AI 写代码、搞科研，中国拿 AI 拍短剧、卖货。

本书借这句话作全书的总比喻。可是借之前，要先把它拆一遍。

一个好比喻，要先承认它对在哪里，再说清它错在哪里。错处要正好落在承重的地方，比喻才帮得上忙。

◆ 二、先说它哪里对

先说它对的地方。

它说出了一个真实的方向：在中国，AI 最先被大规模用起来的地方，是生产注意力。

中文在线 2026 年 6 月的记录里说，国内 Token 消耗中，AI 短剧与视频生成约 55%，电商与营销约 24%。就算这个口径没有公开、就算有一半是计量错觉，钱先去了哪里，方向是看得出的。

2026 年一季度，国内每天新增 AI 短剧约 1500 部，每月上新超过 4 万部，差不多等于 2025 年全年真人短剧的产量。看的人也多：一个头部短剧平台，每天打开它的人上亿，人均一天要看两个小时（第 5 章）。

这是一片很大的烟花。比喻在这一点上没有冤枉人。

它还说出了一点不安：中国很会用别人想出来的东西，用得又快又好。可轮到下一样东西从哪里来，人们心里没底。

这一点不安，也是对的。本书后面的四十多章，都在把这一点不安说清楚。

◆ 三、错的第一处：烟花不丢人

可这个比喻，至少错在三处。

第一处：拿火药做烟花，本身不丢人。

烟花是一门很难的手艺。配方要准，颗粒要匀，装药要稳；一枚烟花在天上开成什么样，是一代代工匠试出来的。烟花作坊养活了硝石、硫磺和木炭的生意，也养活了懂火药脾气的人。没有这些人，后来想造别的东西，也找不到人手。

AI 短剧也一样。宋岚的公司为了一个仿真人镜头，反复抽卡、反复筛选。中文在线解释过，仿真人剧对帧数和 1080p 清晰度要求高，Token 消耗通常是 720p 漫剧的两到三倍。这份苛刻，逼着推理成本一降再降，逼着视频模型一代比一代稳。

钱是真钱，人是真人，本事也是真本事。中文在线旗下的海外短剧平台 FlareFlow，截至 2026 年 5 月底注册用户超过 3700 万。

回到故事本身，也该说句公道话。火药做成烟花，不是古人笨。在那片土里，节庆、祭祀、驱邪，是火药最先能换到钱和体面的地方；造枪炮、开矿山，

要的是另一整套工匠、冶金和需求。火药往哪里走，从来不是一个民族的脾气决定的，是那片土里的钱和路决定的。

今天的 AI 也一样。资本把它押在回款最快的地方，是资本的常态，不是中国人的天性。美国的资本也追回款，只是那片土里，回款快的地方恰好有一大块在写代码、做企业软件。

所以本书不笑话烟花。问题不在做烟花，在做完烟花之后。

◆ 四、错的第二处：两斤火药不能比

第二处，是一个计量的错觉。

一斤火药拿去做烟花，一斤火药拿去造枪炮，算出来都是「一斤」。可这两斤，能说明同一件事吗？

做一挂鞭炮，要的火药很多；改一支枪的膛线，可能一克火药都不用，靠的是图纸和车床。按用量排，鞭炮赢；按改变了什么排，就是另一回事。

Token 也是这样计的。「十分钟」那一节说过，一个视频镜头吃掉的 Token，抵得上文字聊天的千倍万倍；一个短剧项目，几亿 Token 起步〔待核〕。第 3 章会把这杆秤细细拆开。

一个程序员和机器来回改一下午代码，用掉的 Token，可能还不如宋岚的一个镜头。可那个下午改出来的东西，也许会进一套工业软件，用上十年。

所以「中国 Token 烧在短剧」这句话，先要打个折。视频做得多的地方，Token 的大盘一定被撑大。何况美国那个 58%，本书去查过，找不到出处，只能照实写成「找不到出处」（第 4 章）。

用量不可比，这是第二处错。

◆ 五、错的第三处：下一种火药是谁发明的

第三处最要紧。

那个比喻把问题放在「火药用在哪里」。好像只要中国人把 AI 从短剧挪到科研上，问题就解决了。

事情没有这么简单。

真正的问题，是下一种火药是谁发明的。

这一轮 AI 用到的火药，一层一层数下来：1986 年的反向传播论文；2012 年的 AlexNet；2017 年 Google 的 Transformer；专家混合模型的想法，1991 年就有人提出，2017 年在 Google 做成了大规模的稀疏模型。

中国在这条线上也有自己的一笔。DeepSeek 在算力受限的条件下，把这几样东西做到了极致，还添了自己的新招，逼着全世界重算成本（第 14 章细说）。这是真本事，也是这一轮里中国最亮的一笔。

请看清它是哪一种本事：在别人出好的题里，把事做到最好最省。火药用到了极致，下一种火药还没有发明出来。

打个比方。一家饭馆，把别人家传下来的一道菜，做得比原来那家还好吃、还便宜、上菜还快，天天排长队。这是真本事。可这家饭馆的菜单上，没有一道菜是它自己想出来的。哪天顾客的口味换了，换成一道谁也没做过的新菜，它只能等别人先做出来，再去学。

学得快，是一种本事；等别人先做出来，是一种处境。两样可以同时是真的。

故事里的中国人，火药是自己发明的。今天的情形倒过来：中国一直在用别人发明的火药。

所以，问题不在拿火药做烟花，在没有人发明下一种火药。

这才是本书借这个比喻要说的话。

历史上，一种火药换成下一种，常常要很多年，还要换掉一批人。

普朗克晚年说过一个意思，新真理常常要等一代人换过去才算赢，第 31 章会回来讲。魏格纳 1912 年提出大陆漂移，等了半个世纪才被普遍接受（第 36 章）。

新火药出来的那一刻，常常不像火药，像一个怪人的怪话。旁人听了，第一反应多半是摇头。这不怪旁人：旧火药正好用，谁也没有理由先去信一句怪话。

◆ 六、三个关键词：典范

要把「下一种火药」说清楚，得先交代本书最常用的三个词。

第一个词是「典范」。

1962 年，美国科学史家库恩写了一本书，叫《科学革命的结构》。他用了一个词 paradigm，中文常译作「范式」，本书译作「典范」。

用大白话说：典范，是一门科学认定什么是好问题、什么是好答案、怎样算做对的那一整套。

打个比方。一个学校里，什么样的题算好题，什么样的解法算漂亮，老师打分看哪几步，学生从来不用想。可正是这一整套，决定了每一个人怎样学。典范就是一门科学的这一整套。

牛顿以后，物理学家认定，好问题是「这个东西受了什么力」，好答案是一条能算出数的方程。到了量子力学，好问题和好答案都换了。换的不是某一道题的答案，是题本身。

◆ 七、常规科学，和 0 到 1

第二个词是「常规科学」。

库恩说，一门科学的大多数时候，科学家都在一套典范里解题。题是典范给的，什么算解出来了，也是典范定的。科学家的本事，是在这套题里解得又多又好。他把这种时候的科学，叫常规科学。

常规科学一点都不低级。今天绝大多数论文、专利和药物，都是常规科学做出来的。林舟的组用 AlphaFold 预测蛋白结构、用大模型筛分子，做的就是常规科学，而且做得很好。

中国自己也有好例子。2023 年，华为的盘古气象大模型发表在 Nature 上，用 AI 把中期天气预报做得又快又准。题是气象学早就定好的，尺子也是公认的；它在这道题里拿了一流的成绩。这是 AI 进入常规科学的一个样板，第 21 章会细看。

本书不把常规科学当成次等的活。本书只问一件事：当机器越来越会做这一层，做这一层最好的那个国家，下一步做什么。

第三个词是「0 到 1」。

本书说的 0 到 1，指的是换题本身：一门科学认定的好问题和好答案，换了一套。做出一样世界上没有的新产品，未必算；换了题，才算。

这样，三个词的关系就清楚了。常规科学，是在给定的典范里解题。0 到 1，是换一个典范。

第 369 号《典范发生学》说过：AI 能做给定典范下的常规科学。换句话说，AI 正在接手解题；换题，还没有交出去。

下一种火药，指的就是下一套典范。

◆ 八、林舟的一个星期

这三个词，放到林舟的一个星期里看，会清楚得多。

周一上午，组里的博士后小顾拿来一批结果。他用 AlphaFold 预测了三百多个蛋白的结构，挑出二十个口袋形状合适的，再用大模型给五万个候选分子打分。林舟看了一遍，点了点头。题是清楚的：这个靶点，找一个能卡进去的分子。什么算好答案，也是清楚的：打分高、合成得出来、毒性低。

这是常规科学。题是典范给的，尺子也是典范给的。

周三下午，组里吵了一架。一个刚来的博士生说，这个靶点也许本来就选错了；这个病，也许不是一个蛋白卡住了，是好几样东西缠在一起、一起往一个方向推。要是这样，「找一个分子卡进一个口袋」这道题本身就问偏了。

屋里安静了一会儿。

林舟心里知道，这个学生说的不是没有道理。可他也知道，这种想法写不进本子：它没有现成的方法，没有预实验，没有发表记录，评审不知道拿什么尺子打分。他让学生先把手上的筛选做完。

周五晚上，他回家，在餐桌上听宋岚说她这个月又烧了多少 Token。

那个博士生周三下午说的，就是一次可能的 0 到 1：他没有把题解得更好，他怀疑的是题本身。它也许是对的，也许是错的。可在林舟的组里，它没有一个能落脚的地方。

◆ 九、今天，机器站在哪里

把这个星期再看一遍，看机器站在哪里。

周一那一批活，机器已经接走了大半。查结构、筛分子、打分、写报告，过去要一个组忙一个季度，现在小顾一个人加几台机器，几天就够。再过几年，这一段很可能整个交给机器，连「挑哪二十个口袋」也不用人来挑。

周三那一场争吵，机器帮不上太多忙。机器也会想，只是它要先知道题。你告诉它「这个病也许不是一个蛋白的事」，它可以帮你查文献、找证据、搭模型，搭得又快又完整。可「换一道题」这一步，得先有一个人敢说出来，还得有人愿意让他试。

这就是本书一再要回来的那个地方：机器越会解题，出题那一步越显得孤单。

AI 进入科学，最先改变的是周一。周三还是老样子。

◆ 十、一部短剧的一锅粥

典范、常规科学、0 到 1，这几个词有点硬。本书再借几个家常的字，把它们磨软一点。

这套书里有七个常用的小词：形状、路、土、回写、一锅粥、正在成形、定型。你不必记住它们，也不必把它们当术语。本书用得很轻，真正常出场的只有三个：土、路、回写。其余四个，跟着下面这部剧认个脸就够了。

先看宋岚怎样做一部 AI 短剧。

去年冬天，宋岚的组接了一个项目：一部六十集的仿真人古装剧。开会那天，桌上什么都有：三个版本的梗概，二十多张参考图，导演的一段口述，平台给的完播率数据，还有一张算力的预算表。

所有的东西都在桌上，可还搅在一起，谁也说不清这部剧到底是什么样。厨房里管这叫一锅粥：什么都在，还没分开。

接下来两周，组里一遍遍地试。第一版主角的脸太像某个明星，删掉；第二版画面太亮，不像古装；第三版节奏太慢，前三秒留不住人。每一次改，都是照着上一次的毛病改的。一步接一步，前一步的毛病，引出后一步的改法，这就是一条路。顺序不能乱：没有第一版的毛病，就没有第二版的改法。

到第三周，主角的脸稳住了，画面的色调有了，前三集的节奏也顺了。可每做一集，还会冒出新的问题。样子冒头了，还站不稳，这时候叫正在成形。这一段最出东西，也最容易散。

第五周，前十集上线，完播率过了平台的线。后面五十集，照着这个样子往下做就行。稳住了、认得出、传得下去，这就叫定型。那部剧最后长成的样子，就是它的形状。

可为什么长成了这个形状，不是别的？因为这部剧长在一片土里。

平台的算法偏爱前三秒的冲突；观众在地铁上看，一集不能超过两分钟；仿真人每多一秒就多一份钱，导演又偏爱一种冷色调。这些东西缠在一起，像地里的水、肥和气候，本书就叫它土。同一个梗概，放到另一家平台、另一笔预算、另一位导演手里，长出来的不会是同一部剧。

剧上线以后，完播率的数据又回到了组里。宋岚把它写进了下一部剧的开会材料：前三秒怎么起，色调怎么定，抽卡抽到第几遍就该停。这一部剧，回过头改了下一部剧的土和路。本书管这个叫回写，就像庄稼收完，秸秆又埋回地里。

◆ 十一、那部剧缺了什么

把宋岚的例子放在这里，为的不是讲短剧，是看清一件事。

那部剧从一锅粥长到定型，走了一条很漂亮的路。可这条路是在哪片土里走的？是在平台的算法、观众的习惯和完播率这把尺子里走的。

那把尺子，就是这一行的典范：什么是好剧，前三秒要有冲突，两分钟一个钩子，完播率说了算。宋岚的组在这套典范里，一题一题地解，解得又快又好。

这是常规科学的样子，搬到了短剧这一行。

AI 帮她做得更快了。四十多遍抽卡，放在几年前要一个组拍一个月，现在几个晚上就够。机器接手的，正是在给定的尺子里一遍遍试的那一段。

可是有一件事，机器没有替她做，她的组也没有人去做：问一问那把尺子本身对不对。

科学也是这样。林舟的组在蛋白结构这把尺子里做得很好；可蛋白结构为什么是药物研究的好问题，这把尺子是谁定的，下一把尺子会是什么，组里没有人的工作是去想这件事。

这就是回写最难的地方：每一部剧都回写了下一部剧，可很少有人去回写那把尺子。

后面各章，你只要记住三个字的意思就行。

土，是一个人、一件事长在里面的那一整片条件。放到科研上，就是选人的办法、打分的办法、承认的办法。

路，是一件事一步步走过来的先后。

回写，是长出来的东西，回过头改了那片土。

比如，「先做应用」挣来的钱，有没有回写到底层的土里？一个在墙外长出来的人，有没有回写到中文这片土里？全书的很多问题，说到底都是这一个问题：长出来的东西，回不回得来。

另外四个小词，只在要紧处露一下面。碰到了，想想宋岚这部剧就够了。

◆ 十二、一个典范的一锅粥

宋岚那部剧的路，是在定好的尺子里走的。一个新典范的路，要难得多：它连尺子都没有。

深度学习就是一个例子，第 19 章会细讲它在冷宫里的那些年。这里只看它走过的几步。

上世纪八九十年代，神经网络在不少人眼里是一条走不通的路。那时候打分的尺子，是另一套方法定的：理论漂亮、能证明性质，才算好答案。神经网络那边，只有几个想法、几个小实验、几个坚持的人，还是一锅粥。

到 2012 年 AlexNet 在图像识别比赛里大幅领先，它才算正在成形。又过了几年，大家认定了一套新的好问题和好答案：给足数据和算力，让网络自己学，比谁的效果好。这时候它定型了，成了一个新的典范。

这条路上，最要紧的不在哪一次实验，在那几十年里，有一片土允许几个人一直做一件不被看好的事。

本书下篇要问的，就是这片土。

◆ 十三、三层：改进、架构、典范

现在可以把「原创」拆开了。本书把它拆成三层。

第一层，技术改进。在已经定好的典范和架构里，把事做得更快、更省、更准。它常常来自应用的约束：卡不够、钱不够、时间不够，只好把已有的办法磨到极致。DeepSeek 在算力受限的条件下做出的那些工程突破，就是这一层。

第二层，架构跃迁。换一种搭法，让一大片旧题一下子有了新解。它常来自应用现场里一块不考核产品指标的研究飞地。Transformer 出自 Google 做机器翻译的研究者，他们身在应用现场，却有余地试一种新搭法（第 18 章）。

第三层，科学典范。换掉一门科学认定的好问题和好答案。它常来自长期反主流、暂时没有应用的少数人。上一节的深度学习，就是这样等出来的。

一张小表，放在这里，后面还要回来：

层	做什么	常从哪里来	一个例子
技术改进	在给定的题里做得更好	应用的约束	算力受限下把 MoE 做省
架构跃迁	换一种搭法解一大片题	应用现场里的研究飞地	翻译组里长出的 Transformer
科学典范	换掉好问题和好答案	长期反主流的少数人	冷宫里的深度学习

中国在第一层极强，在第二层少，在第三层最空。第一层的强，是四十年真刀真枪练出来的，不打折扣。

应用上的碰撞，撞得出第一层，很少撞得出后两层。中篇二会一层一层拆这张表，第 17 章先把它立起来。

◆ 十四、三层放回林家

这张表，放回林家的一个星期，又是一番样子。

宋岚的组把抽卡从四十遍压到二十遍，是第一层。林舟的组把筛分子的流程从一个季度压到几天，也是第一层。两口子都在第一层做得很好。

假如林舟的组里有一块地方，不考核论文，不考核本子。那个周三下午吵架的博士生，可以在那里试一种全新的建模办法，试一年、两年，失败了也不扣分。那里也许会长出第二层的东西。可这块地方，现在没有。

至于第三层，那个学生说的「这个病也许不是一个蛋白的事」，要是真能长成一套新的好问题和好答案，就是第三层。它需要的不只是一块地方，还要很多年，要有人肯在没有应用、没有掌声的时候一直做下去。

一层比一层难，一层比一层慢，一层比一层不像「成果」。

再把老林拉进来看一眼。八十年代，他在县城的国营厂当会计，第一次用上汉卡，把算盘换成了电脑上的中文。那是一件实打实的生产力：一个厂的账，从一个月做完变成一个星期做完。汉卡和后来的 WPS、用友，在当时那片土里，都算是出了一道新题：中国人怎样用上电脑。

那道题出得好，因为出题的人就站在题旁边。中篇一会回到老林，看这条路后来怎样拐了弯。

所以中国在第一层的强，并不能一年年地累加成第二层、第三层的强。三层不是一级一级的台阶，是三片不同的土。

◆ 十五、先清三个误会

在往下读之前，要先清掉三个误会。

第一个误会：本书是说中国 AI 不行。

不是。中国的工程能力、规模、成本和应用生态，都是真本事。前面说过的 DeepSeek、盘古气象、OpenRouter 上的调用量，都是事实。本书说的是层级：中国最强的那一层，恰好是 AI 最先接手的那一层。这是在说位置，不是在说高低。

第二个误会：本书是说做应用做错了。

也不是。烟花不丢人，先做应用也不丢人。四十年里，应用挣来的钱，养起了电商、支付、物流和云，这些都回流成了真本事。本书问的不是「先」字对不对，是那个「再」字有没有排进日程。

还要补一句：体制自己也看见了问题，这几年一直在推「破四唯」（第 28 章）。本书写的筛子，是一个正在被松动的筛子。本书问的，是松动得够不够快，松开以后，土里能长出什么。

第三个误会：本书是说出路就是出国。

更不是。下篇二会讲很多人出海，可出海是被逼出来的求生存，不是一条好路。海外的土，也不为典范猜想准备；出去的人，根和养分常常被拉开。本书真正关心的，是回流：长在墙外的东西，能不能回到这片土里，改一改这片土。尾篇写的，也是改土，不是搬家。

◆ 十六、小账说明

最后，交代每章末尾的两行小账。

每一章最后，都有一句「带走的话」，接着两行小账，再接一个「想一想」。

第一行小账：看见与漏看。问的是，这一章说的事里，大家看见了什么，漏看了什么。

第二行小账：改进还是出题。问的是，这一章所说的事，是在已有的题里做得更好，还是在出新题。

导论本身的两行小账，就记在本章末尾，你可以先翻过去看一眼样子。

两行小账之后的「想一想」，问的不是书里的知识，是你自己。

读完全书，把四十多笔小账连起来，就是一本总账。

> **问题不在拿火药做烟花，在没有人发明下一种火药。**

带走的话

AI 越来越会解题，中国也最会解题；往后要紧的，是下一道题由谁来出，出了又由谁来认。

小　账

看见了什么，漏看了什么？大家看见了中国拿 AI 做烟花；漏看了用量不可比，更漏看了下一种火药不是从这片土里长出来的。

是在改进，还是在出题？宋岚的那部剧，从一锅粥长到定型，是在平台那把尺子里做得更好，是改进；那把尺子本身，没有人去出新题。

想一想

你这一年做得最好的一件事，是在别人出的题里做得更好，还是你自己出了一道题？要是后者，它被谁认了？



上　篇

烟花：一张图引发的争论

导　语

这一篇八章，从一张在手机上传来传去的图讲起。图上说：中国的 AI 算力，大半烧在短剧和广告上；美国反过来，大半用在写代码和企业自动化上。很多人看完叹一口气，说中国又一次拿火药做了烟花。

第 1 章到第 5 章，把这张图上的数字一个一个拆开：55% 是谁算的、怎么算的；视频和文字的 Token 为什么不能放在一杆秤上称；美国那个 58% 为什么找不到出处；八亿看短剧的人，为什么一个 Token 也没烧。第 6 章到第 8 章，说剩下的那部分真东西：资本按回款的快慢排队，把 AI 先押在了注意力商品上；这笔钱和这身本事，会不会回流到生产力。

这一篇的反差在这里：图想说的是中国人在拿 AI 做什么，它真正说出的，是资本把 AI 押在了哪里。

01

第 1 章

一张流传的图

◆ 一、一个周二的夜里

林舟三十八岁，在上海一家研究院做 AI 制药，带一个十二人的组。

那是一个周二，夜里十一点多。基金本子改到第七稿，他把电脑合上，靠在椅子上刷手机。刷到第三屏，一张图停住了他的拇指。

图很简单，两根柱子。左边一根标着「中国」，大半截涂成红色，写着「短剧、视频、广告、带货」；顶上一小截灰的，写着「编程」。右边一根标着「美国」，颜色正好倒过来：大半截是蓝的「编程与企业自动化」，只有一小截是「娱乐」。

图下面一行字：中国发明了火药，拿去做了烟花。

林舟盯着看了好一会儿。他想起自己组里的服务器，排队跑一个分子对接要等两天；想起上个月为了多申请一点算力，跟所里写了三页说明。

他把图转进了组里的群。没加一个字。

◆ 二、群里的十几分钟

群里本来已经安静了。图一发出去，消息一条接一条地冒出来。

博士后小顾先回：「太真实了。」后面跟了一个苦笑的表情。

一个博三的学生说：「难怪我们排不上卡。卡都去拍霸道总裁了。」

另一个刚进组的硕士生发了一句：「那我毕业去短剧公司是不是工资高一点？」有人回了一串笑。

过了一会儿，组里做数据的那位工程师冒出来，只问了一句：「这图哪来的？有原始出处吗？」

没人回答他。话题很快转到了谁家孩子报了什么暑期班。

第二天早上，组会开始前，几个学生还在聊这张图。有人说要发到朋友圈，有人说已经发了。那位工程师坐在角落里，低头看着自己的电脑，没有加入。

林舟把那段聊天又从头翻了一遍。他注意到，十几分钟里，大家说的都是「对」或者「太对了」，只有一个人问「从哪来」。而这一个问题，沉下去了。

这一章，讲的就是这张图，和它一路被传开的样子。

◆ 三、图上的数字从哪里来

先把能查到的源头摆在桌上。

2026 年 6 月初，一家叫中文在线的公司发布了一份投资者关系活动记录。这种记录，是上市公司接待投资者调研之后，按规定公开的问答纪要。中文在线做网络文学起家，这几年做短剧，也做 AI 短剧，旗下有一个面向海外的短剧平台叫 FlareFlow。

在那份记录里，公司援引了一组「行业 Token 调用数据」：国内 Token 消耗中，AI 短剧与视频生成约占 55%，电商与营销（直播带货、广告）约占 24%，软件开发约占 15%，其他约占 6%。

四个数加起来，正好 100%。

请注意几件事。第一，这组数是一家短剧公司援引的，不是统计部门发布的。第二，「行业数据」从哪个机构来、覆盖哪些平台、按什么时段统计，记录里没有写，原始口径至今没有公开。第三，这组数只说了中国，没有说美国。

图右边那根蓝柱子，是后来别人加上去的。第 4 章专门去找它的出处。

◆ 四、「火药做烟花」是怎么传开的

图能传得这么快，一半靠数字，一半靠那句话。

拿火药做烟花，是一个很老的说法。大意是：火药是中国人发明的，可中国人拿它放爆竹、敬鬼神，别人拿它造枪炮、开矿山。鲁迅在《电的利弊》里说过意思相近的话：外国用火药造子弹御敌，中国却用它做爆竹敬神〔核：大意，《伪自由书》，1933〕。这个说法在中文世界里流传了将近一百年，不少人在中学课堂上就听过。

它当史实看并不全对，宋代的军队已经在用火器，第 8 章还会说到。可它当一句感叹，一直很好用。

它好记，因为它有一个现成的画面：同一包黑色的粉末，一边炸出满天的花，一边炸出一条路。

它扎心，因为它碰到了一种老疼：东西是我们先有的，用处却是别人先想到的。

所以当那张图配上这句话，读的人几乎不用想。画面和疼，都是现成的，数字只是来作证。一个人看到图，觉得「说出了我早就觉得的事」，顺手一转；下一个人也是。

传得越快，越少有人停下来，问那位工程师问的那句话。

◆ 五、79% 是怎样变成 85% 的

传到后来，图上的数字变了。

很多版本写的是：中国 85% 的 Token 烧在短剧和广告上。

照中文在线的原始口径，短剧与视频生成 55%，电商与营销 24%，两项相加是 79%。85% 是从哪里多出来的，没有人说得清。

一种可能是有人把「其他」那 6% 也算进了娱乐，79 加 6，正好 85。另一种可能是转述的时候，图的作者觉得「八成」不够响，「八成五」更像一句话。还有一种可能更平常：某一个人记错了，后面的人照着他的抄。

不管是哪一种，样子都一样：一个数在传的路上，只会往更扎心的方向走，不会往更准确的方向走。

这一点，不怪哪个人。转发的人想说的是一种感受，数字是给感受配的音量。音量调大一点，感受就更清楚一点。至于原来的刻度在哪里，没有人在意。

◆ 六、先把公道还给这张图

说到这里，好像要开始拆这张图了。先停一下。

这张图有很多毛病，后面四章会一样一样摆出来。可是，能让十几个做科研的人在深夜里一齐说「太真实了」，它背后一定碰到了一个真问题。

这个真问题，大致是这样的：中国这几年在 AI 上花了极大的力气，模型做出来了，算力建起来了，用量也冲上去了。国家数据局局长刘烈宏 2026 年 3 月在中国发展高层论坛上说，中国日均 Token 消耗已经超过 140 万亿，比 2024 年初增长了一千多倍。

一千多倍，这是一个实打实的数。

可是这一千多倍里，有多少用在了实验室？有多少用在了新药、新材料、天气、蛋白质上？林舟组里排两天的队，是不是因为别处排着更长的队？

这些问题，没有一张图能回答。可是人们之所以愿意相信那张图，是因为他们心里早就有这个疑问，只是一直没有一个数字来给它作证。

图的数字可能是错的。它碰到的那块疼，是真的。

◆ 七、只看一面会怎样

这张图传开以后，网上很快分成了两拨人。

一拨人骂短剧。他们说，好端端的 AI，被拿去拍「战神归来」「闪婚总裁」，这是糟蹋；中国的科研就是被这种东西挤没的。

另一拨人骂图。他们说，数字是一家短剧公司说的，口径不明，美国那边根本没出处，拿两个不同的东西比，是外行；这种图就是制造焦虑。

两拨人都有对的地方。可是他们都只看了一面。

骂短剧的人，没有问这些数字是怎么算出来的。他们把一个计量上的问题，当成了一个道德问题。

骂图的人，拆完了数字，就以为问题没了。他们没有接着问：数字就算不准，资本为什么先往短剧那边跑？那些钱和那些本事，后来去了哪里？

本书想走第三条路：把图上的数字拆干净，再看拆完以后剩下来的那一块。剩下来的那一块，比图本身更值得想。

◆ 八、一张讲注意力的图，也在抢注意力

还有一件事，值得停下来想。

这张图本身，就是一件注意力商品。

它有一个醒目的颜色对比，一句扎心的话，一个现成的画面。它不需要你点开原文，不需要你算一道加法，三秒钟就能看懂，五秒钟就能转发。推送它的算法，也不在乎它对不对，只在乎你停了多久、转了没有。

一张批评中国把 AI 用来生产注意力的图，自己就是被注意力的那套机器推着跑的。

说不定，这张图的配色和排版，也是哪个工具几秒钟生成的。

这不是要讽刺谁。它只是提醒一件事：我们今天看见什么、相信什么，很大一部分，已经是被同一套机器分发过来的。

在这样的地方，停下来问一句出处，要逆着整套机器的力气。点开原文要多花一分钟，算法不会为这一分钟奖励你。

◆ 九、为什么做科研的人最先点头

那天夜里在林舟群里说「太真实了」的，全是做科研的人。这不是巧合。

做 AI 制药的人，天天和算力打交道。一个蛋白结构预测，一轮分子筛选，一次模型微调，每一步都要卡、要时间、要钱。组里的卡是有数的，所里的卡是排队的，买云上的算力要走报销，报销要对上基金本子里的预算科目。

这些麻烦，每一件都很小，合在一起就是一种日常的憋屈：你知道自己在做一件有用的事，可做这件事的每一步，都要先求人。

于是，当一张图告诉他们「算力都去了别处」，它说中的是这份憋屈，统计上对不对倒在其次。憋屈找到了一个出口，就不会再去查出口的门牌号。

公道地说，这种憋屈也不全是错觉。科研用的算力和商业用的算力，确实不在一个市场里。商业那边，谁出价高，卡就跟谁走；科研这边，钱是按年度批下

来的，批多少、怎么花，早在一年前就写定了。两边抢的是同一种芯片，规矩却完全不同。

这件事，第 6 章会接着说。

◆ 十、火药没有错，烟花也没有错

那句「火药做烟花」，本书还要再借用很多次。所以这里先把它磨一磨。

这个比喻常被用来骂人：你看，好东西被糟蹋了。可是仔细想，烟花本身没有错。过年放一挂鞭，元宵看一场焰火，是实实在在的快乐，也养活了一个行当的人。浏阳的花炮做到今天，工艺一点不比别处差。

火药也没有错。它就是一包粉末，拿去做什么，看人。

那么问题在哪里？本书的看法是：**问题不在拿火药做烟花，在没人发明下一种火药。**

做烟花，是拿一样现成的东西，做出更多、更便宜、更好看的花样。发明火药，是从没有到有。前一件事做到极致，后一件事未必会跟着来。烟花的花样可以年年翻新，配方再精，也还是同一包粉末。

拿这把尺子回看那张图，问题就换了。要问的不再是「为什么 AI 被拿去拍短剧」。要问的是：把 AI 用得最热闹的地方，会不会长出下一代 AI 和下一代科学？这个问题，要到下篇才能答。

◆ 十一、今天与机器：一张图的一辈子

这张图会活多久？

从前，一张传单贴在墙上，雨一淋就没了；一篇报纸文章，过几天就压在箱底。今天，一张图一旦被转开，就会一直在。它被存进一个个聊天记录，被截进

一篇篇文章，被做成短视频的封面。半年以后，有人再刷到它，又转一遍，它又活过来一次。

更要紧的是，它会被机器读进去。大模型从网上学东西，学到的不只是那份投资者记录，也学到了这张图，和图下面成千上万条「太真实了」。

一份记录只有一份，一张图却有一万份。机器读多了，就会觉得那张图说的才是常识。

所以，那位工程师的一句「这图哪来的」，今天比任何时候都更值钱。他查到的东西要是写下来、挂出去，后来的人和后来的机器，就多了一条回到出处的路。第 4 章会看到，他真的去查了。

◆ 十二、放回日子里

回到林舟。

那天快十二点的时候，宋岚从书房出来倒水，看见他还在刷手机，凑过来看了一眼。

宋岚三十五岁，在一家短剧公司做 AI 制片。她看了那张图，笑了一下：「这张图我们公司群里下午就在传了。老板还挺高兴，说我们这一行终于被看见了。」

林舟愣了一下。同一张图，在他的群里是一声叹气，在她的群里是一句夸奖。

「你们一个月烧多少？」他问。

宋岚想了想，报了一个数。林舟没说话。那个数，比他整个组一年的用量还多。

第 3 章会回来算这笔账。这里先给一个小办法，明天就能试：下次你想转一张带数字的图，先替那位工程师问三句。这个数是谁发的，他在这件事里有没有自己的利益？这个百分比的分母是什么，是全国、是一个平台，还是一家公司？拿来对照的另一边，和这一边是不是同一种东西？

三句问完，再决定转不转。多数时候你还是会转，只是你会在转的时候，加上一句「出处待查」。

> **一个数字传得越远，越像一句话，越不像一个数。**

带 走 的 话

一张图让人一齐说「太真实了」，说明它碰到了真疼；数字对不对，要另外查。

小　账

看见了什么，漏看了什么？看见了中国 AI 用量里科研的份额可能很薄；漏看了这组数是谁算的、怎么算的，以及拿来对照的是谁。

是在改进，还是在出题？转发这张图，是在一道早就有答案的题里再喊一遍；那位工程师问「这图哪来的」，才是在出一道新题。

想 一 想

你最近一次转发一个带百分比的说法，是什么时候？你知道那个百分比的分母是什么吗？在你的群里，最后一个问「出处在哪」的人，得到回答了吗？

◆

02

第 2 章

55% 是怎么算出来的

◆ 一、你一定见过这样的数字

报纸上、公众号里，常有这样的句子：「据行业数据显示，超过六成的年轻人……」「业内人士透露，今年市场规模将突破……」

你读的时候点了点头。读完以后，你记住的是那个数，忘掉的是「据」字后面那几个字。

下次再有人说起这件事，你会说「我看到过，六成多」。那个数从你嘴里出去，就不再带着「据行业数据」了。它变成了一个事实。

上一章那张图上的 55%，走的就是这条路。这一章，回到它出生的那份记录，看看它是怎么算出来的，又是怎么一路变了样子。

◆ 二、一份投资者交流记录

先说清楚这份记录是什么。

上市公司接待机构投资者、分析师来调研，事后要按规定公开一份「投资者关系活动记录表」。它像一份会议纪要：投资者问了什么，公司怎么答。记录表要走程序，不过公司在答问里援引的外部数据，交易所不会替它核实。

中文在线 2026 年 6 月初的那份记录里，投资者大意是问：AI 短剧对算力的消耗到底有多大？公司答的时候援引了一组「行业 Token 调用数据」，说国内 Token 消耗中，AI 短剧与视频生成约占 55%，电商与营销约占 24%，软件开发约占 15%，其他约占 6%。

这组数，回答的是「我们这一行用了多少算力」。它被放在一份面向投资者的文件里，用来说明一个行业有多大、多热、多值得看好。

这就是 55% 出生时的样子。

◆ 三、口径未公开，是什么意思

一个百分比要成立，至少要说清三件事：分子是什么，分母是什么，在什么时段里数。

分子好像清楚：短剧与视频生成用掉的 Token。问题是，「视频生成」包不包括电商的商品视频？包不包括广告片？如果包括，它和后面那个「电商与营销 24%」就有重叠；如果不包括，两者的边界划在哪里？

分母更难。「国内 Token 消耗」，是全国所有模型、所有平台的总和吗？谁有这个总账？国家数据局说的日均 140 万亿，是一个总量，可它没有按用途拆开公布。一家公司援引的「行业数据」，分母多半是某几家云平台、某几个模型服务商的调用量。是哪几家，没写。

时段也没写。是一周、一个月，还是一个季度？短剧行业有明显的旺季，春节档和暑期档前后，用量能差出一大截。

三件事都没说清，这个 55% 就只能读成：在某个我们不知道的范围里，某段我们不知道的时间里，有人数出了这么一个比例。

用大白话说：它可能是对的，只是没有办法查它对不对。

◆ 四、发布者站在哪里

还有一件事，得放在桌上说。

援引这组数的，是一家做 AI 短剧的公司。它在一份给投资者看的文件里说：我们这一行，占了全国 Token 消耗的一半以上。

这不是说它在撒谎。公司援引对自己有利的数据，是商业上最平常的事，就像卖房的人会告诉你「这一片学区房三年涨了四成」。

只是读的人要知道，说话的人站在哪里。一家短剧公司援引「短剧占一半以上」，和一家统计机构发布「短剧占一半以上」，分量不一样。前者在说明自己的行业有前途，后者在描述全国的样子。

那张图把前者当成了后者。

说他错之前，先把公道还给他：中文在线在记录里写的是「援引行业数据」，并没有说这是官方统计，也没有拿它和美国比。把这组数从一份投资者文件里拿出来，配上一根美国的柱子，再配上一句「火药做烟花」，是后来的人做的事。

◆ 五、仿真人剧和漫剧，差两三倍

那份记录里，还有一段话很有意思，可惜传开的图上没有。

中文在线解释了为什么短剧这么烧算力。大意是：AI 仿真人剧对帧数和 1080p 清晰度要求高，Token 消耗通常是 720p、偏漫画风格的 AI 漫剧的两到三倍。

这句话说明一件事：在同一个行当里，选什么画风、要多高的清晰度，就能让用量差出两三倍。

那么，如果明年行业里仿真人剧的比例从三成涨到六成，短剧的 Token 占比会一下子跳上去，哪怕拍出来的剧总数一部没变，看的人一个没多。

反过来也一样。如果有一天大家觉得漫剧也挺好看，用量又会掉下来。

所以短剧在 Token 里占多少，有一大块取决于画面有多精细，不取决于这一行有多少人在干、干了多少事。这个数，量的更像是画面的分辨率，不像是一个行业的分量。

◆ 六、宋岚的一天：抽卡

这个「抽卡」，宋岚天天在做。

周四下午，她在公司的剪辑室里，盯着屏幕上的一个镜头。剧本写的是：女主角在雨夜的天台上回头，眼里有泪，嘴角却在笑。

导演要的就是这个「又哭又笑」。

宋岚把提示词写好，点了生成。十秒的镜头出来了。女主角回头了，眼里也有泪，嘴角却是往下撇的，像是要哭出来。

不行，再来。

第二条，嘴角笑了，眼泪没了。第三条，表情都对，背景的霓虹灯在闪，闪得不对劲。第四条，雨的方向和上一个镜头反了。

到第十一条，导演说：「这条可以，但光再暖一点。」

宋岚又生成了六条，挑出一条。这个十秒的镜头，前前后后生成了十七次，最后用掉的只有一次。

她把这叫「抽卡」，行里人都这么叫。像玩手游抽卡一样，你不知道下一张出来的是什么，只能一张一张抽，直到抽到你要的那一张。

一部八十集的短剧，有上千个镜头。每个镜头抽几次到几十次不等。宋岚说，碰上挑剔的导演，一个镜头抽上五十次也是常事。

◆ 七、抽卡烧掉的，是选择的代价

把宋岚的这个下午放回那个 55% 里，能看出一件事。

那些 Token，多数没有变成观众看到的画面。它们烧在被扔掉的那十六条里。

这不是浪费。导演要的那个「又哭又笑」，本来就很难说清楚。真人拍戏，演员也要演好多条；只是从前烧的是胶片、灯光、演员的时间，现在烧的是 Token。

这说明，Token 的数量量出的，很大一部分是「挑」的成本：一个画面要被挑多少次，才能符合一个人心里那个说不清的样子。

一段代码也会被改，不过写代码的人改一次，通常只动几行，费不了多少 Token。一个镜头不满意，是整条十秒重新生成。

同样是「再来一次」，两边的价钱差得很远。下一章，要把这件事算到数量级上。

◆ 八、海外那一边的样子

同一份记录里，还有几个数，能帮着看清这一行的体量。

中文在线旗下的海外短剧平台 FlareFlow，截至 2026 年 5 月底，注册用户超过 3700 万，上线了 5000 多部剧。其中大约 10% 是 AI 短剧。

也就是说，在这家公司自己的海外平台上，九成的剧还是真人拍的。AI 短剧在用户眼里只是一小块；可是在 Token 的账上，它占了大头。

这两个数放在一起，就很能说明问题了。一部真人拍的剧，几乎不烧 Token；一部 AI 生成的剧，从头到尾都在烧。所以在算力账上，AI 短剧的分量被放大了很多倍；在观众的屏幕上，它还是少数。

如果有人只看算力账，会以为短剧行业已经被 AI 接管了。如果只看片单，又会以为 AI 在这一行还是个点缀。两边都是真的，只是量的东西不一样。

◆ 九、只看一面会怎样

拆到这里，很容易滑到另一头去：既然 55% 口径不明、发布者有利益、画风能让用量差几倍，那这个数就是假的，可以扔掉了。

这也是只看一面。

口径不明，不等于数是编的。一家上市公司在正式文件里援引的数，多半有一个真实的来源，只是这个来源没有公开。发布者有利益，也不等于数是歪的。画风能让用量变，恰恰说明短剧确实在大量烧算力。

公道的读法是：短剧和视频生成，很可能是国内 Token 消耗里最大的一块；到底占多大，55% 只是一个没法查证的估计；这一块之所以大，有相当一部分原因在视频本身的计量上，而不全在「中国人把 AI 拿去娱乐」上。

这样读，数没有扔，只是放回了它该在的位子上。

◆ 十、三道转手

回到那个 55%，看它是怎么一路走到林舟手机上的。

第一道，投资者交流。中文在线的答问被写进记录表，挂在交易所的网站上。这时候它还带着「援引行业数据」几个字，旁边还有仿真人剧和漫剧的那段解释。

第二道，媒体。财经媒体和行业媒体摘录记录表，写成报道。标题里留下了数字，「援引」两个字挪到了正文第三段。仿真人剧两三倍那段，篇幅不够，删了。

第三道，自媒体。有人把数字从报道里拿出来，加上一根美国的柱子，画成对比图，配上一句话。到这一步，「援引」没了，「行业数据」没了，「一家短剧公司」也没了。55% 和 24% 相加本是 79%，图上写的却是 85%。

每一道转手的人，都没有故意歪曲。每个人都只是把自己觉得不要紧的部分删掉，把自己觉得要紧的部分放大。

可是三道走完，一个「某家公司在投资者面前援引的、口径不明的行业估计」，就变成了「中国 85% 的 AI 用在了娱乐上」。

一个数每被转一次手，就丢掉一层它从哪里来，多出一层它想说什么。

这三道转手，今天比从前快得多。

从前，一份记录表变成一篇报道，要一个记者去读、去写、去等编辑审。从报道变成一张图，要一个会排版的人花上半天。

今天，三步都能交给机器。一个模型几秒钟就能从一份几十页的记录表里抓出「最有传播力的数字」；另一个模型几秒钟就能画出一张对比图，配好色，加好标题。

机器做摘要的时候，最先删掉的，往往就是那些限定词：「援引」「约」「口径未公开」。因为这些词不影响「意思」，只影响「准不准」。

所以，在一个转述越来越便宜的时代，一个数最要紧的那几个字，反而最容易丢。

反过来，机器也能帮着把丢掉的字找回来。你可以把那张图发给一个模型，问它：这组数最早出自哪份文件？原文怎么说？它未必每次都答得对，可它至少会把你带回第一道转手的那个地方。查一次出处，从前要半天，今天只要一分钟。

机器让转述变便宜了，也让查证变便宜了。用哪一样，在人。

◆ 十一、一个完整的例子：同一句话的三张脸

把三道转手各截一句下来，并排放着看，会更清楚。下面三句，是本书照着这类转述最常见的样子拟的，不是哪一家的原文。

投资者记录里的那一句，大意是：公司援引行业 Token 调用数据，国内 Token 消耗中，AI 短剧与视频生成约占 55%。

财经报道里的那一句，常见的样子是：AI 短剧与视频生成已占国内 Token 消耗的 55%。

自媒体图上的那一句，常见的样子是：中国 85% 的 AI 算力，烧在了短剧和广告上。

三句话，一句比一句短，一句比一句响。

第一句的主语是「公司援引」，第二句的主语是「AI 短剧」，第三句的主语成了「中国」。主语一换，说话的人就从一家公司，变成了一个行业，最后变成了一个国家。

第一句里有一个「约」字，第二句里有一个「已」字，第三句里什么也没有。「约」是在说这个数不太准，「已」是在说这件事已经定了。

这就是转手的样子。没有人改过一个事实，每个人只是换了一个主语、删了一个字。走完三道，意思已经不在原来的地方了。

◆ 十二、放回日子里

那个周四夜里，宋岚回到家，把抽卡的事讲给林舟听。

林舟听完，想了一会儿，说：「那你们的 Token，大半是烧在没用上的镜头里？」

「也不能这么说。」宋岚说，「没有那十六条，就挑不出那一条。」

林舟没再说什么。他想起自己组里筛分子的事：几万个候选分子跑一遍对接，留下来做实验的，不过几十个。他们烧的算力，也大半烧在没被选中的那些上面。

两个人干的是完全不同的活，烧算力的样子却很像：都是为了挑出少数几个，先生成一大堆。

一个小办法，明天就能试：下次读到一个「据行业数据」的百分比，顺手在心里补上三个空——谁统计的，统计了谁，什么时候。补不上的，就在记它的时候连「据」字一起记住。

> **一个数最要紧的，常常是它前面那几个最不起眼的字。**

带 走 的 话

55% 是一家短剧公司援引的估计，口径不明；它量出的，有一大块是画面的精细和挑选的次数。

小　账

看见了什么，漏看了什么？看见了短剧确实烧掉大量 Token；漏看了这个比例由谁统计、分母是什么，以及画风和抽卡能让用量差出几倍。

是在改进，还是在出题？抽卡是在一道导演已经出好的题里找最好的那一张；那个「又哭又笑」，才是导演出的题。

想 一 想

你的工作里，有没有一种「抽卡」——为了挑出一个，先做出一大堆？你心里记着的那些百分比，有几个你还记得它是谁说的？

◆

03

第 3 章

十秒视频和一千行代码

◆ 一、一张水费单

你家每个月都会收到一张水费单。

假设有一户人家，水费单上写着：这个月用水八十吨。你能从这个数看出这家人的日子过得怎样吗?

大概看不出。可是如果有人告诉你，这八十吨里，七十吨是灌后院游泳池的，十吨是做饭、洗澡、洗衣服的。你会不会说：这家人的生活重心是游泳?

多半不会。你会说：游泳池本来就费水。灌满一次，抵得上厨房用一年。

这一章要讲的，就是这件事。拿 Token 的用量去比用途，很像拿用水量去判断一家人的生活重心。游泳池一定压倒厨房，不管这家人一年游几次泳。

◆ 二、Token 是什么

先把 Token 讲清楚。

大模型读和写，都不是一个字一个字来的，是一小块一小块来的。这一小块，就叫 Token。一个英文单词，大约是一到两个 Token；一个汉字，大约是一到两个 Token，看模型怎么切。

文字好办。图像和视频，也要先切成 Token，模型才能处理。一张图片被切成很多个小方块，每一块变成若干个 Token。一段视频，是一帧一帧的图片排起来，再加上前后帧之间的关系，切下来的 Token 就成倍地涨。

所以，Token 是一把统一的尺子：不管是文字、图片还是视频，都换算成 Token，再计价、再统计。

这把尺子统一了单位，却没有统一分量。同样是一个 Token，在一段文字里和在一帧画面里，装的东西完全不一样。

◆ 三、量一量：十秒和一千行

用一组具体的数来比一比。

一篇技术博文做过测算：用字节跳动的 Seedance 2.0 生成一段 10 秒、1080p 的视频，大约要 35 万个 Token；1 分钟 720p 的视频，可以破百万；一个短剧项目，几亿 Token 起步〔待核：单一来源〕。

再看文字这一边。一千行代码，按每行十几个 Token 算，不过一两万 Token。就算让模型反复改、反复读上下文，一次编程会话用掉几万到几十万，已经算多的了。

普通的文字聊天，问一句答一段，一来一回几百到一两千 Token。照那篇博文的说法，文字聊天的用量，只有视频生成的千分之一到万分之一〔待核：单一来源〕。

把这几个数放在一起：十秒视频的 Token，够一个程序员写上万行代码，够一个普通人聊上几百次天〔待核：估算〕。

换成家里的东西想一想：一段十秒的视频，像一口灌满的游泳池；一千行代码，像厨房里一天用的那几盆水。两样都是水，都按吨算。可是没有人会拿吨数，去判断这家人更爱游泳还是更爱做饭。

这就是数量级的差。不是多一倍两倍，是多出三个、四个零。

◆ 四、把它放回那张图

现在回到那张图。

假设全国有一万个程序员，每人每天用 AI 写代码，每人每天烧二十万 Token。合起来是每天 20 亿。

再假设全国有一百个短剧剧组，每个剧组每天生成两百段十秒视频，按每段 35 万 Token 算，再加上抽卡的倍数。合起来，每天轻轻松松上百亿。

一万个人对一百个剧组。人数上，写代码的多一百倍；Token 上，拍短剧的多出好几倍。

这些数是本书为了说明道理随手假设的，不代表真实的行业规模。它说明的那个道理不依赖具体数字：只要一边在生成视频，一边在处理文字，哪怕做视频的人少得多，Token 的账上也一定是视频压倒文字。

游泳池一定压倒厨房。

所以那张图上 55% 对 15% 的差距，有很大一部分，是视频和文字在计量上天然的差距，不是中国人「重娱乐、轻科研」的差距。

◆ 五、这不是中国独有的

还要补一句：这件事在哪里都一样。

美国也有视频生成，也有广告片，也有短视频平台。只要哪里的视频生成用量起来了，Token 的大头就会往那边倒。

一个国家的 Token 结构里视频占得少，可能是因为它的人不爱看视频，也可能只是因为它的视频还没有大规模交给 AI 生成。后者更常见。

中国的短剧行业，是全世界最早大规模把 AI 用进视频生产的行当之一。2026 年一季度，国内每天新增的 AI 短剧大约 1500 部（第 5 章会细说）。一个行当这样大规模地生成视频，它在 Token 账上占大头，几乎是算术上注定的。

这件事说明的，首先是中国的短剧行业动作快、规模大，然后才是别的。

◆ 六、本书的说法

说到这里，可以把本书的看法摆在桌上了：

按用量比用途，就像按用水量判断一家人的生活重心——游泳池一定压倒厨房。

Token 的数量，量的是计算的多少，不是用途的轻重。把不同用途放在同一把用量的尺子上比，量出来的主要是「哪种用途天生费 Token」，而不是「哪种用途更受重视」。

那张图的第一个毛病，就在这里：它把一个计量上的差，读成了一个价值上的差。

这不等于说用量毫无意义。用量能告诉你算力的钱花到了哪里，电烧在了哪里，芯片忙在了哪里。这些都是真的，也都要紧。可是它不能告诉你，一个国家的人拿 AI 在做什么、看重什么。

◆ 七、换几把尺子

那么，有没有更合理的尺子？至少有三把，各自说明不同的事。

第一把，按花费。不数 Token，数钱：每一类用途，一年在 AI 上花了多少钱。视频的 Token 单价低，文字的 Token 单价高，尤其是最强的推理模型，单价能高出很多。按钱来算，视频和编程的差距会缩小很多。这把尺子说明的是：资本把钱押在了哪里。

第二把，按产生的价值。一段十秒的短剧镜头，卖出去能换来多少收入？一千行代码，让一家公司省下多少人工、多卖出多少产品？一次分子筛选，让一款新药提前了多少个月？这把尺子最难量，却最接近「AI 在做有用的事吗」这个问题。

第三把，按使用人数。有多少人在用 AI 写代码，有多少人在用 AI 做视频，有多少人在用 AI 查文献、做实验？这把尺子说明的是：AI 进了多少人的日子，进了哪些人的日子。

三把尺子量出来的样子，很可能和 Token 那把完全不同。按人数，写代码和查资料的人，大概远比做视频的人多。按价值，哪一边更大，眼下谁也说不准。

◆ 八、只看一面会怎样

拆到这里，又要防另一种偏。

既然按 Token 比用途不靠谱，是不是就可以说：中国的 AI 用在科研上的份额其实不低，那张图全是误会？

也不能。

本书只是说，Token 那把尺子量不出这件事。它没有说换一把尺子量出来，结果一定好看。按花费量，短剧和营销照样可能是大头；按价值量，科研那一块可能依然很薄。这些都要另外去查，不能因为拆掉了一个错的证据，就认定反面是真的。

拆掉一个错的数，留下来的是一个空，不是一个反面的答案。

这一点很要紧。第 6 章会接着这个空往下走：就算换成按钱来算，资本的钱也确实先往回款快的地方跑。这才是图背后那件真事。

◆ 九、尺子会回过头改人

还有一件事，比量错更要紧：一把尺子用久了，会回过头来改被量的人。

林舟的所里分算力，看的是去年的用量。于是每到年底，有的组会把一些不急的任务也拿出来跑一跑，把用量冲上去，好让明年多分一点。这不是谁不老实，是尺子在教人怎么做。

一个地方要是拿 Token 消耗当政绩，当「AI 普及率」的证据，最省事的办法，就是多做视频。一个平台要是按 Token 收入给团队记功，最划算的客户，也是视频。

到那一步，Token 的结构就不只是在量用途了，它在推着用途往费 Token 的方向走。第 29 章讲论文计分的时候，会碰到同一个样子：一把尺子被当成目标，它量出来的东西就不再可信。

所以，不用 Token 比用途，不只是为了量得准，也是为了别让一把错的尺子去带路。

◆ 十、为什么偏偏是 Token

既然 Token 这把尺子这么不合用，为什么大家都拿它说事?

一个原因是，它最好数。模型每处理一个 Token，平台的账上就记一笔。用量是现成的，按天、按周、按月，都能拉出来。钱就难一些，要对账，要算折扣，要分清谁付的。价值最难，要等很多年，要有人去追。

一个原因是，它是新东西。从前没有这把尺子，大家对它没有感觉，不知道一个 Token 大概值多少。一个陌生的单位，加上一个很大的数，很容易让人相信它在说一件大事。

还有一个原因：它是卖算力的人最在意的数。云平台按 Token 收钱，芯片公司按 Token 算需求，投资人按 Token 估增长。所以发布 Token 数据最积极的，往往是这些人。他们发布的数，量的就是他们自己生意的大小。

一把尺子最先在谁手里，它量出来的样子，就最像谁的生意。

◆ 十一、今天与机器

还有一个变化，正在发生。

视频生成的 Token 用量，会越来越大。清晰度从 720p 往 1080p、4K 走，时长从几秒往几分钟走，帧率往上走。每一步，Token 都成倍地涨。

文字这一边，也在涨。今天的推理模型，答一个问题之前会先「想」很久，想的过程也要烧 Token。一个复杂的编程任务，交给 AI 代理去跑，它会自己读代码、自己试、自己改，一跑就是几个小时，烧掉的 Token 也不再是小数目。

所以两边的比例，今后会怎样变，谁也说不准。可以确定的只有一件事：只要还拿 Token 当尺子，这个比例就会随着技术的变化上下乱跳，和用途的轻重关系不大。

一把会被技术带着跑的尺子，不适合用来量一个国家的心思。

◆ 十二、一个完整的例子：林舟的组

用林舟的组，把这几把尺子都试一遍。

林舟的组一年用 AI 做的事，大致有三样：用 AlphaFold 一类的工具预测蛋白结构；用大模型读文献、写综述、改基金本子；用生成模型设计候选分子，再拿去做对接筛选。

按 Token 算，这三样加起来，比不上宋岚公司一个月的用量。如果把林舟的组放进那张图，它是那根灰色小柱子里一个看不见的点。

按花费算，组里一年的算力开销，加上云服务和软件许可，是一笔不小的钱，在基金预算里排得上前几位。

按价值算，组里去年有一个候选分子进了临床前研究。如果它最后能变成一种药，它的价值，没有任何一部短剧能比。可是这件事要过很多年才知道。

按人数算，组里十二个人，人人每天都在用 AI。

四把尺子，四个样子。只有 Token 那一把说它「几乎不存在」。

◆ 十三、放回日子里

那个周末，林舟把这四把尺子画在一张纸上，给宋岚看。

宋岚看完，说：「那按你这个算法，我们烧得最多，可能产出的价值最少？」

「不一定。」林舟说，「你们那部剧要是火了，挣的钱也不少。我是说，光看 Token，什么也说明不了。」

宋岚想了想：「那你们所里分算力，按什么分？」

林舟愣住了。他发现，所里分算力，看的恰恰是用量：谁去年用得多，今年就多给一点。

一个小办法，明天就能试：下次你要判断一件事「重不重要」，先问自己，你手里那把尺子量的是什么。量的是数量，还是钱，还是价值，还是人。然后至少再换一把，看看样子变不变。

> **游泳池一定压倒厨房，可是一家人每天都要进厨房。**

带 走 的 话

Token 量的是计算的多少，不是用途的轻重；视频天生费 Token，所以按用量比用途，视频一定赢。

小　账

看见了什么，漏看了什么？看见了视频生成烧掉的 Token 远多于文字；漏看了这是计量单位的差，换成花费、价值、人数，样子会完全不同。

是在改进，还是在出题？换一把尺子，是在旧题里量得更准；问「AI 进了哪些人的日子、长出了什么新东西」，才碰到了新题。

想 一 想

你在工作里被考核的那个数，量的是你做事的多少，还是做成了什么？如果换一把尺子来量你，你在团队里的位置会变吗？

◆

04

第 4 章

找不到出处的 58%

◆ 一、对面那根蓝柱子

那张图，有两根柱子。前两章拆的，都是左边那一根。

这一章，看右边那根蓝色的：美国 58% 的 Token 用于编程开发和企业自动化。

很多人看图的时候，只盯着中国那一根。美国那一根，好像只是一个背景，一个用来衬托的「正常样子」。一张对比图，对比的另一边要是站不住，这张图就只剩下一根柱子，什么也比不出来。

这件事，像极了家长群里常见的一句话：「人家孩子都在学编程了。」说的人很少说清是谁家的孩子，听的人也很少去问。那个「人家」，就站在那里，作为一个让你着急的标准。

美国那根柱子，就是那张图里的「人家孩子」。

◆ 二、去找出处

本书去找过这个 58% 的出处。

它在中文网络上流传很广，很多文章、视频、帖子都引用过。顺着引用往回追，追到最后，大多是「据统计」「有数据显示」「美国的情况则相反」，没有一个指向一份具体的报告、一个具体的机构、一个具体的时段。

虎嗅上有一位作者，专门核查过这组对比。据他的核查，没有找到任何可靠的信源支持「美国 58%」这个数。

美国没有一个统一的部门，按用途统计全国的 Token 消耗。几家大的模型公司，各自知道自己平台上的用量，只是它们不按「编程」「娱乐」这样的分类公开全国口径的比例。

所以，这个数在哪里出生、谁第一个说出来的，至今说不清。本书照实写：找不到出处。

一张对比图，一边是口径不明的估计，一边是找不到出处的数。

◆ 三、常被拿来对照的 OpenRouter

找不到出处，有人会说：总有一些数据能说明美国人在用 AI 写代码吧?

有。最常被拿来对照的，是 OpenRouter。

OpenRouter 是全球最大的模型 API 聚合平台。打个比方，它像一个模型超市：开发者在这里注册一个账号，就能调用几百个不同公司的模型，不用一家一家去签合同。它会定期公布平台上各个模型、各类用途的调用量，所以常被媒体引用。

OpenRouter 和风险投资机构 a16z 合作做过一份研究，结论是：平台上编程类用途的占比，从 2025 年初的约 11%，升到了 50% 以上。

这个数是真的，也有出处。很多人看到它，就觉得「美国一半以上的 AI 用在编程上」得到了印证。

这里有一个大问题。

◆ 四、开发者平台上的用途

OpenRouter 是一个开发者平台。

来这里的，是要调用模型 API 的人：写程序的人，做应用的人，搭工具的人。普通人不会来这里。一个用手机刷视频的人、一个用 AI 写周报的上班族，都不会去 OpenRouter 注册账号。

所以，OpenRouter 上编程占了一半以上，说明的是：在开发者用的平台上，开发者最常拿 AI 写代码。

这句话近乎同义反复。就像说：在五金店里，买锤子和钉子的人最多。

它不能说明全美国的 AI 用途里，编程占一半。更不能拿来和「中国全国的 Token 结构」并排比。

再看用户的分布。OpenRouter 的用户里，美国约占 47%，中国约占 6%。也就是说，这个平台上的数据，大半是美国开发者的样子，中国的样本只是一小块。

拿一个美国开发者平台上的用途，去对一个中国全国口径的估计，比的根本不是同一种东西。

◆ 五、OpenRouter 上的另一件事

说到 OpenRouter，还有一件事，那张图完全没提。

2026 年 2 月 9 日到 15 日那一周，OpenRouter 上中国模型的调用量，第一次超过了美国模型。此后一路领先。到 2026 年 10 月初，据界面新闻的测算，中国模型的周调用量约 86.7 万亿 Token，美国模型约 19 万亿。

请注意，这是在一个以美国开发者为主的平台上。中国用户只占约 6%，可是中国模型的调用量是美国模型的四倍多。

这说明什么？说明大量的美国开发者，在用中国的模型。

他们为什么用？最常听到的理由有三条：便宜，权重开放，够用。便宜，是同样的活，花的钱少很多；权重开放，是可以把模型下载下来，放在自己的服务器上改、上面跑，不用把数据交给别人；够用，是对很多日常的活来说，它和最贵的那几家差得不多。

CNBC 的报道也说到这件事：2026 年 2 月以来，美国企业在 OpenRouter 上，每周都把超过 30% 的 Token 发给中国模型。

这件事，要先把公道还给中国。DeepSeek、通义千问、Kimi 这些开放权重的模型，做得便宜、做得好用，才能在别人的地盘上，被别人的开发者拿去写代码、搭工具。这是一个实打实的工程成就，也是中篇一要写的那种「应用层」的硬本事。

那张图说中国把 AI 用来做烟花。可是在全球开发者的那个超市里，最畅销的火药，有一大半是中国造的。

◆ 六、便宜的和贵的

不过，这件事也有它的另一面，同样要看。

CNBC 的同一篇报道还提到：编程助手、后台自动化这类出错代价高的工作，美国企业仍然愿意为准确付高价，不轻易换成便宜的模型。

也就是说，中国模型的调用量大，有很大一部分是因为便宜：大批量、对准确要求没那么高的活，交给便宜的模型去跑。最要紧、最贵、最不能出错的那部分活，还是留在美国最强的模型上。

按调用量看，中国模型领先；按花费看，可能是另一个样子；按最难的那部分任务看，又是另一个样子。

上一章说的几把尺子，在这里又用上了。同一个平台，换一把尺子，谁领先就换了一个。

这件事和本书的母题有关：做得更便宜、更快、量更大，是在一道已经出好的题里做到最好。第 15 章会回来说这件事。

◆ 七、两边比的是不同的样本

把这一章拆出来的东西排一排：

	中国那根柱子	美国那根柱子
数字	55%＋24%（被说成 85%）	58%
来源	一家短剧公司援引的行业数据	找不到出处
样本	据说是国内全口径，具体范围未公开	常被拿来印证的 OpenRouter 是开发者平台
计量	Token 用量，视频天生费 Token	Token 用量，开发者平台上文字为主

一边是一个国家里所有行业、包括大规模视频生成的估计；一边是一个开发者平台上、以文字为主的用途统计，或者干脆找不到出处。

拿这两样比，就像拿一座城市所有家庭的用水量，去比一个健身房会员的饮水量，然后说：这座城市的人不爱喝水，都拿去灌游泳池了。

◆ 八、一个完整的例子：一晚上的查证

那位工程师是怎么查的？他后来在组会上讲过一遍，大致是这样走的。

第一步，他把图上的两句话分开查。中国那一句，他搜「55%　Token 短剧」，很快找到了几篇财经报道，报道里写着「中文在线在投资者关系活动记录中援引」。他顺着去交易所的网站，翻到了那份记录。原文里有「援引行业数据」几个字，也有仿真人剧和漫剧那段话。这一句，用了二十分钟。

第二步，美国那一句。他搜「美国 58% 编程」，出来几十条，一条一条点开，没有一条给出报告名。他换成英文去搜，也没找到对应的说法。他又问了一个大模型，模型给了他一个看上去像样的研究机构名字，他去那个机构的网站上找，没有这份报告。这一句，用了两个小时，什么也没找到。

第三步，他想到 OpenRouter。他找到了 OpenRouter 和 a16z 的那份研究，看到了编程从约 11% 升到 50% 以上的那张曲线；又找到 OpenRouter 自己公布的用户分布，看到美国约 47%、中国约 6%。到这里他明白了，这是开发者平台的数。

第四步，他在 OpenRouter 的排行榜上，顺手看了一眼各家模型的调用量，发现排在前面的有好几个是中国模型。他又搜到了界面新闻和 CNBC 的报道。

前后三个多小时。他说，最费时间的，是去证明一件东西不存在。

这个例子要记住，后面还要用。找出处这件事，本身就是常规的功夫：有路可循，一步一步走，走得细就能走到。机器可以帮着走，走得比人快。可是「这张图到底问错了什么」，没有路可循，要人自己看出来。

◆ 九、本书的说法

所以，那张图的第二个毛病是：样本错位。

第一个毛病是计量错觉，上一章说过：视频的 Token 密度比文字高几个数量级。第二个毛病是这一章说的：两根柱子量的不是同一群人、同一种东西，其中一根还找不到出处。

两个毛病加在一起，那张图作为证据，几乎站不住。

不过本书在第 1 章说过，它碰到的那块疼是真的。拆掉了图，那块疼还在。下一章，要把图里最后一点真东西找出来。

◆ 十、今天与机器

为什么一个找不到出处的数，能传得这么广？

一个原因是，它听上去「应该是对的」。美国是硅谷的地方，程序员多，企业软件发达，说它一半以上的 AI 用在编程，符合大家心里的那幅画。

一个数，只要符合人心里的那幅画，就不太会有人去查它。

今天的机器，让这件事又多了一层。你问一个大模型「美国 AI 用途中编程占多少」，它可能会给你一个看上去很像样的数，甚至编一个像样的出处。它也是从网上学来的，网上那个 58% 传得多了，它学到的就是那个数。

一个找不到出处的数，被机器学进去，再从机器嘴里说出来，就多了一层「AI 说的」的分量。这样转一圈，它离出处就更远了。

◆ 十一、为什么需要一个「人家」

还要问一句：为什么一张批评中国的图，非要拉上美国？

因为一个数单独摆着，很难让人着急。说「中国 55% 的 Token 用于短剧」，听的人会问：那又怎样？可是旁边站一个「人家」，说「美国 58% 用于编程」，着急就有了方向：我们走反了。

这是一种很老的说话方式。家长说「人家孩子」，老板说「人家公司」，都是这样。它管用，因为它把一个说不清好坏的数，变成了一场比赛的比分。

可是一旦比的是比分，两边的规则就得一样。足球比分不能跟篮球比分比。那张图的美国柱子，连是哪一项运动都说不清。

公道地说，拿美国当参照，也不全是坏事。美国在 AI 的模型、芯片、科研上，确实有很多地方走在前面，这些地方值得认真比。只是要比，就要比同一样东西。比两国的顶尖实验室，各自拿 AI 做出了什么新科学；比两国的大学里，有多少人在用 AI 做研究；比两国的药厂、材料厂、气象部门，各自在 AI 上投了多少钱。这些比法都难，比出来的东西才有用。

中篇二会拿这样的尺子，去比一比真正要紧的那一层。

◆ 十二、一个找不到的数，也是一个信号

最后，把这件事反过来想一想。

一个数找不到出处，本身就是一个消息：在美国，也没有人知道全国的 AI 到底用在了哪里。

中国有国家数据局，会说全国日均 Token 消耗是多少。美国没有一个部门公布这样的总数。两边都不知道全国按用途拆开是什么样子，只是中国这边有一家公司援引了一组数，美国那边连这样的一组数也没有。

所以，那张图比出来的，不是两国 AI 用途的差，是两种「不知道」的差：一边知道一个总数，不知道总数里装着什么；一边连总数都不知道。

这件事说明，今天谁也没有一把好尺子，能量清楚一个国家拿 AI 在做什么。第 3 章说过的那几把尺子，按花费、按价值、按人数，还没有一把被谁认真地造出来。

谁先把这把尺子造出来，谁就先看清自己。这是一道还没人做的题。

◆ 十三、放回日子里

那张图转进林舟的群三天以后，组里那位问「这图哪来的」的工程师，在群里发了一段长消息。

他查了。他说，中国那组数出自一家短剧公司的投资者记录。美国那组数，他找了一晚上没找到出处；能找到的最接近的，是 OpenRouter 的数据，那是开发者平台。他最后一句说：「顺便说一下，OpenRouter 上中国模型调用量已经是美国的四倍了。」

群里安静了一会儿。博士后小顾回了一句：「学到了。」

林舟看着这段消息，想起三天前自己转那张图时，一个字也没加。

一个小办法，明天就能试：一张对比图，先别看两边谁高谁低，先看两边是不是同一种东西。如果一边是全国、一边是一个平台，一边是全行业、一边是开发者，这张图就先放下，不急着转。

> **一张对比图，最先要查的是对面那一边。**

带 走 的 话

美国 58% 找不到出处；常被拿来对照的 OpenRouter 是开发者平台，比的是不同的样本。

小　账

看见了什么，漏看了什么？看见了美国开发者大量用 AI 写代码；漏看了这是开发者平台上的样子，也漏看了中国模型在同一个平台上调用量反超的事。

是在改进，还是在出题？中国模型在 OpenRouter 上靠便宜和好用胜出，是在一道已经出好的题里做到最好；出那道题的，还不是我们。

想 一 想

你心里的「人家孩子」是谁？你查过他真实的样子吗？你最近一次被一个「对比」说服，那两边是同一种东西吗？

◆

05

第 5 章

八亿人和每天一千五百部

◆ 一、地铁上的那一排屏幕

你一定有过这样的时刻。

早上八点的地铁，车厢里挤满了人。你抬头看一眼，一排人低着头，每个人手里一块屏幕。你斜着瞄一眼旁边的那块：一个穿西装的男人正把一份合同摔在桌上，字幕打着「三年之期已到」。

再往旁边看，又是一部。一个女孩在雨里哭，屏幕底下滚着「下一集」。

这些剧，一集一两分钟，一部几十集到一百多集。地铁坐四站，能看完七八集。

这一章，从这一排屏幕讲起。前四章拆的是图上的数字，这一章要问：看这些剧的人，和烧 Token 的人，是不是同一群人？

◆ 二、宋岚那边的地铁

宋岚每天也坐这趟地铁。

她看这一排屏幕的眼光，和别人不一样。别人看的是剧情，她看的是镜头：这个西装男人的手，是不是又多长了一根指头；那个雨里的女孩，头发的走向前后两集对不对得上。有一回，她在车厢里认出了自己公司的剧，一个中年人看得入神，坐过了站。

她没说什么，只是在心里记了一笔：那一集的第三个镜头，光还是偏冷了。

她和那个坐过站的人，一个在生产，一个在看。两个人站在同一节车厢里，离得不到一米。可是在 Token 的账上，一个是全部，一个是零。

◆ 三、八亿人

先看几个数。

据野村证券援引 QuestMobile 的数据，截至 2026 年 7 月，中国微短剧的用户超过 8 亿，大约占网民的三分之二。

红果短剧是字节跳动旗下的免费短剧平台。它的日活跃用户是 1.68 亿，比上一年同期增长了 107%。用户平均每天在上面停留约 125 分钟。

125 分钟，是两个小时零五分。比很多人每天陪孩子说话的时间还长，比很多人每天读书的时间长得多。

国内短剧的市场规模，已经接近 500 亿元。

这些数说明，短剧不是一个小众的东西。它是今天中国人日子里，占时间最多的事情之一。

◆ 四、每天一千五百部

再看供给那一边。

据中文在线同一份投资者记录，2026 年一季度，国内每天新增的 AI 短剧大约 1500 部，每月上新超过 4 万部。一个月的 AI 短剧产量，约等于 2025 年全年的真人短剧产量。

一个月，抵得上从前一年。

这是一个什么样的速度？打个比方。一部真人短剧，从剧本到上线，要找演员、租场地、搭布景、拍、剪，快的也要一两周。一部 AI 短剧，前面说的那篇

技术博文测算，制作周期能从 15 天压到 3 天，5 秒视频的成本约 4.5 到 9 元〔待核〕。

一千五百部，每部几十集，每集一两分钟。每一分钟，都是 AI 一帧一帧生成出来的。

第 3 章说过，一分钟 720p 视频可以破百万 Token〔待核〕。前面那个 55%，大半就烧在这里。

◆ 五、观众不烧 Token

这里有一件很容易被忽略的事。

八亿看短剧的人，一个 Token 也没有烧。

地铁上那个看「三年之期已到」的人，他手机里放出来的，是一段已经生成好、压缩好、存在服务器上的视频。他看一遍，看十遍，都不需要模型再算一次。他花的是流量，是电，是时间，不是 Token。

烧 Token 的，是制作方：是宋岚这样的制片，是她公司里那几排盯着屏幕抽卡的人，是给剧组提供生成服务的那几家公司。

全国做 AI 短剧的从业者，有多少人？本书没有找到可靠的数。就算往多了估，几十万人，和八亿比，也只是一个零头〔待核：从业人数〕。

所以 Token 结构里那 55%，不是八亿人用 AI 用出来的，是几十万人为八亿人生产内容时烧出来的。

◆ 六、125 分钟从哪里来

再回到那 125 分钟。

一个人一天只有二十四个小时。睡掉七八个，上班、上学、通勤掉十来个，吃饭、洗漱又去掉一两个。剩下能自己支配的，不过三四个小时。

红果的人均 125 分钟，就是从这三四个小时里切出来的。它切走的那两个小时，从前可能是看电视、打牌、聊天、发呆，也可能是读一本书、学一样东西、陪孩子下一盘棋。

这就是「注意力商品」这个说法的意思。一部短剧卖的不是剧本身，是人的时间。人在屏幕前多停一分钟，平台就多一分钟可以插广告，多一次机会让人充值看下一集。剧拍得越抓人，切走的时间越多，挣的钱也越多。

注意力有一个和别的东西都不一样的地方：它的总量不会涨。粮食可以多种，房子可以多盖，芯片可以多造，可是全中国人一天的注意力，就是十几亿人乘以那三四个小时。

所以注意力的生意，是一场在一块固定大小的地里抢收成的生意。产量越大，抢得越凶。一天一千五百部，抢的是同一块地。

记住这一点，第 6 章要用它来解释：为什么资本偏爱这一块地，又为什么在这块地里挣到的钱，很难长出别的东西。

◆ 七、本书的说法

把这件事说透：

Token 消耗在哪里，取决于谁在用 AI 生产，不取决于谁在消费。一部剧被八亿人看，和被八个人看，烧的 Token 一样多。

那么，那张图上的 Token 结构，反映的是什么？

它反映的是：做内容的公司愿意为 AI 花多少钱，愿意把多少生产环节交给 AI。也就是说，是钱往哪里去了，是押注押在了哪里。

所以那张图上的 Token 结构，说出的主要是钱押在了哪里，至于中国人拿 AI 在做什么，它说得很少。第 8 章会把这句话收成上篇的总判断。

这是拆完前四章以后，那张图里剩下的最后一点真东西。它不是计量错觉，也不是样本错位。它是真的：在中国，最早、最大规模地为 AI 掏钱的，是生产注意力商品的那一头。

为什么是这一头？第 6 章会讲。简单说一句：它回款最快。

◆ 八、先说它哪里对

说到这里，又容易滑向骂短剧。所以这一节，先把公道还给短剧。

短剧本身，是内容行业里的一场生产力变化。

从前拍一部剧，要几百万、几千万，要一个大剧组，要电视台或大平台买单。只有少数公司、少数人能进这个门。今天，一个小团队、几台电脑、一套 AI 工具，三天就能出一部剧。门槛一下子低了好几个数量级。

这意味着，很多从前进不了门的人，进来了：写网文的作者，学动画的学生，县城里想拍点东西的年轻人。他们不用求一个大公司，就能把自己脑子里的故事变成画面。

FlareFlow 有 3700 多万注册用户，五千多部剧。中国的短剧，正在被卖到海外去，被外国人看。这是中国的内容第一次以这样的规模、这样的速度出海。

还有一层。AI 短剧行业逼着视频生成技术往前跑。导演要的「又哭又笑」，要人物在几十集里长一张不变的脸，要雨的方向前后一致——这些要求，一个一个地压在模型上，模型就一点一点地被磨好。

这是实打实的本事。说它是烟花，它也是一场全世界都没见过的烟花。

◆ 九、只看一面会怎样

可是，把公道还给短剧以后，还是要问一句。

只看生产力那一面，会说：短剧是好事，它让内容生产便宜了一百倍，让更多人进了门，让中国内容出了海。这都对。

可是它便宜了的，是哪一种东西？是注意力商品：一种卖给人时间、换来广告和充值的东西。它进门的人越多，产量越大，每一部剧能分到的注意力就越少。一天一千五百部，一个人一天最多看二十几集。剩下的，没人看。

只看计量那一面，又会说：Token 结构不说明问题，都是视频天生费 Token。这也对。

可是它没有回答：为什么第一批愿意为 AI 大把掏钱的，是短剧和广告，而不是药厂、材料厂、气象局？

两面放在一起，才看得见：短剧是一场真的生产力变化，可它变的是注意力的生产力。资本先押在这里，是因为这里回款最快。至于这里挣到的钱、练出的本事，会不会回流到别处，是另一个问题。

◆ 十、今天与机器

这件事还在加速。

视频生成模型每隔几个月就换一代。每一代，画面更真，成本更低，速度更快。一部剧从三天到一天，从一天到几个小时，大概只是时间问题。

到那时候，每天新增的可能不是一千五百部，是一万五千部。Token 的账上，视频那一块还会更大。

另一边，模型也开始能自己写剧本、自己分镜、自己挑镜头。宋岚的抽卡，也许有一天会交给另一个模型去抽：一个模型生成，另一个模型打分，人只看最后那一条。

这一步一旦走成，内容生产就从「人用 AI 做剧」变成「AI 做剧，人来挑」。生产的人更少，产量更大，Token 更多。

到那时，那张图大概会更扎眼。可是它说明的，仍然只是资本押在了哪里。

◆ 十一、一个完整的例子：一部剧的账

用宋岚手上的一部剧，把这一章的账算一遍。

这部剧八十集，每集一分半钟，讲一个被退婚的女孩继承了一座山庄。剧本是公司里一个签约作者写的，用了两周。分镜交给了 AI，一天。生成画面，宋岚带着四个人，抽卡抽了三天。剪辑、配音、配乐，又是两天。

这部剧一共烧了多少 Token，宋岚没有告诉林舟具体的数。她只说，比林舟的组一年用的还多。

剧上线以后，第一周有两百万次播放。看的人，一个 Token 也没烧。

公司在这部剧上的收入，来自两块：平台的分账，和剧中植入的一个护肤品广告。广告那一块，又是另一家公司用 AI 生成的。

这一部剧，从生产到挣钱，前后不到一个月。

再看林舟组里那个进了临床前研究的分子。从立项到今天，三年。它要变成药，还要五年到十年，还要过三期临床，失败的可能远大于成功。

两笔账放在一起，就看得见资本为什么先选宋岚那一头。不是因为它不重视新药，是因为新药的钱要等十年才知道回不回来，短剧的钱一个月就回来了。

◆ 十二、本事会不会搬家

还有一个问题，要留给第 7 章，这里先提一句。

宋岚这几年练出了一身本事。她知道怎样把一句说不清的要求，写成模型听得懂的提示词；知道怎样让一个人物在八十集里长同一张脸；知道一批画面里，哪一条最可能被导演选中。她公司的技术团队，也练出了本事：怎样把生成的成本压下来，怎样把几百张卡排满，怎样让一个镜头从生成到交付只要几分钟。

这些本事，是真本事。问题是，它们会不会搬家？

一个能让人物在八十集里长同一张脸的团队，能不能帮林舟的组，让一个蛋白在一万次模拟里保持同一种折叠？一个能把几百张卡排满的调度系统，能不能拿去排一个气象模型的计算？

有的能，有的不能。能搬家的那部分，就是回流。

这件事本书在上篇还答不完。第 6 章会讲钱为什么先往这里跑，第 7 章会讲这些钱和本事，是先养后用，还是就此锁死。

◆ 十三、放回日子里

那个周日晚上，林舟和宋岚吃完饭，坐在沙发上。九岁的林豆豆在旁边写作业。

宋岚手机响了一下，是那部山庄剧的后台数据：第二周播放量破了五百万。她笑着把手机递给林舟看。

林舟看了看，说：「八亿人看你们的剧，你们烧的 Token 却跟他们一点关系都没有。」

「本来就没关系啊。」宋岚说，「他们看的是故事。」

林豆豆抬起头：「妈妈，那你们做剧的钱是谁给的？」

宋岚想了一下：「是看剧的人给的，也是打广告的人给的，还有投钱给我们公司的人。」

林豆豆又问：「那爸爸做药的钱是谁给的？」

两个大人对看了一眼，都没马上答上来。

一个小办法，明天就能试：下次看到一个关于「中国人在用 AI 做什么」的数，先分清两件事——这个数量的是谁在生产，还是谁在使用。量生产的，说的是钱押在哪里；量使用的，才说的是人在做什么。

> **八亿人在看，几十万人在烧，钱押在了烧的那一头。**

> 带 走 的 话
>
> **一部剧被八亿人看和被八个人看，烧的 Token 一样多；烧 Token 的是做剧的那一头。**

> 小　账
>
> **看见了什么，漏看了什么？**看见了短剧是一场真的生产力变化，门槛降了好几个数量级；漏看了它变的是注意力的生产力，钱先押在这里，是因为回款最快。
>
> **是在改进，还是在出题？**一天一千五百部，是在「怎样更便宜地抓住人的时间」这道老题里做到了极致；新的题，还没人出。

> 想 一 想
>
> 你一天花在短视频和短剧上的时间，是多少分钟？你工作上用 AI，是在生产东西，还是在消费东西？如果林豆豆问你那个问题，你家里做事的钱是谁给的，你会怎么答？

◆

06

第 6 章

资本按回款速度排队

◆ 一、老板墙上的那张表

周一早上九点半，宋岚走进会议室，老板已经把一张表投在墙上。

表只有三列：投进去多少，多久回来，回来多少。

第一行是上个月上线的一部 AI 仿真人剧，六十集，每集一分多钟。从立项到上线，用了不到三周。投流到第五天，充值就盖过了制作费。

第二行是一个客户找上门的项目：给一家做汽车零件的工厂做培训视频，用 AI 生成装配流程的动画。合同要走三个月，回款要等验收，验收要等对方车间主任点头。

老板指着第二行，只说了一句：「这个，先放一放。」

宋岚没有反对。她知道老板不是看不起工厂。他看的是中间那一列：多久回来。

这一章，讲的就是这一列。

◆ 二、三种钱，回来的速度不一样

先把材料摆在桌上。

第 2 章拆过的那组数，短剧与视频 55%、电商与营销 24%、软件开发 15%、其他 6%，口径没有公开；第 3 章又说过，视频天生费 Token，这个比例被计量方式放大了。

先把放大的部分扣掉，只留下次序。再把这几样按「钱多久能回来」排一排，你会发现排出来的次序，和占比的次序几乎一样。

短剧：一部剧上线，当天就有充值和广告分成。一天新上一千五百部，这一部不行，明天换一部。

电商与营销：一条带货视频、一场直播，转化率当晚就看得见。广告主按点击和成交付钱，钱回来的周期以天计。

软件开发：一个编程助手卖给企业，要试用、要走采购、要接进现有系统，钱回来的周期以月计。

工业、医药、材料这些生产力场景：要改流程、要过验证、要等监管，钱回来的周期以年计。

Token 的去向，差不多就是一张回款速度表。

◆ 三、注意力商品，三样都快

注意力商品，指的是拿来换人的时间和眼睛的东西：一部短剧、一条广告、一段带货视频。它在资本眼里有三样好处。

第一，需求是现成的。第 5 章说过，看短剧的人已经超过八亿。不用教育市场，人已经坐在地铁上、躺在沙发上刷了。

第二，回款快。充值、会员、广告分成，大多按天结算。

第三，失败便宜。一部 AI 短剧砸了，损失的是一笔生成费和几个人几天的工；没有人因此受伤，没有生产线停下，没有病人吃错药。

三样放在一起，就是一门资本最喜欢的生意：试错便宜，反馈快，一天可以赌很多次。

AI 恰好把这三样又往前推了一大步。第 5 章提过那篇博文的测算，制作周期从半个月压到三天〔待核〕。成本降一截，能下注的次数就多一截；周期短一截，回款就快一截。

你可以把它想成一台老虎机，每一次拉杆都便宜，每一次结果都马上出来。这样的机器前面，从来不缺人排队。

◆ 四、生产力场景，三样都慢

再看另一头。

林舟的组做 AI 制药。一个分子从电脑上筛出来，要合成，要做细胞实验，要做动物实验，然后才谈得上临床。业内常说，一款新药从立项到上市要十年上下〔待核〕。AI 能把最前面筛选的那一段压短，压不掉后面那几道关。

工厂也一样。一家工厂要用 AI 排产、做质检，得先把几十年的老流程拆开，把散在各处的数据接起来。还得把老师傅手上的那点感觉写成规则，还得有人为出错签字负责。这是改流程，不是装一个软件。

三样都慢：需求要培养，回款要等，失败的代价大。一个质检模型漏掉一种缺陷，可能是一整批零件召回；一个筛药模型把方向带偏，可能是一个组白干两年。

所以同样一块钱，放在短剧上，一个月就能看见回头；放在工厂里，三年未必回本。

请注意，这里没有谁是坏人。一个要对股东负责的老板，按回款速度给项目排队，是他的本分。宋岚的老板如果把钱先押在工厂培训上，公司撑不到验收那一天，他也就没有机会做第二个工厂了。

◆ 五、投钱的人也在排队

老板排队，投钱给老板的人也在排队。

一家 AI 短剧公司去融资，投资人问的第一句话，常常是「月流水多少」；一家做工业 AI 的公司去融资，投资人问的第一句话，常常是「什么时候能有标杆客户」。前一个问题，下个月就能答；后一个问题，要等一两年。

基金也有自己的期限。一只人民币基金，从募资到退出，常见的年限并不长〔待核：各类基金差别大〕。基金经理要在期限里给出资人交账，他手里的项目，也得在期限里看得见回头。

上市公司还有季报。每三个月，要拿出一张让股东满意的表。一个三年才见效的项目，要连着十二个季度解释「为什么还没有收入」。

这样一层一层排下去，从出资人到基金，从基金到公司，从公司到项目组，每一层都在问同一列：多久回来。

到了宋岚老板那张三列的表，这个问题已经被问过好几遍了。

◆ 六、美国企业为什么肯付高价

那为什么美国那边，看上去不太一样？

这里先打一个预防针。第 4 章说过，美国那个 58% 找不到可靠的出处，本书不用它。能核的只有第 4 章摆过的两件事：OpenRouter 这个以美国开发者为主的平台上，编程用量过了一半；CNBC 报道，美国企业把三成以上的 Token 发给便宜的中国模型，出错代价大的活却仍留给贵的模型。

两件事放在一起看：美国企业也省钱，省的是不要紧的那部分；要紧的那部分，他们肯多付。

设想一家美国公司的工程主管，让编程助手每天替二十个工程师写测试、查漏洞。他算的是一道很直白的题：模型贵一倍，可只要它少放过一个线上故障，这一倍就挣回来了。

他们把 AI 当成替代工程师工时的工具来算账。这笔账算得通，靠三个条件。

一是人贵。一个美国工程师一年的人力成本，在中国可以请好几位同行〔待核〕。AI 每替他省一个小时，省下的钱就多。

二是企业习惯为软件付钱。按年订阅、按人头收费，美国的采购部门做了几十年，一个新工具进来，有现成的预算科目和审批路子。

三是软件早就接进了流程。代码、工单、客服记录、报销单，本来就在系统里流转，AI 接进去，阻力小，效果容易量。

还要补一句公道话：美国企业肯付高价，也不全是因为他们看得远。他们的条件，让这笔账本来就容易算通。换一句话说，美国那边的生产力在前，同样是条件的结果，不是品味的结果。

◆ 七、同一笔账，搬到中国

把这三个条件搬到中国，每一条都打了折扣。

人相对便宜。一家工厂上 AI 质检，要和一位熟练质检工的工资比；一家公司上 AI 写代码，要和一位年轻程序员的工资比。差价小，回本就慢。

企业为软件付费的意愿弱。中国的企业软件市场比美国小得多〔待核：各家研究口径差别大〕。很多企业习惯买硬件、买项目、要定制，不习惯年年为一份订阅付钱。卖软件的公司常常要派人驻场，把产品做成半个外包。

流程接进软件的程度参差不齐。不少中小企业的单据还在微信群和纸面上走。AI 要先把数据接上，才谈得上用；光接数据这一步，就够一个项目做半年。

还有一层，是价格。2024 年 5 月前后，DeepSeek 把模型接口的价格降到一个很低的水平，各家跟进，国内的调用价格一路往下走（DeepSeek 怎样把成

本压下来，第 14 章会讲）。便宜对用的人是好事。可是卖的人赚得薄，就更不愿意为某个回款慢的行业做深做透；「为准确多付钱」的那种账，在国内也就更难长出来。

所以中国的大模型公司，面对的是一个奇怪的市场：C 端有八亿人愿意花时间，B 端有几千万家企业不太愿意花钱。

钱往哪边走，不难猜。

◆ 八、林舟的组为什么买不起好模型

再回到林家。

周三晚上十点，林舟还在研究院。他想让组里的博士后用最好的那几个大模型读文献、写分析代码。他在纸上算了一下：一个月的调用费，大概几千到一万多元。

钱不算多，难在这笔钱没有地方出。

他手上的基金，预算是几年前申请时定的：设备费、材料费、测试化验加工费、劳务费。按 Token 计费的模型调用，该放进哪一栏，财务也说不清。最好的几家海外模型，有的在国内不能直接买，要绕路；绕路的钱，就报不了账。

最后组里用的，是一个有免费额度的国产模型，加上几个学生自己掏钱买的会员。

这个晚上很普通，普通到没有人觉得它是个问题。第二天上午，他去财务处问了一次。财务老师很客气，翻了半天制度，说可以试试算进「测试化验加工费」，可得有正规发票，最好是和有资质的单位签合同。林舟算了算签合同要走的流程，大概要一个多月。

那个月，组里正赶一篇论文的修改意见。最后，读文献的活还是学生一篇一篇自己啃完的。

这件事单拿出来看，只是一桩报销的小麻烦。把它和宋岚公司墙上那张表放在一起，就看得出一件事：短剧公司的 Token 预算按天追加，研究组的 Token 预算按年申请、按科目报销。同一座城市，同一种 Token，一边是流水，一边是审批。

宋岚一个月烧掉的 Token，比林舟的组一年还多。原因不在谁更聪明，在两边的钱，按两种速度流动。

◆ 九、本书的说法

先把结论摆在桌上：

中国的 AI 先走娱乐，是激励结构决定的，不是品味问题。

把同一套条件放在任何一个国家：人力相对便宜，企业不惯为软件付费，八亿人的注意力市场现成摆着，价格战把单价压到很低。在那里，资本也会先往短剧和广告那边排队。换成别的国家的老板坐在宋岚公司的会议室里，大概也会说那句「先放一放」。

这样看，那张流传的图，扣掉计量错觉和样本错位之后剩下的那一半，说的不是中国人爱看什么，说的是钱在怎样排队。

这句话也不是替谁开脱。激励结构是人搭起来的：基金的预算科目、企业的采购习惯、模型的定价方式、资本的退出年限，每一样都可以改。说它是结构，正是说它可以改；说它是品味，反倒什么都改不了。

这一点要紧。它决定了批评该落在哪里：落在人身上，只会变成骂；落在激励结构上，才有东西可改。

◆ 十、只看一面会怎样

只看一面，有两种。

一种只看「烧在短剧上」，于是说：中国 AI 不务正业，拿最先进的技术拍狗血剧。这是把激励结构的结果，算到了人的品味头上。它还漏看了一件事：同样这批钱，养着一大批编剧、剪辑、配音和工程师的饭碗，也在练视频模型的本事。下一章专门讲这一层。

另一种只看「回款快是好事」，于是说：市场选的，就是最好的，不必多想。这漏看的是时间。回款快的地方，练出来的是回款快的本事；要等十年的那一头，如果一直排不上队，十年以后，就没有人会做了。

两种看法在事实上都没有错。错在只看了表上的一列。

◆ 十一、今天与机器：一个完整的例子

AI 让这件事又快了一层。

以前拍一部短剧，要演员、要场地、要摄影组，成本卡住了每天能上新多少。现在成本被压下去，下注的次数暴涨：第 5 章说过，一个月的 AI 短剧，抵得上从前一年的真人短剧。

看一个完整的镜头。一部 AI 仿真人剧里，女主角推门进来，回头看了一眼。导演不满意那一眼的神情，就让模型重新生成，再生成，挑出最抓人的一版。第 2 章讲过宋岚的「抽卡」，一个十秒镜头抽了十七次。

这些 Token，每一个都有人付钱。付钱的人算过账：抽十次卡，换一个更抓人的回眸，多留住几万个观众，划算。账算得清，是因为结果第二天就在后台看得见：哪一集的完播率掉了，哪一个镜头之后有人划走，数据一清二楚。

同样一笔钱，如果拿去给林舟的组读一年文献，换来的也许是一篇论文、一个候选分子，也许什么都没有，而且要等。

资本不会犹豫选哪一个。

◆ 十二、放回日子里

那天开完会，宋岚在电梯里又想起那个工厂培训的项目。

她其实喜欢那个项目。装配动画要准，一颗螺丝拧几圈都不能错，比拍一个回眸难多了。可她也清楚，这种活在她们公司排不上号。

晚上回家，她把这件事讲给林舟听。林舟笑了一下，说他的组也差不多：最想做的那个课题，申请书里写不进去，因为三年内出不了成果。

一个小办法，留给你：下次再看到「中国 AI 都在干什么」这类说法，先别急着下判断，问一句——这笔钱多久能回来？回款的周期，常常比任何口号都更能说明一个行业在做什么。

> **钱不问你想做什么，只问它多久能回来。**

> 带 走 的 话
>
> **短剧的钱按天回来，药厂和工厂的钱按年回来；资本排队，看的就是这一列。**

> 小　账
>
> **看见了什么，漏看了什么？**看见了资本按回款速度排队；漏看了回款快的地方，只练回款快的本事。
>
> **是在改进，还是在出题？**抽卡、降本、提速，全是在别人出好的题里，做得更快更便宜。

> 想 一 想
>
> 你所在的单位，哪一项开支是按天追加的，哪一项是按年审批的？如果把两样对调一下，会怎样？

◆

07

第 7 章

先养后用，还是锁死

◆ 一、林舟的第一块显卡

林舟十六岁那年，攒了两年的压岁钱，买了一块显卡。

那是二十二年前的事。他要的东西很简单：让一款三维游戏不再卡顿，让屏幕上的烟雾和水面好看一点。他妈妈心疼钱，说了一句：「买这个，能当饭吃吗？」

二十多年后的一个晚上，林舟坐在研究院的机房外面，等一批蛋白质结构的预测结果。机房里跑的，是一排一排的计算卡。它们的祖先，就是他十六岁时插进主机箱的那块游戏显卡。

他妈妈那句话，今天可以回答了：能。只是要等二十年，而且吃饭的不是买显卡的那个孩子，是另一个行当。

这一章要问的，就藏在这二十年里：拿娱乐养大的技术，什么时候、怎样才变成生产力？

◆ 二、显卡是被游戏养大的

先说材料。

图形处理器，最早是为了让电脑画三维画面而做的。画面里每一个点的颜色、明暗，都要算；这些计算彼此独立，适合成千上万个小核一起算。谁最需要这种算力？打游戏的人。

于是整个九十年代末到两千年代，推着显卡一代一代变快的，主要是游戏玩家的钱。英伟达在 1999 年把自己的产品叫作 GPU，卖给的就是这些人。

转折出现在 2006 年前后。英伟达推出 CUDA，让程序员可以拿显卡算画面以外的东西。一开始用的人不多，主要是物理模拟和一些科研计算。

2012 年，多伦多大学的 Krizhevsky、Sutskever 和 Hinton 用两块游戏显卡训练了 AlexNet，在 ImageNet 图像识别比赛上大幅领先。深度学习从此走出冷宫——第 19 章会细讲那段冷宫。

把这条路连起来看：游戏玩家花了十几年的钱，养大了一种算力；这种算力后来成了 AI 的底座。今天全世界最贵的 AI 训练集群，追根溯源，是被一代一代打游戏的少年喂大的。

◆ 三、带宽、支付、视频编码

显卡不是孤例。

家里的宽带，最早被谁拉起来的？很大一部分是看视频、下电影、打网游的人〔待核〕。他们嫌慢，愿意为更快付钱，运营商才有动力铺光纤。今天远程会诊、在线课堂、工业云，走的都是这些光纤。

网上支付，在中国最早的大规模练兵，有一块是网游点卡和各种虚拟物品〔待核〕。小额、高频、要防盗刷，这些本事后来都用在了淘宝和支付宝上。第 10 章、第 11 章会接着讲这段。

视频编码也是。为了让一部电影在更窄的带宽里传得更清楚，工程师一代一代改压缩算法。推动它的是影碟、网络视频、直播；今天的医疗影像传输、卫星遥感回传，用的也是同一类技术的后代〔待核：具体对应〕。

所以先说公道话：娱乐常常是技术的第一推动。它需求大、付钱快、容忍失败，正好替一项还很粗糙的技术付学费。很多技术，如果一开始就要面对工厂和医院的苛刻要求，可能根本活不到长大的那一天。

◆ 四、为短剧而卷的视频模型

现在看今天。

国内几家做视频生成的大模型，几乎都被短剧和广告推着往前跑：快手的可灵、字节的 Seedance、阿里的通义万相。它们比的是什么？

比人物一致：同一个女主角，第一集和第六十集要是同一张脸。

比动作合理：推门、转身、倒水，手不能多出一根指头，杯子里的水不能往上流。

比镜头可控：导演说要一个从左往右的推镜，模型就得照做。

比时长和清晰度：从几秒到几十秒，从 720p 到 1080p。

这些要求，听起来全是娱乐的要求。可你仔细看，每一条背后都是一种更一般的本事。记住一个东西在不同时刻是同一个东西；知道物体在三维空间里怎么动；听懂人用语言下的指令，再把它变成连续的画面。

短剧在替视频模型交学费，交的正是这几门课的学费。

◆ 五、这些本事能流到哪里

这些本事，可以流到很多地方去。下面几处，是可能，不是已经发生的事实。

工业仿真：工厂里的缺陷样本少，坏件不常有。视频模型如果能按要求生成「这种划痕在这种光线下的样子」，就能给质检模型造训练材料。

医疗影像：病例少的罕见病，图像也少。会生成、会理解连续画面的模型，可以帮着补数据、做动态影像的分析，当然要过严格的临床验证。

教育：一个物理过程、一段细胞分裂，用准确的动画讲给九岁的孩子看，比一页课本好懂得多。

机器人视觉：机器人要在真实世界里走路、抓东西，得先在「想象」里练很多遍。能生成合理物理画面的模型，就是一座练功房。

这一串可能，和显卡的故事很像：先被娱乐养大，后来被别的行当用上。

◆ 六、本书的说法：关键在回流

先把结论摆在桌上：

娱乐养出来的本事回流到生产力，就是「先养后用」；不回流，就是锁死在注意力经济里。

这一句里，要紧的是「回流」两个字。

先后本身不是问题。显卡先给游戏，再给 AI；宽带先给视频，再给远程医疗。先后顺序是自然的，甚至是好的。问题出在后半截：养大以后，有没有人把它接到另一个行当里去，有没有钱和人愿意走那段回款慢的路。

回流不会自己发生。英伟达押 CUDA 的那些年，科研计算这块市场小得可怜，是一家公司愿意在主业之外养一个没什么回报的方向，才有后来的接口。

火药也是这样。拿火药做烟花，一点错都没有，烟花好看，过年要放。问题在烟花放了一千年，火药一直在做烟花，没有人去问它还能推什么。

◆ 七、谁来付那段慢钱

回流要有人付钱，付的是一段慢钱。

在显卡的故事里，付慢钱的是一家公司自己。在别的故事里，付慢钱的人各不一样。有的是大公司里一块不考核收入的研究部门，有的是政府的长期科研经费。还有的，是大学里一个愿意花十年做冷门方向的实验室。

在中国，这几种付慢钱的人都有。大厂有自己的研究院，国家有重点研发计划，高校有实验室。他们做出了不少真东西，后面几章会提到。

可是你把这些钱放回第 6 章那张表里看，会发现它们也常常被要求「快」。大厂研究院要对业务部门交代，科研项目要按年度考核、按期结题，实验室要按年数论文。慢钱被一层一层的考核，又催成了快钱。

所以回流难，不全难在没有钱，也难在钱不肯慢。第 43 章会回到这件事，讲一块免于指标的飞地是什么样子。

◆ 八、Token 越多，越像通货膨胀

不回流会怎样？

第 1 章提过国家数据局给的那个数：日均 Token 消耗超过 140 万亿，两年多涨了一千多倍。

这个数字很大，也确实说明了中国的工程和规模。可是一个数字涨得快，不等于它代表的东西涨得快。

打个比方。一个国家印的钞票多了一千倍，商品没有多一千倍，每张钞票就不值钱了。Token 也有点像。如果新增的 Token 绝大部分去生产同一类注意力商品——同一种霸总、同一种逆袭、同一种回眸——那么 Token 越多，每一个 Token 换来的新东西就越少。

第 5 章说过，每月 4 万部的 AI 短剧，已经约等于去年一整年的真人短剧产量。剧多了，人每天的时间不会多。注意力是有上限的，八亿人一天也只有二十四小时。

在一个有上限的池子里不断加 Token，加出来的主要是竞争，不是新东西。这就是锁死的样子：钱在转，Token 在烧，本事在练，可练出来的本事只在一个圈子里打转。

◆ 九、只看一面会怎样

只看「锁死」那一面，会把短剧说成一无是处，把视频模型的进步说成浪费。这不公道。它们是真进步，而且这类进步在许多国家都是从娱乐里起步的。

只看「先养」那一面，会说：放心，养大了自然会用。这就把回流当成了必然。显卡的故事被讲得太顺了，好像游戏养大的东西注定会变成 AI 的底座。可是同一时期，也有很多被娱乐养大的技术，最后只留在了娱乐里。

看一件事是「先养后用」还是「锁死」，不能看它出发时的样子，要看它后来流向了哪里。这需要证据，需要一年一年地看。

还有一种情形，比锁死更隐蔽：本事流出去了，可只流到了另一个注意力行当。视频模型从短剧流到广告，从广告流到直播，从直播流到游戏宣传片。每一步都像是在「外溢」，其实一直在同一个池子里打转。

所以看回流，不只看它流没流，还要看它流到了哪里：是流到了另一种更便宜地生产注意力的地方，还是流到了要为准确负责的地方。

◆ 十、几个看得见的信号

回流看不见摸不着吗？其实有几个信号，普通人也能看。

第一个信号，看客户名单。做视频模型的公司，年报和发布会上，有没有出现工厂、医院、机器人公司这类付费客户，还是只有短剧平台和广告公司。

第二个信号，看钱的去向。短剧挣来的钱，有没有一部分进了三年以上才见效的研发项目，比如面向物理仿真、医疗的专门模型。

第三个信号，看人的流向。做视频生成的工程师，有没有人去了机器人、制造、医疗器械公司，带着他们练出来的本事。

第四个信号，看科研引用。国产的开放权重视频模型，有没有被科研团队拿去做科学问题，比如细胞运动、流体、天气的生成与预测，论文里能不能看到它们的名字。

第五个信号，看价格。有没有人开始为「准确」付高价，而不只为「便宜」和「快」付钱。

五个信号里，有两三个亮起来，回流就在发生。全都不亮，那就是锁死。

◆ 十一、开放权重是一条现成的渠道

说到回流，中国手里其实已经有一条现成的渠道。

第 4 章说过，在 OpenRouter 上，中国模型的周调用量从 2026 年 2 月起超过了美国模型，而大量调用它们的，是美国企业。很多中国大模型是开放权重的：模型的参数公开放出来，谁都可以下载，拿到自己的机器上改、上自己的数据去练。

这条渠道的意思是：一个在短剧和对话里练出来的模型，不必等国内的工厂和医院来买。世界各地的研究组和工程师，已经可以拿它去试别的事。

这是真回流的一种可能，而且是中国做得很好的一件事。

不过渠道开着，不等于水在流。别人拿去做了什么，做成的东西回不回到中国的研究和产业里，又是另一件事。第 41 章讲回流通道的时候，还会碰到这个问题。

◆ 十二、今天与机器：一个完整的例子

把一个镜头从头到尾跟一遍。

宋岚的组做了一部古装 AI 短剧。为了让女主角在六十集里保持同一张脸、同一件衣服的同一道褶子，他们和模型厂商一起，花了几周调人物一致性的参数，抽了上万次卡。

这份功夫，在短剧里只换来一个结果：观众没有出戏。

可是同样的本事——让一个东西在长时间、多角度的画面里保持是它自己——在别处是硬功夫。一台机器人要认出桌上那只杯子，从不同角度看都是同一只；一套手术录像分析系统，要在几十分钟的画面里一直跟住同一根血管。

这份本事会不会流过去，取决于有没有人去接。模型公司愿不愿意开一条回款慢的产品线，机器人公司愿不愿意付钱，研究组有没有预算去试。

上一章说过，林舟的组连调用费都报不了账。回流的那一头，接的人手里也得有钱。

◆ 十三、放回日子里

那天晚饭，宋岚随口说起，他们用的模型现在能让一个角色六十集不变脸了。

林舟停下筷子，问：「它能不能生成细胞在显微镜底下分裂的视频？要准的那种。」

宋岚想了想，说：「能生成好看的，准不准，我不知道。没人跟我们提过这个需求。」

没人提过。这四个字，正是回流还没打通的样子。

一个小办法，留给你：下次看到一家 AI 公司的发布会，别只看它生成的视频有多惊艳。翻到客户和合作那一页，数一数有几个不是娱乐和广告的名字。这个数字一年一年变化，就是回流的温度计。

> **拿火药做烟花不是错，错在一千年只做烟花。**

带 走 的 话

先养后用和锁死，起点一模一样；分开它们的，是养大以后有没有人把本事接到别的行当里去。

小　账

看见了什么，漏看了什么？看见了娱乐替新技术交学费；漏看了回流不会自己发生，要有人付那段慢钱。

是在改进，还是在出题？视频模型的一致性、可控性是在已有题里做得更好；把它接进机器人和医学，才开始碰到新题。

想 一 想

你用过的哪一样东西，最早是为了好玩而做的，后来成了你干活的工具？中间是谁把它接过来的？

◆

08

第 8 章

本篇总账：烟花的账

◆ 一、除夕夜的烟花

除夕夜，林家开车到上海郊区老林租的那块菜地边上，借邻居的院子吃年夜饭。

快到零点，林豆豆拉着爷爷去院子外面看烟花。远处的天空一片一片地亮，红的、金的、绿的，一朵刚散，下一朵就接上。

林豆豆问：「爷爷，烟花是谁发明的？」

老林说：「火药是我们中国人发明的。」

林豆豆又问：「那火药后来还干了什么？」

老林想了半天，没有答上来。

上篇七章，讲的都是一场烟花：一张图，几个百分比，八亿观众，每天一千五百部剧，一个月烧掉的海量 Token。这一章把账算一算：七章里，看见了什么，漏看了什么，剩下的那件真事是什么。

◆ 二、七章的看见与漏看

先把七章并排放：

章	看见了	漏看了
第 1 章　一张流传的图	中国 AI 用量里科研的份额可能很薄，这张图戳中了很多人的担心	图里的两组数，量的不是同一样东西

章	看见了	漏看了
第 2 章　55% 是怎么算出来的	数据来自一家短剧公司的投资者记录，有真实的行业观察	原始口径没有公开；「85%」是加总再往上说的版本
第 3 章　十秒视频和一千行代码	视频生成确实烧 Token	视频 Token 的密度比文字高几个数量级，按 Token 比用途，比的是秤，不是货
第 4 章　找不到出处的 58%	美国在编程与企业自动化上投入大，有可核的旁证	58% 这个数找不到出处；拿全国总量比开发者平台，样本错了位
第 5 章　八亿人和每天一千五百部	短剧市场是真的，规模大，养活了很多人	注意力有上限，剧越多，每部越难
第 6 章　资本按回款速度排队	先走娱乐由激励结构决定：回款快、需求现成、失败便宜	回款快的地方只练回款快的本事
第 7 章　先养后用，还是锁死	娱乐常常替新技术交学费，视频模型在长本事	回流不会自己发生；不回流，Token 越多越像通货膨胀

这张表从上往下读，是一个漏斗。前四章把图上的水分挤掉：一半是计量错觉，一半是样本错位。后三章看剩下的那一点真货。

◆ 三、挤掉的那一半

先说挤掉的。

那张图之所以流传得这么广，是因为它把一个复杂的局面，压成了一句好记的话：中国人用 AI 拍短剧，美国人用 AI 写代码。

可这句话里，两边的秤不一样。一边是一家短剧公司援引的全国行业数据，一边是一个找不到出处的数字。一边按 Token 算，可第 3 章算过，十秒视频烧的 Token，抵得上写上万行代码。拿 Token 量去比用途，好比拿重量去比一车棉花和一车铁谁更值钱。

还要扣掉一层。OpenRouter 上，中国模型的周调用量从 2026 年 2 月起超过美国模型，到 10 月初约 86.7 万亿 Token 对 19 万亿（界面新闻测算）；在这

个以编程为主的平台上，大量调用中国模型的，恰恰是美国企业。中国的模型，正在别人的生产力场景里干活。

所以「中国 AI 只会拍短剧」这句话，站不住。

◆ 四、剩下的那一半

挤掉水分以后，还剩一样东西，挤不掉。

在国内，用 AI 的钱，确实大量先流向了注意力商品。这不是计量错觉造的，第 6 章的那张回款速度表，在任何口径下都成立：短剧和广告的钱按天回来，工厂和药厂的钱按年回来；企业不惯为软件付费，人力相对便宜，模型价格又被压得很低。

这件真事，和中国人爱看什么没有关系，和中国的工程师会做什么也没有关系。它是一张激励结构的照片。

打个比方。一条河，上游下了一场大雨，下游的水就涨。你站在下游，看见水往哪边涨，就知道雨下在哪里。Token 的流向，就是下游的水位；资本押注的方向，就是上游的那场雨。

看水位的时候，要先校准水尺。前四章做的，就是校准水尺。校准完了，水位仍然偏向一边，那就是雨真的下在了那一边。

这张照片里，还有一个没有答案的问题：注意力挣来的钱、练出来的本事，会不会回流到生产力？第 7 章给了五个信号，现在下判断还太早。

◆ 五、火药与烟花

回到除夕夜那句没有答上来的话。

这个故事常被讲成一句感叹：我们发明了火药，却只拿来放烟花。第 1 章说过，这句感叹当史实看未必准确，火药在中国也早早上了战场。本书只借它当一个比喻。

比喻要用对地方。拿火药做烟花，一点错都没有。烟花有人要看，有人愿意付钱，做烟花的人练出了配方、引信、外壳的手艺。上篇七章说的短剧，就是今天的烟花：好看，挣钱，养人，也练手艺。

问题不在做烟花。问题在两件事。

一件是，这门手艺有没有流到烟花以外的地方去，这是第 7 章说的回流。

另一件更深：有没有人去发明下一种火药。烟花做得再好，也是在火药已经有了之后的事。中国 AI 今天做得最好的，是把已有的模型、已有的方法，用到最便宜、最快、最大规模；可是模型的架构、训练的方法、这一整套做法本身，大多是别人先出的题。

后一件事，上篇只能提一句。它要靠后面几篇慢慢拆。

◆ 六、这张表的两种读法

同一张表，有两种读法。

一种是从左往右读：每一章先看见了什么，然后漏看了什么。这样读，你会觉得上篇在一章一章地纠错，把一张流传的图拆成了碎片。

另一种是从上往下，只读「漏看了」那一列：两组数不是同一样东西；口径没有公开；秤不一样；出处找不到；注意力有上限；回款快的地方只练回款快的本事；回流不会自己发生。

前四条，是关于数据的；后三条，是关于钱的。

关于数据的漏看，挤掉就好了，挤掉以后，图上那种「中国人只会拍短剧」的惊慌，就没有了根。关于钱的漏看，挤不掉，它们指向的是一个结构，要一篇一篇往下拆。

上篇的工作，就是把这两类漏看分开。分开以后，才知道哪些担心可以放下，哪些担心要带着往下走。

◆ 七、本篇总判断

把七章的账合在一起，本篇的判断是：

Token 结构说出的，不是中国人在做什么，而是资本把 AI 押在了哪里。

这句话有两层意思。

第一层是替中国说公道话：那张图不能拿来证明中国人短视、爱娱乐、不务正业。它照出来的是钱的排队方式，换一套条件，排队的方式就会变。

第二层是不放过问题：钱押在哪里，本事就长在哪里。押在回款最快的地方，长出来的就是把已有的东西做得更快更便宜的本事。这本事很了不起，可它不是出新题的本事。

◆ 八、公道要先还

在往下走之前，先把公道还足。

中国在 AI 上做成的事，并不少。工程能力强，能把一个模型从实验室推到几亿人手里；规模大，日均 140 万亿 Token 的调用量背后，是机房、电力、网络和运维的硬功夫；成本低，同样的能力，价格常常是海外的几分之一。开放权重的中国模型，在全球最大的模型接口平台上，调用量已经排到前面。

还有科学本身。华为的盘古气象模型 2023 年发表在 Nature 上，中期天气预报算得比传统方法快得多，第 21 章会细讲。DeepSeek 在算力受限的条件下把模型的成本压到极低，让全世界重新算了一遍训练的账，第 14 章会细讲。

这些成就成立，靠的是一个条件：题目已经出好了。天气怎么预报、模型怎么训练、什么叫好结果，这一整套判断标准是现成的。中国在现成的题里，做到了世界前列。

还有一层公道要还给短剧本身。短剧公司把 AI 用到了生产的每一个环节：写本子、画分镜、生成画面、配音、剪辑、投流。这是 AI 在中国最早、最大规

模地真正改了一个行业的流程。这件事本身，就是一种生产力。只是它改的，是生产注意力的流程。

停在哪里，后面再说。

◆ 九、带着走的三个问题

上篇结束，有三个问题，要带进后面几篇。

第一个问题：注意力的钱，以前回流过吗？回流到了哪里？这是中篇一要回看的。四十年里，计算机和互联网各走过一遍，那两遍的账，比今天的任何预测都实在。

第二个问题：就算钱回流了，它能把中国带到哪一层？是把已有的东西做得更好，还是出一道新题？这是中篇一后半和中篇二要拆的。第 17 章会把这件事分成三层：改进、架构、典范。

第三个问题：如果出新题的那一层要靠人，那个人从哪里来？这是下篇的事。

三个问题，一个比一个难。上篇只负责把它们从一张图里拣出来。

◆ 十、上篇没有说的几件事

账算完，还要把边界划清楚。上篇没有说下面几件事。

上篇没有说短剧不该做。短剧是一门正当的生意，养活了很多人，也让很多普通人在一天的疲惫之后有个消遣。

上篇也没有说中国的 AI 公司选错了路。在给定的条件里，它们做的是最合理的选择，而且做得很好。

上篇更没有说美国就是对的。美国的生产力在前，同样是条件的结果；它的模型被大量用在广告推荐和内容生成上，也是事实，只是那张流传的图里没有画。

上篇只说了一件事：钱押在哪里，本事就长在哪里。这件事放在哪个国家都成立，放在中国，就有了中国的形状。

◆ 十一、这件事，四十年前就发生过

上篇看的是今天。可是「钱先流向注意力，后来才流向生产力」，这个形状并不新。

四十年前，计算机进中国；二十多年前，互联网进中国。每一次，都有人问同一个问题：这东西最先用来干什么，挣来的钱后来去了哪里？

有意思的是，两次的答案不一样。计算机进来，最先火的是记账和排版，生产力在前；互联网进来，最先火的是聊天和网游，注意力在前。

这说明了一件事：先走哪一头，不是谁的天性，是当时的条件。条件变了，顺序就会变。

中篇一，就从一个县城的会计和他的第一块汉卡讲起。

◆ 十二、放回日子里

那天夜里回家的车上，林豆豆睡着了。宋岚开车，林舟坐在旁边刷手机，刷到的又是那张图。

他第一次没有转发。他在下面打了一行字，又删掉了，最后只留了一句：「先问这笔钱多久能回来。」

那天夜里，林舟还想起除夕时女儿问的那句话：火药后来还干了什么。他想，等她再大一点，可以和她一起去查一查。查的不只是火药的历史，还有一件更难的事：今天的这一把火药，有没有人在想它还能推什么。

一个小办法，留给你：下次有人拿一张图、一个百分比来说「中国就是如何如何」，你可以按上篇的顺序问三句：量的是不是同一样东西？比的是不是同一批人？扣掉这两样，剩下的那件真事是什么？

烟花照亮的，是放烟花的人把钱押在了哪里。

带 走 的 话

一张图里的百分比，先扣掉秤的差别和样本的差别，剩下的那一点，才是值得想的真事。

小　账

看见了什么，漏看了什么？看见了资本押注的方向；漏看了押注的方向是条件造的，条件会变。

是在改进，还是在出题？上篇所说的一切，几乎都在已有的题里做得更多更快；下一种火药，还没有人在做。

想 一 想

你最近一次转发的「中国如何如何」的图表，量的是什么，比的是谁？扣掉之后，你还会转吗？

◆



中　篇　一

「先做应用」：四十年的老路

导　语

这一篇八章，回看四十年。第 9 章到第 11 章看计算机和互联网进中国的那几十年：计算机进来，先火的是记账和排版；互联网进来，先火的是聊天和网游；后来电商、支付、物流、云一样一样长了起来。第 12 章到第 15 章拆一句人人都说过的话：「先做应用」。拆它里面那个「再」字，拆解题和出题的区别，拆撞墙之后改的到底是什么，拆改进为什么不是 0 到 1。第 16 章把应用的账排成一张表。

这一篇要让你看见两个反差。一个是：先走注意力还是先走生产力，不是中国人的天性，是条件；计算机那一回，顺序就是反过来的。另一个是：注意力挣来的钱确实回流了，可它只回到了应用这一层，很少流到操作系统、芯片、工业软件这些出题的地方。

09

第 9 章

汉卡与财务软件：计算机是一个反例

◆ 一、县城会计的算盘

1986 年前后的一个冬天，老林二十六岁，在一个县城的国营农机厂当会计。

每到月底，财务科的四张桌子并在一起，四把算盘一起响。工资表、材料账、成本账，全靠手抄，抄错一位，整张表要从头对。老林的手指上，常年有一块被算盘珠磨出来的硬皮。

那年厂里分来一台微型计算机，放在厂长办公室隔壁的小屋里，屋里装了窗帘，进门要换拖鞋。厂长说，这是给财务科用的。

老林第一次坐在它面前，屏幕上是一行一行的英文和数字。他一个字也不认识。

半年以后，厂里给这台机器插上了一块卡，屏幕上第一次出现了汉字。

这块卡，叫汉卡。

◆ 二、一块卡让机器认识汉字

先说材料。

计算机刚进中国的时候，遇到的第一道坎，是汉字。英文只有二十六个字母，一个键对一个字；汉字常用的就有几千个，怎么输入，怎么存，怎么显示，怎么打印，每一步都是难题。当时的微机内存很小，光是把一套汉字字库装进去，就能把机器塞满。

联想汉卡，是 1980 年代中期中国科学院计算所的一批人做出来的。它把汉字处理的一部分活，做到了一块插在主机里的卡上，让一台普通的微机能流畅地处理汉字。后来的联想集团，最早就是靠这块卡起的家。

同一时期，王选带着团队做汉字激光照排。报社和印刷厂从此告别了铅字排版，很多人后来回忆这件事，都用一个说法：「告别铅与火」。下一节把它走一遍。

1988 年，求伯君写出了 WPS。这是一个能在中文环境下打字、排版的文字处理软件，很快在机关和单位里流行开来。

同一年，用友创立；1993 年，金蝶创立。两家都从财务软件做起，把会计的账本搬进计算机。

◆ 三、一个完整的例子：报社告别铅字

把王选那件事，走一遍。

铅字排版的年代，一家报社的排字车间里，立着一排一排的字架。排字工人手拿稿子，从字架上一个一个捡出铅字，排进版里，再浇成铅版，上机印刷。一个熟练工一天能捡多少字，报纸就能出多少版。车间里热，铅有毒，工人的手常年发黑。

当时国外已经有几代照排机。据后来的记载，王选做了一个大胆的决定：跳过国外正在用的第二代、第三代，直接做当时还没有商品化的第四代激光照排。他要解决的最难一关，是汉字太多、字形太大，机器存不下。他用轮廓加参数的办法描述字形，把要存的信息量压到原来的很小一部分。

1980 年代后期，有报社开始用上这套系统。到了九十年代，全国的报社和印刷厂，大多换了过来。字架拆了，铅炉熄了。

这件事里，买单的是谁？是报社、出版社、印刷厂。它们买，是因为排版快了、印得好了、成本降了。每一笔钱，都花在干活上。

◆ 四、最先火的是记账和排版

把这几件事排在一起，你会看见一个形状。

计算机进中国，最先在什么地方站住了脚？在报社的排版车间，在机关的打字室，在工厂的财务科。它最先干的，是记账、打字、排版。这些全是干活的事。

网吧、电脑游戏，当然也有，可那是后来的事。九十年代中后期，网吧才在县城里多起来，那是第 10 章要讲的。

所以，在计算机这一回，顺序是生产力在前、娱乐在后。

这和上篇看到的那个形状，正好相反。上篇里，AI 的钱先流向短剧和广告；计算机这一回，钱先流向了财务科和印刷厂。

◆ 五、老林学 WPS

1990 年前后，老林被厂里派到地区的培训班，学了两个星期的计算机。

培训班教的，一是五笔字型，二是 WPS。老师让大家背字根，「王旁青头戈五一」，一屋子三四十岁的会计、文书、打字员，像小学生一样念口诀。

老林记得最清楚的，是第一次用 WPS 把一份工作总结打出来。以前这种东西要誊三遍，现在改一个字，按几下键就好，打印出来的字比手写的还整齐。

他回到厂里，成了财务科第一个会用电脑的人。

又过了几年，厂里装了用友的财务软件。月底结账，从四把算盘响三天，变成一个人在电脑前坐一个下午。老林手指上那块硬皮，慢慢也就消了。

这个故事很平常。那几年，全国成千上万的会计、打字员、编辑，都经历过差不多的事。

◆ 六、为什么那一回顺序反过来

为什么计算机那一回，生产力在前？

不是因为八十年代的人更勤劳、更有远见。看条件。

第一个条件，谁在买机器。八十年代的计算机很贵，一台微机的价钱，抵得上一个普通职工好多年的工资。私人买不起，买它的是单位：报社、机关、工厂、银行。单位买它，是为了干活。

第二个条件，有没有一个能直接变现的大众市场。那时候没有互联网，没有网上支付，机器和机器之间不连着。一个人在电脑上玩游戏，没有办法让别人为此付钱给谁。注意力即使有，也没有一条路把它换成钱。

第三个条件，单位愿不愿意为软件付钱。当时的机关和国营企业，有采购计算机的专项安排，汉字系统、财务软件跟着机器一起进来。很多单位买的是一整套，盒子里有硬件，也有软件。

还有一个条件，是会计电算化这件事本身被推着走。1980 年代后期到 1990 年代，财政部门推动会计电算化，给财务软件做评审。用友、金蝶早期的成长，和这条路分不开。

四个条件放在一起，钱就只能往生产力那边流。

◆ 七、本书的说法

先把结论摆在桌上：

先走注意力还是先走生产力，不是一个民族的天性，是当时的条件。

计算机那一回，条件是：买家是单位，没有能直接变现的大众市场，单位愿意为干活的软件付钱。于是生产力在前。

互联网那一回，条件反过来：上网的是个人，网络和支付让 C 端可以直接变现，B 端企业不太愿意为软件付钱，人力又相对便宜。于是注意力在前。

AI 这一回，条件和互联网那一回很像。第 6 章那张回款速度表，就是这几个条件的今天版。

这个反例有用。它告诉你：如果条件变了，顺序是可以变的。上篇讲的「先烟花」，不是写在中国人骨头里的东西。

◆ 八、只看一面会怎样

只看计算机这一回，会得出一个太乐观的结论：你看，中国人本来就重实干，先做生产力。这漏看了，那一回生产力在前，有一大半是因为没有别的路可走。

只看互联网那一回，会得出一个太悲观的结论：中国人就是爱娱乐、短视。这漏看了计算机那一回的反例，也漏看了后来淘宝、支付宝、阿里云这些生产力的东西，正是从注意力的钱里长出来的。第 11 章会细讲。

两个结论都把条件的结果，说成了人的本性。

把两回放在一起，才看得清条件这一层。

还要说一句：这个反例里，中国的工程师并不被动。汉字进计算机，是一道外国人没有、也不必解的题。中国的工程师接下这道题，在很短的时间里给出了好几种解法：字库、输入法、照排、文字处理。这份本事，正是后来几十年中国在应用层一路领先的源头。

◆ 九、反例里也有一样东西没变

不过，计算机这一回也不全是好消息。

看那几件成就：汉卡、激光照排、WPS、财务软件。它们了不起。可它们是在什么样的题目里做到的？

计算机本身，不是中国造的；操作系统，后来是 DOS 和 Windows；芯片，是别人的。汉卡和 WPS，做的是让别人造的机器认识汉字、处理中文；财务软件，做的是把中国的账，记进别人的机器里。

王选是一个要单独说的人。他的照排系统，在汉字字形压缩和输出上，有自己的原创方法，在当时走在了前面。可就算是他，面对的题目也是清楚的：怎么让汉字进入计算机印刷。题是现成的，他用一种别人没想到的方法，把它解得又好又省。

所以计算机这个反例，只反了一半。它反的是顺序：生产力确实在前。它没有反的是层级：中国做得最好的，仍是把别人的机器用到中国的场景里。

这就是「先做应用」最早的样子。第 12 章要拆的那个「再」字，在这里第一次出现。

◆ 十、反例的边界：盗版与 Word

这个反例还有一条边界，要说清楚。

计算机那一回，生产力在前，靠的是单位肯为软件付钱。可这个条件，到九十年代中后期就松了。

个人电脑便宜下来，进了更多的办公室和家庭。随它一起进来的，还有大量盗版软件。一张光盘，装着好几套正版要卖很贵的软件，花很少的钱就能买到。

WPS 在那几年吃了大亏。九十年代中期，微软的 Word 进入中国，借着 Windows 的普及很快占了上风；金山另起炉灶做的新一代办公软件，据说投入很大，市场反响不好。WPS 后来能活下来，靠的是很多年的坚持和后来的免费策略。

这段往事说明了一件事：条件一变，顺序就会跟着变。

单位肯付钱的时候，软件公司做生产力；盗版一来，个人和企业都不再为软件付钱，做软件挣不到钱，钱和人就往别处走。几年以后，网吧和网游起来了，那是下一章的事。

所以计算机这个反例，不只是告诉你「顺序可以反过来」，还告诉你「它是怎样又被反回去的」。

◆ 十一、今天与机器

今天的 AI，和那时候的计算机，有一处很像。

都是一样从外面来的新机器，最先的难题是让它适应中文、适应中国的场景。八十年代是汉字输入和汉字字库，今天是中文语料、中文理解、中文的行业知识。

那时候，中国的工程师做汉卡，做得又快又好；今天，中国的工程师把大模型做到中文能力强、价格低、部署快，也做得又快又好。

那时候缺的，是操作系统和芯片；今天缺的，第 17 章会说，是架构和典范。

还有一处不一样，值得留心。八十年代，单位为计算机付钱，是因为它替人干活。今天，单位为大模型付钱的意愿，反倒不如八十年代那么强。

原因之一是人力。八十年代一台计算机替下的是好几个会计的工时；今天的企业算账时，常常觉得多招一个年轻人更省事。

一台机器进来，你怎么用它，取决于条件；你能不能造下一台，取决于另一套东西。

◆ 十二、放回日子里

去年春节，林舟给老林装了一个手机上的大模型应用。老林让它帮着算一笔退休工资的账，它算得飞快，一条一条讲得清清楚楚。

老林看了很久，说：「当年我学五笔，背了一个月的字根。」

停了一下，他又说：「可那台电脑，从来不是我们造的。」

一个小办法，留给你：问一问家里的长辈，他们第一次用电脑是在什么地方、干什么用的。你会发现，很多五六十岁的人，第一次碰电脑是在单位，是为了干活；而很多二三十岁的人，第一次碰电脑是在网吧，是为了玩。这两个回答之间，就是条件的变化。

> **顺序跟着条件走，层级却四十年没动。**

带 走 的 话

计算机进中国先做记账和排版，说明先走注意力不是天性；可它做得最好的，仍是把别人的机器用到中文里。

小　账

看见了什么，漏看了什么？看见了顺序是条件造的；漏看了顺序变了，层级可能一点没变。

是在改进，还是在出题？汉卡、WPS、财务软件，是在别人造好的机器上，把中文的题解得又好又快。

想 一 想

你第一次碰电脑是在什么地方、为了什么？你孩子第一次碰 AI，又是为了什么？

◆

10

第 10 章

网吧和网游的第一桶金

◆ 一、老林辞职那年

1998 年，老林三十八岁，厂里效益越来越差，工资时有时无。

那年秋天，他做了一件全家都反对的事：从厂里出来，把攒下的钱和找亲戚借的钱凑在一起，在县城中学对面租了两间门面，开了一家网吧。

门面里摆了二十台电脑。电话线接进来，拨号上网，速度慢得要命。一个小时收几块钱。

开张头一个月，来的人不多，大多是来打单机游戏的中学生。老林每天晚上守到十二点，算一算当天的流水，心里没底。

第二年春天，事情变了。来的人越来越多，他们不玩单机游戏了。他们在屏幕上挂着一个企鹅头像，一个晚上和天南海北的陌生人聊天。

◆ 二、一只企鹅把人聚起来

先说材料。

1998 年底到 1999 年，腾讯推出了 OICQ，后来改名 QQ。它做的事情很简单：让两个人在网上即时说话。

这件事看起来小，可是它把人聚了起来。一个县城的中学生，第一次可以和远方的人说话；一个刚进城的打工者，可以在网吧里和老家的同学联系。在网吧里，一个账号就是一个人在网上的身份。

QQ 一开始是免费的。腾讯很长一段时间都在为「怎么挣钱」发愁，据说一度想把公司卖掉，没卖成。

后来腾讯找到的路，是增值服务：会员、QQ 秀、虚拟形象、靓号、游戏。用户为了让自己的头像好看一点、号码好记一点，愿意付一点钱。一个人付得不多，可是人多。

这条路，靠的是注意力：用户在 QQ 上花的时间越多，就越愿意为在那里的「样子」付钱。

◆ 三、《传奇》和网吧的通宵

再说网游。

2001 年，盛大代理了一款韩国网络游戏《传奇》。这款游戏在中国火到什么程度？很多网吧通宵爆满，有人肯为游戏里的一件装备，在现实里花上不少钱。

老林的网吧，也是在那几年真正赚了钱。晚上十点以后的通宵场，几块钱包一夜，二十台机器坐得满满的，一大半人在打《传奇》。他后来又加了二十台机器。

盛大靠《传奇》，几年间成了一家大公司，陈天桥一度成了中国首富。

网易也有一段相近的故事。2000 年代初，互联网泡沫破裂，网易一度处境艰难，股价跌得很低。后来网易靠自己做的网络游戏翻了身，《大话西游》《梦幻西游》都是那几年的产品。

腾讯、盛大、网易，这几家公司那几年最重要的钱，都是从注意力里挣来的：聊天、虚拟形象、网游。

◆ 四、为什么这一回注意力在前

和上一章对照，这一回的条件全变了。

第一，买家变了。九十年代末，上网的大多是个人，是网吧里的年轻人。他们不是为了干活上网，是为了聊天、为了玩。

第二，有了一条把注意力换成钱的路。网络把人连在一起，游戏点卡、QQ 会员、短信增值服务，让一个人可以为几块钱的东西付钱。几千万人一起付，就是一大笔钱。

第三，企业那一头不肯付钱，也付不起太多钱。九十年代末到 2000 年代初，国内企业为软件付费的习惯很弱，盗版软件到处都是。做企业软件，要上门、要定制、要催款，钱来得又慢又难。

第四，人力相对便宜。一个企业要提高效率，加几个人常常比买一套系统便宜。

四个条件放在一起，第 6 章那张回款速度表，在二十多年前就已经排好了：C 端的注意力，按天回款；B 端的生产力，按年回款，还常常收不回来。

◆ 五、网吧里的一个晚上

2002 年冬天的一个晚上，老林照例守着收银台。

十一点，通宵场开始。四十台机器，三十八台亮着。一大半屏幕上是《传奇》的画面，剩下的挂着 QQ，有人在看一部下载好的电影。烟味很重，键盘噼里啪啦地响。

角落里有一个戴眼镜的高中生，没有打游戏。他在一个论坛上，和人讨论怎么写一个网页。他问了老林好几次，能不能把那台机器的浏览器换个新版本。

老林后来常想起这个孩子。四十个人里，三十九个在花时间，一个在学本事。网吧的流水，是那三十九个人付的；那一个人学到的东西，也是那三十九个人付的电费和网费养着的。

这就是那几年中国互联网的一个小样本：绝大部分钱，来自娱乐；一小部分本事，在娱乐的屋檐底下悄悄长出来。

◆ 六、钱确实先从注意力挣到

这里要说一句老实话：中国互联网的第一桶金，确实是从注意力里挣到的。

这不是坏话。那是一个很实在的结果。

一个县城的中学生，花几块钱在网吧坐一个晚上；一个年轻人，花十块钱买一张点卡；一个上班族，每个月为 QQ 会员付一点钱。这些小钱，汇成了中国最早一批互联网公司的命根子。

老林的网吧，也是这样活下来的。他后来说，那几年是他一辈子挣钱最轻松的几年：不用催款，不用看谁的脸色，钱每天晚上都在抽屉里。

你看，这和第 6 章宋岚公司墙上那张表，几乎一模一样：钱按天回来，不必等。

◆ 七、公道：这些钱后来去了哪里

说到这里，要把公道还足。

这些从注意力里挣来的钱，没有白白散掉。

第一，它们养出了一支工程师队伍。为了让几千万人同时在线聊天，腾讯的工程师要解决服务器、带宽、稳定性这些硬问题；为了让几十万人同时在一个游

戏世界里打怪，盛大和网易的工程师要搭大规模的服务器集群，要防外挂、防盗号。这些本事，在当时的国内，很少有别的地方能练到。

第二，它们养出了一批基础设施。机房、带宽、在线支付的早期形态、大规模用户账号的管理办法，都是在这几年里先被注意力生意拉起来的。

第三，它们养出了一种组织能力。怎样做一个几千万人用的产品，怎样快速改版、快速试错，怎样运营一个巨大的社群。后来的电商、支付、外卖、打车，很多带头的人，都是从这几家公司出来的。

这就是第 7 章说的「先养」。注意力挣的钱，确实养大了一些东西。这几样东西，后来都成了中国 AI 的家底。今天国内几家大模型的训练集群、推理服务、海量用户的运营，很多带头的工程师，是在这些公司里练出来的。没有二十多年前的网游服务器和聊天软件，就没有今天日均 140 万亿 Token 的调用规模。

◆ 八、本书的说法

先把结论摆在桌上：

中国互联网的第一桶金从注意力里挣到，它确实回流了，可它流进的是应用这一层。

QQ 挣的钱，养出了微信；网游挣的钱，养出了大规模服务器的本事；网吧和点卡练出来的小额支付，后来接上了淘宝和支付宝。这些都是真回流。

可是你看它们流进了哪里：更好的产品，更大的平台，更强的运营。它们没有流进操作系统、浏览器内核、数据库、芯片。那些东西，当时用的是别人的，后来很多年也还是别人的。

这一句里，是「先养」有了，「后用」也有了，可「用」停在了应用这一层。第 11 章会把这条回流的路走完一遍。

还要看一眼回流的速度。从 1999 年的 QQ 到 2011 年的微信，中间隔了十来年；从网游的服务器到云，也差不多。这十年里，挣钱的主业一直在跑，回流的那一小股，是在主业的屋檐下一点一点长起来的。它不是一个计划，是一串顺手。

顺手能走到的地方，就是离主业最近的地方。离主业最远的那一层，没有人顺手去做。

◆ 九、只看一面会怎样

只看「第一桶金来自网游」，会说：中国互联网是靠游戏和娱乐起家的，不务正业。这不公道。那些钱后来养出的，是全世界数一数二的电商、支付和物流网络。

只看「后来都回流了」，会说：你看，注意力最后都变成了生产力，今天的 AI 短剧也一样，放心。这又漏看了一层：回流到应用层，和回流到出题的那一层，是两件事。

还有一种看法要拆一拆：说那几年的网游是「害人」的，网吧是「毒瘤」。当年确实有未成年人沉迷、有人通宵出事，这些都是真的，后来的管理也是必要的。可这和本章要算的账是两本账。本章只算钱从哪里来、流到哪里去。

◆ 十、今天与机器

把今天的 AI 和二十多年前的互联网放在一起，像得惊人。

那时候，QQ 把人聚起来；今天，短剧平台把第 5 章说过的八亿人聚起来。

那时候，网游让人为虚拟装备付钱；今天，短剧让人为下一集付钱。

那时候，企业软件不好卖，人力便宜；今天，企业为生产力工具付费的意愿仍然弱，人力仍然相对便宜。

那时候，注意力挣的钱回流到了应用层；今天，短剧挣的钱会流到哪里，还要看第 7 章说的那五个信号。

有一处不一样：二十多年前，中国离世界前沿很远，先在应用层追上来，是合理的路。今天，在 AI 的很多应用上，中国已经站在前沿。追的路走完了，下一步该往哪里走，第 12 章会问。

还有一点，和今天的 AI 很像：那几年的注意力生意，几乎没有门槛。开一家网吧，要的只是几十台电脑和一根电话线；代理一款游戏，要的只是和韩国的公司谈下一份合同。门槛低，进来的人就多；进来的人多，比的就是谁更快、更便宜、更会运营。今天每天上新一千五百部 AI 短剧，也是同一个样子。

◆ 十一、一个完整的例子

把老林的网吧，从头到尾走一遍。

1998 年，二十台机器，拨号上网，来的人打单机游戏。第二年，QQ 来了，人多了起来。2001 年以后，《传奇》来了，通宵场满了，机器加到四十台。

老林用这些钱，供林舟读完了高中，又读了大学。林舟后来考上研究生，读博士，进了研究院，做 AI 制药。

你可以说，林舟今天在研究院里用的那套本事，有一小块，是网吧里那些通宵打《传奇》的年轻人付的学费。

这不是比喻，是一个家庭里的真实流向：注意力挣来的钱，变成了下一代的教育。

可是林舟进了研究院之后，做的是什么？他用的 AlphaFold 是别人造的，他用的大模型的架构是别人先出的，他写的基金本子，按的也是一套给定好了的评价标准。第 20 章会细讲这件事。

钱回流了，人也回流了。回到的，仍然是在别人出好的题里，做得更好的那一层。

◆ 十二、宋岚和老林的一次对话

去年中秋，宋岚陪老林在阳台上喝茶，讲起公司里的事：每天上新多少部剧，抽多少次卡，一个月烧多少 Token。

老林听完，笑了。

「跟我那个网吧一样。」他说，「人坐在那里，时间就是钱。」

宋岚问他：「那你后悔开网吧吗？」

老林想了想，说不后悔。那几年要是不开网吧，林舟的书都念不下去。可他又补了一句：「那时候我总觉得，开几年网吧，攒点钱，以后做点别的。后来家家都有电脑了，网吧关了，那个『别的』也没做成。」

那个「别的」，就是第 12 章要拆的那个「再」字。

◆ 十三、放回日子里

2008 年前后，老林把网吧关了。家家都有了电脑和宽带，网吧的生意一年不如一年。

关门那天，他把四十台旧电脑一台一台搬上车，卖给了收旧货的人。林舟那时候在读大学，放假回家帮忙搬。

老林说了一句：「这些机器，养活了我们家十年。」

林舟没说话。他那时候还不知道，自己后来的一辈子，都要在另一种机器上，替别人出好的题找答案。

一个小办法，留给你：想一想你今天用得最多的那个手机应用，它最早靠什么挣到第一笔钱？那笔钱后来又去了哪里？顺着这条线往下摸，你会摸到一张回流的地图。

注意力的钱会回流，可它认得的路，只通到应用这一层。

带 走 的 话

网吧和网游的小钱是真钱，养大了一代公司和工程师；可回流靠的是顺手，顺手只到得了离主业最近的地方。

小　账

看见了什么，漏看了什么？看见了注意力的钱确实回流；漏看了回流到哪一层，比回流本身更要紧。

是在改进，还是在出题？QQ、网游、增值服务，都是在别人定义好的网络和机器上，把产品做得更好、更会挣钱。

想 一 想

你家里有没有一笔钱，是从别人的娱乐里挣来的，后来供了谁读书？那个人今天做的事，是在解题，还是在出题？

◆

11

第 11 章

从电商到支付：钱回流到了哪一层

◆ 一、老林手机里的那几个图标

老林六十六岁，搬到上海和儿子住，已经第三年了。

那天傍晚，他下楼买了一个煎饼，三块五。他掏出手机，对着摊位上那张塑料二维码扫了一下，「嘀」的一声，钱就过去了。摊主连头都没抬。

回到家，他把手机递给林舟：「你帮我看看，这个月我到底花了多少。」

林舟一滑屏幕，愣了一下。老人的手机里，排着支付宝、淘宝、美团、滴滴、一个快递查询，还有一个看病挂号的小程序。买菜、打车、寄东西、交水电费、给孙女买练习册，全在这一块玻璃里。

「爸，你这个手机里，装着半座城。」

老林笑了。他想起九十年代末开网吧的那几年。那时候收钱是一张一张的纸币，晚上关门要数硬币，数到手指发黑；进货要坐长途车去省城，网线坏了要等三天。二十多年过去，钱、货、人、车，都被这几个图标接住了。

这一章要问的，就藏在这几个图标后面：网吧、网游、电商挣来的那些钱，后来流到了哪里？

◆ 二、2003 年的淘宝，2004 年的支付宝

第 10 章说过，网吧和网游给中国互联网挣下了第一桶金。这一章接着往下走。

2003 年，淘宝上线。2004 年，支付宝出现。这一步，常被人轻轻带过，其实它解决的是一件很硬的事：两个从没见过面的人，隔着几千里，怎么放心地做一笔买卖？

支付宝最早的办法很朴素，叫「担保交易」。买家先把钱交给平台，平台替他拿着；货到了，买家点头，钱才转给卖家。陌生人之间那道不信任的墙，就这样被一个中间人搭了座桥。

桥搭起来以后，后面的东西一样一样长出来。快递网络从几个城市铺到县城和村口；电商平台给卖家打分、给买家记信用；手机普及以后，扫码付款从商场走到了菜市场和煎饼摊。再往后，外卖和网约车把城市里的人、车、饭菜也排进了同一张调度表。

2009 年，阿里云开始做。一家卖货的公司，自己去造算力和存储，这在当时很多人看不懂。据报道，它在内部被质疑了好几年。后来，「云」成了中国互联网的一块地基，也成了人们说「产业互联网」时最常举的例子。

◆ 三、回流确实发生了

上篇留下一个问题：注意力挣来的钱，会不会回流到生产力？

说公道话，在这一段历史里，回流是真的发生了。

门户和网游的钱，养大了第一批互联网公司；电商的钱，养出了支付和物流；支付和物流的钱，又养出了云。一条路，一站接一站，没有断。

这条路上长出来的东西，放在全世界看，也是少见的。一个县城的小摊贩，用一张打印的二维码收款。一个村里的老人，在手机上给城里的孙子寄一箱橘子。一个外卖骑手，靠一部手机串起一天几十单。这样一整套东西，很多发达国家到今天也没有铺得这么密、这么便宜。

它们早已不是「应用」两个字能装下的了。它们是城市的水管和电线，一天不通，整座城都会觉得别扭。

所以，说中国的互联网「只会做注意力」，是不公道的。注意力的钱，确实回到了实打实的东西上。

◆ 四、钱停在了哪一层

可是，把这条路再往下看一层，就会看见一件事。

钱回流到的，是同一种东西：能被规模和流量喂大的东西。

支付是这样，用的人越多，每一笔的成本越低，于是更多的人来用。物流是这样，单量越大，网点越密，送得越快，单量就更大。云也是这样，客户越多，机器摊得越薄，价格越低，客户就更多。外卖、网约车，全是这个样子。

这一类东西，有一个共同的脾气：流量是它的饭。喂得越多，它长得越快。

再往下一层，是另一类东西。电脑和手机里的操作系统。芯片，以及芯片下面的指令集和制造工艺。设计芯片要用的那套软件，行里叫 EDA。造飞机、造汽车、造机床要用的设计和仿真软件，统称工业软件。

这一层，中国四十年里很少有钱回流过去。手机里的系统，长期是别人的；芯片设计软件，长期由几家美国和欧洲公司把着；工业软件，大头也在海外。近几年开始补，补得很辛苦，而且多半是在被卡住之后才补的。第 14 章会回到这一点。

◆ 五、底层为什么喂不大

这里要问一句：钱那么多，为什么偏偏不往底层去？

不是因为没人看见底层重要。是因为底层不吃流量这口饭。

拿 EDA 来说。全世界用它的人，是芯片设计工程师，人数有限。多来一万个用户，它不会因此变好；它要变好，靠的是几十年里一点一点攒下来的算法、工艺数据和踩过的坑。操作系统也类似：用户多当然要紧，可一个系统的内核和生态，是十几年、几十年慢慢长出来的，不是一个季度的推广能堆出来的。

所以底层要的，是另外三样东西：长周期，少数人，还有愿意为不确定性买单的资本。

长周期，是说十年不一定见分晓。少数人，是说真正懂行、能把一个内核想透的人，一个国家也就那么一小撮。愿意为不确定性买单，是说投进去的钱，很可能一分都收不回来，而出钱的人事先就认了。

这三样，恰好都是流量的反面。流量要快，底层要慢；流量要多，底层要少；流量要确定的回报，底层只给不确定。

于是，上篇那条规律要修正一下。它不是「注意力会回流到生产力」那么简单。修正以后，可以写成一句话：

注意力→应用层基础设施，但很少→底层原创。

◆ 六、资本也有一只钟

还有一样东西，把水往坡底推：钱自己也有一只钟。

一笔风险投资，通常要在七到十年里退出。投资人拿的是别人的钱，到期要还本、要分红。所以他看一个项目，第一眼看的往往不是它对不对，是它在钟走完之前能不能长大。

能被流量喂大的东西，正好跟这只钟合拍。用户一个季度翻一倍，估值一年翻几倍，几轮融资之后上市或者被收购，钟还没走完，账就结了。

底层原创和这只钟不合拍。一个操作系统内核，十年后也许刚刚好用；一款工业仿真软件，要先在几家工厂里磨上好几年，才敢说自己可靠。钟走完的时候，它可能还在半山腰。

这不怪投资人。他的钟是别人给他上的弦。可是一整个国家的钱，如果几乎都揣着同一种钟，那么钟走不完的事，就很少有人去做。

第 9 章说过，八十年代生产力走在前面，原因之一正是买单的人不一样：那时掏钱的是要把账算对、把字排好的单位。买单的人一换，水流的方向就跟着换。

可是这一回，揣着同一种钟的，不只是风投。银行的贷款按年还息，上市公司按季度交财报，地方的项目按任期验收。钟的样子各不相同，走完一圈的时间却都差不多，都比一个内核长成的时间短。

顺序是条件定的，不是天性。第 16 章算总账时，还会用到这句话。

◆ 七、补底层的那几年

公道还要再还一笔：这几年，钱确实开始往坡那一边走了。

国产操作系统在电脑和手机上都有了自己的版本，有的已经装进了大批手机和电脑。国产芯片设计软件、工业软件，也有一批公司在做，有的在某几个环节里已经能用。国家设了专门的产业基金，地方政府也拿出真金白银。

这些都是真的，而且做得很辛苦。

可是请看它们开始的时间。大多数是在 2018 年以后，在一轮一轮被卡之后。不是应用正赚钱时主动翻过去的，是坡底的湖被人截了水，才不得不去修水泵。

被逼出来的修，和主动的修，修出来的东西不一样。被逼的时候，目标很清楚：把别人有的那一样，先做一个自己的。这仍是在一道别人出好的题里作答。第 13 章会把这件事说透。

◆ 八、只看一面会怎样

这条规律，很容易被看成一面。

只看前半句的人会说：你看，回流成功了，从网游到云，一站一站都到了；再等一等，钱自然会流到芯片和系统上去。

只看后半句的人会说：说到底全是虚的，几十年忙下来，连一个像样的系统都没有。

两种说法都只看了一面。

打个比方。钱像水，水往低处流。流量造出的那片地，是一道顺坡：哪里用户多、回款快，水就往哪里淌。应用层的基础设施，正好在坡底，所以水全汇过去了，汇成了湖。底层原创，在坡的另一边，是一道上坡。水再多，也不会自己往上坡走。

所以，「再等等就会到」是一厢情愿。水不会因为湖满了就翻山。要让水上坡，得有人修一台水泵，还得有人肯一直交电费。

反过来，「全是虚的」也不对。湖是真的，湖边的庄稼也是真的。一个国家能修出这样一片湖，本身就是本事。

◆ 九、今天与机器

到了 AI 这一波，这条坡又出现了一次。

电商、支付、云和短剧挣来的钱，这几年大量流进了大模型：建机房，买显卡，训模型，降价抢用户。国家数据局局长刘烈宏 2026 年 3 月在中国发展高层论坛上说，中国日均 Token 消耗已经超过 140 万亿，比 2024 年初涨了一千多倍。这是一片新的湖，而且涨得极快。

请注意这片湖坐落在哪里。

模型跑在显卡上，显卡的架构和配套的编程平台，长期是别人的。模型本身的骨架，是 2017 年那篇论文提出的 Transformer，第 18 章会细讲它的来历。中国公司在这片地上做得极好：模型便宜，开放，调用量大。

你仔细看，钱又一次停在了「能被规模和流量喂大」的那一层。大模型的推理服务，像当年的云一样，用得越多越便宜，越便宜用得越多。它是 AI 时代的应用层基础设施。

而更下面的那几层——芯片架构、编程平台、模型骨架本身——还是那道上坡。

◆ 十、一笔扫码支付走过的楼层

把老林那三块五的煎饼钱，从头到尾走一遍，就看得更清楚。

老林按下手机屏幕上那个蓝色的图标。这是第一层，应用，中国造，而且全世界做得最好的之一。

钱从他的账户扣出去，经过清算，记进摊主的账上。这是第二层，支付网络，中国造，规模是世界级的。

这笔交易被记录、核对、备份，存在一座远处的机房里。这是第三层，云，中国造，一部分核心系统是自己写的。

再往下走。那座机房里的服务器，跑的操作系统，内核多半来自开源的 Linux，那是芬兰人林纳斯在 1991 年开的头。服务器里的芯片，指令集来自美国或英国的公司。设计这些芯片的软件，出自几家欧美公司。造这些芯片的光刻机，最先进的那几台，来自荷兰。

老林的手机也一样。屏幕上的支付图标是中国的，图标下面那套系统，是别人定的。

一笔三块五的钱，走过六七层楼。上面几层，灯火通明，写满了中文；下面几层，图纸是别人画的。

这不丢人。全世界没有哪一个国家，一个人盖得起整栋楼。可是你要知道，自己住在哪几层。

◆ 十一、放回日子里

那天晚上，林舟把手机还给老林，顺口把那几层楼讲了一遍。

老林听完，沉默了一会儿，说：「我开网吧那会儿，就琢磨过这个。电脑是人家的，系统是人家的，游戏一大半也是人家的，我挣的是电费和网费的差价。」

他停了一下：「没想到二十多年了，楼是盖高了，地下室还是那几间。」

这句话不用加什么评语。

明天能试一件事：打开你的手机，挑一个你每天都用的图标，往下数三层——它跑在什么系统上，系统跑在什么芯片上，芯片是用什么软件设计的。数到哪一层，你开始答不上来，或者答案开始变成外国公司的名字，那里就是你脚下这座楼的分界线。

再问一句：你自己的工作，在第几层？

钱会回流，可是它顺着流量的坡走；坡底是湖，坡那边的山，要有人修水泵。

带 走 的 话

注意力的钱确实回流了，回到了能被流量喂大的那一层；底层原创不吃流量，水不会自己翻山。

小　账

看见了什么，漏看了什么？看见了回流真实发生、长成了城市的水管电线；漏看了它只沿着流量的坡流，从没翻到底层那一边。

是在改进，还是在出题？支付、物流、云，是在别人定好的楼板上把上面几层盖得最好；地下室的图纸，还是别人出的题。

想 一 想

你手机里每天用得最多的三个东西，往下数三层，哪一层开始不是中国人画的图？你自己的工作，吃的是流量这口饭，还是要靠长周期、少数人和肯赔钱的人？

◆

12

第 12 章

那个「再」字

◆ 一、你一定有过这样的时刻

这样的话，你大概说过：「先把这个月忙完，再好好休息。」

或者：「先挣点钱，再去学我真正想学的东西。」「先把孩子带大，再说自己的事。」「先把房贷还上，再考虑换工作。」

说的时候，你是真心的。那个「再」后面，排着一件你在乎的事。

过了一个月，又有下一个月要忙。钱挣到一点，又有下一笔要挣。孩子大了，父母又老了。那件事一直排在「再」后面，一年，两年，十年。

有一天你忽然发现，「先」已经变成了「一直」。那个「再」，从来没进过你的日程本。

这一章，讲的就是这个字。不是你一个人的「再」，是一个国家四十年的「再」。

◆ 二、一句人人都会说的话

问一个中国的工程师、企业家、科研管理者：我们为什么不先做原创？

十个人里，大概有八个会给出差不多的回答：「先做应用。」

这句话后面，通常还跟着几句：我们起步晚，底子薄，钱要先挣回来；应用离市场近，能养活队伍；先把应用做起来，有了钱、有了人、有了数据，再回头做底层。

这番话，四十年来几乎没变过。八十年代说，九十年代说，互联网来了说，移动互联网来了说，大模型来了还在说。

请注意这句话里的那个字。

「先做应用」，三个字说出来是一个完整的句子。不过，「先」这个字，天生不是一个人站着的。有「先」，就该有「后」；说「先」，就是在许一个诺：后面还有一件事，会排上来。

这件事，就是「再做原创」。

◆ 三、先说它哪里对

说这句话错之前，先把公道还给它。

「先做应用」不是偷懒，也不是短视。在 1980 年代的中国，它几乎是唯一走得通的路。

第 9 章说过，那时候一台电脑抵得上一个工人好多年的工资。科研经费少得可怜，懂计算机的人屈指可数。要从零开始造一个操作系统、一颗芯片，钱从哪里来？人从哪里来？就算造出来，谁来用？

所以先做应用，是在一片很薄的土里，先种能当年收成的庄稼。收成换成钱，钱换成人，人换成更厚的土。第 9 章说过的汉卡、WPS 和财务软件，第 10、11 章说过的网游、电商、支付和云，都是这样一季一季种出来的。

这条路的成绩，摆在那里。一个四十年前几乎没有计算机产业的国家，今天有全世界最密的移动支付、最快的物流，有规模最大的短剧和外卖市场之一。它还有在 OpenRouter 上调用量领先的开放权重模型。

所以「先做应用」作为一个「先」，是对的。它完全成立，只要一个条件：后面那个「再」真的会来。

◆ 四、「再」在哪里

把四十年排成一张表，看看那个「再」每一次被排到了哪里。

那一轮浪潮	「先」做成了什么	当时说要「再」做的	后来
1980—1990 年代，个人电脑	汉卡、中文排版、WPS、财务软件	自己的操作系统、CPU	应用转向 Windows 平台，系统与芯片长期依赖外方
1990 年代末—2000 年代，互联网	门户、即时通信、网游	自己的基础软件、数据库	先赚流量钱，数据库多年后才在被卡中补
2000—2010 年代，电商与移动	电商、支付、物流、云	手机操作系统、芯片设计工具	应用登顶，系统与 EDA 依旧在外
2010 年代，深度学习	人脸识别、推荐算法、安防	框架、芯片、基础算法	框架多用外方开源，芯片在被卡后加速补
2020 年代，大模型	便宜、开放、调用量大的模型；AI 短剧	新架构、新的科学典范	正在发生

这张表的最后一列，说得很宽。每一行底下，都有具体的人、具体的项目在底层做过努力，有的还做得很漂亮。本书不是说「一点都没做」。本书说的是：作为一个国家的激励结构，「再」从来没有被正式排上日程——没有一轮浪潮，是在应用还赚钱的时候，主动停下来把底层做完的。

补底层的时刻，几乎都在被卡住之后。不是「再」来了，是墙来了。

于是「先做应用」悄悄变了一个字：变成了「一直做应用」。

◆ 五、应用越成功，越没理由停

为什么「再」排不上?

最容易想到的答案是：钱不够，人不够。可是到了 2010 年代，钱和人都已经很多了。这个答案解释得了八十年代，解释不了后来这十几年。

更准确的答案，反过来：**不是应用失败了才没轮到底层，是应用越成功，越没有理由停下来**。

把一个做应用的团队放在桌上看。它有现金流，每个月都有回款。它有KPI，日活、转化率、营收，每一项都能画成一条往上走的线。它有掌声，媒体报道、投资人追捧、领导视察时站在大屏前听汇报。

再把一个做底层的团队放在旁边。它有什么？有一张五年、十年的路线图，路线图上每一站都可能走不通。它没有回款，只有花钱。它的 KPI 很难写，因为「我们离一个好用的内核更近了一点」画不成一条线。它也很少有掌声，因为做了五年，外人看不出它和五年前有什么两样。

现在让一个管着钱的人，在这两个团队之间分预算。

他不是坏人，他也很清楚底层重要。可是他要对上面交代，要对股东交代，要对明年的考核交代。一边是能写进报告的数字，一边是一片看不清的雾。他每一次都会把多一点的钱，分给前一个。

一次这样分，没什么。四十年每一次都这样分，「再」就永远在下一年。

应用的成功，本身就是推迟「再」的力量。它越成功，越能证明「先做应用」是对的，也就越证明，现在还不到停下来的时候。

◆ 六、只看一面会怎样

这里有两种常见的误读，都只看了一面。

第一种误读：怪人。说企业家短视，说科研人员不肯坐冷板凳，说管理者只看政绩。这种说法让人解气，却解释不了一件事：为什么换了一批又一批人，结果几乎一样？四十年里，聪明、有抱负、真心想做底层的人，中国一点也不少。人换了，「再」还是没来。那就不是人的问题，是人站着的那片土。

第二种误读：怪钱。说只要国家多投钱，「再」自然就来了。钱投进一套只奖励应用的考核里，会被这套考核重新塑形。一笔写着「底层原创」的经费，到了执行的时候，要按年交成果，按期结题，按论文和专利数验收。到最后，它也被做成了一个「应用」：一个能按时交差的项目。

钱和人都是被土推着走的。土不变，「再」不会来。

◆ 七、林舟的基金本子

把这件事放到一个具体的人身上。

林舟三十八岁，在一家研究院做 AI 制药，带一个十二人的组。那个周四晚上，他在书房里写一份基金申请书，写到十一点半。

申请指南第一页，有一行字他看过很多遍：「面向重大应用需求」。往下翻，评审要点里写着：「成果转化前景」「与产业结合程度」。考核那一栏，列着论文、专利、合作企业的数量，还有「是否形成可转化的候选分子」。

林舟心里其实有一个想法，已经揣了两年。

他发现，组里用的几种分子筛选模型，都默认一个前提：一个药分子好不好，主要看它和靶点蛋白结合得紧不紧。可是在几个失败的项目里，他隐约觉得，问题不在结合得紧不紧，在分子进了细胞以后，和一大群别的东西怎么缠在一起。他想换一个问法：不问「它粘不粘得住」，问「它进去以后，细胞里那一整片东西会怎么动」。

这个想法，可能是对的，也可能是错的。要验证它，至少要三五年，要重新攒数据，要写新的模型，前两年大概什么候选分子也交不出来。

他把这个想法写了半页，看了看申请指南，又删掉了。

他换成了另一份方案：用现有的筛选模型，加上一个更大的分子库和一个新的打分函数，把某一类靶点的命中率提高若干个百分点。这份方案，指南里的每一条都对得上；三年内交得出论文，交得出候选分子，说不定还能和一家药企签个合作。

写完已经过了十二点。他在文件名后面加了「终稿」两个字，又在电脑桌面上新建了一个文件夹，把那半页删掉的想法存了进去。文件夹的名字，叫「以后再做」。

那个文件夹里，已经有三个文档了。

◆ 八、一个「再」字怎样在每一轮被推迟

林舟的那个文件夹，是一个国家的缩影。

每一轮技术浪潮来的时候，推迟「再」的方式，都差不多分三步。

第一步，新浪潮刚来，别人定义了题目：个人电脑、互联网、移动互联网、深度学习、大模型。我们离题目的源头远，所以说「先做应用」——这时候，这句话是对的。

第二步，应用做成了，而且做得极好，成本最低、速度最快、规模最大。这时候本该排「再」了，可是应用正在赚钱，正在被表扬，正在被写进下一年的计划。没人愿意在最顺的时候停车。

第三步，还没等停车，下一轮浪潮又来了。新的题目又是别人定义的。于是又一次「先做应用」，上一轮没做的「再」，被新一轮的「先」盖住了。

三步走完，循环回到原处。每一轮都是真心的「先」，每一轮的「再」都被下一轮的「先」吞掉。

所以这个字的问题，不在说这句话的人不真诚。在于这句话的结构本身：它只给「先」配了现金流、考核和掌声，从没给「再」配过任何一样东西。一个只有起点、没有排期的承诺，在四十年里，就成了一个习惯。

◆ 九、一个排得上日程的「再」长什么样

说了这么多推迟，也要说一句：一个真能排上日程的「再」，要配哪几样东西？

把林舟的那个文件夹反过来想，大概是四样。

第一样，一笔不跟着应用考核走的钱。它不按年交成果，不按论文数验收，事先就说好，五年里可以什么也交不出来。

第二样，一个日期。不是「条件成熟以后」，是写在纸上的某年某月。到了那天，要么开始，要么写明为什么还不开始。

第三样，一个对「再」负责的人。他的考核不看应用做得多好，只看「再」有没有真的发生。

第四样，失败的许可。底层的事，十件里能成一两件就算好。如果每一件失败都要有人担责，就没有人敢开第一件。

这四样，单拿出来都不难。难的是，它们和「先」那一套考核放在同一张桌上时，常常被「先」挤下去。

这四样也不是纸上谈兵。第 16 章会举几处例外：华为的海思，还有北斗、龙芯。它们熬了十几、二十年，背后恰好都配齐了这几样里的大半——一笔不看当年回款的钱，一个肯为它负责很多年的人或机构，还有被允许的失败。哪里配齐了，「再」就真的来过。

公道地说，体制自己也看见了考核这一头的问题，2018 年 10 月科技部等五部门发文「破四唯」，就是在这个方向上动土，第 28 章会细看。

◆ 十、一个人身上的「再」

这个字，不只排在一个国家的日程上，也排在每一个做研究的人的履历上。

一个年轻人读完博士，进了研究院。他心里有一个想做的问题，可是先要评上副研究员。评副研究员要论文，论文要发在好期刊上，好期刊要的是能在两三年里做完、做得干净的题目。于是他说：先评上职称，再做自己想做的。

评上副研究员，接着要评研究员，要拿更大的基金，要带更大的组。每一站都说得通，每一站都有下一站。等他终于站到可以随意挑题的位子上，往往已经四十多岁，手下有十几个人要养，组里每年要交的成果比当年更多。

那个年轻时想做的问题，还在。只是他已经很多年没打开过它了。

林舟今年三十八岁，正走在这条路的中间。他那个叫「以后再做」的文件夹，就是这条路上的一块里程碑。

一个国家的「再」，就是由几十万个这样的文件夹叠起来的。第 28 章会回到这几道关。

◆ 十一、今天与机器

到了 AI 进入科学这一轮，「再」字遇上了一件以前没遇过的事。

以前，「先做应用」挣来的那些本事——工程、规模、成本、速度——至少是自己的。今天拿它们换不来底层，明天它们还在，还能换别的。

这一次不一样。中篇二会细讲：AI 正在接管的，恰恰是「在给定的题里做得最好、最快、最便宜」这一层。也就是「先」这一层。第 369 号《典范发生学》说过，常规科学正被机器接过去。

这意味着，「先」挣来的那份本事，这一次可能会贬值。「先」这一层被机器做了，而「再」那一层从来没练过。

以前推迟「再」，代价是慢一点。这一次推迟，代价可能是两头落空。

◆ 十二、放回日子里

回到林舟那个叫「以后再做」的文件夹。

本书不劝他明天就辞职，去赌那半页想法。他有组员要发工资，有孩子要上学，有考核要过。叫一个人一个人去扛整片土的分量，是不公道的。

可是有一件小事，他明天就能做。

把那个文件夹打开，在那半页想法下面，写一个日期。不是「以后」，是一个具体的日子，比如明年的三月一日。再写一句：到那一天，如果我还没开始，我要写下为什么。

这件事很小，可它把一个没有排期的「再」，变成了一个有排期的「再」。

你也可以试一试。你的日子里，也有一个那样的文件夹：一件你说了很多年「以后再做」的事。今晚可以把它找出来，给它写一个日期。

「先」字从不一个人站着；一个只配了掌声、没配排期的「再」，四十年里就会变成「一直」。

带 走 的 话

「先做应用」本身没有错，错在只给「先」配了现金流和掌声，从没给「再」排过一天日程。

小　账

看见了什么，漏看了什么？看见了在薄土上先种能当年收成的庄稼，四十年收成极好；漏看了收成越好，越没人肯停下来翻一次土。

是在改进，还是在出题？每一轮的「先」都是在别人出的题里做到最好；那个被推迟的「再」，原本是轮到自己出题的那一天。

想 一 想

你说了最久的那一句「以后再做」，是哪一件事？是什么在每一次把它往后推——缺钱、缺时间，还是眼下的事做得太顺？如果给它定一个日期，你敢定在哪一天？

13

第 13 章

解题的能力和出题的能力

◆ 一、林豆豆的奥数班

林豆豆九岁，读小学四年级，每周六上午去上奥数班。

那个周六晚上，宋岚在厨房洗碗，豆豆趴在餐桌上订正一张卷子。卷子上是一道「鸡兔同笼」的变形题：笼子里有鸡、兔和一种三条腿的怪鸟，给了头数和脚数，还多给了一个条件。

豆豆用了不到五分钟，就写出了答案。她做过这一类题，已经有几十道了。老师教过一个口诀，先假设全是鸡，再一步步把多出来的脚分给兔子和怪鸟。

宋岚擦着手走过来，看了一眼，说：「全对。厉害。」

豆豆没抬头，忽然问了一句：「妈妈，这道题是谁想出来的？」

宋岚愣了一下：「出题的老师吧。」

「那出题的老师，是怎么想出这道题的？」

宋岚答不上来。她想了想，反问：「你要不要自己出一道？」

豆豆咬着铅笔，盯着本子看了很久。她写了一行，划掉；又写了一行，又划掉。最后她说：「我只会改数字。把鸡改成鸭子，把三十五个头改成四十个头。」

她会解这一类题的所有变形，却没见过一道自己出的题。

◆ 二、两种本事

这一章要分开两种本事。它们常常被叫成同一个名字，其实是两回事。

一种叫解题。题目已经有人定好了：问什么、用什么条件、怎样算答对，都写在卷子上。解题的人要做的，是在这道题里，做得最好、最快、最便宜。

另一种叫出题。没有卷子。出题的人要决定：这件事里，什么值得问？用哪些条件来问？什么样的答案才算答对？一道好题出来以后，成千上万的人可以去解它；在它出来之前，世界上并没有这道题。

在小学的卷子上，这两种本事的差别很小。到了一个国家的产业和科学里，这两种本事的差别，就是全书反复回到的那一件事。

第 12 章说的「先做应用」，用这两个词说，就是：做应用，是接受别人定义好的题，在题里做到最好、最快、最便宜。原创，是出题。

◆ 三、把四十年的题目摆出来

把前几章说过的事，用这两个词重新排一遍。

操作系统是一道题。它问的是：一台机器上，怎样让很多程序共用一块芯片、一片内存、一块屏幕，又互不打架？这道题，是美国的大学和公司在几十年前出的。我们在这道题出完以后，做了全世界最好用的一批 App。

搜索是一道题，社交网络是一道题，智能手机上的「应用商店」也是一道题。每一道题出来以后，中国都很快做出了更便宜、更贴近本地用户、更会运营的那一份答卷。

Transformer 也是一道题。2017 年，Google 的一组研究者在论文《Attention Is All You Need》里，把「怎样让机器理解一段话」重新问了一遍：不用一个字一个字往后读，而是让每个字一次看见所有的字。它的来历，第

18 章会细讲。这道题出完以后，中国做出了全世界最便宜的一批模型，也做出了每天上新一千五百部的 AI 短剧。

这张单子还能列很长。它的样子是一样的：题，从别处来；最好的答卷，常常出自中国。

◆ 四、先说解题的本事

说出题之前，先把公道还给解题。

解题能力是真本事，世界上没有几个国家解得这么好。

这句话要放在最前面。中国的工程师，能把一个别人发明的东西，在几个月里做成成本只有原来几分之一的版本；能把一套在实验室里勉强跑通的方法，做成能服务几亿人的系统；能在算力被限制、显卡被卡的情况下，把模型训练的成本压到让世界吃惊的程度，第 14 章会细讲这件事。

这些都不是「抄」能做到的。抄只能抄出一个一样的东西；中国做出来的，常常是一个更好、更快、更便宜的东西。要做到这一点，要懂原理，要懂工程，要懂用户，还要有一整套肯吃苦的组织。

全世界许多国家，有的会出题、不太会解，有的既不会出也不会解。会解，而且解到这个程度的，屈指可数。

所以本书说中国的困境，不是说中国人笨，也不是说中国人只会模仿。恰恰相反：中国把解题这门本事，练到了全世界的前排。

◆ 五、解题不会自动长出出题

可是有一件事，很多人默认了，却从来没被证明过：解题解得多了，会不会自己长出出题的能力？

直觉上，好像会。豆豆做了几十道鸡兔同笼，难道不是离出一道新题更近了吗?

答案是：近了一点，却不会自己跨过去。

原因在于，两种本事练的是两块不同的肌肉。

解题练的是：在一个已经定好的框里，找最短的路。题做得越多，你越熟悉这个框，越知道哪条路最短。框本身，你从来没碰过。你越熟悉它，越会把它当成天然就在那里的东西。

出题练的是另一件事：在一片还没有框的地方，决定框画在哪里。它要求你先看见「这里有一件事，没人问过」，然后忍住不去马上解它，先把它问对。

一个人做了一万道鸡兔同笼，他对鸡兔同笼的框会熟悉到骨头里。也正因为熟悉到骨头里，他很难想到：也许「头数和脚数」这个问法本身，就可以换一种问法。

解题解得越好，框越像墙。

◆ 六、考场上的金牌

这件事，在中国的考场上看得最清楚。

中国队在国际数学奥林匹克上，多次拿过团体总分第一。参赛的孩子，是从几十万学奥数的学生里一层层选出来的。他们能在几个小时里，解出全世界最难的一批竞赛题。

这是解题能力的顶点，没有人能说它不是本事。

可是请看那几道竞赛题是怎么来的。它们由各国的数学家提出，经过一个委员会层层筛选，最后摆上考桌。每一道题，都有出题人事先确认过的标准答案。考场上的孩子要做的，是在规定的时间里找到它。

一个数学家一辈子的工作，长得很不一样。没有人给他确认过答案，甚至没有人确认过他的问题值得问。他可能花十年追一个问题，最后发现它问错了；也可能问出一个问题，让后面一百年的人去解。

从考场到这样的一生，中间不是一条更长的跑道，是一次换道。

金牌说明一个孩子解得好。它不说明，也说明不了，他能不能出题。很多金牌得主后来成了极好的数学家，那是因为他们在别处又长出了另一块肌肉，不是因为金牌自己长成了它。

◆ 七、出题要的那片土

出题既然不能靠多做来练，它靠什么？

说不全，可以说几样看得见的。

一样是闲。不是无所事事，是有一段时间，不必马上交出答案。豆豆的奥数班，每一分钟都排着要解的题；一个人如果从小到大每一分钟都排着题，他就没有空去想「这道题为什么是这样问的」。

一样是被允许问傻问题。豆豆那句「这道题是谁想出来的」，在很多课堂上会被当成打岔。问的人被打断几次，就不问了。

还有一样，是有人肯把一个半生不熟的问题当回事。一道新题刚冒头的时候，往往说不清楚，像胡说。旁边要是有一个人说「你接着说」，它才有机会长成形。

这几样，都不是多刷题能刷出来的。它们是土。第 46 章讲教育时，会回到这片土。

◆ 八、只看一面会怎样

这件事，也有两种只看一面的说法。

一种说：出题才是高级的，解题是低级的。这不对。没有解题，出题就落不了地。Transformer 这道题出完以后，如果没有一大批人把它解成能用的系统，它就只是一篇论文。一个只会出题、不会解题的地方，会被自己的题目饿死。

另一种说：先把题解好，出题是将来的事，到时候自然就会了。这就是第 12 章那个「再」字。它默认解题和出题在同一条路上，只是一前一后。可是它们不在同一条路上。一条路走得再远，也不会自动走到另一条路上去。

两种本事都要。差别只在于：解题可以靠多做来练，出题不能。出题要另一片土。

◆ 九、题目是从哪里来的

还可以换一个角度，看一看一道题是从哪里来的。

一道题，往往来自一个解不开的别扭。第 15 章会说到，天文学家对着那一堆越套越多的小圆圈，心里越来越别扭。孟德尔数豌豆，数来数去，后代里总冒出一个三比一，他也别扭：这个比例凭什么这样整齐？别扭本身不是答案，可它是一道新题的种子。

解题的训练，恰好会把别扭磨掉。一道题做不出来，老师会说：再想想，用上学过的方法。做出来了，别扭就消失了。一千道题做下来，一个人学会的是：每一个别扭，都能被一个已知的方法化掉。

这当然是好事。可是如果有一种别扭，任何已知的方法都化不掉呢？被训练得很好的人，第一反应常常是：一定是我还不够熟练。他很少想到，也许是题目本身有问题。

豆豆那句「这道题是谁想出来的」，就是一点没被磨掉的别扭。

◆ 十、今天与机器

现在来一件新事。

宋岚的短剧公司里，有一个年轻的编剧，前几天给她看了一段 AI 写的剧本大纲。十分钟，十个版本，个个能拍。

宋岚翻了翻，说了一句很准的话：「它写得都很顺，只是都像上个月爆过的那几部。」

这句话说出了 AI 和解题的关系。今天的大模型，最擅长的就是在一个已有的框里，找出最顺、最像样、最合格的那条路。它是全世界最勤快的解题者。

第 369 号《典范发生学》说过，答案、过程、甚至提问，都正在变得不再稀缺。机器会解题，而且越来越会；它也能在已有的框里提出很多像样的问题。

这对一个最擅长解题的国家意味着什么？意味着「解得最好、最快、最便宜」这件事，正在从一个国家的长处，变成一台机器的日常。中篇二会把这件事一层一层拆开。

剩下那件机器还做不好的事，是在没有框的地方画框——出一道连它都没见过的题。

◆ 十一、一个完整的例子：两道药物题

回到林舟的工作，看看解题和出题在一个真实的行当里长什么样。

林舟组里每天做的，是一道已经出好的题：给定一个靶点蛋白，从几亿个分子里，找出和它结合得最紧的那几个。

这道题的框，是几十年前定下的。它默认：药之所以有效，是因为分子像钥匙一样插进蛋白的锁孔。AlphaFold 出来以后，锁孔的形状更容易算了；大模

型出来以后，钥匙更容易筛了。林舟的组把这道题解得很好，命中率一年比一年高。

第 12 章里，林舟揣着另一个想法：药效的关键，也许不在钥匙和锁孔，而在分子进了细胞以后，和一大群别的东西怎么缠在一起。

这就是一道新题。不是「怎样把钥匙筛得更准」，而是「药效这件事，是不是根本不该用钥匙和锁来问」。

请看这两道题的差别：

前一道，题目清楚，答案可以打分，今年做可以今年交差，机器可以帮你做得更快。

后一道，题目还说不清楚，答案不知道怎样打分，五年后也不一定有结果，机器可以帮你算，却不能替你决定这样问对不对。

林舟能把前一道解到全国前列。后一道，他连写进申请书都不敢。

这不是林舟的问题。他的整个职业生涯，从高考、读博、发论文到评职称，一路被训练的，都是前一种本事。第 27、28 章会回来讲这条路。

◆ 十二、放回日子里

那个周六晚上，宋岚没有让豆豆继续订正卷子。

她拿过一张白纸，说：「我们今晚不做题，玩一个游戏。你看窗外那栋楼，有多少个窗户亮着灯？」

豆豆数了一会儿：「四十七个。」

「好。现在你想一个问题，要用到这个四十七，不过不许问『有多少个』。」

豆豆想了很久。她先问：「如果每个亮灯的窗户后面有三个人，一共有多少人？」宋岚笑着摇头：「这还是『有多少个』。」

又过了好一会儿，豆豆说：「为什么十一点的时候，这栋楼亮着的灯，比九点的时候少一半，七楼那一排却一个都没灭？」

宋岚也答不出来。她知道，这是豆豆今晚出的第一道题。

一个小办法，你今晚也能试：和孩子，或者和你自己，做一个只出题、不解题的游戏。规则只有一条：出的题，不许是你做过的任何一类题的变形。

> **解题解得再多，框也只会越来越像墙；出题，是第一次决定墙画在哪里。**

带 走 的 话

解题是真本事，中国解得极好；可是解题不会自己长出出题，两件事要两片土。

小　账

看见了什么，漏看了什么？看见了中国把别人出的每一道题都解成了最好、最快、最便宜的答卷；漏看了解得越熟，越把题目的框当成了天生的墙。

是在改进，还是在出题？这一章本身就在问这件事——四十年的成绩几乎全在解题，第一次自己画框的时刻，还很少。

想 一 想

你今天做的工作里，哪一道「题」是你自己出的，哪些是别人早就定好、你只负责解得更好的？上一次你问出一个连自己都没见过的问题，是什么时候？

◆

14

第 14 章

撞墙之后改的是什么

◆ 一、2025 年初的那个周一

2025 年 1 月底，一个周一的早上，宋岚在公司茶水间听见两个同事在争。

一个说：「你看见没有，英伟达的股票一天跌了一大截。」另一个说：「就因为杭州那家公司的一个模型？」

那家公司叫 DeepSeek，中文名深度求索。几天前，它发布了一个推理模型 R1，效果和美国最好的几家相近，而且权重公开，谁都可以下载。更让人吃惊的，是它此前公布的另一个数字：它的基础模型 V3，最后一轮训练花在显卡上的钱，按它自己的算账口径，只有几百万美元。这个口径不含此前的试验和人工，可它仍然让同行吃了一惊。

据报道，那一天英伟达的市值蒸发了数千亿美元。硅谷的人开始问：花几十亿美元建的那些机房，是不是建多了？

那天晚上，宋岚回家问林舟：「你们圈子里怎么看？」

林舟说：「了不起。真的了不起。」

他顿了一下，又说了一句：「不过，你要看它改的是什么。」

这一章，就是要把林舟后半句话说清楚。

◆ 二、墙是怎么来的

先说墙。

从 2022 年起，美国一轮一轮收紧对中国的高端芯片出口。最强的那几款训练用显卡，中国公司买不到了；能买到的，是专门为中国市场降了规格的版本。据 DeepSeek 自己的技术报告，V3 是在这种降了规格的显卡上训练出来的。

这就是墙：算力被卡，而且钱也不能像美国的大公司那样随便烧。

撞上墙的人，通常有两种反应。一种是停下来，等墙倒；一种是逼着自己，在墙底下找一条别人没走过的缝。

DeepSeek 选了第二种。

◆ 三、它做成了什么

说它停在哪里之前，先老老实实把它做成的事摆出来。这一节要写足，因为它值得。

第一件，叫 MoE，「混合专家」。

打个比方。一个普通的大模型，像一家所有员工都要参与每一件事的公司：来一个问题，全公司几千亿个参数一起算一遍。MoE 的办法是把公司分成很多小科室，每个科室是一个「专家」；来一个问题，一个调度员只叫其中几个科室来干活，其他科室歇着。这样，模型的总参数可以很大，每一次真正动起来的却只是一小部分，算力就省下来了。

DeepSeek 在 MoE 上做了好几处细活：把专家切得更细、数量更多；留出几个「共享专家」管人人都要的常识；还琢磨出不额外加罚分、也能让各科室忙闲均匀的办法。据它的技术报告，V3 总参数约 6710 亿，每次只调动约 370 亿。

第二件，叫 MLA，「多头潜在注意力」，在 2024 年的 DeepSeek-V2 里提出。

这一处更技术，可以这样理解：大模型在读一段很长的话时，要把前面每一个字的「笔记」都存在显存里，好让后面的字随时回头查。话越长，笔记越厚，显存就越不够用。MLA 的办法，是把这些笔记压缩成一份很薄的摘要存起来，用的时候再展开。显存省下一大块，长文本就跑得动、跑得便宜了。

此外还有一长串工程活：用更低精度的数字来训练，自己改写显卡之间的通信，把计算和传输叠在一起做。每一项单看都不惊人，叠在一起，把同样的能力压到了一个很低的成本上。

这些成绩是实打实的。它改变的不只是一家公司的账本，是全世界做大模型的成本结构。此后，便宜而且开放的模型成了一条主路；第 4 章说过，中国模型在 OpenRouter 上的调用量，2026 年 2 月起超过了美国模型。这条路，DeepSeek 是重要的开路人之一。

◆ 四、把年份排一排

现在，按林舟说的，看它改的是什么。

先排年份。

MoE 的想法，不是 2024 年才有的。1991 年，Jacobs、Jordan 等几位研究者就发表了「让几个专家网络分工、由一个门控网络来分配」的论文。那时候的网络很小，这个想法在一个很窄的圈子里流传。

2017 年，Google 的 Shazeer 等人把它做成了大规模稀疏 MoE：在一个大网络里放上成千上万个专家，每次只激活少数几个，参数规模一下子推到了以前不敢想的程度。此后 Google 又在这条线上做了好几代。

所以，「把模型拆成很多专家、每次只叫几个」，这道题在 DeepSeek 之前，已经被出过两次。DeepSeek 做的，是把这道题解得更细、更省、更稳。

MLA 呢？它是在 Transformer 里面省显存。Transformer 是 2017 年 Google 那篇论文出的题，第 18 章会细讲。注意力机制要存笔记，笔记占显存，这是 Transformer 这道题自带的一个代价。MLA 是在这道题里，把这个代价压到了很低。

请注意它的位置：在 Transformer 内部。它没有问「是不是不该用注意力」，它问的是「注意力怎样才能更省」。

◆ 五、它自己怎么说

还有一件事要说公道：DeepSeek 自己，并没有把这些办法说成新的题。

据公开的技术报告，它在写 MoE 时，引的正是前人的那一串工作；写 MLA 时，说的也是在注意力机制里压缩缓存。它的创始人梁文锋接受过少数几次采访。据报道，他的话大意是：中国公司不能永远做跟随者，总要有人站到技术的前沿去做原创。这说明，做这件事的人，自己心里是清楚的：这一步解得极好，还不是出题。

把它拔高成「出题」的，多半是旁观者。旁观者需要一个好消息，于是把一次顶尖的改进，讲成了一次 0 到 1。

这种拔高，看上去是在夸人，其实帮了倒忙。它让人觉得「再」字已经到了，可以松口气了。

◆ 六、本书的说法

把这几件事放在一起，本书的说法是：

撞墙之后，改的是墙的绕法，不是题。

题，是「用一个 Transformer 式的大网络，读大量文字，学会预测下一个字」。这道题在 2017 年到 2020 年之间定型，此后全世界都在里面做。

DeepSeek 在算力被卡的墙下，把这道题解到了几乎榨干的程度：同样的题，花更少的钱，跑出差不多的分数。

这是了不起的。用第 13 章的话说，这是解题能力的一次顶峰表演。

可是请看这面墙逼出了什么。墙卡的是算力，所以逼出来的全是省算力的办法：少激活一点参数，少占一点显存，少传一点数据。墙给了一个约束，人在约束里找到了最好的绕法。

墙没有逼人去问：预测下一个字，是不是理解世界的那条对的路？

这不是因为 DeepSeek 的人想不到。是因为墙压下来的方向，本来就不在那里。约束告诉你的永远是「更省」，它从不告诉你「换题」。

◆ 七、墙有没有可能逼出新题

这里要防一个误会：本书不是说，约束永远只能逼出改进。

约束有时候也会逼出新题。只是它要打在一个特别的地方：打在目标上，而不是打在手段上。

打个比方。一个厨师被告诉，「煤气费涨了，你做同样的菜要少用三成火」，他会去研究更省火的锅、更短的工序。这是打在手段上，逼出来的是改进。

如果他被告诉的是，「从今天起，这家店再也没有明火」，事情就不一样了。他不能只是省火，他得重新问：没有明火，我们还做不做炒菜？做什么菜？一家饭馆是靠什么让人回来的？这一下打在了目标上，才有可能逼出一道新题。

算力被卡这面墙，大体是前一种。它说的是「你只能用更少的显卡」，而不是「你永远不能这样训练模型」。所以它逼出的，几乎全是更省的锅和更短的工序。

要让一面墙逼出新题，墙得高到让人对原来的目标起疑。而一个还在赚钱、还在被表扬的目标，很少有人愿意对它起疑。这又回到了第 12 章那个「再」字。

◆ 八、只看一面会怎样

这件事，国内国外各有一种只看一面的说法。

一种说：DeepSeek 证明了中国能做原创，从此不再只是跟随。这句话里，「原创」两个字用宽了。MoE 和 MLA 都是真新的东西，里面有不少别人没想到的细招，说它们有原创的成分，是公道的。可是它们是在一道已有的题里的原创，不是出了一道新题。把这两种原创混为一谈，会让人以为路已经走到了，「再」字已经来了。

另一种说：它不过是在别人的架构上抠细节。这句话更不公道。抠细节抠到改变全世界的成本结构，这件事本身就证明，「细节」有时候就是一个产业的命门。把它说成「不过」，是看不见工程的分量。

两种说法都错在同一处：把「改进」和「出题」放在一把尺子上量，要么拔高，要么踩低。其实它们是两把尺子。在改进这把尺子上，DeepSeek 站在世界的最前排；在出题这把尺子上，它还没有上场——它也没说过自己上了场。

◆ 九、成本结构变了以后

说「改变了世界的成本结构」，不是一句客气话。可以看几件后来的事。

便宜而开放的模型一出来，全世界的小公司和个人开发者，第一次用得起接近前沿的能力。据 CNBC 报道，2026 年 2 月以来，美国企业在 OpenRouter 上每周都把超过三成的 Token 发给中国模型。到 2026 年 10 月初，按界面新闻的

测算，中国模型在这个平台上的周调用量约 86.7 万亿 Token，美国模型约 19 万亿。

这些数字要读得小心。OpenRouter 的用户里，美国将近一半，中国只占百分之六左右，它不是全世界的缩影。同一篇报道也说，编程助手、后台自动化这类出了错代价很大的活，企业仍肯为准确多付钱，不轻易换便宜的模型。

不过方向是清楚的：在「同样的事，更便宜地做」这一条上，中国模型拿下了一大块地。

这块地有多大，正好说明改进的力量有多大。它也说明，这块地的规矩是什么：谁更便宜，谁就多拿。一旦有人更便宜，地就换主人。

◆ 十、今天与机器

还有一件事，比 DeepSeek 本身更值得停下来想。

DeepSeek 做的那一类事——在一个固定的架构里，找出最省的训练办法、最省的存储办法、最省的通信办法——恰好是今天的 AI 开始能帮人做的事。

举一个公开的例子。2025 年 5 月，Google DeepMind 公布了一个叫 AlphaEvolve 的系统：让大模型自己写代码、自己打分、一轮一轮改。据它的介绍，这套系统替 Google 改写过训练自家模型时用的计算内核，让它跑得更快。这类任务的样子很清楚：目标定好了，就是更快更省；可以反复试，可以自动打分。这正是机器最擅长的「解题」。

所以，中篇二会回来问一个让人不太舒服的问题：DeepSeek 这一次表演的、中国最引以为豪的那种本事，会不会正是 AI 最先接过去的那一种?

如果是，那么在墙下面逼出来的这份绝活，它的价钱，会一年一年往下走。

◆ 十一、一个完整的例子：林舟组里的那面墙

把这一章的道理，放到林舟的组里走一遍。

去年，林舟的组也撞了一面墙。研究院的算力配额被砍了三成，他们原来的分子筛选流程，一轮要跑两周，现在跑不完了。

组里有个博士后，叫小顾，二十九岁，脑子极快。他花了一个月，把筛选流程拆开看：哪些分子其实一看就不行，不必进大模型细算；哪些计算可以用一个小模型先粗算，再交给大模型精算；数据在几台机器之间怎么搬最省时间。

一个月后，筛选从两周压到了四天，命中率还略有提高。

林舟在组会上表扬了他，研究院还把这件事写进了年度简报。小顾很高兴，这也是他应得的。

可是那天晚上，林舟一个人在办公室里，想到了第 12 章那个叫「以后再做」的文件夹。

小顾改的，是「怎样更快地找到和靶点结合最紧的分子」。墙卡的是算力，所以他找到的是更省算力的绕法。他没有、也不可能因为这面墙，去问那半页纸上的问题：「药效是不是不该用钥匙和锁来问？」

墙让组里的每个人都变得更聪明了。可是它让每个人都更聪明地在同一道题里跑。

◆ 十二、放回日子里

第二天，林舟在组会结束前，留了五分钟。

他在白板上画了一条线，左边写「墙让我们省了什么」，右边写「墙让我们问了什么新问题」。他请每个人往两边各写一条。

左边很快写满了：省算力，省时间，省存储，省人工标注。

右边空了很久。最后小顾走上去，写了一行字：「如果算力永远只有现在这么多，我们还要不要筛几亿个分子？」

林舟看着这行字，没说话。这行字还不是一道新题，却第一次把墙的方向，从「怎样更省」掰到了「是不是该这样做」。

明天能试：你最近撞上的那面墙——预算砍了、人手少了、时间紧了——在一张纸上画同样的两栏。左边写它逼你省下了什么，右边写它逼你问出了什么新问题。如果右边一直是空的，你就知道，这面墙只改了你的绕法。

> **墙只告诉你「更省」，从不告诉你「换题」；绕墙绕得再漂亮，题还是原来那道。**

带 走 的 话

墙卡的是手段，逼出来的就只能是更省的手段；要逼出新题，得有东西打在目标上。

小　账

看见了什么，漏看了什么？看见了算力被卡时逼出的 MoE、MLA 与一整套工程，世界为之改账；漏看了这些办法都长在 1991 年和 2017 年出过的题里。

是在改进，还是在出题？这是改进的顶峰——同一个目标，更省更快；题，仍是「用大网络预测下一个字」。

想 一 想

你最近撞上的那面墙，逼你想出的办法，是让你「更省」，还是让你「换一个问法」？有没有哪一次，一个约束让你放弃了原来的问题，而不只是换了一条路去解它？

◆

15

第 15 章

改进不是 0 到 1

◆ 一、一把越磨越快的刀

你家里大概有一把用了很多年的菜刀。

刚买来的时候，它切土豆丝还算利落。后来你学会了怎样磨它：先在粗石上磨出刃口，再在细石上推几十下，最后在报纸上蹭一蹭。磨得好的那几天，它切出来的土豆丝能透光。

你还能把它磨得更快。换一块更细的磨石，换一个更稳的角度，学一学日本师傅的手法。刀会一点一点变好。

可是不管怎样磨，它都是一把刀。

有一天你在饭馆后厨看见，师傅把一大盆土豆倒进一台机器，十秒钟，出来一盆土豆丝。那台机器不是一把更快的刀。它根本没在「磨刀」这条路上。

改进和 0 到 1 的差别，就在这一把刀和那一台机器之间。

◆ 二、王德生博士的一句判断

本书写作之初，王德生博士口述过一句判断，这一章就从它展开。他说，大意是：

应用上的碰撞，涌现的是技术改进，不是架构与科学新典范的 0 到 1。

这句话要一个词一个词地看。

「应用上的碰撞」，是说在产品、市场、用户那里，人和问题撞在一起：用户嫌慢，老板嫌贵，算力不够，竞争对手在逼。中国四十年的绝大多数创造，都是在这种碰撞里长出来的。

「涌现的是技术改进」，是说这种碰撞，确实会逼出新东西，而且常常很巧。第 14 章的 MoE 和 MLA，第 11 章的扫码支付和担保交易，都是这样长出来的。

「不是架构与科学新典范的 0 到 1」，是说这种碰撞，再激烈，也逼不出另一种东西：一个新的架构，或者一门科学新的典范。

这里的「典范」，导论讲过：一门科学认定什么是好问题、什么是好答案、怎样算做对的那一整套。0 到 1，就是换掉这一整套。

◆ 三、库恩的一个说法

这件事，库恩在 1962 年的《科学革命的结构》里，已经用另一种话说过。

他把一门科学的日常工作叫作常规科学。在他看来，常规科学做的事，很像解谜题：题目和规则是典范给定的，一个好的科学家，是能在规则里把谜解开的人。解不开，人们先怪解题的人手艺不够，不怪题。

只有当解不开的谜越堆越多，多到连最好的手艺都化不掉，才会有人回过头去怀疑规则本身。那时候，才有典范的更替。

把这段话放在王德生博士那句判断旁边，就看得出两者在说同一件事的两头。库恩说的是科学内部：解谜不会自己变成换典范。本书说的是科学外面的那片土：应用的碰撞，只会不断地递来新的谜题，不会递来一个「规则错了」的信号。

◆ 四、为什么碰撞只出改进

这句判断不是随口一说。它背后有一个很简单、却常被忽略的道理。

请看应用的约束长什么样。

用户嫌慢：要的是同一个功能，更快。老板嫌贵：要的是同一个效果，更便宜。算力被卡：要的是同一个模型，更省。竞争对手在逼：要的是同一个产品，比他好一点。

把这些约束放在一起，它们的形状是一样的：同一个目标，更便宜、更快。

目标是固定的，变的只是手段。

在一个目标固定的问题里，你能找到的，只能是到达这个目标的更好的路。路可以找得极巧，巧到别人想不到；可是它终点不变。

这就是改进。用典范的话说，它是典范内部的优化：一门科学认定的「好答案」没变，你只是找到了一个更好的好答案。

0 到 1 要动的，是另一样东西：题目本身。它说的不是「这个问题我找到了更好的答案」，而是「原来问的那个问题，问错了」。

一个只问「更快、更便宜」的约束，永远不会告诉你「问错了」。因为它从一开始就把题目当成定好的。

◆ 五、几个从「问错了」开始的故事

把科学史上几次 0 到 1 拿出来看，它们几乎都是从「问错了」开始的。

几百年里，天文学家在问：怎样用更多的小圆圈，把行星的路线算得更准？他们一直在改进，圆圈套圆圈，算得越来越细。后来有人说：也许不该问「地球周围的圆圈怎么套」，该问「如果太阳在中间呢」。题一换，一大堆圆圈就不需要了。

十九世纪末，物理学家在问：光靠一种叫「以太」的东西传播，地球在以太里跑，怎样才能把这股「以太风」测出来？实验做得越来越精，就是测不出来。

1905 年，爱因斯坦换了一个问法：如果根本不必有以太，光速对谁量都一样呢？题一换，「怎样测以太风」这个问题就作废了。

深度学习也是。几十年里，人工智能的主流在问：怎样把人的知识写成规则，教给机器？少数人换了一个问法：能不能不写规则，让机器从大量例子里自己长出本事？1986 年反向传播的论文，2012 年的 AlexNet，第 19 章会细讲这段冷宫里的路。

这几个故事有一个共同点：它们都不是在应用的压力下逼出来的。没有哪个用户、哪个老板、哪个市场，向托勒密的后人要过「把太阳放到中间」，也没有人向爱因斯坦订过「取消以太」。

◆ 六、一张对照表

把改进和 0 到 1 并排放，各问各的问题：

	改进	0 到 1
它问什么	这道题，怎样解得更好	这道题，是不是问错了
目标	固定	被改写
约束从哪里来	用户、成本、算力、对手	一个说不出用处的困惑
怎样算成功	更快、更省、分数更高	一批旧问题不再需要回答
多久见分晓	一个季度到几年	十年到几十年，常常更长
谁来认	市场、指标、排行榜	起初几乎没人认
失败的样子	做得不够好	被当成胡说
典型的人	一个团队，越大越强	一个人或几个人，常在边上
例子	MoE 的细化、MLA、扫码支付	日心说、狭义相对论、深度学习、Transformer

这张表里，最要紧的是第一行和最后一行。

第一行说，两者问的是不同的问题。不是一个问得深、一个问得浅，是它们根本问在两处。

最后一行里，Transformer 被放在了右边。这一格要说一句：Transformer 是架构的跃迁，它离应用比日心说近得多，它出自 Google 的机器翻译组。第 18 章会专门讲，它是怎样在一个应用部门里，靠一块免于产品指标的「飞地」长出来的。它恰好说明，0 到 1 需要的不是离应用远，是有一块地方，暂时不被「更快更便宜」那一套考核管着。

◆ 七、怎样认出一次 0 到 1

表上那几行，还可以变成三个随身带的问题。下一次你听到「突破」两个字，可以拿它们去量一量。

第一问：它让哪些旧问题不必再回答了？日心说出来以后，「再加几个小圆圈」这个问题没人问了。深度学习出来以后，「这条规则该怎么写」在很多地方不必问了。一次改进，会把旧问题答得更好；一次 0 到 1，会让一批旧问题直接作废。

第二问：它改了打分的办法吗？改进沿用原来的分数，只是分数更高。0 到 1 常常连「怎样算好」都要重新定。哥白尼的日心说刚出来时，算行星位置并不比托勒密的老办法准多少。按当时的打分办法，它拿不出更高的分数；它赢的，是另一种好：用更少的圆圈说清同一片天。

第三问：它刚出来的时候，像不像胡说？改进刚出来，内行一看就懂，最多说一句「巧」。0 到 1 刚出来，内行常常第一眼觉得不对劲。这一问要小心用：像胡说的，大多数真是胡说。它只说明，0 到 1 很难靠内行的第一眼来认。第 36 章会回来讲，这件事要另一种评审。

三问都答「没有」，那是一次改进，也许是极好的改进。三问里有一问答「有」，就值得多看一眼。

拿这三问量一量第 14 章的 MoE 和 MLA。它们没有让哪个旧问题作废，打分办法照旧，内行一看就懂、一看就服。所以它们是改进，是顶尖的改进。

◆ 八、改进的价值，一点都不小

说到这里，要再还一次公道。

改进不是二等的事。人类今天手上的几乎每一样东西，从一粒药到一部手机，都是被改进了成千上万次才变得便宜、可靠、人人用得上的。0 到 1 出了一道题，可是要让这道题变成千家万户桌上的东西，靠的是成千上万次改进。

没有改进，0 到 1 只是一个想法。

所以本书说「改进不是 0 到 1」，不是说改进不重要，更不是说做改进的人不如做 0 到 1 的人。这句话只说一件事：两样东西不在一条路上，改进做得再多，也不会自己变成 0 到 1。

一把刀磨一万次，不会变成那台机器。不过那台机器造出来以后，也要有人一遍一遍地改它的刀片。

◆ 九、只看一面会怎样

这一章的道理，最容易被读成两种样子。

一种读法：既然改进出不了 0 到 1，那就别做改进了，全力去做原创。这是走了另一个极端。一个国家要是放下自己最擅长的那件事，去赌一件没把握的事，多半两头都落空。而且，0 到 1 从来不是靠「全力去做」做出来的。它需要的是一小块不受「更快更便宜」管的地方，和少数几个能待在那里的人。

另一种读法：既然改进也这么重要，那继续做改进就好了，0 到 1 是别人的事。这一种读法，在 AI 时代以前或许还说得过去。第 12 章说过，「再」字推迟的代价，这一次变了。第 14 章说过，改进这件事，AI 正一点一点接过去。

两种读法都想把两件事合成一件。本书的意思恰好是：它们是两件事，要两片土、两种考核、两种人。

◆ 十、把中国四十年放到表上

拿那张对照表，把前几章说过的事，一样一样放上去。

汉卡和 WPS，是在别人的电脑上，把中文这道题解好了。扫码支付和担保交易，是把「陌生人怎样放心买卖」这道题，用一个极巧的办法解得极便宜。云，是把别人出的「算力能不能像水电一样买」，解成了中国自己的一份。DeepSeek，是把 Transformer 这道题，在墙下榨到了极致。

每一样，都在表的左边。每一样，都是左边那一栏里的佼佼者。

右边那一栏里，中国四十年来写进去的，不是没有，是很少。王选的激光照排算得上半个：他没有去改进铅字，跳过了当时国外流行的几代照排机。可第 9 章也说过，他面对的那道题本身是清楚的，就是让汉字进入计算机印刷。连这样的半个，一只手大概也数得过来。

这张表不是用来羞人的。它只是把一件很多人心里隐约知道的事，摆到了桌上：左边的格子满了，右边的格子空着。

◆ 十一、今天与机器

第 369 号《典范发生学》有一个比方：车替代了腿，不是让腿去学一种更快的跑法，而是交通本身升级了。第 22 章会把这个比方细细展开，这里只借它的一角。

改进，是让腿跑得更快。训练更科学，鞋子更轻，跑道更平，一百米从十一秒跑到十秒。这是一件真本事，而且每进一步都难。

AI 进入科学，是来了一辆车。它在「让腿跑得更快」这件事上，很快会比任何一条腿都快。在给定的典范里找最好的答案，在固定的目标下找最省的手段，这正是机器越来越拿手的事。

在这种时候，一个国家要是把全部力气还放在练腿上，它练得再好，也是在和车比赛跑。

车来了以后，剩下那件人还做得了、也必须做的事，是决定往哪里去。要不要修一条新路？原来那个目的地，是不是本身就选错了？这就是 0 到 1，第 369 号把它叫作典范猜想。

◆ 十二、一个完整的例子：林舟的两份报告

把这一章收在林舟身上。

年底，研究院让每个组交一份年度总结，还要附一页「下一年创新点」。

林舟先写了总结。筛选速度提高了三倍，命中率提高了几个百分点，发了几篇论文，和一家药企签了合作。这一份写得很顺，每一条都能打分。

然后他写「创新点」那一页。他按惯例列了三条：换一个更大的分子库，加一个更好的打分函数，接入一个新的大模型。

写完以后，他照着这一章那张表，把这三条一条一条对了一遍。三条都在左边那一栏：目标固定，问的是怎样解得更好。

他想了一会儿，在页底加了第四条，只有一句话：「药效是否不该用『结合强度』来问？拟用一名学生、十分之一的算力，试验一年，不设论文指标。」

他知道这一条多半不会被批。可是这一回，他没有把它存进那个叫「以后再做」的文件夹。

◆ 十三、放回日子里

你也可以给自己的工作，做一次林舟那样的对照。

今晚可以试：拿出你明年的计划，或者你手上最要紧的一件事，问自己两个问题。第一，它的目标是不是早就定好了，我要做的只是更快、更省、更好？第二，有没有一个可能，这个目标本身就问错了？

第一个问题答「是」，没关系，大多数好工作都是这样。要紧的是第二个问题：你有没有给它留过一页纸、一个下午、一点点算力。

> **「更便宜、更快」从不说「问错了」；0 到 1 的第一句话，偏偏就是这一句。**

带 走 的 话

应用的碰撞只给一个固定的目标逼出更好的手段；0 到 1 要先承认，原来那个目标问错了。

小　账

看见了什么，漏看了什么？看见了应用碰撞逼出的改进巧妙、真实、改变世界；漏看了它的约束形式本身，就把题目钉死在原处。

是在改进，还是在出题？这一章给了一把尺子——目标固定、只换手段的，是改进；动到题目本身的，才是出题。

想 一 想

你手上最要紧的那件事，它的目标是谁定的，定了多久了？如果有人告诉你「这个目标问错了」，你第一反应是反驳他，还是想听他说完？

◆

16

第 16 章

本篇总账：应用的账

◆ 一、抽屉里的一张旧发票

老林的抽屉里压着一张发黄的发票。

那是 1990 年代初，县城国营厂给财务科买一套财务软件开的票。经手人一栏，写着他的名字。厂子早没了，发票他一直没扔。

一个周六晚上，一家人吃完饭，宋岚在厨房切水果。老林把发票翻出来，摊在饭桌上，说：「那年我们几个会计，围着一台电脑看它出报表，跟看电视一样。」

林舟笑着问：「那时候你们觉得，以后会怎样？」

老林想了想，说：「觉得先把账算上电脑，以后什么都会有。」

「以后是什么？」

老林张了张嘴，没说出来。他夹了一筷子菜，过了一会儿才说：「反正后来是都有了。淘宝有了，手机付钱有了，你们那个什么云也有了。」

林豆豆在旁边插嘴：「爷爷，电脑是谁发明的？」

老林摇摇头：「反正不是我们。」

这一章，就从这句说不出来的「以后」开始算账。中篇一走了七章，从汉卡写到 DeepSeek。现在把七章的看见和漏看并排放，看一看四十年里，那个「以后」到底排在了哪一天。

◆ 二、七章排成一张表

先把表摆在桌上。每一行，左边是那一章看见的，右边是它漏看的。

章	看见了	漏看了
第 9 章　汉卡与财务软件	计算机进中国时，生产力走在前面：汉卡、WPS、用友、金蝶都是干活用的；顺序由条件定，不是天性	它们解的是「中文怎么进电脑」「账怎么上电脑」，底下的操作系统和芯片仍是别人的
第 10 章　网吧和网游的第一桶金	注意力挣来的钱是真钱，养活了一代公司，也练出了一代工程师	钱回流到了下一个应用，很少往下走一层
第 11 章　从电商到支付	钱确实回流了，回流成支付、物流、云这些应用层的基础设施	再往下，到操作系统、芯片、EDA、工业软件，路就断了
第 12 章　那个「再」字	「先做应用」是务实的选择，在钱少、底子薄的年代几乎是唯一的选择	「先」字后面那个「再」，没有年份、没有预算、没有一个对它负责的人
第 13 章　解题的能力和出题的能力	在给定的题里，中国能做到最好、最快、最便宜	出题是另一种能力，不会因为解题解熟了就自己长出来
第 14 章　撞墙之后改的是什么	算力被卡时，DeepSeek 把 MoE、MLA 做到了极致，改了全世界的成本账	墙卡的是手段，改的是绕墙的法子；MoE 的想法出自 1991 年，Transformer 出自 2017 年，题都是别人出的
第 15 章　改进不是 0 到 1	应用的碰撞逼出的改进巧妙、真实，没有它，任何新题都落不了地	碰撞只给固定的目标找更好的手段，从不说「问错了」；改进做得再多，也不会自己变成 0 到 1

这张表的每一行都不是骂人。左边那一列，每一格都是一代人一题一题做出来的。

你读这张表，最好先遮住右边，只读左边。读完了，再把手挪开。

◆ 三、先把左边那一列读完

说漏看之前，先把看见读足。

四十年里，中国在应用这一层做成的事，放在世界上也少见。从激光照排、WPS 和财务软件，到 QQ、网游、淘宝、支付宝，再到阿里云，前面几章一样一样看过。

这些事有一个共同的样子：题摆在那里，有人在题里做到了最好。中文怎么进电脑，账怎么算得快，几亿人怎么在网上买东西、付钱、收货。每一道题都很难，每一道都被做出来了。

到了今天，这个样子还在。第 14 章说过，DeepSeek 在算力被卡的条件下，把稀疏专家模型和注意力的压缩做到了一个新的高度；开放权重的中国模型，在 OpenRouter 上的调用量从 2026 年 2 月起超过了美国模型。

这不是运气。这是规模、成本、工程纪律和一大批肯熬夜的人，一起长出来的能力。

你身边大概就有这样的人：一个凌晨两点还在改上线脚本的工程师，一个为了把物流成本压低一毛钱跑遍三个仓库的经理。左边那一列，是他们写的。

所以左边那一列，本书一个字都不打折。

◆ 四、再把右边那一列连起来

右边那一列，单看一格，像是各有各的缘故。连起来读，说的是同一件事。

每一次，钱和能力都回流了。网游的钱养出了电商，电商的钱养出了支付，支付的数据养出了云。回流是真的。

不过每一次，回流到的都是下一个应用，或者应用底下那一层的基础设施。很少再往下，到操作系统、芯片、EDA、工业软件。更少再往下，到那些还没有应用、也不知道有没有应用的题。

还有一处更要紧。七章里，题几乎都是别人出的。图形界面是别人出的题，搜索是别人出的题，大模型的那套结构也是别人出的题。中国在这些题里做得越来越好，有时候做得最好。出题的那一下，却很少发生在这片土里。

所以右边那一列，漏看的不是能力，是一个日程。

◆ 五、本篇的总判断

两列并排放，中篇一的总判断是：

「先做应用，再做原创」里的「再」，四十年没排进过日程。

这句话要说得准一点。

它不是说四十年里没人想做原创。想的人很多，喊的人也很多，有的人还真做了一辈子。它说的是：在整个组织方式里，「再」一直是一个副词，没有变成一行日程。

一行日程是什么样子？有一个年份，有一笔不对当年回款负责的钱，有一个对它负责的人，到了日子有人来问：做到哪一步了？

「先」字有这些。每一个季度的财报，每一轮融资，每一次考核，都在替「先」字排日程。「再」字没有。于是「先」一直在先，「再」一直在后面，像老林饭桌上那句说不出来的「以后」。

你可以拿自己的日历试一试。翻开下个月，看看有几格写着「先」要做的事，有几格写着「再」要做的事。多半，后一种一格也没有。

◆ 六、为什么「再」排不进去

这里要多问一句：为什么四十年里，「再」字一直排不进去？

不是因为没人看见。早在 1980 年代，就有人在说要有自己的操作系统、自己的芯片。也不是因为没有钱。到了 2010 年代，钱多得找不到地方投。

本书的回答是：「先」和「再」用的是两种钱，两种考核，两种人，而这片土里只长得出一种。

「先」用的是要回款的钱，考核的是今年做到了多少，要的是能把给定的题做快做好的人。「再」用的是可能十年都不回款的钱，考核的是这件事有没有往前走，要的是愿意在没人信的路上走很久的人。

前一种，市场会自动长出来，评价体系也会自动奖励。后一种，要有人刻意去养，像在一块地里专门留出不种庄稼的一角。

四十年里，每一次留出那一角的提议，都要和当年最赚钱的那一茬庄稼抢地。每一次，庄稼都赢了。这不怪庄稼。

◆ 七、「再」字的三种说法

「再」字在日子里，很少以「再」的样子出现。它通常有三种说法。

第一种说法是「以后再说」。这是老林饭桌上那一种。它听起来像一个约定，其实没有日期。说的人是真心的，听的人也信，可是谁也不记在本子上。

第二种说法是「等条件成熟」。这一种更体面。钱够了，人齐了，市场大了，就去做。问题是，条件从来不会自己宣布它成熟了。每一年都有更成熟的事要先做，于是条件永远差一点。

第三种说法是「先活下来」。这一种最难反驳。一家公司、一个课题组、一个刚回国的年轻人，都要先活下来。活下来是对的。可是活下来以后，活法往往已经定型了：靠什么活下来，就接着靠什么活。

三种说法放在一起，你会发现它们有一个共同的语法：都把「再」放在一个没有写明的将来。

将来是一个不需要负责的时态。一行日程却要现在就写下一个日子。中篇一七章里，「再」字一直待在第一种时态里，从来没进过第二种。

◆ 八、表里装不下的例外

这张表也有它装不下的东西，本书要先认下来。

四十年里，有一些事不在「先应用、后回流」这条路上。华为很早就自己做芯片设计，海思熬了很多年才见到回报。北斗是一套从头搭起来的导航系统。高铁在引进以后，长出了一整套自己的工程体系。中科院也做过自己的处理器，龙芯走了二十多年。

这些例外有一个共同点：它们背后，都有一笔不靠当年回款的钱，和一个肯为它负责很多年的人或机构。换句话说，它们恰好是「再」字被排进了日程的那几处。

所以例外不推翻总判断，反倒替它作证。哪里给「再」排了日程，哪里就长出了东西；哪里没排，「再」就一直是一句话。

还要说一句：这些例外大多仍在第二层以内，在别人已经定好的题里，自己造出一套能用的家伙。这本身就很难，很了不起。至于出一道全新的题，那是中篇二要讲的另一层。

你要是在这些例外里做过事，大概最清楚：那几年最难的，不是技术，是每年都要重新说服一遍「为什么还不放弃」。

◆ 九、只看一面会怎样

这张表很容易被读歪，有两种歪法。

只看左边一列，会读出自满：我们什么都能做出来，原创也就是时间问题。这一读，把「再」当成了会自己到来的东西。第 12 章说过，「再」不会自己到来，它要被排进日程，要有钱、有人、有人问。

只看右边一列，会读出自轻：中国人只会抄，天生不会出题。这一读更错。第 9 章那个反例就摆在那里：计算机刚进来的时候，生产力就走在前面。顺序是条件定的，不是天性定的。条件一换，顺序就会换。

两种歪法有一个共同点：都把四十年读成了一种性格。本书要读的是一种土。性格改不了，土是可以改的。尾篇会回到这件事。

◆ 十、今天与机器：「再」要重算

「先做应用」这条路，四十年里一直走得通。它靠的是一个没说出口的前提：应用这一层是值钱的，而且会一直值钱。

只要这一层值钱，「再」推迟一年，损失的只是时间。先做的那一层还在挣钱，还在练兵，练出来的人将来也许会往下走。推迟是可以承受的。

AI 来了，这个前提开始松动。

请注意 AI 接手的是哪一层。写代码，调参数，读文献，做报表，生成画面，筛分子。这些活有一个共同的样子：题已经给定了，要在题里做得更好、更快、更便宜。这正是应用和改进那一层，正是中国四十年练得最熟的那一层。

第 4 章引过 OpenRouter 和 a16z 的一个数：这个平台上编程类的用量，从 2025 年初约 11% 升到了 50% 以上。那是开发者平台的样本，不代表全国。可它至少说明，写代码这门应用层里最硬的手艺，已经有一大块交给了机器。

于是「再」字要被重新算一遍。从前推迟「再」，先做的那一层还在升值；今天推迟「再」，先做的那一层本身在贬值。等的时间不再是中性的，等一年，手里的牌就薄一层。

这一笔，中篇二要接着算。

◆ 十一、一个完整的例子

举一个设想的例子：一家做软件外包的公司。

这样的公司，二十年前在很多城市里都有。接的是国外客户的活：需求文档写好了发过来，这边的工程师照着写代码、测试、交付。题是别人出的，做得又快又好，价格还便宜。

二十年里，这家公司练出了几千个工程师。当时有一种说法：先接外包，练好手艺，再做自己的产品。

后来有一部分人真的出去做了自己的产品，做成了应用层的公司。这是回流，是左边那一列。

更多的人留在原地，继续接下一份需求文档。「再做自己的产品」那句话，一直挂在年会的横幅上。

到了今天，一份需求文档交给编程助手，一个下午就能出第一版。这家公司最值钱的那一层手艺，正在变成一个按 Token 计价的东西。它当年推迟的那个「再」，现在不是晚了几年的问题，是那张等着兑现的票，面额在缩水。

这家公司没有做错什么。它在每一年都做了那一年最对的事。错位出在日程上：每一年最对的事加起来，没有一年是留给「再」的。

◆ 十二、放回日子里

回到老林的饭桌。

那晚豆豆睡下以后，林舟在书房里坐了很久。他想起自己的组：十二个人，三个基金本子，一年要出的论文篇数写在年初的计划表上。计划表里有「先」，每一行都是「先」。他想了一路，没找到哪一行是「再」。

明天能试：拿一张纸，把你手上的工作分成两栏。左边写「在别人给的题里做得更好」的事，右边写「在出一道新题」的事。右边那一栏，哪怕只有一行，给它写上一个日期，写上一个人的名字，哪怕这个人就是你自己。

一个小办法：到了那个日期，去问一句：做到哪一步了？

> **一个「再」字，没有日期，就只是一句客气话。**

> 带 走 的 话
>
> **四十年里钱回流过很多次，每次都流进下一个应用；从前等一年只是慢，如今等一年，手里的牌就薄一层。**

> 小　账
>
> **看见了什么，漏看了什么？**看见了四十年应用层的真本事，漏看了「再」字从来没有一行日程。
>
> **是在改进，还是在出题？**七章所记的成就，几乎都是在别人出好的题里做得更好；出题的那一下，很少发生在这片土里。

> 想 一 想
>
> 你家里或你单位里，有没有一句挂了很多年的「以后再说」？它有过一个日期吗？如果今天给它写一个日期，你会写哪一天？

◆



中　篇　二

层级：AI 接管了哪一层

导　语

中篇一算完了「先做应用」的账，留下一个问题：AI 来了以后，应用那一层还值钱吗？要回答它，先得把「新东西」分成几层。这一篇八章。第 17 章立三层：技术改进、架构跃迁、科学典范，并说清每一层从哪里来。第 18 到 21 章各举一层的来历和中国的位置：Transformer 出自一个翻译组，深度学习在冷宫里熬了几十年，AlphaFold 和「用 AlphaFold」之间隔着一层，再看中国前沿成绩里的题目是新是旧。第 22、23 章说 AI 接管了什么，以及连提问也不再稀缺。第 24 章把这一篇排成一张总账。

这一篇要让你看见的反差是：AI 接管的正是中国最强的那一层，留下的正是中国最空的那一层。

17

第 17 章

三层：改进、架构、典范

◆ 一、一个周四的晚上十一点

周四晚上十一点，林舟的组里还亮着三盏灯。

博后小顾刚跑完一轮虚拟筛选。几十万个候选分子，先用 AlphaFold 预测出靶蛋白的结构，再一个一个往口袋里对接，打分，排序。这一轮他改了打分函数里的两个权重，又换了一个更快的采样办法。前一百名里，能和已知活性数据对上的，比上个月多了一截。

他把结果截图发进组群：「林老师，这版好像能写进本子里。」

林舟在家里看到这条消息，回了一个大拇指。过了一会儿，他又打了一行字：「明天把对比实验补齐，周一组会过一下。」

放下手机，他坐在沙发上没动。宋岚在旁边看剪辑样片，问他怎么了。

「没什么。就是想，这一版要是投出去，审稿人会问什么。」

他心里其实知道审稿人会问什么：比基线高了多少，换了几个数据集，消融实验做没做。没有人会问：这一套办法本身，是谁出的题？

这一章要问的就是这件事：同样叫「新」，小顾今晚做出来的新，和当年 AlphaFold 本身的新，是不是同一种新？

◆ 二、同一个「新」字，三种重量

先把结论摆在桌上：科学和技术里的「新」，至少有三种重量。

第一种，是技术改进。题和办法都已经给定了，在里面做得更好、更快、更省。小顾改权重、换采样，就是这一种。

第二种，是架构跃迁。题还是那道题，办法的骨架整个换掉了。2017 年，翻译还是那件事，Transformer 把原来那种一个字一个字往后传的结构整个拿掉了，只留下注意力。

第三种，是科学典范。题本身变了，什么算好问题、什么算好答案、怎样算做对，都换了一套。深度学习就是这样一套：它让「让机器自己从大量数据里学出表示」，从一个被嘲笑的想法，变成了整个领域默认的做法。

三种都叫「新」，在论文里都写作「我们提出了」。它们的重量却差着数量级，来历也完全不同。

打个比方。你家的厨房，换一把更快的刀，是改进；把灶台从烧煤改成电磁炉，整副做饭的家伙都换了，是架构；而发现食物可以加热了再吃，是典范。你今天做的每一顿饭，都站在那第三种「新」上，只是没人再记得它是新的。

◆ 三、改进从约束里来

第一层从哪里来？从应用的约束里来。

算力不够，就逼出压缩的办法；成本太高，就逼出加速的办法；工期太紧，就逼出流水线。约束越硬，改进越狠。

中国在这一层极强。第 14 章写过 DeepSeek，在被卡算力的条件下把稀疏专家和注意力压缩做到极致。短剧公司把一部剧的制作周期大幅压短。药企把一轮筛选从几个月压到几周。这些都是改进，都是真本事。

改进有一个特点：它离钱最近。改进一出来，下个季度就能看到成本降了多少、速度快了多少。所以它最容易拿到钱，也最容易被考核。

它还有一个特点：它是一群人一起做得最好的事。约束摆在那里，几十个人分头去挤，每人挤出一点，合起来就是一大截。中国的工程团队，最擅长的就是这种合起来的一大截。

◆ 四、架构从飞地里来

第二层从哪里来？也从应用现场里来，不过来自现场里一块特殊的地。

那块地有一个条件：在一段时间里，不对产品指标负责。人还在应用的现场，天天碰得到真实的难处；可是不用每个季度交一个提升了几个点的数，可以把整副骨架拆掉重来。本书把这块地叫「飞地」。

Transformer 出自 Google 的机器翻译研究，翻译是一个再实在不过的应用。写出它的那几个人，在当时有一块飞地：撞上老结构的瓶颈时，他们可以不去修老结构，而是换一个。第 18 章细讲。

这一层，中国偶有。有的研究院、有的实验室，做出过让同行改用的新结构。不过飞地很难持续。一块地免于指标三年，到了第四年，考核就会回来问：这三年给业务带来了什么？

◆ 五、典范从冷宫里来

第三层从哪里来？从一个更冷的地方。

那里没有应用，没有经费，没有掌声，常常还逆着主流。一小群人在那里待很多年，有时候是几十年，做一件同行觉得走不通的事。

深度学习就是从那里长出来的。神经网络被冷落了几十年，Hinton 和少数几个人一直没放手。直到 2012 年，一场图像识别比赛，整个领域才回过头来。第 19 章细讲。

更早，地质学里的大陆漂移也走过这样一条冷路，第 36 章细讲。

这一层，在中国几乎缺席。不是没有聪明人，是很少有一个地方，容许一个人十几年、几十年拿不出成果，还能继续待下去。

◆ 六、一张表

把三层排成一张表：

层级	从哪里来	例子	中国的位置
技术改进	应用的约束：算力、成本、工期	MoE 与 MLA 做到极致；推理加速；筛选提速	极强，常在世界前列
架构跃迁	应用现场里一块免于产品指标的研究飞地	Transformer（出自翻译研究）；AlphaFold2 的整体结构	偶有，难持续
科学典范	长期反主流、没应用、没经费、没掌声的少数人	深度学习；板块构造	几乎缺席

这张表要读两遍。

第一遍从上往下读：越往下，离钱越远，周期越长，人越少，也越难被考核。

第二遍从下往上读：上一层的题，是下一层给的。没有深度学习这个典范，就没有 Transformer 这个架构；没有 Transformer，就没有今天这些改进可做。改进是在别人搭好的楼里装修，装得再好，楼是别人的。

所以这一章的判断是：

三层不是一架梯子的三级，是三块不同的地；改进做得再多，也不会自己长成典范。

◆ 七、三层要的是三种人

三块地不只要三种土，还要三种人。

改进那一层，要的是手快、功底扎实、能在约束里找到最后一点空间的人。这种人做事有章法，交得出数，带得了团队。他们在任何一个组织里，都是最被倚重的人。

架构那一层，要的人有点不一样。他也要有功底，可是更要有一种不安分：用着一副骨架，心里老在想它哪里根本不对。他要在应用的现场，也要能暂时不看指标。这种人在组织里，常常被看作「想法很多、落地太慢」。

典范那一层，要的人更少见。他认定的东西，大多数同行不信；他要在没有掌声的地方待很多年；他得忍受被当成怪人，甚至被当成走错了路的人。这种人在任何一个按年考核的地方，都很难待下去。

请注意，这三种人不是三个等级。一个人可以在改进那一层做得极好，而一辈子都不想去冷宫。这没有任何问题。

问题在于比例和许可。一个领域里，要是每一千个人中有一个愿意去冷宫，还有一块地能让他待下去，这个领域的下一个典范就有可能从这里长出来。要是那一个人一冒头，就被劝回了第一层，这个领域就只能等别人的典范。

下篇一会细讲，这种人是怎样被一道道筛子劝回去的。

◆ 八、三层之间怎样回写

三块地不是互不往来。它们之间有一条回写的路，方向要看清。

一个典范立住以后，会给出一大批题。这些题里，有的要换一副骨架才解得开，于是长出架构；架构一出来，又冒出成千上万个可以改进的地方，于是长出改进。这是从下往上的那一路，库恩把它叫作常规科学：在一个典范里一题一题地解。

还有一条从上往下的路。改进做得足够多、足够细，会撞上一些怎么改也改不掉的怪事。库恩把这些叫作反常。反常积得多了，有人会回头问：是不是题本身出错了？典范的松动，常常就从这里开始。

所以改进不是没用。它是一台很灵敏的探测器，能把典范里的裂缝一条条照出来。

问题在于，照出裂缝以后，谁去接。一个只奖励改进的地方，看到裂缝的第一反应是绕过去，换个数据集，换个评测，再挤出一个点。裂缝被照出来了，又被绕过去了。

你在自己的工作里，大概也碰到过这样的裂缝：一个怎么调都调不好的怪现象，最后在报告里写成「误差」。它可能真的是误差，也可能是一道新题在敲门。

◆ 九、只看一面会怎样

这张表也容易读歪。

一种歪法，是看不起第一层。改进不是低级的活。没有改进，典范永远停在论文里，走不进工厂和医院。很多年里，中国工程师把别人的想法做到能用、好用、便宜到人人用得起，这件事本身就改变了世界。

另一种歪法，是以为三层可以一级一级往上爬：改进做多了，自然就到架构；架构做多了，自然就到典范。中篇一已经算过这笔账，四十年的改进，没有自己爬上去。

公道的读法是：三层都要，各有各的土。问题不在中国第一层太强，在后两层的土太薄。

你要是在第一层做得很好，不必觉得低人一等。你只需要知道，自己脚下是哪一块地，旁边那两块地，要的是什么样的土。

◆ 十、今天与机器

这张表放在今天，有一处要特别看。

AI 正在接手的，首先是第一层。改权重、换采样、调结构、跑消融实验、写对比表，这些活，越来越多地交给机器去做。小顾花了三个晚上做出的那一版，再过一两年，也许一个智能体跑一夜就能做出十版。

第 22 章会细讲常规科学怎样交给机器，第 23 章会讲连提问也不再稀缺。这里先记下一个形状：AI 接手的顺序，是从表的最上面一行往下压。最上面一行，正好是中国最满的那一行；最下面一行，正好是中国最空的那一行。

第 369 号《典范发生学》说过，AI 能做给定典范下的常规科学，人被推到典范发生那一头去。放进这张表里看，就是人被从第一行往第三行推。推的力气越来越大，第三行的地却还没整出来。

◆ 十一、一个完整的例子

把三层放进同一个领域里，看得更清楚。就拿天气预报。

几十年来，天气预报靠的是把大气的物理方程放进超级计算机里一步一步算。改进这一层的活多得很：网格加密，算法提速，观测数据同化得更好。每一点改进，都让预报准一点、快一点。

2023 年，华为的盘古气象模型发表在 Nature 上。它不再一步一步解物理方程，而是让一个深度网络从几十年的历史资料里学出天气怎样演变。预报速度比传统办法快了几个数量级，这是实打实的成绩。

它落在表里哪一格？本书把它放在第一层的最上沿。对气象这一行来说，从解方程换成学数据，看上去像换了一副骨架；可是网络的骨架是借来的，什么算好预报、拿什么尺子打分，也都没动。第 21 章会细算这笔账。

它站在什么上面？站在深度学习这个典范上，也站在 Transformer 一类的架构上。这两样都不是它造的。

所以盘古气象是一份公道的样本：中国能在第一层做出世界级的东西，前提是上面两层已经有人铺好了。下一个让天气预报换一套典范的想法，会从哪里来？这个问题，表里那一格还空着。

◆ 十二、放回日子里

回到那个周四的晚上。

第二天上午，小顾把对比实验补齐了。周一组会上，林舟让他讲完，然后多问了一个平时不问的问题：「这一版里，哪一处是我们自己出的题？」

会议室里安静了一会儿。小顾想了想，说：「好像没有。题是别人的，结构是别人的，我们改的是参数和流程。」

林舟点点头，没有批评他。他说：「这没错，本子也照这样写。我只是想让大家知道，我们站在哪一层。」

明天能试：拿出你最近做的一件得意的事，问它三个问题。它是在别人的题和办法里做得更好吗？它换掉了办法的骨架吗？它换掉了题本身吗？答案多半是第一个，这很正常。知道自己站在哪一层，才谈得上往哪一层走。

> **三种「新」都写作「我们提出了」，重量却差着几个数量级。**

带 走 的 话

改进来自约束，架构来自飞地，典范来自冷宫；三层各有各的土，不是一架梯子。

小　账

看见了什么，漏看了什么？看见了改进这一层的真本事，漏看了另外两层需要的是完全不同的土。

是在改进，还是在出题？小顾那一版是在别人的题里做得更好；这一章自己要做的，是把「新」分层，好让人看清哪一层没人在出题。

想 一 想

你这一行里，最近十年最大的一次「新」，属于哪一层？它是在哪一块地上长出来的？你所在的地方，有没有一块哪怕很小的飞地？

◆

18

第 18 章

Transformer 出自一个翻译组

◆ 一、你一定用过的翻译框

你一定有过这样的时刻：一封英文邮件，一页说明书，一段看不懂的论文摘要，你把它贴进翻译框，按一下，中文就出来了。

十几年前，这样出来的中文常常前言不搭后语。一个长句子翻到后半截，前半截说了什么，机器好像已经忘了。今天，翻出来的东西通顺得多，有时候比人译得还顺。

这中间的变化，大半要算到一篇 2017 年的论文头上。题目只有一句话：《Attention Is All You Need》，意思大致是「注意力就够了」。

今天的大模型，从写作文的到筛分子的，骨架几乎都是这篇论文搭起来的那一种，叫 Transformer。

这一章要讲一件常被忽略的事：这副改变了整个行当的骨架，不是从一个远离应用的象牙塔里来的。它出自一个做机器翻译的研究组。翻译，是再实在不过的一个应用。

◆ 二、翻译把人逼到了墙角

先看当时的墙。

2017 年以前，机器翻译的主流办法，是一种叫循环神经网络的结构，简称 RNN。它读句子的样子，很像一个人用手指着字往下念。读完第一个字，记下一点什么，带着这点记忆去读第二个字，再带着去读第三个。

这个办法有两处难处。

一处是记性。句子一长，前面的字传到后面，已经淡得差不多了。翻译偏偏最怕这个：主语在句首，动词在句尾，中间隔着二十个字。

另一处是快慢。一个字一个字往后传，前一个没算完，后一个就不能开始。显卡最擅长的是很多事一起算，这种结构却逼着它排队。数据越多，排的队越长。

这两处难处，都是翻译这个应用推到人面前的。要翻得准，要训得快，要喂进去更多的语料，每一条都撞在同一堵墙上。

◆ 三、不修墙，换房子

面对这堵墙，有两种做法。

一种是修墙。给 RNN 加一个更好的记忆门，加一道注意力当拐杖，把排队的环节压得更短。当时很多组都在这样做，也做出了好结果。这是改进，是第 17 章说的第一层。

另一种是换房子。既然一个字一个字往后传是病根，那就不传了。让句子里每一个字，一下子同时去看所有别的字，自己算出该多看谁、少看谁。这就是「自注意力」。

《Attention Is All You Need》走的是第二条路。它把循环的那一套整个拿掉，只留下注意力，再把这一层一层叠起来。这样一来，记性的难处少了一大半，所有的字也可以一起算，显卡终于不用排队了。

据后来的报道，组里先有人提出：能不能干脆不要循环，只靠注意力？当时不少人觉得这条路走不通。论文在当年的翻译评测上，用更少的训练成本，拿到了更好的成绩。

请注意这件事的形状：题还是翻译，评测还是那几套英德、英法的数据。办法的骨架，却整个换了。

◆ 四、那块飞地

为什么这个组能换房子，而很多组只能修墙？

本书的回答是：那个组在当时有一块飞地。

这块飞地有三个条件。第一，人在应用的现场，天天碰得到翻译的难处，知道墙在哪里。第二，他们不对产品指标负责，没有人要求他们这个季度把线上翻译的质量提几个点。第三，他们有足够的算力和时间，去试一条可能走不通的路。

三个条件缺一个，都换不了房子。

只有第一条，会修墙：天天被指标追着，最划算的永远是在现有结构上再挤一点。只有第二条，会飘：离开了真实的难处，换出来的房子没人住。只有前两条、没有第三条，会停在纸上：想法有了，跑不起来。

所以架构跃迁可以发生在应用现场。它的条件不是离开应用，是在应用里面，划出一块暂时不交指标的地。

◆ 五、要是他们在一个按季度交数的组里

做一个反面的想象。

假如 2016 年，同样这几个人，坐在一个按季度交数的翻译产品组里。季度目标是：线上翻译的质量分提高一个点。

最划算的做法是什么？是在现有的 RNN 上再加一道注意力，再多喂一批语料，再调一调解码的办法。这样做，一个季度大概真能交出一个点。

要是有人说：我们把循环整个拿掉试试？组长会算一笔账：拿掉以后，新结构要从头训，头几个月的分数多半比现在低。这个季度的一个点，就交不出来了。

于是那个想法，被写进了「下个季度再看」。下个季度，又有下个季度的一个点。

你在工作里大概见过这样的「下个季度再看」。它和第 12 章那个「再」字，是同一个字的两种写法。

◆ 六、先把公道还给 Google

说到这里，要先还 Google 一笔公道。

后来的事情，很多人都知道一些：论文的八位作者，后来陆续离开了 Google，有的去创办了自己的公司。最早把这副骨架做成聊天产品、做出巨大声势的，也不是 Google，而是另一家公司。

有人据此说，Google 没守住自己的发明。这话有一半对。不过另一半更要紧：Google 养出了那块飞地，让这篇论文能写出来、发出来，而且公开发表，谁都能用。

这块地是花钱养出来的，而且这笔钱在当时看不到回报。一家公司肯为看不到回报的研究开一块地，本身就不寻常。贝尔实验室当年出了晶体管和信息论，也是同一种不寻常。尾篇第 43 章会回到这块地。

◆ 七、一副骨架之后的十年

Transformer 发表以后，接下来的几年，几乎整个行当都搬了家。

2018 年，Google 又用这副骨架做出了 BERT，用来读懂文字。同一年前后，另一家公司用它做出了最早的 GPT。后来的大模型，越做越大，可是骨架没有大变。图像、语音、蛋白质，一个个领域也陆续改用它。

这十年里，全世界有成千上万个团队在这副骨架上做改进。让注意力算得更省，让模型训得更稳，让上下文拉得更长。中国团队在这里做得非常出色，第 14 章写过的 MLA，就是在注意力这一处做出的世界级改进。

请把这件事和第 17 章那张表对一对。一副骨架出来，给后面十年的人出了成千上万道题。出骨架的是一个组；解题的是整个行当。

你要是在这十年里做过大模型，你做的每一道题，题面上都印着那篇 2017 年的论文。这不丢人。只是要记得，题面是谁印的。

◆ 八、中国的大厂有没有这块地

现在问一个公道的问题：中国的大厂，有没有这块飞地？

先说有的一面。阿里、腾讯、百度、华为、字节，都建过研究院或实验室，招了很多一流的研究者，发了大量顶会论文。华为的盘古气象模型 2023 年发表在 Nature 上，是实打实的成绩。DeepSeek 背后的团队，也有很强的研究气质，肯做长线的事。

再说难的一面。多数研究院的考核，最终要落到业务上：给哪条产品线带来了什么，降了多少成本，增了多少收入。研究者可以发论文，年终却要回答「对业务有什么用」。一块地头两年免于指标，第三年考核就回来了。

还有一种更隐蔽的压力：方向由谁来定。研究院的方向，常常跟着公司的战略走。战略今年押视频，明年押智能体，研究方向也跟着换。换一次，前面积下的那一点，就散了一次。

所以本书的判断是：中国的大厂有研究院，有过飞地，也出过好东西；可是很少有一块地，能让一群人在三五年里不交指标，专门去换一副骨架。不是没有人想换，是那块地太容易被收回去。

◆ 九、怎样认出一块飞地

一块飞地，从外面看不太出来。它不挂牌子，也不一定叫研究院。认它，可以看三个信号。

第一个信号：有人可以说「这一年我不交数」，而且不会因此被清退。不是口头上允许，是年底考核真的不看那个数。

第二个信号：失败可以被讲出来。一条路走了一年没走通，可以在组会上从头讲一遍，讲完了大家记下这条路走不通，而不是记下这个人不行。

第三个信号：方向不跟着每年的战略换。今年押视频、明年押智能体，那是业务的事；飞地里那件事，三年里只问一个问题：往前走了没有？

三个信号都有的地方，很少。有一两个的，已经不错。一个都没有的地方，叫研究院也好，叫实验室也好，做的多半还是修墙。

你可以拿自己所在的地方对一对。对完了不必灰心，大多数地方本来就该修墙。只是要知道，换房子的事，不会在这里发生。

◆ 十、只看一面会怎样

这一章容易读出两种偏。

一种偏，是以为只要钱够、人够，架构跃迁就会来。钱和人是必要的，只是不够。同样的钱，放在一个每季度考核的组里，买到的多半是更好的修墙。

另一种偏，是以为飞地就是离开应用，躲进书斋。Transformer 的故事正好反过来：飞地在应用里面。人离开了翻译的难处，就不知道该换哪副骨架。

两种偏都漏看了一件事：架构跃迁要的是一种很具体的组合，在难处旁边，又不被难处的指标拴住。

◆ 十一、今天与机器

这块飞地，在 AI 时代会变得更重要，也更难养。

更重要，是因为修墙的活正在交给机器。在现有结构上调一调、挤一挤，是 AI 最擅长的事。第 22 章会细讲。人要是还只会修墙，很快就没有墙可修。

更难养，是因为今天的研究，离钱比 2017 年近得多。一个新结构出来，几个月就能变成产品，产品又立刻变成指标。指标回来得越快，飞地被收回得越快。

于是有一个不太舒服的局面：最需要飞地的时候，飞地最容易被收走。

◆ 十二、一个完整的例子

举一个设想的例子。一家大厂的语音组，负责语音识别。

这个组每个季度都有一个数：识别错误率要降多少。组里最聪明的一个年轻人，发现现有结构在嘈杂环境里有一个绕不过去的毛病。他想试一种完全不同的结构，可能三年都赶不上现有系统，也可能三年后整个行当都改用它。

在一个没有飞地的组里，组长会对他说：先把这季度的数做完，那个想法业余时间试试。业余时间是晚上十一点以后。三个月后，那个想法还停在一页草稿上。

在一个有飞地的组里，组长会说：给你两个人、一年的算力，这一年不考核错误率，年底你给大家讲一讲走到了哪里，走不通也讲。

两种组都不坏。第一种组，每个季度都交出了好数。第二种组，大概十次里有九次什么也换不出来。而那一次换出来的，会让第一种组接下来十年的数，都建在它上面。

◆ 十三、放回日子里

回到林舟的组。

那周的组会以后，林舟算了一笔账。十二个人，如果拿出一个人，一年不写本子、不赶论文，专门去想现有筛选办法的骨架哪里不对，组里的论文会少几篇？

他算出来大概少两篇。他又算了一下，少两篇，明年的基金评审会不会受影响。算到这里，他停下了。

明天能试：在你的团队里，找出一件「修墙修了很多年」的事。问一句：要是不修了，换一副骨架，会是什么样？哪怕只是在白板上画半个小时，不交任何指标。

> **架构跃迁不在书斋里，在难处旁边那块暂时不交数的地上。**

带 走 的 话

翻译把人逼到墙角，飞地让人有胆子换房子；没有飞地，再难的应用也只逼得出更好的修墙。

小　账

看见了什么，漏看了什么？ 看见了 Transformer 是从应用里长出来的，漏看了它要的是应用里那块不交指标的地。

是在改进，还是在出题？ 修 RNN 是在旧题里做得更好；Transformer 没有换题，却换了解题的整副骨架，给后来所有人出了新的题。

想 一 想

你所在的地方，一个想法从提出到被要求「对业务有用」，中间隔多久？如果给你一年不交数，你最想拆掉哪一副骨架？

◆

19

第 19 章

冷宫里的三十年

◆ 一、一张不肯撤的海报

林豆豆的学校办科技节。她和同学做了一张海报，讲「机器是怎么认出猫的」。

海报是她和 AI 一起做的。左边画一只猫，右边画了很多小圆圈，用线连起来，写着「神经网络」。

周日晚上，她拿着海报来问林舟：「爸爸，这个神经网络是谁发明的？」

林舟想了想，说：「很多人。不过有一个人，在别人都不信的时候，一直没撤。」

「没撤什么？」

「没撤他那张海报。」

林豆豆没听懂。林舟也没再解释。他想起自己读博士时听过的一个故事。一位研究者，几十年里在会议上讲神经网络，台下的人越来越少；有些年头，连论文都很难发出去。

这一章，讲的就是那几十年。

◆ 二、被冷落的一条路

神经网络的想法很早就有。大意是：不给机器写死规则，让它模仿大脑里神经元连成网的样子，从例子里自己学。

1950 年代末，这条路热过一阵。到了 1969 年，Minsky 和 Papert 写了一本叫《感知机》的书，指出当时那种简单网络连一些很基本的问题都解不了。这本书影响很大，很多经费和人从这条路上撤走了。

后来的几十年，人工智能的主流几度换手，大多数时候都不在神经网络这一边。这条路被冷落，有起有伏，冷的时候居多。

Geoffrey Hinton 就是那几个没撤的人之一。他在英国读博士，研究的就是这条被冷落的路。后来他去了加拿大，在多伦多大学一待就是很多年。

◆ 三、几个年份

把几个年份排在一起：

1986 年，Rumelhart、Hinton、Williams 发表了反向传播的论文。大意是：让网络从输出的错误往回算，一层一层修正连线的权重。这个办法让多层网络能被训练起来。

论文发表以后，神经网络并没有就此翻身。数据不够，算力不够，别的办法在很多任务上做得更好。冷宫的门开了一条缝，又关上了。

2012 年，Hinton 和他的两个学生做出了 AlexNet，在 ImageNet 图像识别比赛上，把错误率一下子压低了一大截。整个领域这才回过头来。

2018 年，Hinton 与 Yoshua Bengio、Yann LeCun 一起获得图灵奖。

2024 年，Hinton 与 John Hopfield 一起获得诺贝尔物理学奖。

从反向传播到 AlexNet，是二十六年。从他读博士研究神经网络，到 AlexNet，是三十多年。

◆ 四、典范是在没应用的时候长的

请注意这张年表里，最长的那一段是什么样子。

那一段里，深度学习没有应用。没有产品等着它，没有工厂用它，没有哪个季度的财报因为它变好看。它只是一小群人认定的一条路。

这一小群人在那几十年里做的事，就是第 17 章说的第三层：一套新的典范，什么算好问题、什么算好答案、怎样算做对，都换了一套。从前的主流认为，智能要靠人把规则和知识写进去。这一小群人认为，智能可以从大量例子里自己长出来。

这样一套典范，是在没有应用的时候长出来的。等到应用来了，它已经长了几十年。

本书要把这句话说得重一点：

深度学习这个典范，是在冷宫里长成的；应用不是它的土，是它的果。

◆ 五、先把公道还给冷宫之外

说冷宫之前，先还一笔公道。

冷宫不是一个人熬出来的。Hinton 在加拿大能熬下去，靠的是一套肯为长线研究出钱的安排。据报道，加拿大高等研究院（CIFAR）在 2000 年代资助了一个神经计算的项目，把 Hinton、LeCun、Bengio 这些人拢在一起，定期开会。这笔钱不大，不过它不问「明年能用在哪里」。

还有一笔公道要还给冷宫里的反对者。当年不信神经网络的人，不是傻子。在数据和算力都不够的年代，别的办法确实做得更好。他们的判断，在当时的条件下是合理的。

所以冷宫的故事，不是「坚持就会胜利」。很多人在冷宫里熬了一辈子，什么也没等来。冷宫的价值在于：它让那极少数对的人，有机会熬到条件到来的那一天。

◆ 六、要是他在一个按年考核的地方

现在做一个思想实验。

假如 1990 年代的 Hinton，不在多伦多，而是在一个每年考核论文篇数与成果转化的体系里，会怎样？

第一年，他要报一个课题。课题书上要写清预期成果、应用前景、考核指标。他写「让机器从例子里自己学出表示」，评审会问：做出来能用在哪里？三年内能转化吗？

第二年，他发了两篇论文，却发在不那么热门的地方，因为热门的地方不太收神经网络的稿。年终考核，他排在后面。

第三年，他的学生毕业找工作，发现做神经网络的岗位很少。学生改了方向。

第五年，他自己也许还在坚持，组里却只剩他一个人。到了要评职称、要续课题的时候，最理性的选择是：换一个热门方向，先保住位置，以后再说。

又是那个「以后再说」。

这不是在说哪一个体系坏。按年考核的体系，在常规科学里非常有效：它让每一个人都在已知的题里拼命往前跑。它只是对冷宫不友好。冷宫要的，是一个人好几年拿不出东西，也不被清退。

◆ 七、冷宫和走错路，只隔一层纸

这里要打一个预防针。

冷宫里的人，和走错了路的人，在当时看起来是一样的。都是多年没有成果，都是同行不信，都是自己坚持说「这条路是对的」。

Hinton 在 1990 年代，站在会场上讲神经网络的样子，和一个坚持了三十年永动机的人，从外面看，分得开吗？

很难分开。分开他们的，是后来的事：条件到来以后，一条路走通了，另一条永远走不通。可是后来的事，在当时是看不见的。

所以一个容得下冷宫的地方，一定也会养着一些走错路的人。这是冷宫的代价。一个想要冷宫、又不想付这个代价的地方，最后两样都得不到。

这件事，下篇二第 35 章还要回来细说：一张只看身份、不看内容的标签，会把冷宫里的人和走错路的人一起归类掉。

◆ 八、「板凳要坐十年冷」

中国学界并不缺这种说法。

史学家范文澜有一句常被引用的话：「板凳要坐十年冷，文章不写一句空。」很多老一辈的学者，把它抄下来挂在书房里。

这句话说的，正是冷宫的样子。十年不求热闹，一句不说空话。它在中文里流传了几十年，人人都会背。

可是请把这句话放进今天的一张考核表里，看看会怎样。十年冷板凳，对应的是十年的年度考核；一句不写空，对应的是每年规定的论文篇数。表格里没有一栏，能填「这一年我坐在冷板凳上」。

于是这句话从一条做学问的规矩，慢慢变成了一句只在退休致辞里说的话。说的人是真心的，听的人也感动，第二天照样填表。

这和第 12 章那个「再」字，是同一个形状：一句人人都认的话，没有排进日程。

◆ 九、中国的土里有没有冷宫

那么，中国的土里有没有冷宫？

公道地说，有过，也在长。有的院所给过学者很长的周期。有的基金开始设「非共识」项目，专门资助评审意见不一致的申请〔待核〕。这说明体制自己也看见了问题，评价上的松动，第 28 章会细看。

不过冷宫要的不只是一笔钱。它要的是一整套许可。许可一个人十年不出热门论文；许可他的学生找得到工作；许可他在评审会上被大多数人否定以后，还能拿到下一笔不大的钱。

这一整套许可，在一个由篇数、影响因子、人才帽子、年度考核编织起来的网里，很难长出来。每一根线都不坏，合在一起，就织成了一张没有冷宫的网。

下篇一会细讲这张网怎样筛人。这里只先记下一个形状：网织得越密，常规科学跑得越快，冷宫越难有立足之地。

◆ 十、今天与机器

AI 来了，冷宫的问题变得更尖。

第一，常规科学正在交给机器。在给定的典范里解题，AI 越来越在行。人要是只会在给定的典范里跑，跑得再快，也快不过机器。人还剩下的，越来越像冷宫里的那种活：认定一条别人不信的路，在没有应用的时候让它长出来。

第二，AI 让冷宫里的人不再那么孤单。从前一个人在冷宫里，没有学生、没有算力、没有同行，几乎寸步难行。今天，一个人加一台 AI，能做从前一个小组才做得完的事。资源的墙在变矮。

可是另一堵墙还在：一个人认定的那条路，有没有人肯让他走下去。第 34 章会细讲这堵墙。

◆ 十一、一个完整的例子

再看一个离我们更近的例子：一个想做「慢题」的年轻人。

设想一位三十二岁的青年研究员，刚回国，拿到一个不大的启动经费。他认定，现有的分子模拟办法有一个根上的假定是错的，要换一套描述办法。走通了，整个计算化学都要改写；走不通，就是五六年什么也没有。

他所在的单位，对新人有一个考核：五年内要评上副高，否则不再续聘。评副高看的是论文篇数和课题级别。

有一个晚上，最能看清他的处境。

那是他回国后的第一个十二月，单位要交年度考核表。表上有五栏：发表论文、主持课题、获得奖励、人才培养、其他。他在电脑前坐到半夜。第一栏填了一篇，是读博士后时的旧稿；第二栏填了那笔启动经费；后面三栏，空着。

他在「其他」一栏里打了一行字：「正在尝试一种新的分子描述办法，目前尚无可发表的结果。」打完，他看了很久，又一个字一个字删掉了。这一行写上去不会加分，只会让人记住，他今年交了半张白卷。

其实这一年，是他想得最多、走得最远的一年。他推翻了自己的两个草稿，找到了旧办法在哪一步偷偷做了假定。这些，表上没有一栏装得下。他甚至说不

清该把它算作「发表」还是「其他」：它还没有变成任何一页能交给别人看的东西，只是他脑子里一张越画越清楚的草图。

第二天，他把表交了上去。科室主任看了一眼，说了一句很客气的话：「第一年嘛，明年多发几篇。」

也是那天晚上，他认真算了一回自己的路。

他面前有两条路。

一条是把那个想法放进抽屉，先做几个用现有办法就能出结果的课题，五年发十几篇，评上副高，再把抽屉打开。

另一条是现在就做那个想法。五年后，也许他已经走出了半条路，也许还在原地；论文篇数却不够，合同到期了。

大多数聪明人会选第一条。这是理性的选择，也是负责任的选择，他还要养家。

五年后，他评上了副高。他打开抽屉，发现那个想法已经有点陌生了。他这五年练出来的，是怎样把一个现成的办法用得又快又好。那个抽屉，他又关上了。

这个例子里，没有一个人做错事。考核的人在守规矩，他自己在为家人负责。一间冷宫，就这样在没有任何人反对的情况下，没有长出来。

◆ 十二、放回日子里

那晚林豆豆睡了以后，林舟又想起那张海报。

他想，假如自己的组里出了一个想走冷宫路线的人，他会怎么办？他会劝这个人先发两篇论文、先评上副研究员、以后再说。他会这样劝，是因为他真心为这个人好。

他意识到，这张没有冷宫的网，有一根线就织在他自己手里。

明天能试：如果你带学生、带团队，找一个你觉得「太偏」的想法，问提出它的人一句：要走通它，你需要多少年？然后认真听完，不急着劝他换方向。

冷宫不是坚持就会胜利的地方，是让极少数对的人熬到条件到来的地方。

带 走 的 话

深度学习在没有应用的几十年里长成；一张只认年度成果的网，织得再好，也织不出一间冷宫。

小　账

看见了什么，漏看了什么？看见了按年考核让常规科学跑得快，漏看了典范要的是一个能好几年不出成果的地方。

是在改进，还是在出题？Hinton 那几十年不是在旧题里做得更好，是在出一道当时没人认的新题。

想 一 想

你身边有没有一个被认为「太偏」的人？如果他是对的，需要多少年才看得出来？你所在的地方，容得下这么多年吗？

20

第 20 章

AlphaFold 和「用 AlphaFold」

◆ 一、本子上的那一行

林舟的基金本子写到第三页，有一行他改了五遍。

那一行写的是技术路线：「基于 AlphaFold 预测的靶蛋白结构，结合大模型对候选分子进行虚拟筛选。」

第一遍他写的是「采用」，第二遍改成「基于」，第三遍加上「改进的」，第四遍删掉「改进的」，第五遍又加了回去。

宋岚端着一杯水路过，看了一眼屏幕：「一个词改这么多遍？」

「评审要看这个。说『采用』，显得没新意；说『改进的』，又得说清楚改进了哪里。」

宋岚笑了：「我们剪片子也这样。素材都是 AI 生成的，可是标题得说是我们原创的。」

林舟也笑了，笑完以后，他盯着那一行看了很久。

这一章就从这一行讲起：用 AlphaFold，和造出 AlphaFold，中间隔着什么？

◆ 二、五十年的难题

先说 AlphaFold 解的是什么题。

蛋白质是一串氨基酸连成的链。这条链会在几乎一瞬间折叠成一个三维的形状，而这个形状决定了它在身体里做什么。知道了形状，才好设计药物去对准它。

难的是，从一串氨基酸的顺序，推出它折成什么形状，难极了。过去要用 X 射线晶体学、冷冻电镜这些实验办法去测，一个结构常常要花几个月甚至几年。

1994 年起，有一个叫 CASP 的比赛，每两年一届，让各个团队预测一批还没公布的蛋白质结构，再拿实验结果来比。几十年里，成绩一点一点往上挪，离实验的精度一直差着一截。

2020 年，CASP14。DeepMind 的 AlphaFold2 交出的预测，在很多目标上已经接近实验的精度。比赛组织者说，这个困扰了生物学五十年的问题，在很大程度上被解决了。

2024 年，DeepMind 的 Demis Hassabis 与 John Jumper，和做蛋白质设计的 David Baker，一起获得诺贝尔化学奖。

◆ 三、AlphaFold 背后是什么

AlphaFold 不只是一个软件。

它背后，是 DeepMind 自己定义的一道题和一套解法。他们把结构预测当成一个能端到端学习的问题。进化留在序列里的痕迹、氨基酸两两之间的关系、三维空间里的几何，都被焊进了一副专门设计的网络结构里。

这副结构，在第 17 章那张表里属于第二层，是架构跃迁。而它所依靠的那一整套「从数据里学出表示」的做法，属于第三层，是深度学习那个典范。

所以 AlphaFold 是一座楼：地基是深度学习的典范，楼体是 DeepMind 设计的架构，楼里装的是五十年 CASP 积下来的题和数据。

DeepMind 在 2021 年公开了代码，又和欧洲生物信息研究所一起，把预测出来的大量结构放进一个公开的数据库，后来扩展到两亿多个。全世界的研究者，都可以免费走进这座楼。

◆ 四、另一位得奖人走的路

同一年和他们分享诺贝尔奖的 David　Baker，走的是另一条路，值得放在旁边看。

Baker 在华盛顿大学做了几十年蛋白质的计算研究。他的组做了一个叫 Rosetta 的软件。它最早是为了预测结构，后来越来越多地用来反过来设计。先想好要一个什么形状、做什么事的蛋白质，再去算出能折成这个形状的氨基酸序列。

这件事在很长时间里被看作难以做成。自然界的蛋白质，是几十亿年进化出来的；人从头设计一个，听起来像是要和进化比赛。

Baker 的组做成了一批自然界没有的蛋白质。AlphaFold2 出来以后，他们很快做出了自己的结构预测工具 RoseTTAFold，又把深度学习用到设计上。

请注意他的形状。他用了别人的典范，也在 AlphaFold 出来以后立刻接住了它。不过他手里一直握着一道自己的题：从头设计。这道题是他几十年前就认定的，那时候没有多少人觉得能做成。

所以「用」和「造」不是两种人。同一个人可以用得很勤，同时造着自己的东西。区别在于，他手里有没有一道自己的题。

◆ 五、走进楼里的人

很多人走进了这座楼，包括中国的研究者，而且走得很勤。

这几年，国内大量的课题、论文、基金本子里，都能看到这样一行：用 AlphaFold 预测某某蛋白的结构，然后做某某事。筛分子，找靶点，解释突变，设计抗体。

林舟的组正是这样。他们的日常，就是第 17 章开头那个晚上：用 AlphaFold 出结构，用大模型筛分子，改打分函数，补对比实验，发论文，写本子。

这是在别人的典范里做应用。题是别人出的，楼是别人盖的。林舟的组在楼里装修，而且装得很好。

◆ 六、用好别人的工具是正当的科研

这里要先把公道写足。

用好别人的工具，是正当的科研，而且是科研的大部分。牛顿说过站在巨人的肩膀上，大意如此。没有人从烧砖开始盖房子。

用 AlphaFold 做出好成果的中国团队，不在少数。有的找到了新的药物靶点，有的解释了一些疾病突变的来由，有的把筛选的效率提了好几倍。这些成果会变成新药、新的诊断办法，最后落到病人身上。

更要紧的是，在一个典范定下来以后，大量的工作本来就是常规科学：把典范用到一个又一个具体的问题上。库恩 1962 年在《科学革命的结构》里说过，常规科学是科学的常态。它不低人一等，没有它，典范只是一纸空文。

所以林舟的组没有做错什么。他们做的，是每一个领域里都需要有人做的那一部分。

◆ 七、可是用与造之间隔着一层

公道写完，再说隔着的那一层。

用 AlphaFold 和造 AlphaFold，差的不是聪明程度，也不只是算力。差的是谁在出题。

造的人，要先认定一件当时没人做成的事是能做成的；要决定把结构预测当成什么样的问题；要设计一副前所未有的结构；要在几年里一次次失败，还能继续拿到资源。

用的人，题已经出好了：结构可以预测。剩下的问题是：拿这个结构去做什么。

这两种活，要的土完全不同。造，要第 18 章说的飞地，甚至第 19 章说的冷宫。用，要的是快、准、多、便宜，要的是第 17 章那张表最上面那一行的本事。

中国的科研体系，非常擅长培养后一种本事。一个课题，三年结题；一篇论文，几个月见刊；一个本子，写清楚基于什么、改进了什么、预期发几篇。这套节奏，正好配得上「用」。

「造」却不在这个节奏里。一个人要是在本子上写：「本课题拟重新定义某某问题，预期三年内可能没有可发表的成果」，这个本子大概很难通过。

◆ 八、宋岚的那句玩笑

回头看宋岚那句玩笑：「素材都是 AI 生成的，可是标题得说是我们原创的。」

这句话放在林舟的本子上，一样成立。结构是 AlphaFold 预测的，分子是大模型筛的，本子上却要写清楚「本课题的创新点」。于是「创新点」常常落在流程上：换了一种组合，加了一步过滤，在一个新的蛋白上跑了一遍。

这些创新点是真的，只是它们都在别人盖好的楼里。评审也知道，申请人也知道，大家默认这样写。

请注意这个默认。它不是哪一个人的不诚实，是整个评价体系对「新」字的一种约定：在别人的楼里换一间屋子装修，也算新。

这个约定在 AI 以前很好用。它让成千上万个组都有活干、有本子可写、有论文可发。它的代价是，「造楼」和「装修」用同一个词来记，账上分不出来。

你下次读到「本研究首次」这几个字，可以停下来问一句：首次的是楼，还是屋子？

◆ 九、只看一面会怎样

这一章也有两种歪读法。

一种是轻视「用」：觉得用别人工具的都是二流。这一读不公道，也不合事实。世界上大多数好的科研，都是在用别人的工具。

另一种是满足于「用」：觉得工具既然是公开的，谁用都一样，用得好就是本事，造不造无所谓。这一读漏看了一件事：造的人，同时握着下一道题。AlphaFold 之后，DeepMind 又往蛋白质与别的分子的相互作用走，又往更大的生物学问题走。下一座楼盖在哪里，是造的人先知道。

用的人永远晚一步。晚一步的代价，在 AI 以前不算太大。在 AI 以后，会变得很大。

◆ 十、今天与机器

为什么在 AI 以后，晚一步的代价会变大？

因为「用」这一层，正在被机器接手。用 AlphaFold 出结构，用模型筛分子，改打分函数，跑对比实验，写论文初稿，这一整条流水线上的大部分环节，智能体都已经能做，或者很快能做。

据报道，已经有团队在做能自己设计实验、自己跑筛选、自己写报告的科研智能体〔待核〕。第 22 章会细讲。

那时候，林舟的组十二个人一年做的事，也许一个人加几个智能体一个月就能做完。组还在，组最值钱的那一层手艺，却在变成一个按 Token 计价的东西。第 16 章那家外包公司的形状，会在实验室里再现一遍。

而楼，还是别人盖的。

◆ 十一、一个完整的例子

举一个设想的例子，把两种路并排放。

两个组都做同一种难治的病，都有很好的人。

第一个组的路是：拿 AlphaFold 预测这种病相关的几十个蛋白，挑出最可能成药的几个，筛几百万个分子，选出几十个去做实验。三年里，发了八篇论文，有两个分子进了临床前研究。这是一份漂亮的成绩单。

第二个组的路是：他们觉得，这种病的难处不在某一个蛋白的形状，而在蛋白之间一种还没人描述过的动态关系。现有的工具都是为静态结构设计的，用不上。他们花了三年，试着定义这个新问题，做一个新的模型。三年里，只发了两篇论文，有一篇还被很多人质疑。

在今天的评价体系里，第一个组赢得毫无悬念。

假如第二个组是对的，十年以后，第一个组那样的工作，都会搬到第二个组定义的新问题上去做。而那时候，做这些工作的，也许已经主要是机器。

本书不是说第二个组一定对。十个这样的组里，大概九个是错的。本书要说的是：在一个只奖励第一种路的体系里，那第十个对的组，连出现的机会都很少。

◆ 十二、要是这座楼要在这里盖

最后做一个想象：要是 AlphaFold 这样的楼，要在中国的一家机构里盖，需要什么？

据公开报道，DeepMind 做蛋白质结构预测，从起步到 CASP14 夺冠，前后用了好几年，中间还有一个第一代 AlphaFold，在 2018 年的 CASP13 上拿了第一，可是离实验精度还差得远。

换成一张中国的课题书来写，这几年大概是这样：第一个三年，结题时排名第一，却没解决问题；要续第二个三年，评审会问，第一期已经第一了，第二期的新意在哪里？申请人回答：要把整个结构推倒重来。

这个回答，在评审会上很难过关。推倒重来，听起来像第一期白做了。

请注意，这里难的不是钱。几年的算力和一个几十人的组，中国很多机构都拿得出来。难的是许可：许可一个已经拿了第一的组，说「我们的办法根本不对，要换」；许可它在下一个三年里，可能交不出任何一篇像样的论文。

你在自己的单位里，见过一个已经做得不错的项目，被允许推倒重来吗？

◆ 十三、放回日子里

回到林舟本子上那一行。

那天晚上，他最后没有再改那个词。他另开了一个文档，标题叫「如果不用 AlphaFold」。他写了一段：在他这个方向上，现有工具最根本的盲区在哪里？要是自己出一道题，会是什么题？

他写了半页，就写不下去了。不是因为想不出来，是因为他发现，自己已经很多年没有用这种方式想过问题了。每一个问题，他想的第一步都是：用什么工具，基于什么，改进哪里。

他把那个文档存下来，没有删。

明天能试：在你最常用的那件工具旁边，写一行字：「这件工具假定了什么？」然后试着写出它没法回答的一个问题。写不出来也没关系，写得出来，就是一道题的开头。

> **用好别人的楼是本事；可是下一座楼盖在哪里，是造楼的人先知道。**

带 走 的 话

用 AlphaFold 是正当的科研，也出了好成果；可是用与造之间隔着一层，那一层叫「谁在出题」。

小　账

看见了什么，漏看了什么？ 看见了用好别人的工具是科研的常态，漏看了「用」这一层正是 AI 最先接手的一层。

是在改进，还是在出题？ 用 AlphaFold 筛分子，是在别人的典范和架构里做得更好；AlphaFold 本身，是在一道五十年的老题上换了解法，又握住了下一道题。

想 一 想

你每天用得最多的那件工具，是谁造的？它假定了什么？如果它明天被机器替你用了，你还剩下什么？

◆

21

第 21 章

前沿的成绩，旧的题目

◆ 一、你一定有过这样的时刻

周五晚上，你打开手机里的天气，看周末两天会不会下雨。

屏幕上一排小图标：周六多云，周日中雨，下雨的概率是百分之七十。你看了一眼，把周日的郊游改到了下周。

你大概从没想过，这一排小图标背后，站着一台什么样的机器。

几十年来，它是一台巨大的计算机，日夜不停地解一组描写空气怎样流动、水汽怎样凝结的方程。今天，这台机器旁边多了一个新来的帮手：一个用几十年的天气记录练出来的 AI 模型。它不解方程，它看过太多次天气，于是学会了「下一步大概会怎样」。

这一章讲的，就是这个新帮手里最出名的一个中国名字。它做成了前沿的事，题目却是一道很老的题。

◆ 二、一篇登上《自然》的论文

2023 年 7 月，华为云的毕恺峰等人在《自然》（Nature）杂志上发表了盘古气象大模型的论文。

先把它做了什么摆在桌上。

它拿欧洲中期天气预报中心整理的全球再分析资料来练。据论文介绍，用的是 1979 年到 2017 年的 39 年数据。它把地球大气切成很细的格子，横向大约四

分之一经纬度一格，纵向分成十几个气压层。练好以后，给它此刻的大气状态，它就能推出一小时、几小时、一天直到一周以后的样子。

论文报告的结果有三条，都很硬。

第一，在中期预报上，它的准确度超过了当时世界上最好的传统数值预报系统，也就是欧洲中期天气预报中心的业务系统。

第二，速度快得惊人。传统方法要在超级计算机上算几个小时的事，它在一块图形卡上几秒钟就能算完，论文说快了一万倍以上。

第三，它追踪台风路径，也比传统方法准。

这是中国团队在 AI for Science 里最响亮的一次亮相。中国的论文登上《自然》不算稀奇；稀奇的是，它在一道全世界最好的机构比拼了半个世纪的题目上，拿到了第一。

◆ 三、先把公道还足

说它停在哪里之前，先说它对在哪里。这一章要还的公道，比别的章多一些。

第一笔公道：这不是跟风。2022 年到 2023 年，全世界有好几支队伍在做同一件事。英伟达有 FourCastNet，谷歌 DeepMind 有 GraphCast；中国自己还有上海人工智能实验室的「风乌」、复旦大学的「伏羲」。这是一场真正的赛跑，盘古是跑在最前面的几个之一。

第二笔公道：它的设计有自己的东西。大气是三维的，高空和地面的天气连在一起；它没有把天气当成一张张平面图片来处理，而是做了一个按地球的三维形状设计的网络。为了不让误差在一步步往后推的时候越滚越大，它还练了几个不同步长的模型，一小时的、六小时的、一天的，搭起来用。这些是工程上的真功夫。

第三笔公道：它有用。气象部门真的拿它来试，预报员真的拿它来参考〔待核：业务试用的范围〕。几秒钟出一次全球预报，意味着可以一下子跑出很多个略有不同的预报，看哪种天气更可能来。台风来的那几天，这种速度能救命。

第四笔公道：它说明中国能做到前沿。只要有一样东西摆在那里，中国的团队就跑得极快：海量的数据，清楚的题目，明确的打分标准。

这三样东西，天气预报全有。数据是现成的，几十年整理好；题目是现成的，给定今天，推出明天；打分也是现成的，明天一到，对错立见。

◆ 四、它完全成立的条件

公道还要再往前走一步：给出盘古这一类成绩完全成立的条件。

条件有三个，缺一个都不行。

一，要有一份足够长、足够全、足够干净的历史记录。天气有，几十年的全球大气，每隔几小时一张图，没有缺页。

二，题目要能写成「给定此刻，推出下一刻」。天气能，大气今天的样子，大体决定了明天的样子。

三，对错要来得快，来得清楚。天气来得最快，明天一到，预报准不准，谁都看得见。

在这三个条件下，中国的团队几乎没有短板。会收数据，会搭网络，会调到最好，会在最短的时间里把论文写出来、把系统推上线。盘古不是运气，是这三个条件遇上了一支擅长在这种条件下冲刺的队伍。

请记住这三个条件。下一章会看到，满足这三个条件的题，正是机器最喜欢的题。

◆ 五、三样现成的东西

把这三样东西再看一遍，你会发现它们有一个共同点：都是别人准备好的。

数据是谁准备的？那 39 年的再分析资料，是传统数值预报的老功夫做出来的。全世界的气象站、探空气球、卫星观测，加上数值模型几十年的同化计算，才把零零散散的观测，补成一张张完整的全球大气图。AI 学的，是这套老方法替它铺好的路。

题目是谁定的？「把大气切成格子，给定此刻的格子，推出下一刻的格子」——这个提法，从二十世纪中叶数值天气预报诞生起，就是这个行当的老规矩。

打分是谁定的？均方根误差、距平相关系数，这些尺子也是气象学几十年用下来的。

盘古气象做的，是在这三样现成的东西上，换一种工具，把成绩往上提了一截。

用大白话说：它是一个极好的新学生，参加了一场出了几十年的老考试，考了第一。

◆ 六、本书的说法

第 17 章把科学里的进步分成三层：技术改进、架构跃迁、科学典范。盘古气象站在哪一层？

先看它的网络。它用的是 Transformer 一路的结构，这一路出自 2017 年谷歌翻译组的论文，第 18 章讲过。盘古把它改成了适合地球大气的样子，这是漂亮的改进，还算不上一次架构跃迁。

再看它的题目。它没有改变「天气预报是什么」。好的预报，仍然是格点上的误差小；好的问题，仍然是「明天的格子长什么样」。这一行认定什么是好问题、什么是好答案的那一整套老规矩，也就是典范，一点没动。

所以本书的说法是：盘古气象是「用新工具把旧问题做得更好」的一个顶级样本。

它的成绩是前沿的，它的题目是旧的。

这句话不是贬它。在旧题里拿第一，需要真本事；全世界最好的几支队伍都在抢这块第一，抢到了就是抢到了。本书只是要把它放在正确的那一层上看：这一层有多高，下一章会说，这一层也是 AI 最先接管的地方。

◆ 七、只看一面会怎样

只看成绩，会得出一个乐观的结论：中国的 AI for Science 已经站上世界前沿，剩下的只是多几个盘古。

只看题目，会得出一个刻薄的结论：这不过是拿别人的数据、别人的架构、别人的题目刷了一次榜。

两个结论都不对。

乐观的那一个，漏看了题目是谁出的。一个国家能出一百个盘古，也可能一道新题都没出。成绩可以一项一项加起来，出题的能力不能这样加。

刻薄的那一个，漏看了在旧题里拿第一有多难。没有工程上的硬功夫，没有对气象这一行的真懂，拿不到这个第一。把真本事说成刷榜，是对做事的人不公道。

本书两个都不取。它站在两者中间说一句：这一层，中国真能做到最好；可是这一层，不是 AI 时代最稀缺的那一层。

◆ 八、今天与机器

还有一件事，更值得停下来想。

盘古气象做的事，本身就是「让机器接手人做了几十年的活」。从前，预报的进步靠一代代气象学家改方程、改参数、改算法，一点一点抠出精度。现在，一个模型看过 39 年的天气，就把这些年抠出来的精度追平、超过了。

也就是说，盘古证明了一件事：在一道定义清楚、数据充足的老题上，机器可以比几代人的精细手艺做得更好。

这件事反过来也照着中国自己。

中国最擅长的，恰好就是这种题：题目清楚、数据海量、打分明确，比的是谁更快、更准、更便宜。盘古证明中国能在这种题上赢；同一个盘古也证明，这种题，机器正在越做越好。

今天赢下这一类题的，是「中国团队＋AI」。明天赢下这一类题的，可能只需要「AI＋很少的人」。中国最擅长的那一层，正在变成谁都能做好的那一层。

◆ 九、一个完整的例子：气象学那一次换题

天气预报这一行，历史上真的换过一次题。拿它来对照，就看得出「做好旧题」和「出新题」差在哪里。

二十世纪初，天气预报大体是看图说话：把各地的气压、风向画在一张图上，凭经验认出「这种形状的图，后面常常跟着下雨」。好的预报员，是看过最多张图、记性最好的那个人。

后来有人提出了一个新的想法：天气不该靠认图，该靠算。空气的流动服从物理定律，把定律写成方程，把今天的观测代进去，一步步算出明天。据气象史的通常讲法，英国人理查森在第一次世界大战期间试着手算过一次，结果错得离谱；到了二十世纪中叶有了电子计算机，这条路才走通。

请注意这次变化变的是什么。它没有把认图认得更准，而是换了「什么才算好的预报」：好的预报，从「经验丰富」变成了「方程算得对」。好问题也跟着变了：从「这张图像哪一张旧图」，变成了「这个方程哪一项没写对」。

这才是典范的更换。

现在回头看盘古。它是数值预报这个典范养大的孩子：用数值方法做出来的数据练，用数值预报的尺子打分，去赢数值预报的老冠军。它跑得比老冠军快，可它跑在老冠军修好的跑道上。

有人会说，AI 天气预报会不会本身就是又一次换题：从「算方程」回到了一种更高级的「认图」？这个问题很好，也还没有答案。如果有一天，有人借着 AI，重新定义了「什么才算好的天气预报」，那才是这一行的下一次换题。比如不再预报格点，而是直接预报一座城市会不会内涝、一片农田要不要抢收。这个人在哪里，第 25 章开始找。

◆ 十、为什么偏偏是天气

你也许会问：AI for Science 那么多方向，为什么最先打出名堂的，是天气、是蛋白质结构?

答案就藏在上面那三个条件里。

蛋白质结构，有几十年攒下的结构数据库；题目清楚，给定一串氨基酸，推出它折成的样子；对错也清楚，有实验测出来的结构可以对照。第 20 章讲过的 AlphaFold，正是在这样一道题上赢的。

天气也是一样。两道题都老，都清楚，都有一份很长的历史记录等着被学。

反过来看，那些 AI 还没打出名堂的方向，往往缺的正是这三样。数据少，或者数据很脏；题目说不清，不知道该推出什么；对错来得很慢，十年后才知道当初对不对。

这不是 AI 的错，也不是这些方向的错。它只说明一件事：AI 最先接管的，是那些题目早就被前人定义清楚了的地方。

一道题被定义清楚，本身就是前人几十年的功劳。AI 收的，是这份功劳的利息。

◆ 十一、停在哪里

把这一章的账收一下。

盘古气象停在了一个很高的地方：旧题里的第一名。它没有停错，它本来就是去拿这个第一的。

问题不在它，在一个国家把它当成了什么。

如果把盘古当成「中国 AI for Science 已经到了前沿」的证据，就会接着要一百个盘古：一百个在旧题上刷新纪录的项目，一百篇登上好期刊的论文。这些都会有，也都值得有。

可是 AI 越强，旧题上的第一名就越容易拿，拿到的分量也越轻。

这里要把「轻」字说清楚。不是说盘古这样的成果没价值，是说同样一份成绩，在一个越来越多人都能拿到的地方，能换来的东西越来越少。

五年前，一个旧题上的世界第一，可以换来一个实验室、一批人、十年的经费。五年后，同样的第一，也许只能换来一篇论文和一次新闻发布。因为到那时，很多支队伍加上更强的 AI，都能在几个月里追平它。

一个国家如果把「多拿几个旧题第一」当成 AI for Science 的路线图，就是把力气押在一块正在缩水的地上。

火药做成了一束极好看的烟花。这束烟花本身没错；只是火药是别人发明的，下一种火药，还没有人去想。

◆ 十二、放回日子里

回到周五晚上那排小图标。

下一次你打开天气，看到「周日中雨，百分之七十」，可以多想一步。这个数字背后，有几十年的观测、几代人的方程、一个新来的 AI。还有一个从没变过的题目：给定今天，推出明天。

明天能试：找一件你在工作里做得最好的事，问自己两个问题。这件事的题目是谁出的？打分的尺子是谁定的？

如果两个答案都是「早就有了」，那你做的是旧题里的第一名。这很好，可它也是机器最先学会的那一类。

> **一个极好的学生，参加了一场出了几十年的老考试，考了第一。**

> 带 走 的 话
>
> **在旧题里拿第一要真本事；可是题越旧、越清楚，机器就越先学会，第一名也就越轻。**

> 小　账
>
> **看见了什么，漏看了什么？**看见了中国在数据多、题目清的地方能做到世界最好；漏看了那三样现成的东西，都是别人准备好的。
>
> **是在改进，还是在出题？**盘古气象是在一道老题里做得更好；天气预报这一行的下一次换题，还没有人出。

> 想 一 想
>
> 你手上做得最好的那件事，题目、数据、打分，有哪一样是你自己定的？如果一样都不是，它离被机器接手还有多远？

◆

22

第 22 章

常规科学交给机器

◆ 一、林舟的一个晚上

晚上十一点，林舟还在研究院的办公室里。

他的组有十二个人，做 AI 制药。这个月要交一份基金本子，还有两篇论文在赶。屏幕上开着四个窗口。一个是 AlphaFold 预测的蛋白结构，一个是大模型帮他整理的近三年文献综述；一个是学生发来的分子筛选结果，还有一个是基金本子的第三稿。

他翻着那份文献综述，忽然有点愣。

这份综述是下午让机器写的，四十分钟，八十多篇文献，分好了类，标出了哪几个方向做得最多、哪几个还空着。他记得自己读博士的时候，做一份这样的综述，要在图书馆坐两个月。

更让他愣的是最后一段。机器在「尚待研究的问题」下面，列了七个可以申请基金的方向，每个都写得像模像样。

他看了一会儿，挑了第三个，抄进了基金本子。

关掉电脑的时候，他心里有一句话没说出来：这些年，自己到底在做什么？

◆ 二、库恩说的常规科学

要回答林舟那句没说出来的话，得先借一本书。

1962 年，美国科学史家库恩出版了《科学革命的结构》。这本书里有两个词，后来传遍了整个学术界：典范和常规科学。

典范，本书导论讲过，就是一门科学认定什么是好问题、什么是好答案、怎样算做对的那一整套。

常规科学，是在一个给定的典范里做事。库恩打了个比方，大意是：常规科学像解谜。拼图的图案是定好的，规则是定好的，每块拼图该放在哪里也有标准；科学家的本事，是把这幅图拼得更快、更全、更漂亮。

他说，绝大多数科学家，绝大多数时间，做的都是常规科学。这不是贬低。正是常规科学把一个典范的潜力挖到底，把它的细节填满，也把它解不开的那几个谜一个个摆到桌面上。没有常规科学，就没有后来的典范更换。

请把「解谜」两个字记住。它是这一章的钥匙。

◆ 三、常规科学，一天里有哪些活

把「解谜」换成林舟的日子，常规科学的活大致有这几样。

读文献，知道这一行做到了哪里。定题目，在已有的空白里挑一块。设计实验，选模型、调参数、定对照。跑实验，筛分子、测数据。分析结果，画图、做统计。写论文，按期刊的格式讲清楚自己做了什么。申请基金，把下一步的谜说成值得花钱去解的谜。

这七样活，养活了全世界几百万科研人员。

现在一样一样看过去：哪一样，AI 还没碰？

读文献，它已经在做，林舟下午刚用过。定题目，它在列「尚待研究的问题」。

设计实验，已经有实验室让 AI 设计、让机器人去做，做完了 AI 自己看结果、改方案、再做一轮〔待核：自动实验室的进展〕。

调参数、搜网络结构，机器早就比人做得好，这一行叫自动机器学习，已经做了好些年。写论文初稿，大模型写得又快又顺。据报道，2024 年就有团队做

出「AI 科学家」，从想法到实验到论文一条龙；后来还有一篇这样写出的论文，据说过了一个研讨会的同行评审〔待核〕。

七样活，没有一样是安全的。

◆ 四、解谜解出来的那几个怪数

库恩还说了一件事，第 17 章提过一句，常常被人忘掉。

常规科学解着解着，总会碰到几个解不开的谜。数据和理论对不上，实验一再出怪结果。库恩管这叫反常。起初，大家会把它当成测量误差，或者当成还没找到的那一步，放到一边，继续解别的谜。

反常越攒越多，有一天，一个典范就撑不住了。这时候，有人站出来说：不是谜没解好，是这幅拼图的图案本身画错了。这就是库恩说的科学革命。

请注意这里的分工。把反常一个个找出来，是常规科学的功劳，机器在这一段上会越来越强：它读得多、算得快，最先发现哪几个数对不上。

说出「图案本身画错了」，却是另一件事。

机器可以列出一百个对不上的数。哪一个是测量误差，哪一个是整幅图要换的信号？在旧图案里看，它们长得一模一样。

这一段，库恩那本书里没有机器，今天的机器也还做不了。这一点，第 23 章和第 25 章还会回来。

◆ 五、本书的说法

第 369 号《典范发生学》说过一个判断：AI 能做给定典范下的常规科学。

这句话放在库恩的框里，意思就很清楚了。常规科学是解谜；解谜有图案、有规则、有标准。凡是有图案、有规则、有标准的事，正是机器最擅长学会的事。

上一章的盘古气象，是一个样本。它是在「数值天气预报」这个给定的典范里，把谜解得更好。它证明了常规科学可以由机器来做，还能做得比人更好。

这一章把样本推成一条判断：常规科学，正在整体地交给机器。

不是哪一样活被机器替了，是这一整层活，从读到写，从想到做，都在被机器一段一段接过去。

◆ 六、中国的家底

到这里，要把一笔账算到中国自己头上。

中国科研这三四十年的家底，是什么？

人多。中国的研发人员总数多年位居世界前列〔待核：具体排名与口径〕。一个课题组十几个人、几十个人，很平常。

实验快。别人一个月做完的实验，中国的组两周做完；别人一年发一篇，中国的组一年发三篇。很多外国同行都佩服这个速度。

论文多。中国发表的科学论文总量，这些年已经站到了世界最前面的位置〔待核：统计口径与年份〕。

执行力强。题目一定，人一排，钱一到，就能干出来。高铁、桥梁、大型装置，都是这样干出来的。

这四样是真本事，是几代人一点一点攒下来的。说它们之前，先把这句话说在前面：没有这份家底，就没有今天的中国科研。

可是把这四样放回库恩的框里看，它们是什么？

人多，是解谜的人手。实验快，是解谜的速度。论文多，是解完的谜的数量。执行力强，是一个谜定下来之后解开它的效率。

中国最大的科研资产，恰好是常规科学的生产要素；而常规科学，正是 AI 最先接管的那一层。

◆ 七、贬值是怎样发生的

「贬值」两个字听起来刺耳，其实说的是一件很普通的事。

一样东西，从前只有你多，现在谁都能有，它就不值钱了。

从前，一个组有十二个人，就比只有三个人的组多做四倍的活。现在，三个人加一套好用的 AI，可能比十二个人还快。人多这件事，就不再是优势。

从前，实验快是因为人肯熬、组织得好。现在，自动化实验室一天二十四小时不停，人再肯熬，也熬不过机器。

从前，论文多说明产出多。现在，一篇像样的论文初稿，机器几个小时就能写出来。论文多这件事，越来越说明不了什么。

执行力也一样。题目定了以后怎样最快做出来，这本身就是一道解谜题；机器越会解谜，人在这一段上的执行力就越不稀奇。

这四样家底不会消失，它们只是越来越像所有人都有的东西。

打个比方。从前一座城里会写字的人少，能替人写信的先生，在街口摆一张桌子就有饭吃。后来人人识字，先生还会写字，写得还是比别人好，可是没有人再排队请他写信了。

他的本事没退步，是他的本事不再稀缺了。

中国科研的那四样家底，今天就站在那张桌子后面。

◆ 八、只看一面会怎样

只看「机器能做」这一面，会以为科学家要失业了，科研要变成按一下按钮的事。

这不对。第一，机器接手常规科学，不是一夜之间的事，很多实验仍然要人的手、人的眼。第二，常规科学被机器接走，科学不会变小，反而会变大：同样的人，能解开多得多的谜。

只看「中国很强」这一面，会以为 AI 只会让中国更强：中国人多、组织好，再加上 AI，岂不是如虎添翼？

这也只对了一半。AI 确实让中国的组跑得更快。可是它让所有人都跑得更快，让三个人的小组和十二个人的大组跑得差不多快。从前拉开差距的东西，正在被拉平。

还有第三种看法，也要提一句：既然常规科学交给了机器，那中国把机器用得最好，不就又赢了？

这一半是对的。中国用 AI 的速度和规模，确实可能是世界上最快的。第 20 章讲过，用 AlphaFold，中国的组用得又快又好。

不过「用机器把常规科学做得最好」，本身也是一道常规科学的题。人人都在用，比的是谁用得更熟。这一场比赛，中国很可能赢；只是赢了以后，手里拿到的，还是常规科学那一层的奖杯。

所以本书两面都看。机器接手常规科学，不是坏事；坏在一个国家把全部的家底，都押在了正在被拉平的那一层上。

◆ 九、车替腿

第 369 号《典范发生学》打过一个比方：车替代了腿，不是让腿去学一种新的跑法，是整个交通升了一级。这个比方，那本书讲透了，这里只点一句：今天中国科研最引以为傲的，是一双全世界最好的腿。

第 27 章会回来，讲这双腿是怎样被一道道关选出来的。

◆ 十、宋岚那一边的同一件事

这件事不只发生在研究院里。

同一个晚上，宋岚的短剧公司也在开会。从前，一部短剧要一个编剧组写本子、一个摄制组拍、一个后期组剪。现在，本子让大模型出初稿，画面让视频模型一条条生成，剪辑也有工具帮着排。

她的公司没有因此变小，反而一个月出的剧更多了。可是公司里最吃香的人变了：从前是手最快的剪辑师，现在是能判断「这一条观众会不会划走」的那个人。

两个行当，同一个方向。题目定好以后怎样做出来的那一段，正在交给机器；值钱的，越来越只剩「做什么」和「这样做对不对」的判断。

区别只在一处。宋岚那一行的题目，是观众的手指每天替她出的；林舟那一行的题目，要有人去出新题。没人出，就只剩旧题越做越快。

◆ 十一、这里要先打一个预防针

说「常规科学交给机器」，不是说明天早上，所有实验室的人都可以回家了。

有三处，机器一时还接不过去。

一处在手上。很多实验，要人去养细胞、喂小鼠、调仪器。自动化实验室在铺开，可是离铺满还远，有些活很长时间里仍然要人的手。

一处在脏数据上。天气和蛋白质结构，有干净的历史记录；很多领域没有。数据少、数据乱、数据彼此对不上的地方，机器的本事会打很大的折扣。

一处在担责上。一个新药能不能上人体，一座桥能不能通车，最后要有人签字。机器可以算，签字的还得是人。

这三处是真的，也是中国的科研人员今后很多年的饭碗所在。本书不想把话说过头。

只是请注意这三处的性质：它们都是常规科学里「还没被机器够着」的那一段，不是机器永远够不着的那一段。手上的活在被机器人接，脏数据在被更好的方法洗，签字之前的那一大段判断，也在一点点交出去。

◆ 十二、一个完整的例子：林舟组里的一年

拿林舟的组走一遍，看贬值是怎样落到一群具体的人身上。

早些年，林舟的组靠什么赢别人？靠的是十二个人一起筛分子。一个博士生一周能手工处理几百个分子结构，十二个人一个月就是上万个。同行三个人的组，追不上这个量。

后来，AlphaFold 和各种预测工具铺开了。蛋白结构不用再一个个去猜，筛分子的量一下子翻了几十倍。林舟的组还是领先，因为他们最先用上，用得最熟。

今年，情况变了。工具人人都有，大模型帮着写代码、跑流程、看结果。上个月，林舟在一个会上碰到一个只有三个人的组，做的方向和他差不多，进度比他还快。那个组的负责人说，他们三个人主要做两件事：想下一步该问什么，判断机器给的结果靠不靠谱。

林舟回来想了很久。

他的组十二个人里，有九个在做的事，机器都能做。剩下三个，做的也多半是常规科学里比较难的那一段。他自己呢？每天写本子、改论文、开组会，也是在一个给定的方向里，把谜解得更快。

他不是做得不好。他是把一件正在被机器接手的事，做到了全组最好。

◆ 十三、放回日子里

回到林舟关电脑的那个晚上。

他抄进基金本子的第三个方向，是机器列出来的。这件事本身没什么不对，工具就是拿来用的。可是那一刻他愣住，愣在一个很实在的地方：连「下一步做什么」这件事，机器也开始给建议了。

明天能试：把你这一周的工作列成一张单子，每一项后面标一个字，「解」或者「出」。「解」，是题目早就定好，你在里面做得更好；「出」，是你改变了「什么才算一道好题」。

大多数人的单子上，会全是「解」。这不丢人，库恩说过，绝大多数科学家绝大多数时间都在解谜。

只是要知道：标着「解」的那些行，正是机器正在一行一行接过去的地方。

> **解谜的人手、速度和数量，正在变成谁都有的东西。**

带 走 的 话

人多、实验快、论文多、执行力强是真本事；可它们是解谜的本事，而解谜正在交给机器。

小　账

看见了什么，漏看了什么？看见了 AI 会让中国的组跑得更快；漏看了它也让所有人跑得一样快，从前的差距正在被拉平。

是在改进，还是在出题？常规科学是在已有的题里做得更好；被机器接走的，正是这一整层。

想 一 想

你这一周做的事里，有多少是在别人定好的谜里解得更快？如果明天有一台机器能把这些都做了，你还剩下什么是它做不了的？

◆

23

第 23 章

连提问也不再稀缺

◆ 一、林豆豆的作业

星期三晚上，林豆豆在写科学课的作业。

题目是：「观察家里的一种植物，提出三个你想研究的问题。」

她对着阳台上那盆绿萝看了五分钟，只想出一个：「绿萝为什么是绿的？」她觉得这个问题太简单，老师一定不满意。

于是她拿起平板，问了家里的 AI：「关于绿萝，可以提出哪些好问题？」

三秒钟，屏幕上出来十二个问题。绿萝的叶子为什么朝着窗户长？浇水多了为什么会发黄？把它放进不同颜色的灯光下，长得一样快吗？一片叶子剪下来插进水里，多久能生根？……每个问题后面，还附了一个可以在家里做的小实验。

林豆豆挑了三个最好听的，抄进了作业本。

宋岚路过，看了一眼，心里咯噔一下：老师要的，不就是「会提问」吗？这道题，机器替她答了。

◆ 二、本书前面说过的话

这本书前面，好几处说过一个意思：要会提问。

第 13 章讲解题的能力和出题的能力，说中国的教育和科研，最会的是解题，最缺的是出题。第 15 章讲改进不是 0 到 1，说应用里撞出来的，是在已有的

题里做得更好，而不是出一道新题。这些说法，落到教育里，常常变成一句口号：要培养会提问的孩子。

这句口号，过去几十年是对的。

一个只会做题的孩子，长大了只会做别人交代的事；一个会提问的孩子，才可能自己找到值得做的事。在 AI 来之前，提问确实比答题稀缺。

林豆豆的作业，却把这句话推到了一个尴尬的地方：机器三秒钟提了十二个问题，个个比她自己想的好。

如果「会提问」就是终点，那机器已经比大多数孩子、甚至比大多数大人，更会提问了。

◆ 三、第 369 号更进一步的口径

第 369 号《典范发生学》在这里比本书前面走得更远。它的口径是：答案、过程、提问，都不再是竞争点。

答案不是竞争点，这个容易懂。查一个事实、解一道方程，机器早就比人快。

过程不是竞争点，上一章讲过。读文献、设计实验、调参数、写初稿，这些「怎样解」的功夫，机器正在一段段接过去。

提问也不是竞争点，这一句最难接受。

不过你回头看林舟抄进基金本子的那七个方向，看林豆豆抄进作业本的那十二个问题：AI 能提出大量好问题。而且它提得又多又快，还能把每个问题的来龙去脉、可行性、前人做到哪里，一并给你。

所以，提问也开始变得不稀缺了。

◆ 四、什么样的问题，机器提得出来

先别急着下结论。要看清机器提的是哪一类问题。

林豆豆那十二个问题，都很好。可它们有一个共同点：都在一个早就定好的框里。

这个框是什么？是中学生物课的那一套：植物有光合作用，有向光性，需要水和养分，可以无性繁殖。所有的问题，都是在这套框里找一个还没被这个孩子亲手验证过的点。

林舟的七个基金方向，也是这样。它们都在「AI 筛分子、找靶点」这个框里，从文献的空白处挑出来。机器读过所有的文献，最清楚哪里空着。

也就是说，机器提出的好问题，是某个典范之下的好问题。

什么叫「好问题」，这个标准本身是现成的。机器做的，是在现成的标准下，把好问题找得又多又全。

◆ 五、本书的说法

这样一看，就能把第 369 号的话接下去了。

真正稀缺的，不是提出好问题，是猜出一个新的「什么才算好问题」。第 369 号管这个叫典范猜想。

用大白话说：好问题有一个标准；在标准里找问题，机器会；换一个标准，机器还不会。

打个比方。一家餐馆，厨师很会做菜，也很会想新菜：同样的川菜规矩，今天创一道麻辣小龙虾，明天创一道藤椒鱼。这是在一套菜系里提出好问题。另一件完全不同的事，是有人第一次说「菜不一定要热的」「菜可以是一口一口端上来的」——他改的不是哪道菜，是「什么才算一桌好菜」。

前一种，机器正在学会，而且学得很好。后一种，才是第 369 号说的典范猜想。

请注意两种提问的差别在哪里。前一种问的是「在这套规矩里，还能做出什么」；后一种问的是「这套规矩，是不是该换了」。前一种越问越多，规矩越来越满；后一种只问一次，规矩就变了一套。

机器读过人类写下的所有菜谱，它最会的就是前一种。后一种要的，偏偏是一个菜谱里还没有的念头。

◆ 六、只看一面会怎样

只看「机器也会提问」这一面，会觉得人再也没有用处了。

这不对。机器能提出的，是现有标准下的好问题；标准本身怎样变，它要靠人来猜。而且它提的问题越多，挑哪一个、为什么挑这一个的判断，反而越重要。

只看「人还是更会提问」这一面，会以为教育照旧就行：继续培养会提问的孩子，就够了。

这也不对。如果「会提问」只是会在现成的框里找问题，那这个本事正在贬值，速度不比答题慢多少。把「会提问」当成教育的终点，等于培养一种机器正在学会的能力。

本书两面都看：人还有用处，可是这个用处比「会提问」窄得多，也难得多。

窄在哪里？窄在它只剩一件事：在机器给出的一千个问题面前，说出「这一类问题都问偏了」，并且说得出偏在哪里。

难在哪里？难在这句话说出来的时候，多半没人信，连说的人自己也不太有把握。这一层，正是第 25 章要找的那个人站的地方。

◆ 七、今天与机器

这件事对中国，有一层特别的意思。

中国的教育和科研，这些年也在追「会提问」。课堂上鼓励孩子举手问为什么，基金评审看重「科学问题凝练」，研究生培养强调「自己找题目」。这些努力是真诚的，也做出了成绩。评价上的「破四唯」，也是体制自己在往「问题」上转，第 28 章会细看。

可是中国追上「会提问」的那一天，机器也追上了。

这像一场赛跑：中国在加速追前面那个人，跑着跑着，前面那个人上了一辆车。问题不是你追得慢，是你追的那个位置，本身在往后退。

所以，「会提问」这条路要走，也要知道它不是终点。下一个位置更远，也更窄：在没有人告诉你什么是好问题的时候，猜出一个新的标准来。

◆ 八、一个完整的例子：深度学习那一次换标准

第 19 章讲过深度学习在冷宫里的几十年。拿它来看一次「好问题」的标准是怎样被换掉的。

很长一段时间里，人工智能这一行认定的好问题是：怎样把人的知识写成规则，让机器照着规则推理？好答案是：规则写得越全、越精，机器就越聪明。在这套标准下，好问题有成千上万：怎样表示常识？怎样组织专家的经验？怎样让推理更快？

少数人猜了另一个标准：好问题不是「怎样写规则」，是「怎样让机器自己从数据里长出规则」。在当年的主流看来，这个问题提得不好，甚至不像一个问题：数据那么少，算力那么弱，网络那么难练。

后来的事，第 19 章讲过。那几位少数人赢了，赢在他们换了「什么才算好问题」的标准，而不只是答对了旧标准里的题。

请注意：在旧标准里，机器（如果那时有今天这样的机器）能提出一千个好问题，个个都是「怎样把规则写得更好」。它提不出那一个「别写规则了」。

那一个问题，在旧标准里不像好问题。它只有在新标准立起来以后，回头看，才是好问题。

这就是典范猜想难的地方：它在被猜出来的那一刻，常常看起来像一个坏问题。

◆ 九、两千多年前的一口气问

中国人最早的一次「大提问」，常被举出来的是屈原的《天问》。

据通常的讲法，它一口气问了一百七十多个问题：天地是怎样开头的，日月挂在哪里，鲧为什么治不好水。一句一句问下去，问得又多又急。

这是一篇了不起的文字。只是细看会发现，它问的，大多还在当时那一套天地、神话、历史的框里。它问的是「这件事为什么会这样」，很少问「我们为什么要这样问天」。

这不是屈原的短处，是任何一个时代的提问都有的样子：问题再多，也多半长在那个时代认定的「什么值得问」里面。

今天的机器，就像一位读过所有书的屈原。它能一口气问出一千个好问题，个个有根有据。可是这一千个问题，仍然长在今天认定的「什么值得问」里面。

◆ 十、挑问题的那只手

还有一件事，容易被漏看。

机器提出的问题越多，挑的那一下就越重。

从前，一个研究者一年能想出三五个值得做的题目，挑起来不难。现在，机器一个下午给他七十个，个个写得像模像样。七十个里挑哪一个，凭什么挑，挑错了一年就白干。

挑问题，靠的不是会不会提问，是心里有没有一个「什么才算好问题」的标准。没有自己的标准，就只能用机器的标准挑；用机器的标准挑出来的，又回到了旧框里。

所以，提问不稀缺了，挑问题的标准反而更稀缺。而标准再往前走一步，就是猜一个新的标准。

林舟抄进基金本子的那个第三方向，就是这样挑出来的：哪个最像能拿到钱，就挑哪个。这是评审的标准，也是旧框的标准。

◆ 十一、典范猜想交出去以后

猜出一个新标准，只是开头。猜想之后的路怎样走，第 369 号《典范发生学》讲透了，这里只看一眼它和本书相关的那一处。

从前，一个新猜想要变成一个典范，要走很长的路。要有人把它写成理论，要有人设计实验去检验，要有人一代代把它教给学生，要有人在它下面做出成千上万个常规科学的成果。这条路，常常要走几十年，要一大群人。

今天，这条路的很大一段，可以交给机器。猜想一旦说清楚，机器可以替它推出后果，找出能检验它的实验。它还能在猜想下面提出成百上千个「新标准下的好问题」，甚至替它解开其中的一大半。

这就是为什么，第 369 号把典范发生说成「人—AI 复合体」的事：人出那一猜，机器把一猜铺成一片地。

这里也藏着一个危险，第 37 章会专门讲：机器铺地铺得太快、太顺，一个错的猜想也会被铺成一片像模像样的地。

◆ 十二、「会提问」的口号要改写

回到教育。

「要培养会提问的孩子」，这句话要改写。改成什么，本书还不敢说定，只先试一个说法：要培养一个敢说「这一类问题本身不对」的孩子。

这两句话差得很远。

会提问的孩子，问的是「绿萝为什么朝着窗户长」。这是好问题，机器也会问。

敢说「这一类问题本身不对」的孩子，可能会问：「为什么科学课只让我们看植物怎样长，不让我们看它怎样和别的东西一起长？」这个问题可能不成熟，可能很幼稚，也可能根本不对。可是它在改「什么才算一个好的植物问题」。

前一种孩子，学校很会培养，机器也很会帮忙。后一种孩子，学校很少鼓励，考试里没有这一项，打分的尺子也量不出来。

这里也要把公道还给学校。一个班四十多个孩子，一个老师，一套统一的考试，要它去鼓励每个孩子「说这一类问题不对」，几乎做不到。学校守着统一的标准，是因为它要对四十多个孩子都公平。

问题不在老师，在那把尺子只量得出一种本事。

第 46 章会回来，讲教育怎样让本来没有的材料发生。

◆ 十三、放回日子里

回到星期三那个晚上。

宋岚没有骂林豆豆。她坐下来，指着屏幕上那十二个问题，问了女儿一句：

「这十二个问题里，有没有一个，你觉得问得不对？」

林豆豆看了很久，指着第七个：「把绿萝放进不同颜色的灯光下，长得一样快吗？」她说：「可是绿萝自己又不选灯，它在我们家阳台上，只有一种光。为什么要问它在别的灯下面怎么样？」

宋岚没听懂女儿的意思，也不知道这个问题对不对。可她觉得，这一句，比抄进作业本的三句都值钱。

明天能试：下一次孩子（或你自己）用 AI 找问题，不要只挑最好听的那一个。多问一句：「这些问题里，哪一个你觉得问错了？为什么错？」

挑出错的问题，比挑出好的问题，离典范猜想更近一步。

> **在标准里找问题，机器会；换一个标准，机器还不会。**

带 走 的 话

机器能提出一千个好问题，却提不出那一个让「好问题」换了标准的问题；它在被提出时常像个坏问题。

小　账

看见了什么，漏看了什么？看见了「会提问」比「会答题」稀缺；漏看了机器追上提问的速度，不比追上答题慢多少。

是在改进，还是在出题？在一个典范里提出好问题，仍是改进；猜出一个新的「什么才算好问题」，才是出题。

想 一 想

你最近一次问出的好问题，是在一个现成的框里找到的，还是改了那个框？你身边有没有一个孩子，问过一个让你觉得「问得不对」的问题？你当时是怎样回他的？

◆

24

第 24 章

本篇总账：层级的账

◆ 一、一张写满的草稿纸

考试做到最后一道大题，草稿纸上常常密密麻麻：有算对的，有划掉的，有中途改道的。交卷之前，有人会把草稿纸翻过来，在背面写一行，把整道题的思路理一遍。

这一章就是那一行。

中篇二八章，从第 17 章分出三层，到第 23 章说连提问也不再稀缺，每一章都在问同一件事：AI 进入科学，接管的是哪一层？

现在把前七章的账并排放。

◆ 二、三层再说一遍

第 17 章把科学里的进步分成三层。

第一层，技术改进。在已有的架构和典范里，把东西做得更快、更省、更准。它来自应用里的约束：算力不够，就逼出省算力的办法；成本太高，就逼出降成本的办法。中国在这一层极强。

第二层，架构跃迁。换一种搭法，让一大片问题一下子变得可做。它常常来自应用现场里的一块飞地：离产品很近，却暂时不用对产品指标负责。

第三层，科学典范。换掉「什么才算好问题、好答案」的那一整套。它常常来自长期反主流、没有应用、在冷宫里坐了很多年的少数人。

三层不是高低贵贱，是三种不同的活，长在三种不同的土里。

改进的土，是一个逼得很紧的现场：成本、工期、算力，样样卡着。架构的土，是一个现场旁边的小院子：离问题很近，又允许几年不出产品。典范的土，是一块荒地：没有现场，没有指标，常常连同行的认可也没有，只有少数人凭着一个猜想熬下去。

中国这几十年，把第一种土铺得又大又密。第二种土有一些，不多。第三种土，几乎没有被当成一块要专门留出来的地。

◆ 三、看见与漏看

下面这张表，把本篇各章的看见与漏看并排放。

章	这一章看见了	这一章说出的漏看
第 17 章　三层：改进、架构、典范	进步不是一种东西，至少分三层，各有各的来处	把三层混成一个「创新」，就看不出中国强在哪层、空在哪层
第 18 章　Transformer 出自一个翻译组	架构跃迁可以出自应用现场，翻译组里就长出了新架构	它要一块暂时不对产品指标负责的飞地；只有应用、没有飞地，长不出来
第 19 章　冷宫里的三十年	典范来自长期反主流的少数人，深度学习等了几十年	主流的评价和资助，恰恰是把这些人挤进冷宫的那只手
第 20 章　AlphaFold 和「用 AlphaFold」	中国用 AlphaFold 用得极快、极好，应用跟进一流	「用」和「造」是两件事；会用，不等于能出下一个
第 21 章　前沿的成绩，旧的题目	盘古气象证明，数据多、题目清的地方，中国能做到世界最好	数据、题目、打分三样都是现成的；成绩前沿，题目是旧的
第 22 章　常规科学交给机器	人多、实验快、论文多、执行力强，是中国的真家底	这些恰是常规科学的生产要素，而常规科学正在交给机器
第 23 章　连提问也不再稀缺	「要会提问」曾经是对的，也是中国在追的方向	机器在标准里提问已经很会；稀缺的是换标准的那一猜

把左边一列连起来读，是中国做对、做成的事：应用跟得快，工程做得好，旧题能拿第一，家底厚，教育和科研也在往「问题」上转。

把右边一列连起来读，是另一张图：每一处做成的事，都站在别人准备好的题目、别人搭好的架构、别人定下的标准上。

◆ 四、一条线从上往下压

再换一个方向，照第 17 章那张表的顺序，从第一行往下排。

第一行是技术改进。AI 在这一行，几乎已经是主力：搜网络结构、调参数、写代码、跑实验，机器做得比人好。

紧挨着它的，是常规科学里「怎样解」的那一段：读文献、设计实验、分析结果、写论文。AI 正在一段一段接过去，第 22 章讲过。

再往下，是常规科学里「问什么」的那一段：在一个典范里找好问题。第 23 章说，AI 也开始会了。

最底下一格，是科学典范本身：猜出一个新的「什么才算好问题」。这一格，AI 还做不了。

表上的第二行架构跃迁，这条线暂且不算进去。第 18 章说过，它要一块飞地，机器一时接不全，中国也只是偶有。

这条线从上往下，就是 AI 接管的方向。它从第一行压下来，一段一段往下吃。

而中国的家底，从上往下数，厚在前面三段，薄在最底下那一格。

请注意这两条线的方向。机器是从上往下一段段吃下去，家底是从上往下一段段薄下去。机器每往下吃一段，中国就有一段厚家底变得不再稀缺。

等机器吃到「问什么」那一段，也就是今天正在发生的「连提问也不再稀缺」，中国手里还稀缺的东西，就只剩最薄的那一格了。

这不是哪一个人做错了什么。几十年里，下面三层是最有回报、最能出成绩、最能让一个国家追上来的地方，把力气押在那里，是一个很合理的选择。只是 AI 改了这道题的条件。

◆ 五、本篇总判断

两张图叠在一起，本篇的总判断就出来了。

AI 接管的正是中国最强的那一层，留下的正是中国最空的那一层。

这句话有两半。

前一半，说的是贬值。中国几十年攒下的本事，人手、速度、规模、执行力、在旧题里拿第一的能力，都长在常规科学这一层。AI 来了，这一层的东西越来越像谁都有的东西。

这里的「贬值」，不是一夜之间清零，是一年一年地变便宜。中国在这一层还会赢很多比赛，还会发很多好论文，盘古这样的成绩还会再来。可是每一次赢，换回来的东西都会比上一次少一点。

后一半，说的是空缺。AI 还做不了的那一层，典范猜想，恰好是中国几十年里长得最少的那一层。不是没有人，是那一层需要的土、需要的时间、需要的宽容，中国的科研组织方式很少准备。

◆ 六、先清两个误会

这个判断很容易被听歪，先清两个误会。

第一个误会：这是在说中国科研不行。不是。中国在常规科学这一层，是世界上最强的几家之一，这一篇每一章都先把这一点写足了。判断说的是位置，不是高低：强的那一层正在贬值，空的那一层正在变成唯一值钱的地方。

一个人跑得最快，路却被改成了水路，他还是跑得最快，只是这件事不再像从前那样管用。

第二个误会：这是在说中国人不会想。也不是。屈原问过天，王阳明在龙场想通过，陈省身、杨振宁、李政道都在世界上做出过开一个方向的工作。会想的人，中国从来不缺。

缺的是另一样东西：让这种人从小长下去、长出来以后被认下来的那片土。

这句话，下篇一和下篇二要花十八章来拆。第 25 章从那个人开始，第 33 章从土开始。

◆ 七、一个晚上的三个房间

把这一篇的判断，放进林家的一个晚上看。

同一个晚上，林家有三个房间亮着灯。

书房里，林舟在改基金本子。他抄进去的那个方向，是机器列出来的；他熬夜做的事，是常规科学里最好的那一种。

客厅里，宋岚在看样片。她的公司一个月出的剧越来越多，最值钱的人，从手快的剪辑师变成了会判断观众的人。

儿童房里，林豆豆在写作业。她让机器替她提了十二个问题，又在妈妈的追问下，挑出了一个「问得不对」的问题。

三个房间里，机器都在做同一件事：把定好了的题做得更快。三个房间里，只有一个地方，冒出过一点点换题的苗头——在那个九岁孩子的一句话里。

这一点苗头，能不能长下去？这是下一篇要问的。

◆ 八、这个判断怎样才算错

判断能被推翻，才配写在纸上。

这个判断有两个推翻条件。

第一个：如果有一天，AI 能自己做出典范猜想，自己换掉「什么才算好问题」的标准，那么最上面那一层也会被接管。那时，「中国最空」就不再是要命的事——大家一起空。第 369 号对这一点有自己的看法，本书在第 48 章也会回来。

第二个：如果有人能找出这几十年里，中国本土长出、被世界认可的几个典范级的猜想，而且数量不少，那么「最空」两个字就要收窄。本书愿意被这样推翻，也欢迎读者把例子寄来。

还有一处，本书自己也拿不太准：三层之间的界线，并不总是那么清楚。一项工作是改进还是架构，常常要过了很多年才看得出；一个架构会不会长成新的典范，当时谁也说不准。表里给每一章分了层，这是为了看清，不是说界线已经画死了。

在这两个条件出现之前，本书先站在这个判断上往下走。

◆ 九、放回日子里

回到你自己的那张草稿纸。

明天能试：拿一张纸，画三条横线，分成三层。最下面写「改进」，中间写「架构」，最上面写「典范」。把你这一年做过的、最得意的三件事，各自写进它该在的那一层。

大多数人，会把三件事都写在最下面那一层。这很正常。

再问自己一句：最上面那一层，空着的那一格，是谁该来填?

◆ 十、给三种读者的一句话

这一篇写得有点硬，最后给三种读者各留一句话。

如果你在做科研：你做得最好的那件事，多半在常规科学那一层。这不是你的错，那是这一行几十年认定的好本事。只是从今天起，要留一点时间，给那件「不知道该不该做」的事。

如果你在管科研：一个组里人多、实验快、论文多，今后不再能说明这个组有多强。要找的，是组里有没有人在问「我们这一类题是不是问偏了」，以及有没有地方让他问。

如果你是家长：孩子会刷题、会提问，都很好；可是这两样，机器都在学。多留意孩子哪一次说「这道题出得不对」，别急着纠正他。

◆ 十一、指向下一篇

本篇讲完了「哪一层」。下一篇要问的是「哪个人」。

如果最上面那一层只能由人来猜，那么来猜的这个人，是什么样的人？他从哪里来？为什么中国找他那么难？

下篇一的回答，会让人不太舒服：那个人不是没被发现，是在很早的时候，就被一道道筛子筛掉了。

> **机器从表的第一行往下压，家底从第一行往下薄。**

> 带 走 的 话
>
> **AI 从表的第一行往下接管，中国的家底也厚在上面几段；两张图一叠，最强的先贬值，最空的还空着。**

> 小　账
>
> **看见了什么，漏看了什么？**看见了中国在改进与常规科学上的真本事；漏看了这些本事正好落在机器接管的那条线上。
>
> **是在改进，还是在出题？**本篇七章的成绩，几乎都是在已有的题里做得更好；出新题的那一层，至今空着。

> 想 一 想
>
> 你画的那三层里，你最得意的事落在哪一层？如果机器下个月就能做到那一层，你打算往上走一层，还是换一件事做？

◆



下　篇　一

筛子：那个人从哪里来

导　语

这一篇八章，从「哪一层」转到「哪个人」。第 25、26 章说复合体里那个人是谁，为什么他比天才还稀缺；第 27 到 29 章说中国选人的三道筛子，以及论文计分遇上 AI 以后会变成什么样；第 30、31 章说团队与方向，为什么共同体出击本身就不利于 0 到 1；第 32 章把本篇的账并排一张表。

上一篇说，AI 留下的是典范猜想那一层。这一篇问：来猜的那个人，在今天的中国，是怎样被找到的，又是怎样被错过的。

这一篇要让你看见一个反差：复合体里的那个人，比天才还稀缺；而中国的筛子，最会筛掉的，恰恰就是他。

25

第 25 章

复合体里的那个人

◆ 一、林舟想起的一个同学

那天开完会回来，林舟一直想着会上碰到的那个三个人的小组。第 22 章写过那次会。

那个组的负责人说，他们主要做两件事：想下一步该问什么，判断机器给的结果靠不靠谱。林舟想，这两件事，自己的组也在做，只是被淹没在十二个人的日常里了。

晚上睡不着，他忽然想起了一个大学同学，姓许。

许同学是个怪人。上课常常走神，作业交得乱七八糟，考试成绩在班里中下游。可是他会在实验课上突然问一个谁也答不上来的问题，比如「为什么我们只测这三个指标，不测别的」。老师常常被问住，有时候也不太高兴。

后来许同学考研没考上，去了一家小公司，再后来，林舟就没了他的消息。

林舟躺在床上想：如果今天有一台机器，能替许同学把作业做好、把实验做完、把论文写好，他会不会是那个三个人小组里最值钱的那一个？

这一章讲的，就是这个问题。

◆ 二、0 到 1 交给谁

第 369 号《典范发生学》有一个判断：典范发生，是人—AI 复合体的使命。

这句话分两半。

前一半说，0 到 1——从没有到有的那一次换题——不再是某个天才独自完成的事，也不是 AI 独自完成的事，是一个人和一台 AI 合在一起完成的事。

后一半说，这个复合体里的分工很清楚：人先作典范猜想，猜出一个新的「什么才算好问题」；然后交给 AI，去把这个猜想打造成一个典范，再打造这个典范下的整套系统。

上一篇讲了 AI 能做什么：常规科学，甚至常规科学里的提问。这一章要讲复合体里剩下的那一半：那个人。

◆ 三、王德生博士的一句口述

王德生博士在口述这本书的命题时，有一句话说得很短：复合体是人—AI，这里的人比少数天才更稀缺。

第一次听到这句话，很多人会觉得说反了。天才已经是最稀缺的人了，还有什么人比天才更稀缺?

要回答这个问题，先得看清「稀缺」有两种。

◆ 四、天才的稀缺：能力的稀缺

先说天才。

天才为什么稀缺？因为他的能力比别人高出一大截。同一道题，别人想三天，他想三个小时；同一条路，别人走到一半，他走到尽头，还多走了几步。

这种稀缺，是「能力的稀缺」。它有一个特点：天才和普通人走在同一条路上，只是他走得更远、更快。

所以天才是可以比的。你可以拿一个尺子量：谁的数学更好，谁的实验更巧，谁的论文更多更好。天才站在尺子的最右端，稀少，可是看得见。

这也是为什么，过去的天才大多被认出来了。爱因斯坦在专利局里，也有人读他的论文；拉马努金写信给哈代，哈代读懂了。他们的功底摆在那里，功底就是他们被认出的凭据。

第 26 章会把这一点展开：天才的猜想能力和他的常规功底，是长在一起的。

◆ 五、复合体里的人：位置的稀缺

复合体里的那个人，稀缺在另一处。

他不一定比别人更能干。他的功底，可能很一般；他的作业，可能像许同学那样乱七八糟。因为在复合体里，功底那一部分，交给了 AI。

那他靠什么？靠一个位置。

他必须恰好站在 AI 做不到的那一点上。

这一点在哪里？上一篇讲过：读文献、设计实验、写论文，AI 能做；在给定的标准里提出好问题，AI 也开始会了。AI 做不到的，是猜出一个新的标准，以及判断这个新标准值不值得走下去。

复合体里的那个人，就要站在这一点上。

这和天才很不一样。天才是在同一条路上走得更远；复合体里的人，是站在路的外面，一个机器还走不到的地方。

◆ 六、那一点为什么越来越窄

这里还有一层，更值得停下来想。

AI 做得到的事，是在扩张的。

五年前，AI 还不太会写论文；今天，它能写一篇像样的初稿。三年前，AI 还不太会读文献；今天，它四十分钟读完八十多篇。一年前，AI 提出的研究问题还很稚嫩；今天，林舟把它的建议直接抄进了基金本子。第 22 章写过那个晚上。

AI 做得到的那一片，每年都在往外长。

那么，AI 做不到的那一点，就在每年往里缩。

打个比方。一片海滩，潮水在涨。从前，海滩很宽，站在上面的人很多：会算的人、会写的人、会做实验的人、会提问的人，都有地方站。潮水一年一年涨上来，先淹掉会算的那一片，再淹掉会写的、会做实验的，现在正在淹会提问的那一片。

剩下还露在水面上的，只有一小块礁石。

复合体里的那个人，就是要站在这块礁石上的人。

而礁石，还在变小。

◆ 七、本书的说法

把两种稀缺并排放，就能看懂那句口述了。

天才的稀缺，是在一条大家都走的路上，走得比谁都远。这条路越宽，走的人越多，天才越容易被看见。

复合体里的人的稀缺，是要站在一个大家都不走、机器也走不到的地方。这个地方越来越窄，能站上去的人越来越少，而且站上去的人，往往看起来不像一个能人。

所以本书的说法是：复合体里的那个人，比天才更稀缺，不是因为他更聪明，是因为他要站的那个位置更窄，而且还在变窄。

天才的稀缺，是能力分布的尾巴；复合体里的人的稀缺，是一个正在缩小的位置。

◆ 八、只看一面会怎样

只看「能力」这一面，会以为复合体里的人就是天才，找人的办法照旧：考试、论文、头衔，挑最能干的那几个。

这不对。最能干的人，往往是常规科学里最好的人。他们的本事，正是 AI 正在接管的那一片。拿考常规科学的尺子去找复合体里的人，找到的多半是一双最好的腿，不是那个站在礁石上的人。

只看「位置」这一面，又会走到另一头：以为谁都可以，随便找一个怪人、一个考不好的人、一个不守规矩的人，就是复合体里的那个人。

这也不对。站在礁石上，不等于站在海滩外面的随便哪里。那一点是有方向的：它要正好在 AI 做不到、而又值得做的那个地方。一个人只是怪，不等于他站对了地方。第 37 章会回来讲：一个人加一台 AI，最容易的事，就是把一个错的猜想打造得越来越像对的。

本书两头都不取：复合体里的人，不是最能干的人，也不是随便哪个怪人，是恰好站在那一点上的人。

◆ 九、历史上的一个影子

历史上有没有过这样的人？没有 AI 的时代，很难找到完全对得上的例子。可是有一个影子，值得看一眼。

先看一张桌子。十九世纪二三十年代，伦敦皇家研究所的实验室里，桌上摆着一块磁铁、一张纸、一撮铁屑。法拉第把铁屑撒上去，轻轻敲一敲纸，铁屑排成一圈一圈的弧线。

法拉第是铁匠的儿子，十几岁在书店当装订学徒，没上过几年学，也几乎不会高等数学。他从这些弧线里猜出一个念头：电和磁的力，不是隔着空气直接扑过去的，是沿着空间里一条条看不见的「力线」传过去的。

这个念头，在当时那些擅长计算的物理学家看来，有点像外行人的比方。它画得出来，算不出来。法拉第自己也写不出它的方程。

请注意他的位置。他不是在当时数学物理的路上走得最远的人；论计算的功底，他比不过那一代的数学家。他是站在路外面，看着一桌铁屑，换了一个问法：不问两个东西之间隔空有多大的力，问力在空间里怎样铺开。

他缺的那一块，后来由另一个人补上。1850 年代起，年轻的麦克斯韦把法拉第的力线一点一点翻成数学，到 1860 年代写成了电磁场的方程组。猜在一个人身上，算在另一个人身上，中间隔了二三十年。

今天，算的那一块，正好是 AI 能补的。设想一个法拉第加一台 AI：他出那一猜，机器替他去搭模型、写方程、对数据。那二三十年的等待，也许会短很多。第 36 章要讲的魏格纳，也是一个缺了「算」的猜想者，他等得更久。

这只是一个影子，不是证据。它只说明一件事：站在路外的那一猜，和它背后的功底，本来就可以分开；从前分开了，猜的人要等另一个人来补，今天它们可以由人和机器分着扛。

◆ 十、今天与机器

还有一件事，这个位置是会动的。

去年站在礁石上的人，今年可能已经被潮水淹了。一年前，「会提问」还在礁石上；今年，它正在被淹。一个人今天站对了位置，明天不一定还对。

所以，复合体里的那个人，不能只是「站在那里」，还要一直看着潮水在哪里，一直往机器还够不着的地方挪。

这和天才又不一样。一个天才练成了一身功夫，这身功夫可以用一辈子。复合体里的人没有可以用一辈子的功夫，他的那一点本事，每年都要重新找。

这也是为什么，这种人很难培养。你没法给他一套教材，因为教材上写得出来的东西，机器明年就会了。

更难的是，他也很难被认出来。

认一个天才，可以看他做出了什么。认复合体里的这个人，看他做出了什么往往不够，因为做出来的那一大半，是机器干的。真正属于他的，只有那一猜，和他后来一次次「这条路对、那条路不对」的判断。这两样，都不太能写进简历，也不太能拿到评审会上去比。

一个人的价值，落在一份简历量不出来的地方。这件事本身，就是一道筛子的难题。第 27 章到第 29 章，要把这道难题拆开来看。

◆ 十一、一个完整的例子：两种考法

拿一场想象中的考试，把两种稀缺看清楚。

第一种考法：给一百个学生同一道难题，看谁做得最快、最好。最好的那一两个，就是天才。这种考法，中国最熟悉，也做得最好。

第二种考法：给一百个学生同一台很强的 AI，告诉他们，这台 AI 能把所有的题都做出来。然后问：「你想让它去做什么？」

大多数学生会让 AI 去做老师布置过的题，或者去做一道比老师布置的更难的题。这些都在已有的框里，AI 会做得很好。

有极少数学生，会说一些奇怪的话：「我想让它去证明我们一直在做的这一类题，本身问错了。」或者：「我想让它去找一类从来没人当成问题的问题。」

第一种考法里的天才，在第二种考法里不一定出彩，他太会做题了，可能想不到要换题。第二种考法里的那极少数人，在第一种考法里可能排在中下游，像许同学那样。

第二种考法，今天的中国几乎没有。所有的筛子，用的都是第一种考法。第 27 章接着讲这件事。

◆ 十二、放回日子里

回到林舟那个睡不着的晚上。

他想起许同学，不是因为许同学是个天才。许同学不是。他想起许同学，是因为许同学总是站在课堂的边上，问一些不在课本里的问题。那些问题，当年让老师不太高兴，让考试成绩不太好看，也让他没能考上研究生。

林舟忽然明白了一件事：在今天的复合体里，最值钱的，可能不是自己这种把作业做得最好的人。

他也明白了另一件事，更让他不安：就算许同学今天就坐在他面前，他也不知道该怎样把这个人招进组里。

组里招人，看的是学历、论文、推荐信。许同学这三样都拿不出来。基金本子上要写团队成员的成果，许同学也写不进去。他能做的，最多是请许同学吃一顿饭，听他说说那些不在课本里的问题。

明天能试：在你的组里、你的班里、你的公司里，找一个「总问不在课本里的问题」的人。下次开会，不要急着把他的问题拉回正题，给他五分钟，让他把那个问题说完。然后问一句：「如果机器能把我们现在的活都做了，你这个问题，值不值得让它去试一试？」

天才走在大家的路上，走得更远；那个人站在路外，站在机器够不着的地方。

带 走 的 话

天才稀缺在能力，复合体里的人稀缺在位置；机器做得到的越多，他能站的那一点就越窄。

小　账

看见了什么，漏看了什么？看见了天才稀缺，也看见了 AI 接手常规功底；漏看了功底交出去以后，留给人的那一点在一年年缩小。

是在改进，还是在出题？天才多半是在已有的路上走得更远；复合体里的那个人，要在路外出一道机器还出不了的题。

想 一 想

你身边有没有一个像许同学那样的人，成绩不好，却总问课本外的问题？你当时是怎样看他的？如果今天再遇见他，你会怎样看？

◆

26

第 26 章

剥掉功底之后还剩什么

◆ 一、林舟的那个晚上

周四晚上十点，研究院的楼里只剩林舟一个人。

他面前开着两个窗口。左边是一份三个月前的组会纪要：一个靶点，两万个候选分子，三个博士生分头做对接、打分、跑分子动力学。计划是用一个季度，筛出二十个值得合成的。

右边是一个下午的结果。他把同一个任务交给一套 AI 智能体：它自己调了 AlphaFold 的结构，自己写了筛选流程，自己跑完，自己画好图。最后，它还附了一段像模像样的方法学描述。

二十个分子，有十四个和组里三个月筛出来的重合。剩下六个，他看了半天，挑不出毛病。

他把椅子往后推了推，盯着天花板。一个问题冒出来，赶不走：

「要是组里的活全交给它，我还剩什么？」

这一章，就从这个问题往下挖。它不只是林舟一个人的问题。前一章说过，0 到 1 要交给人和 AI 合起来的那个复合体；这一章要问的是，复合体里那个人，到底要带着什么进去。

◆ 二、爱因斯坦也要自己学张量

先回到一百多年前。

1907 年前后，爱因斯坦有了一个后来他自己很看重的念头，大意是：一个从高处自由落下的人，感觉不到自己的重量。从这个念头出发，引力和加速度可能是一回事。这是一个典范级的猜想，它要改写的是牛顿以来人们对引力的整套看法。

猜想有了，路还很远。要把「引力就是时空弯曲」写成方程，需要一套当时物理学家几乎都不会的数学：黎曼几何和张量分析。

据后来的记载，1912 年爱因斯坦回到苏黎世，去找他的大学同学格罗斯曼。格罗斯曼是数学家，他告诉爱因斯坦，黎曼那一路的几何正好合用，又陪他一起啃。两人在 1913 年合写过一篇论文，方程还没写对。爱因斯坦又花了两年多，改了又改，到 1915 年才写出引力场方程。

请注意这段经历的形状。猜想在前，功底在后；可是没有那几年的功底，猜想就只是一个好念头，进不了物理学。爱因斯坦在 1912 年给索末菲的信里说过大意如此的话：他从前小看了数学，这一回才知道它的分量〔核：大意〕。

他是天才。可天才也得自己一行一行地推，推错了再推。

◆ 三、Hinton 也要自己调网络

再看离今天近一点的人。

1980 年代，神经网络在主流人工智能界并不吃香。Hinton 和同事们一边坚持「让机器从数据里自己学」这个方向，一边一行一行地写程序、调参数、试结构。1986 年那篇反向传播的论文，是他和 Rumelhart、Williams 合写的；在那之前和之后，是几十年的实验、失败、再实验。

到 2012 年，他的两个学生用深度卷积网络在 ImageNet 比赛里大幅领先，深度学习才被整个行当认真看待。

这里也是同一个形状。「神经网络能学会东西」这个猜想，他在冷宫里坚持了很久；能让别人回头看他，靠的是一次次跑出来的结果。第 19 章讲过那段冷宫，这里只取一点：他的坚持，是穿着功底的衣服被人认出来的。

AlphaFold 也有同一个影子，第 20 章讲过它，这里只取一点。它能被结构生物学界很快接住，部分原因是这一行里早有一大批做了几十年实验的人。他们懂行，会验，能一眼看出机器预测的结构靠不靠谱。

猜想可以来自一个新来的团队；认得出它的，仍是一群有功底的人。

◆ 四、功底的两份差事

把这两个人放在一起，功底在他们身上干着两份差事。

第一份是地基。一个猜想要变成能检验、能计算、能做实验的东西，中间要铺一条很长的路。爱因斯坦要学张量，Hinton 要调网络，都是在铺这条路。

第二份是凭据。一个新典范刚提出来时，几乎总是看起来不对。同行凭什么愿意多看它一眼？很多时候，凭的是提出者在旧典范里的功夫：这个人会做难的题，他算的东西经得起查，所以他说的怪话也许值得听一听。1905 年爱因斯坦还是专利局的职员，没有教职，他那几篇论文能被物理学界接住，部分原因是论文本身的计算干净、经得起核。

用大白话说：功底一边帮天才把猜想落地，一边替天才向同行担保。

这两份差事，在过去是长在一起的。一个人在常规科学里磨得够久，磨出了手艺，也磨出了名声；他在磨的时候看见了旧典范的裂缝，于是提出了新典范；他的手艺让他能把新典范做出来，他的名声让别人肯看。猜想、功底、凭据，三样是一根藤上的。

◆ 五、本书的说法

现在，AI 把这根藤剪开了。

第 22 章说过，给定典范下的常规科学，正在交给机器。林舟那个下午看到的，就是这一刀：对接、打分、跑流程、写方法学，正是他和他的学生多年磨出来的功底，机器一个下午做完了。

功底交出去以后，留给复合体里那个人的，只剩两样：

一样是纯粹的典范猜想。不是在旧题里找一个更好的答案，是提出一个新的「什么算好问题、什么算好答案」。

另一样是判断。AI 会顺着这个猜想去打造系统：建模、计算、设计实验、写成论文。那个人要判断，机器打造出来的东西，是不是还在他猜的那个典范里，有没有悄悄滑回旧典范。

这两样都很难。更难的是，它们现在要光着身子站出来。没有功底给它铺路，也没有功底替它担保。

所以本书说：

剥掉功底之后还剩东西的人，比天才更少。

天才是猜想和功底长在一起的人，这种人历史上本来就少。复合体要的那个人，要在没有功底托底的情况下，单凭猜想和判断站住。过去，一个人的猜想能力常常是在磨功底的过程中长出来的；现在那条磨的路被机器走掉了一大半，猜想要从哪片土里长？这是一层难。即便长出来了，他拿什么让人相信？这是第二层难。

◆ 六、凭据丢了，谁来担保

这里还藏着一层，比「猜想从哪片土里长」更扎手。

过去，一个人在一个行当里磨了二十年，他的名字就是一张担保单。他说出一句怪话，同行会想：这个人算东西从来不出错，他说的怪话，也许有它的道理。这张担保单，是功底一笔一笔攒出来的。

现在想一想复合体里那个人。他的推导是机器做的，他的代码是机器写的，他的论文有一半是机器起草的。同行看着他的成果，心里会冒出一个很自然的疑问：这里面哪一块是他的？

功底越是交给机器，这个疑问越大。一个人的手艺被机器替下来以后，他的名字就不再自动替他担保。剩下能替他担保的，只有他猜对过几次。可是典范猜想的对错，往往要十年、二十年才见分晓。

于是会出现一个空档：在他第一次猜对、被人看见之前，他几乎没有任何凭据。这一段空档，过去由功底填上；今天没有东西填。

这一层，第 36 章和第 45 章会接着讲：谁来判一个新典范，怎样给它一个新的评审办法。这里只记一笔：复合体里那个人难，不只难在长出来，还难在长出来以后，没有一张现成的担保单。

◆ 七、只看一面会怎样

这里容易走偏到两头。

一头说：功底没用了，以后只要会想就行。这不对。判断「机器做出来的东西还在不在这个典范里」，本身就要懂行。一个不知道旧典范长什么样的人，看不出机器什么时候滑了回去。功底没有消失，它从「自己做」变成了「看得懂别人做的」，从手上的活变成了眼里的尺子。

另一头说：功底还是一切，先把基本功练扎实，别的以后再说。这也不对。如果一个人带进复合体的只是功底，那他只是一台慢一点的机器。机器一年比一年快，他一年比一年老；他的位置，迟早被那台机器自己坐满。

打个比方。一个老木匠，手上的活被机器替下来以后，他还能做什么？他能看一眼机器出来的椅子，说：这条腿的榫，迟早要松。他说得出来，是因为他亲手做过几百把椅子，松过几十把。可他的徒弟的徒弟，从来没亲手做过，他们靠什么看出那条榫要松？

这是功底留下的一个难题，本书解不开，只能记在这里：判断要靠懂行，懂行过去靠自己做；自己做的机会越来越少，懂行从哪里来？

两头各对一半。功底还要，可它不再是答案；它只是让人能看懂、能判断的那副眼镜。真正要带进去的，是功底剥掉之后还剩的那一块。

◆ 八、今天与机器：它会顺着你走，也会悄悄拐弯

为什么判断这么要紧？因为机器有一个天然的倾向。

今天的大模型，是从已经发表的文献里学出来的。旧典范在文献里写了几十年、几百万篇；新典范可能只在一个人的脑子里，加上几页草稿。你把一个新猜想交给它，它会认真地帮你建模、推导、写代码。可是在每一个你没说清楚的地方，它会用它最熟悉的东西去补。它最熟悉的，正是旧典范。

结果可能是这样：一套漂亮的系统，图表齐全，逻辑通顺，可它的核心假设在第三步已经换成了旧的那一套。表面上它在验证你的猜想，实际上它在用新名字重复旧结论。

第 37 章会回来讲这件事更危险的一面：机器顺着一个人的典范打造系统，打造得越完整，就越像被证实了。那里讲的是寡头的自我证实；这里只说一件小

事：复合体里的那个人，要时时问一句「它还在我的路上吗」。这一句，机器自己不会问。

◆ 九、一个完整的例子：林舟组里那张打分表

下面这个例子是虚构的，用来把上面几节捏到一起。

林舟心里一直有一个说不太清的想法。组里筛分子，用的是行当里通行的那一套：看分子和蛋白口袋结合得紧不紧，打一个分，分高的留下。他越做越觉得，有一类病，药效也许不全在「结合得紧」，而在分子进去以后，蛋白的整个形状在一段时间里怎么变。他没有证据，只有一些对不上的实验数据和一种别扭的感觉。

这算不算一个新典范的苗头，本书不判断。它至少有一个典范猜想的样子：它要改的不是哪个分子更好，是「什么叫好分子」。

那天下午之后，林舟试着把这个想法讲给 AI 智能体，让它设计一套新的筛选流程。机器很快交回一份方案：先模拟蛋白在分子进入后的构象变化，再据此给分子排序。看起来正是他要的。

他往下翻，翻到排序的那一步，停住了。机器在「据此排序」的地方，用的还是结合能。构象变化被算出来了，可最后决定谁留下的，仍是旧的那张打分表。新想法被请进门，坐在客厅里喝茶，做决定的还是原来那个人。

林舟改了三次提示，才让机器真的换掉那张表。改的过程里，他用到的不是对接的手艺，那部分机器比他快；他用到的是另一种东西：他知道旧打分表长什么样，所以认得出它换了衣服回来；他心里有那个别扭的感觉，所以知道新排序应该在哪里和旧的分岔。

这一番来回，前后不过两个小时。如果在过去，同样的事要一个博士生改三个月的代码，他未必有耐心等到结果，多半会在第二个月就退回旧打分表。机器把这三个月压成了两个小时，于是他那一点别扭的感觉，第一次有机会被真正试一试。

试出来的结果并不漂亮。新排序挑出来的分子，有一半看上去很怪，组里任何一个学生都会把它们划掉。他不知道这些怪分子是对的，还是机器又在别的什么地方拐了弯。

可是他知道，这个问题是他的。机器能把它跑完，不能替他决定该不该跑、跑出来以后该信几分。

那个晚上，他第一次觉得问题有了一点答案。

◆ 十、放回日子里

回到周四晚上十点的林舟。

他没有关电脑，打开一个空白文档，把自己这一周干的事一件件列下来。改了两个学生的论文，审了一份基金本子，开了三个会。调了一个模型的参数，跟合作方吵了一架。还有半小时，他盯着一组对不上的数据发呆。

然后他在每一件后面标一个字。机器现在就能做的，标「交」；机器还做不了、可是三年内大概能做的，标「等」；他想不出机器怎么做的，标「留」。

标完一看，「留」的只有两行：盯着那组对不上的数据发呆的半小时，和那天下午认出旧打分表换了衣服回来的那一下。

他有点难受，也有点踏实。难受的是，一周五十多个小时，「留」下的不到一个小时。踏实的是，这一个小时，他知道自己该把它养大。

明天能试：把你这一周的工作也这样列一遍，标「交」「等」「留」。不必给别人看。看一看你的「留」有多少，它们是从哪里长出来的。

机器拿走功底以后，你身上还剩下的那一块，才是你。

带 走 的 话

过去，功底替天才铺路也替他担保；功底交给机器以后，猜想和判断要光着身子站住。

小　账

看见了什么，漏看了什么？看见了天才身上功底的两份差事；容易漏看的是，判断机器有没有滑回旧典范，仍要懂行。

是在改进，还是在出题？交给机器的那部分功底是改进；剥掉之后还剩的那一块，才是出题的人。

想 一 想

你这一周的工作里，标「留」的有几行？它们是在磨功底时长出来的，还是另有来路？

◆

27

第 27 章

全世界最好的腿部选拔

◆ 一、林豆豆的错题本

林豆豆有一本错题本，封面上贴着一张小猫贴纸。

晚上九点半，她还趴在书桌上，用红笔把今天做错的三道奥数题抄上去，在旁边写上正确的解法，再用蓝笔写一行「错因」。第一道写的是「审题不清」，第二道写的是「计算粗心」，第三道她想了很久，写的是「不会」。

宋岚走进来，看了一眼：「不会的那道，明天问老师。先把会的做对，会的做对就有分。」

豆豆点点头，又抬头问：「妈妈，这道题为什么要这样出？」

宋岚愣了一下，笑着摸摸她的头：「考试又不考为什么。快睡吧。」

这个晚上，中国有几千万个孩子在做差不多的事。这一章，讲的就是这件事背后的那一把尺子。

◆ 二、它量什么：准、快、少错

高考是全世界规模最大、效率最高的选拔之一。1977 年恢复以来，它每年把上千万考生排成一列，排得清清楚楚，几乎不靠任何人的面子。

它量的东西很明确，可以压成三个词：答得准、做得快、错得少。

一道题，有一个标准答案，按步骤给分。时间有限，题量很大，谁在规定时间里拿到的标准答案多、丢的步骤分少，谁排在前面。豆豆的错题本，就是这套尺子往下传到小学四年级的样子：把「错」一个一个找出来，一个一个消灭。

这套尺子养出来的能力，第 13 章叫它「解题的能力」：题是别人出好的，答案是事先定好的，你要在题里做到最准、最快、最稳。

用第 369 号《典范发生学》的比方，这是在选腿。谁的腿跑得快、跑得稳、不摔跤，谁就被选出来。而且中国把这件事做成了全世界最好的腿部选拔。

◆ 三、先把公道还给它

说它只量一种东西之前，先把它的公道说足。

高考的公平是真的。在一个人情、关系、出身都很重的社会里，一张谁都不认识谁的卷子，一个只看分数的录取线，是一扇对所有人都开着的门。四十多年里，无数农家孩子、小城孩子、父母不识字的孩子，从这扇门走了出来。

王德生博士本人也是农家孩子，父亲种地，他复读一年才考上大学。对这样的孩子来说，高考不是一道筛子，是一条出路。没有这条路，后来的一切都谈不上。

而且这把尺子量得准。答得准、做得快、错得少，是常规科学最需要的能力。中国四十年的工程奇迹、制造业、基础设施，背后都有这一批批被选出来、又被训练得极好的人。第 21 章说过，中国科研在前沿成绩上一点不弱，那些成绩，很多正是这把尺子选出来的人做成的。

还有一层公道常被忘记：高考不只选人，也给了一代代人一个共同的底子。几亿人读过同样的课本，会解同样的方程，认得同样的化学式。这个底子，让中国在需要大规模工程师、技术工人和研究人员的时候，能一下子拿出来。很多国家没有这个底子。

所以，本书不说高考错了。它做的事，它做得很好。

◆ 四、公平和单一，是同一枚硬币

问题出在一个容易被忽略的地方：高考为了公平，必须有标准答案。

一道题要能被几十万阅卷老师打出几乎相同的分数，它就必须事先有一个公认的对。公认的对，从哪里来？从已经定型的知识里来，也就是从给定的典范里来。一道「这个问题本身问得对不对」的题，没法公平打分；一个「也许课本错了」的答案，在阅卷标准里只能算错。

所以，高考的公平和它的单一，不是两件事，是一枚硬币的两面。它越想公平，就越要量那种有标准答案的东西；它量得越准，就越只量这一样东西。

这不是谁的设计失误。在一把要给上千万人排队的尺子上，这几乎是唯一可行的形状。

作文题看起来是个例外，它没有唯一的答案。可是你问一问身边的高三老师，他们会告诉你，作文也有一套得分的路数：开头怎么点题，中间用几个例子，结尾怎么升华。阅卷老师几分钟看一篇，他要找的，是那套路数在不在。一篇真正想得与众不同的作文，在几分钟里，常常不如一篇路数整齐的作文吃香。

所以，连最该留给「怪想法」的那一道题，也被尺子慢慢拉成了有标准答案的样子。

◆ 五、本书的说法：越会选腿，越早筛掉那种人

现在回到复合体里那个人。

第 26 章说过，他要带进去的，是剥掉功底之后还剩的那一块：典范猜想，加上判断。这种人在少年时，常常是什么样子？

本书不敢画像，只能说几种常见的苗头。他会问「这道题为什么要这样出」。他做题慢，因为他总想换一种路走。他错得多，而且错得有意思：不是粗心，是他在用一个课本上没有的想法。他对已经会做的题没兴趣，对没有答案的问题着迷。

把这几条和高考量的三样放在一起看：准、快、少错。几乎每一条都是反着的。

所以，一个越会选好腿的体系，越会在越早的时候，把那种人排到后面。不是因为有人讨厌他，是因为他的长处在尺子上没有刻度，他的短处在尺子上清清楚楚。

更要紧的是，十二年刷题不只是在选人，也是在改人。一个本来爱问「为什么」的孩子，问一次被说「考试不考」，问十次被说「先把会的做对」，到了十八岁，他多半已经不问了。筛子筛掉的，不只是那种孩子，还有每个孩子身上的那一块。

◆ 六、筛子提前到了九岁

这把尺子还有一个变化，这些年越来越明显：它往下走了。

高考在十八岁。为了十八岁那一天，中考变得很要紧；为了中考，初中要进好的；为了好初中，小学就要攒奥数、英语、各种竞赛的成绩。一层压一层，尺子的影子一直落到了林豆豆的四年级书桌上。

豆豆班上的家长群，每天晚上都很热闹。有人发一张练习卷，有人问哪个机构的思维课好，有人晒孩子拿的证书。没有人在群里问：「你家孩子最近问了什么怪问题？」

这样一来，筛子不只在十八岁那天筛一次。从九岁起，孩子每天都在被同一把尺子量。准、快、少错，一年量三百多天，量九年。

一个爱问「为什么」的孩子，在十八岁那天被筛掉一次，也许还有别的路。可是从九岁起每天被量一次，他身上那一块，多半等不到十八岁。

还有一个新的东西进了这张书桌：AI。

豆豆会用 AI 写作业。她把一道应用题拍下来，机器一秒钟给出步骤和答案。宋岚发现以后，没收了平板，又有点犹豫：她自己白天就是靠 AI 出剧的，凭什么不让孩子用？

这件事难，难在两头都有道理。不让用，孩子以后要和会用 AI 的人站在一起；让用，错题本就成了抄题本。可是不管让不让用，有一件事是一样的：考试量的那三样，机器都比孩子强。一个孩子花九年练机器最擅长的事，练得再好，也只是在和一台不会累的对手比耐力。

◆ 七、只看一面会怎样

看到这里，有人会说：那就废掉高考，改成综合评价。

这一步要很小心。综合评价在别的地方也有，它常常把一扇公平的门，换成了一扇看家庭、看关系、看资源的门。农家孩子最先被挤出去。为了一种还说不清的能力，拆掉一扇真实的出路，代价会落在最没有退路的人身上。

也要看到，高考自己一直在改。新高考给了学生选科的余地；2020 年起的强基计划，在总分之外给基础学科留了一条路。这些改法都很小心，先护住公平，再一点点加别的刻度。这份小心，本身就是一种公道。

也有人会说：高考没问题，孩子有好奇心，课外自己保护就行了。

这也只看了一面。十二年里，一个孩子醒着的时间大半交给了刷题；一个家庭的钱、时间、焦虑，大半跟着分数走。课外那一点空，挡不住整套激励往一个方向推。

两头都不对，说明这不是「要不要高考」的问题。高考可以照旧量腿；问题是，整个社会除了这把尺子，几乎没有第二把。

◆ 八、今天与机器：最好的腿，遇上了车

还有一件事，这几年越来越扎眼。

据报道，近几年每到高考结束，就有人拿当年的数学卷去考大模型，成绩一年比一年好〔待核：具体分数各家测评不一〕。不必信哪一个具体分数，趋势已经很清楚：准、快、少错，正是机器最擅长的三样。

这就出现了一个很别扭的局面。一个社会花了最大的力气、最公平的办法，选出最会做标准答案的人；而标准答案，恰好是机器最先拿走的东西。

第 22 章说过车替腿的比方。拿它来看，今天的高考，是在车已经开上路的时候，仍在全力举办一场全世界最好的赛跑。跑得最快的孩子当然了不起。可那个在跑道边上发呆、想着「为什么一定要用腿」的孩子，没有被记下名字。

还要看到另一面。一个在跑道上跑赢了的孩子，十八岁进了好大学，二十八岁读完博士，三十八岁像林舟一样带起了组。他这一路的每一步都是对的。走到三十八岁，他才发现，自己最拿手的那一样，机器一个下午就做完了。

这不是他的错，他照着规则一路赢了上来。只是这条规则定下的时候，还没有车。

◆ 九、一个完整的例子：鸡兔同笼

豆豆那天那道「不会」的题，是一道鸡兔同笼的变形。

老师教过一个好用的办法，叫抬腿法：让每只兔子抬起两条腿，笼子里每只动物就都只有两条腿着地。总腿数减一减、除一除，兔子的只数就出来了。这个方法快、准、不容易错，是一个好办法。

第二天，豆豆没有问那道题怎么做，她问了另一个问题：

「要是笼子里还有三条腿的动物呢？」

老师笑了，说这个考试不考，先把两种动物的做熟。

这个回答没有错。在那一天，在那间教室里，它是对的。可是豆豆问的那一句，恰好是一个小小的出题动作：她不满足于在已有的题里做得更好，她想动一动题本身。如果有人顺着她往下走，她可能会自己发现，抬腿法背后是一个更一般的想法，三种动物要多一个条件才算得出来，再往后就是方程。那是她自己走出来的一条路。

可是这条路在错题本上没有位置。错题本只有两栏：错在哪，对的是什么。没有一栏写「我想换一道题」。

◆ 十、林舟也是这把尺子的赢家

那天晚上，宋岚把豆豆的问题讲给林舟听。林舟听完，半天没说话。

他想起自己。他是县里那年的理科前几名，一路考上来，硕士、博士、出国做博士后、回国带组。他一直觉得自己是这套尺子的受益者，这也是事实。

可是第 26 章那个晚上之后，他开始怀疑另一件事：他是不是也曾经是那个问「三条腿」的孩子，只是问着问着，就不问了？

他不记得了。这大概就是答案。

他倒是记得另一件事。高三那年，有一道物理压轴题，他想出了一个和标准答案完全不同的解法，步骤更短。他很得意，交上去，得了一半的分。老师在卷子边上写了一行字：「思路新颖，但不符合评分步骤。」

他那天有点委屈。第二天，他把标准步骤抄了三遍。高考那天，他用的是标准步骤，拿了满分。

那张卷子他早就丢了。那行字，他一直记得。

◆ 十一、放回日子里

本书不劝任何家庭放弃高考，也不劝孩子少刷题。路要走，门要进。

一个小办法，只花几分钟：在错题本旁边，再放一个小本子，叫「怪问题本」。孩子问出一个考试不考的问题，不急着回答，也不急着说「先做会的」，让她自己写进去。写一句就行，比如「要是有三条腿的动物呢」。周末吃饭的时候，随便挑一个，全家一起想五分钟，想不出来也没关系。

这个本子不会让分数变高。它只做一件事：让那一块不被磨掉。

今晚可以试：孩子问你「为什么」的时候，先别说「考试不考」，问回去一句「你觉得呢？」

> **一把只量腿的尺子量得再准，也量不出那个想换掉腿的人。**

带走的话

高考的公平是真的，它的单一也是真的；两样是同一枚硬币，所以只能在它旁边再放一把尺子。

小　账

看见了什么，漏看了什么？看见了高考给农家孩子的出路和它量得准；漏看的是十二年刷题不只选人，也在改人。

是在改进，还是在出题？错题本教孩子在已有的题里做得更好；那本「怪问题本」，才是在练出题。

想一想

你家孩子最近一次问「为什么」，你是怎么回答的？你自己上一次问一个考试不考的问题，是多少年前？

◆

28

第 28 章

博士、基金、职称：三道筛子

◆ 一、开春写本子的夜

每年开春，林舟都要过一段不太像人的日子。

国家自然科学基金的申请截止日一到，研究院里的灯就亮到很晚。他要写自己的面上项目，还要帮组里两个年轻人改青年基金。三份本子，每份几十页，里面有一栏他每次写到都要停很久：研究基础与工作条件。

这一栏要写：你做过什么，发过什么，有什么初步数据，说明你有能力做成你申请的事。

今年他心里那个「蛋白形状怎么变」的想法，第 26 章写过，他很想写进本子里。可是写到研究基础那一栏，他发现自己几乎什么都填不了。没有发过相关的文章，没有成套的初步数据，只有一些对不上的旧实验和一种别扭的感觉。

他犹豫了两个晚上，最后还是把那个想法删了，换回自己做了八年的那个方向。那个方向，研究基础可以写满三页。

这一章讲的，是林舟一路走过来的三道关。

◆ 二、第一道：在导师的题里做

第一道筛子是博士训练。

在中国的大多数实验室里，博士生的题目来自导师的项目。项目有任务书，任务书有考核指标，学生的课题就是任务书里的一块。这样做有它的道理：学生有经费，有设备，有人带，几年内能做出结果，能毕业。

毕业要求通常写着要发多少篇论文、发在什么级别的期刊上。一个博士生在第三年还没有一篇像样的文章，焦虑是很实在的。

于是，一个理性的学生会怎么选？选最稳的题，用最熟的方法，在导师已经铺好的路上往前走一步。这一步走得越扎实，他就越像一个好博士。

这里也有中国特有的一层：导师的项目，多半本身就是在追国际前沿的热点。热点是别人先提出来的典范，中国的实验室接过来，在它下面做得更快、更多、更细。学生跟着导师，导师跟着热点，一层套一层，离那个出题的位置越来越远。

这道筛子训练的，正是第 13 章说的解题能力：题是导师出的，你要在题里做到最好。五年下来，他学会了怎样在一个给定的问题里做得又快又对，也慢慢学会了，不要碰一个没人给过的问题。

◆ 三、第二道：可行性与前期基础

第二道筛子是基金。

基金评审要回答一个很合理的问题：这笔钱给出去，能不能做成？于是本子里要写可行性分析，要写研究基础，要写团队构成。评审专家看的是：这个人有没有做过相关的事，有没有初步数据，有没有一个看起来能干活的团队。

对常规科学来说，这是一套好办法。它把钱交给最可能做成事的人，避免浪费。

请注意它对新典范意味着什么。一个新典范，按定义，就是之前没人这样做过。它不可能有「前期基础」，因为前期还没有发生；它的可行性也很难写，因为连「什么算做成」都还在被它重新定义。一个典范猜想走进基金评审，就像一个刚出生的孩子被要求出示工作经历。

所以林舟删掉那个想法，不是胆小。在这套规则下，删掉是唯一会被奖励的选择。

◆ 四、第三道：按论文和项目计分

第三道筛子是职称和各种人才称号，行当里叫「帽子」。

评职称、评帽子，各单位有各自的细则，大体都绕着几样东西转：发了多少论文，发在哪里，拿了多少项目，经费多少，获过什么奖。每一样都能折成分数，分数加起来，排出先后。

这样做也有它的道理。在评审人和被评人往往互相认识的环境里，一套能数的指标，能挡住不少人情。

不过，数什么，就长什么。一个四十岁的研究员，在三年一次的考核和五年一次的职称面前，会把时间投到最能变成分数的地方。一个可能要做十年、可能什么都做不出来的问题，在分数表上是零。

三道筛子连起来看：第一道教你在别人的题里做，第二道要你证明已经做过，第三道按做出来的数量给你位置。每一道都在问同一个问题：你在已知的题里做得怎么样？

◆ 五、先说公道：体制自己也看见了

说到这里，要先把公道还回去。

这三道筛子，帮中国在几十年里建起了世界上规模最大的科研队伍之一，论文数量、专利数量都排到了世界前列。从一穷二白到今天，没有一套能快速分配资源、能大规模考核的制度，这件事做不成。

体制自己也看见了问题。2018 年 10 月，科技部、教育部、人力资源和社会保障部、中国科学院、中国工程院五家联合发出通知，清理「唯论文、唯职称、唯学历、唯奖项」。这就是常说的「破四唯」。

2020 年，科技部又出了破除科技评价中「唯论文」导向的若干措施。核心一条是推行代表作评价：评一个人，看他拿出来的几篇最好的工作，不再数总篇

数。同一年，教育部和科技部也发文规范论文指标的使用。据公开报道，国家自然科学基金这几年还设了专门资助原创探索的项目类别，申请时更看重想法本身。

这些都是真的努力，不是摆设。本书要说的，不是没有人在改。

◆ 六、本书的说法：筛子的形状没变

本书要说的是：改的多半是筛子上的刻度，不是筛子的形状。

「破四唯」减少了论文数量的权重，可评审时仍要拿出东西来评；代表作评价把十篇换成几篇，可那几篇仍是已经发表的；原创探索项目放宽了前期基础的要求，可评审专家仍是旧典范里最好的那批人，他们判断一个想法有没有道理，用的仍是旧典范里的尺子。

形状是什么？是「凭已经做出来的东西，决定谁能做接下来的事」。这个形状对常规科学几乎是完美的；对典范猜想，几乎是天然的门槛。因为典范猜想恰恰是一个人还没做出来、而且旧尺子还量不准的东西。

三道筛子还会叠起来。一个人要在三道关上都不被筛掉，就得在博士阶段做稳妥的题，在拿基金时写满前期基础，在评职称时攒够分数。等他四十五岁，终于有了一点自由的时候，那个当年想问的问题，他多半已经不记得了。

所以，复合体里那个人不是没有被发现，是被一道道筛子反向选掉了。每一道都不坏，三道连起来，方向就定了。

◆ 七、评审的人，也是好腿

还有一层，藏得更深：三道筛子的守门人是谁？

是同行。博士答辩请的是本领域的专家，基金评审请的是本领域的专家，职称评审也是。同行评议是现代科学最好的发明之一，它让懂行的人来判断懂不懂行，挡住了外行，也挡住了大量的胡说。

可是同行之所以成为专家，是因为他们在现行典范里做得最好。他们是上一轮筛子筛出来的最好的腿。让最好的腿来评一个「也许不用腿」的方案，他们会怎么判？

不是他们心眼小。一个人用了三十年证明某条路走得通，当有人说这条路也许从一开始就绕了远，他第一个反应是去找这句话哪里错了。他多半找得到，因为一个刚提出的新典范，处处都有毛病。

你可以设身处地想一想。一位专家一个晚上要看十几份本子，每份几十页。翻到一份他看不懂的，他手里没有现成的尺子去量它。给高分，万一项目做砸了，他说不清当初凭什么；给低分，几乎不会有人来问。这不是哪一个人的胆子，是这个位置本身的处境。

所以三道筛子不只是规则，它还是一代代人。每一代被筛出来的最好的腿，变成下一代的守门人；他们手里拿的，是筛出自己的那把尺子。尺子就这样一代一代传下去，越传越稳。

◆ 八、只看一面会怎样

一种说法是：那就别评了，给钱让大家自由探索。

这一步也走不通。没有评价，资源就会流向会说话、有关系的人，结果比现在更糟。而且绝大多数科研本来就是常规科学，它们需要考核，需要有人对钱负责。

另一种说法是：那些真有想法的人，自己会冲出来。

历史上确实有冲出来的人。可是一个靠「真有本事就会冲出来」运转的体系，只能收下那些在三道筛子之外另有退路的人：有家底的，有海外机会的，运气特别好的。第 38 章会讲，很多人最后是去墙外才长成的。冲出来的人越少，越说明不是他们不够好，是路太窄。

两种说法都只看了一面。需要的不是拆掉筛子，而是在筛子旁边，开一条不按这个形状评的路。尾篇第 43 章和第 45 章会回来讲，那条路可能长什么样。

◆ 九、今天与机器：AI 让前期基础更便宜了

AI 的到来，会不会让这件事变好？

一种乐观的看法是：AI 能很快做出初步数据，前期基础的门槛就低了，新想法也能写满本子。

细想一下，AI 最快能做出初步数据的，是哪一类想法？是那些已经有成熟方法、成熟数据、成熟模型可以调用的想法，也就是常规科学里的想法。一个真正的新典范，连该算什么都还没定，AI 能帮的反而少。

于是，AI 把常规想法的前期基础变得更便宜，把新典范的前期基础留在原地。两者之间的差距，不是缩小了，是拉大了。评审专家面对两份本子：一份有漂亮的初步数据，一份只有一个怪念头。他会选哪一份？

第 22 章里，林舟已经把大模型列的方向抄进了基金本子。这本身不是坏事，省下的时间可以拿去想问题。可是如果省下的时间又被拿去写第二份、第三份本子，筛子就转得更快了。

评审那一头也一样。本子越写越漂亮，评审人越难从文字上分出高下，于是更依赖那几样硬东西：发过什么，拿过什么，在哪个单位，有没有帽子。机器让文字变得便宜，反倒让履历变得更贵。履历是什么？是在旧典范里攒下的东西。

◆ 十、一个完整的例子：小唐的选题

林舟组里有个博士二年级的学生，大家叫他小唐。

有一天组会散了，小唐留下来，有点不好意思地开口。他说读文献时发现，有一类老药的副作用，好像和它在体内停留的时间有一种很奇怪的关系。这件事，没人认真研究过。他想换个方向，做这个。

林舟听完，心里一动。这个想法和他自己那个「形状怎么变」的念头，隐约是一个方向。

可是他张嘴说出来的是：「想法挺好。先把手上这个课题做完，发两篇再说。你明年要中期考核。」

小唐点点头，出去了。

林舟一个人坐在办公室里，想起自己博士二年级的时候，导师也对他说过几乎一模一样的话。他当时也点了点头。

那句话没有错。在那个位置上，任何一个负责任的导师都会这样说。三道筛子最厉害的地方就在这里：它不需要一个坏人，它只需要每一个人都负责任。

◆ 十一、放回日子里

林舟后来做了一件很小的事。

他没有让小唐换题，中期考核是真的，毕业要求也是真的。他跟小唐约好：每周五下午四点到五点，这一个小时不谈课题，不谈论文，只谈那个老药的怪现象。不记进度，不写周报，不进任何考核。小唐可以拿 AI 去查文献、跑数据，有了任何发现，先只讲给他一个人听。

头两周，小唐有点放不开，每次都带着一页提纲来，像是在汇报。第三周，他什么都没带，进门就说：「林老师，我昨晚想到一件事，可能很蠢。」那一个小时，两个人在白板上画到快六点。

林舟后来才明白，这一个小时最要紧的，不是讨论出了什么，是小唐可以说「可能很蠢」。在三道筛子的世界里，没有一张表格收这种话。

一个小时很少。可它是这间实验室里，唯一一块不按三道筛子的形状长的地。

明天能试：如果你带学生、带团队，或者只是管你自己，在下周的日程里划出一个小时，写上「不计分」。在这一个小时里，只做一件不会变成任何指标的事。

> **三道关都守得很尽责，那个想法还是在第一道关前就被自己删掉了。**

带走的话

博士、基金、职称三道关各有道理，连起来问的都是同一件事：你在已知的题里做得怎样。

小　账

看见了什么，漏看了什么？看见了「破四唯」等改革和筛子挡人情的作用；漏看的是刻度改了，「凭已做出的东西定谁能做下一件」的形状没改。

是在改进，还是在出题？三道筛子奖励的，都是在已有的题里做得更好；一个新题，在这里没有前期基础可写。

想一想

你上一次删掉一个想法，是因为它不对，还是因为它写不进表格？你有没有一个「不计分」的小时？

◆

29

第 29 章

论文计分遇上 AI

◆ 一、饭桌上的一句话

周日晚上，林家吃饭。豆豆在讲学校里的事，林舟在想下周要交的两篇稿子。

宋岚夹了一筷子菜，随口问：「你们组今年打算发几篇？」

林舟说，按考核，组里要有十篇左右，最好有两篇进一区。

宋岚笑了：「跟我们出剧差不多嘛。我们是一个月上新多少部，你们是一年发多少篇。我们用 AI 一天能出好几集，你们现在用 AI，是不是也快多了？」

林舟想反驳，张了张嘴，没说出话。

她说的不是全部，可也不全是玩笑。这一章，就从这句话往下看。

◆ 二、写一篇论文，便宜了多少

先说一个大家都看得到的变化：AI 让产出一篇论文的成本骤降。

过去，一篇论文从数据到成稿，中间有很多费工夫的环节：查文献，整理数据，写分析代码，画图，写引言和讨论，英文润色，按期刊格式改。一个博士生做完这些，几个月是常事。

今天，这里的每一步都有 AI 能帮上忙。查文献可以让模型先读一遍、出个摘要；分析代码可以让它写；图可以让它画；英文可以让它改。一个熟练的人，把这些串起来，速度可以快上好几倍。

这里没有夸张。林舟组里就有学生，用两周时间写完了过去要两个月的稿子。

宋岚那边是同一个样子。第 5 章说过那个每天一千五百部的数字：成本一降，产量就翻着跟头往上走。

论文这边没有这样惊人的数字，至少本书没有找到可靠的统计。可是方向是一样的：一样东西的生产成本降到原来的零头，生产它的人如果按件计酬，件数就会往上走。

成本降下来，本身是好事。问题是，降下来的是哪一块的成本。

◆ 三、理性的选择

把一件事拆成两半看。

一半是「写出一篇论文」的成本：整理、分析、成文、投稿。这一半，AI 让它便宜了很多。

另一半是「想清楚一个更难的问题」的成本：在一个没人走过的方向上发呆、试错、走死胡同、推翻自己。这一半，AI 几乎没有让它便宜。它要的还是年头，还是风险，还是那种可能五年没有任何结果的耐心。

现在把考核规则放进来：按论文数量、期刊影响因子、项目数目计分。

一个理性的研究者面对这两半，会怎么分配他的时间？他会把 AI 省下来的时间，投到变得更便宜的那一半去。同样一年，过去写五篇，现在可以写十篇；而追问一个难题，过去要十年，现在还是十年。在分数表上，前者翻了一倍，后者还是零。

这不是谁的品德问题。一个规则把便宜的东西和昂贵的东西用同一种分数去数，人就会往便宜的那边挤。AI 只是让两边的价格差变得更大了。

◆ 四、科学版的注意力经济

宋岚那句话，说中了一层。

上篇讲过，中国的 AI 先被用在了回款最快的地方：短剧、广告、直播带货。第 6 章把这件事叫做「资本按回款速度排队」。短剧卖的是注意力，一部剧的价值，最后折成播放量、充值、广告收入。AI 来了，生产一部剧的成本降到原来的零头，于是一个月上新几万部。

论文在今天的评价体系里，也有一点像这样的商品。它的一部分价值，不在于它回答了什么，在于它能折成多少分数：一篇一区论文能换来考核、项目、帽子。本书把这种东西叫「指标商品」。

AI 来了，生产指标商品的成本也降到了零头。于是，同样的事情在另一个行当里发生了一遍：AI 先被用来更便宜地生产指标商品。

这里说「指标商品」，不是说论文没有用。一篇论文可以同时是知识，也是商品。问题在于两样东西的比例。当一个体系主要按件数和期刊级别去数，论文里「商品」那一部分就会越长越大，「知识」那一部分，未必跟着长。

这是科学版的注意力经济。短剧抢的是观众的眼睛，指标商品抢的是评审的眼睛。两边都不坏，两边都很有效率；两边也都有同一个问题：它们挣到的东西，会不会回流到下一种火药上去?

◆ 五、先说公道：AI 帮了很多老实人

这里也要先把公道说足。

AI 写作，首先帮的是很多老实做研究的人。一个英文不好的中国学者，有了好的数据和真的发现，过去常常因为语言被审稿人退回；现在，他能把文章写得和母语作者一样清楚。这是一种真实的公平。

AI 也让很多正经的工作变快了：整理数据少出错，代码更规范，图更清楚。省下来的时间，有人真的拿去想了更难的问题。

还有，论文多不一定就是坏事。常规科学本来就需要大量的论文，把一个典范里的细节一块块填满。中国在很多领域的论文产出，对全世界的知识是实打实的贡献。

所以，本书不说用 AI 写论文是错的。它说的是另一件事。

◆ 六、论文工厂与代写

在公道的另一边，也有被反复报道的问题。

这些年，国内外媒体和学术期刊都报道过「论文工厂」：有组织地批量生产、买卖论文，有的连数据都是编的。一些出版社也公开撤过成批的稿件。AI 出现以后，又多了一层担心：AI 代写、AI 生成的图片和数据，让造假变得更便宜、更难查。各家期刊和出版社都在加紧检测〔待核：具体做法各家不一〕。

本书不在这里报数字，这些数字各家口径不同，变化也很快。只说一个常识：当一样东西能换分数，就会有人专门生产它；当生产它的成本降下来，专门生产它的人就会更多。

还有一种更安静的变化，不上新闻。一个研究者用 AI 把一项工作拆成三篇发，每篇都是真的，数据也都是真的，只是切得更薄。过去这样切要花不少工夫，现在几乎不费力。没有人造假，可是一个问题被切成了三份指标商品。

论文工厂是这件事最难看的一端。可是要看清，它和那个用 AI 一年写十篇正经论文的研究者，站在同一条坡上，只是站的位置不同。推着他们往下走的，是同一股力。

◆ 七、本书的说法：AI 是放大器

把这几节放在一起看：

AI 没有在评价体系里放进任何新东西，它只是把原本就有的东西放大了。

一个看重论文数量的体系，遇上 AI，会得到更多论文。一个看重影响因子的体系，遇上 AI，会得到更多瞄准影响因子的论文。一个原本就有论文工厂的环境，遇上 AI，会得到更便宜的论文工厂。

反过来，如果一个体系原本就看重难问题、长周期、敢失败，AI 也会放大这些：研究者省下来的时间，会被投到难问题上去。

所以，AI 是一面放大镜。用它照一照，一个评价体系的形状会变得非常清楚。今天照出来的形状，和第 28 章那三道筛子的形状，是一样的。

◆ 八、审稿人那一头

放大镜还照到了另一头：读论文、审论文的人。

论文多了，审稿人并没有多。一位研究者一年要审的稿子，可能比过去多出不少，他能花在每一篇上的时间，就少了。据报道，已经有审稿人用 AI 帮忙读稿、写审稿意见，期刊也在为此定规矩〔待核〕。

于是会出现一幅有点好笑的图景：一边是机器帮忙写的稿子，一边是机器帮忙写的审稿意见，中间站着两个忙得喘不过气的人。稿子越写越像稿子，意见越写越像意见。

这样的循环，最擅长通过的是什么样的论文？是格式规范、方法常见、结论稳妥、和已有文献对得上的论文。也就是常规科学里的论文。

一篇写着怪想法的论文，在这个循环里会遇到双重的不利。写它的人没法让机器替他写，因为机器只会把它往熟悉的样子改；审它的人没有时间细看，而一个怪想法恰恰最需要细看。

所以，论文计分遇上 AI，不只是论文多了。它把整条评审的路都调得更顺了，顺的方向，是朝着旧典范的。

还有一件小事值得记一笔。林舟组里一个学生说，他用 AI 改完一篇稿子，读起来比原来顺多了，可他自己也说不清，原来那个有点拧巴的论点去哪儿了。机器把拧巴的地方抚平了。有时候，一个新想法就藏在那个拧巴的地方。

◆ 九、只看一面会怎样

一种反应是：禁止用 AI 写论文。

这挡不住，也不该挡。它会先挡住那些老实标注、老实使用的人，把好处留给不标注的人。而且它没有碰到真正的问题：就算没有 AI，按论文计分的规则还是会推人去写更多论文，只是慢一点。

另一种反应是：AI 让科研效率大增，论文多是好事。

这也只看了一面。效率增加的是常规科学那一层的效率。第 22 章说过，那一层正在交给机器。一个人用 AI 一年写十篇常规论文，看上去是在加速，实际上是在和机器抢一份正在被机器接走的活。等到机器自己就能写这十篇的时候，他这一年的十篇，值多少？

两头都没有碰到规则本身。规则不变，换什么工具，人都会往同一个方向挤。

◆ 十、一个完整的例子：十篇，还是一个问题

年底，林舟坐下来给组里排明年的计划。

他算了一笔账。组里十二个人，用上现在的 AI 工具，按目前的速度，明年发十二到十五篇不难，比今年多一半。考核会很好看，后年申请大项目，前期基础也能写得很厚。

他又算了另一笔账。如果抽出三个人、一半的时间，跟他一起追那个「蛋白形状怎么变」的问题，再加上小唐那个老药的怪现象，明年大概只能发八篇。追的那个问题，两年内可能什么都出不来，也可能被证明根本不对。

第一笔账，每一项都写得进考核表。第二笔账，写得进考核表的只有那个「八」。

他盯着两笔账看了很久。他知道，研究院里几乎每一个组长，在这个晚上，都会选第一笔。他也知道，第 26 章那个下午之后，第一笔账里的十五篇，有一大半，三年后机器自己就能写。

这一次，他没有马上做决定。

他想起宋岚说过的一件事。她的公司去年也算过一笔类似的账：用 AI 一个月出三十部剧，还是抽一个小组，花半年磨一部真正不一样的。公司选了三十部，这在商业上完全说得通。宋岚后来说，三十部里，她现在一部的名字都想不起来。

她说这话时没有后悔的意思，只是有点空。林舟听着，觉得那份空，他也认得。

◆ 十一、放回日子里

本书不告诉林舟该选哪一笔。他有十二个人要养，有考核要过，这些都是真的。

有一件事，谁都可以明天就做：把 AI 帮你省下的时间，单独记一笔账。

一个小办法：每周末花五分钟写两行。第一行，这周 AI 帮你省了多少小时。第二行，这些小时最后花在了哪里。连着记一个月，看一看，省下来的时间，是流回了更多同样的活，还是流到了一个更难的问题上。

如果你是带组的人，还可以再走一步：每年从 AI 省下的时间里，拿出一块，明着划给一个不计分的问题。写进组里的计划，不写进考核表。一年以后，看看这一块长出了什么。长不出来也是一个结果，它告诉你，那个问题也许不对，或者那一块还不够大。

这笔账不进任何考核。它只是让你看见，那面放大镜照出来的，是你的哪一种样子。

> **放大镜不挑东西，它照见什么，就放大什么。**

带 走 的 话

AI 让写论文便宜，没让想难题便宜；按论文计分的规则下，省下的时间会流回更多论文。

小　账

看见了什么，漏看了什么？ 看见了 AI 帮老实人写作的公平；漏看的是它也放大了指标商品，和短剧是同一种经济。

是在改进，还是在出题？ 用 AI 一年写十篇是在已有的题里做得更快；省下的时间流不到出题那一边，规则就没变。

想 一 想

这个月 AI 帮你省下的时间，流到了哪里？如果没有考核表，你明年最想追的是哪一个问题？

◆

30

第 30 章

大团队发展，小团队颠覆

◆ 一、一个四十人的会

周三上午，研究院的大会议室坐满了人。

这是院里一个重点平台项目的启动会。林舟的十二人组是其中一块，另外还有三个组、两家合作医院、一家做算力的企业。项目的名字很长，核心意思是：用大模型加自动化实验，系统地做一批靶点的药物发现。

会开了三个小时。每一块汇报分工，每一块讲里程碑，最后对齐考核指标。主持人说，这是共同体出击，要集中力量办大事。

散会的时候，林舟在走廊里想起第 28 章那个周五下午的一小时。那一小时，只有他和小唐两个人。这个上午，有四十个人。

他说不清哪里不对。这一章，替他把这一点说清楚。

◆ 二、团队出击，本身是反 0 到 1 的

本书有一个口述命题，大意是：今天中国的 AI for Science，是团队出击、共同体出击，这本身是反 0 到 1 的。

这句话乍听有点刺耳。集中力量办大事，是中国过去几十年做成很多事的办法：两弹一星、高铁、大飞机、超算，背后都是大团队、大协同。今天的 AI for Science，各地建平台、组联盟、拉共同体，也是同一个思路。

所以先要问清楚：这句话说的是团队没用吗？

不是。它说的是，团队擅长的那件事，和 0 到 1 需要的那件事，方向是反的。下面几节，就一层一层把这句话拆开。

◆ 三、六千五百万份记录

先看一份很有分量的证据。

2019 年，Wu、Wang、Evans 三位学者在 Nature 上发表了一项研究。他们分析了约 6500 万篇论文、专利和软件项目，问了一个很朴素的问题：团队的大小，和成果的性质有没有关系？

他们用了一种量法，大意是：一篇成果出来以后，后来的人引用它的时候，还引不引它所引的那些前人。如果后来的人只引它、不再引它的前人，说明它把前面的路改道了，这叫颠覆；如果后来的人连它带它的前人一起引，说明它是在前面的路上往前走了一步，这叫发展〔核：大意〕。

结论很清楚：大团队的成果更倾向于发展已有的科学和技术；小团队的成果更倾向于颠覆。而且团队越大，这个倾向越明显。他们还发现，小团队更常去翻那些更老、更冷门的旧想法〔核：大意〕。

这项研究不说大团队不好。大团队做出了大量扎实的、被广泛引用的成果。它只是说：两种团队擅长的，是两件不同的事。

这里也要打一个预防针。一项统计说的是很多份记录的平均倾向，不是每一个团队的命运。有的大团队也做出过翻天覆地的东西，有的小团队一辈子只在旧路上挪步。本书借这项研究，只借它的方向，不拿它去给哪一个具体的团队下判语。

还有一点，这项研究量的「颠覆」，是一种用引用关系算出来的近似。它抓得住「后来的人改道了没有」，抓不住改道改得对不对。所以它更像一盏指示灯，告诉你往哪里看，不替你下结论。

◆ 四、为什么：共识的平均

为什么会这样？本书的说法落在一个很日常的地方：团队靠共识运转。

一个四十人的项目，要往前走，就要定方向、定分工、定里程碑。每一个决定都要大多数人点头，至少要主要负责人都点头。点头的依据是什么？是大家共同认可的那套「什么算好问题、什么算好答案」，也就是现行的典范。

不过，一个典范猜想刚提出时，几乎一定看起来是错的。第 19 章讲过的神经网络是这样，第 25 章讲过的法拉第力线也是这样：一个能画不能算的念头，在擅长计算的同行眼里，先是一个外行的比方。第 31 章还会引普朗克的一句话，说新东西往往要等一代人。

把这个猜想放进一个四十人的会议室，会发生什么？大多数人会凭现行典范判断它，判断的结果多半是「不靠谱」。即使有一两个人心动，平均下来，项目的方向也会落回大家都能点头的地方。

平均化的结果，一定落在旧典范里。这不是谁在打压，是算术。

◆ 五、先说公道：团队原本的价值

说团队不擅长 0 到 1 之前，先把它擅长的那一段说足。

曼哈顿计划是一个大团队。它要做的事，在它开始之前，物理已经给出了典范：核裂变在 1938 年前后被发现，链式反应的道理也已经清楚。曼哈顿计划做的，是把一个已经成立的典范，变成一套能运转的系统：分离同位素、设计装置、协调几万人。

人类基因组计划也是一个大团队。从 1990 年左右启动，到 2003 年前后宣布完成，几个国家、几千名科学家，把一个已经清楚的问题——人的基因组序列是什么——系统地做完。

这一段，本书叫它「典范之后的系统打造」。这一段需要的正是大团队：分工、协同、标准化、海量的重复劳动、严格的进度管理。

中国在这一段上非常强。上篇和中篇说过的那些成绩，盘古气象模型发表在 Nature 上，DeepSeek 在算力受限下把模型训练的工程做到极致，都离不开一个高效的团队。中国的组织方式，恰恰最适合这一段。

再说一层：大团队还是一所学校。一个年轻人进了一个好的大团队，能在几年里见识最先进的设备、最规范的流程、最严格的质量标准。中国这些年出来的大量工程和研究骨干，很多是在这样的团队里练出来的。这份训练，是实打实的。

所以，团队有价值，而且价值很大。问题出在下面这一节。

◆ 六、本书的说法：双重错位

把前面几节放在一起，能看到两层错位。

第一层：团队擅长的，是典范之后的系统打造；它不擅长的，是典范本身的发生。用大团队去攻 0 到 1，就像用一台联合收割机去育种：收割机很了不起，可它的长处在收，不在长。

第二层更要紧：团队擅长的那一段，正在被 AI 拿走。

第 26 章林舟的那个下午，就是这件事的一个小样子：三个博士生一个季度的系统打造，一套智能体一个下午做完。第 369 号《典范发生学》说过，人给出典范猜想之后，典范和典范下的系统打造，可以交给 AI。今天大团队里大部分人干的活——分工跑流程、协调数据、标准化地重复——正是 AI 最先学会的东西。

所以：

团队不擅长 0 到 1，而它擅长的典范之后那一段，正在交给 AI；中国把最多的人，放在了两头都不对的位置上。

一个越来越大的共同体，一边用共识把新典范平均掉，一边在做一件机器会越做越好的事。

◆ 七、方向是怎样被定下来的

再往深处走一步：一个大团队的方向，实际上是怎么定的？

四十人的启动会上，没有人投票决定「我们要用哪一种典范」。方向在会前就定好了：写在项目指南里，写在任务书里，写在牵头人过去十年的成果里。牵头人通常是行当里资历最深、成果最多的人；他之所以能牵头，正是因为他在现行典范里做得最好。

所以，大团队的方向，往往是按资历定的。资历是什么？是一个人在旧典范里攒下的年头和成绩。

这件事本身很合理。一个几千万、上亿的项目，交给一个做过、做成过的人，是对钱负责。可是它和第 28 章那三道筛子是同一个形状：凭已经做出来的东西，决定谁来定下一件事的方向。

团队里年轻的人，离旧典范的根扎得最浅，有时候最容易看见裂缝。可他们在团队里的位置，是执行者。他们看见的裂缝，要一层一层往上报；每报一层，就被平均一次。到了牵头人那里，往往只剩下一句「这个想法有意思，以后再说」。

那个「以后」，和中篇一讲过的那个「再」字，是同一个字。

还有一点要说清：团队越大，越需要统一的流程、统一的数据格式、统一的考核。统一是大团队能运转的前提，也正好把每一个人都钉在同一个典范里。一个组想换一种量法，就要说服整个平台改格式；这几乎不可能。

◆ 八、只看一面会怎样

一种反应是：那就解散大团队，全部改成小组。

这走不通，也不该走。大量的常规科学、大量的工程，仍然需要大团队，至少在 AI 完全接手之前需要。而且一个小组本身也不会自动颠覆，Wu 等人的研究说的是倾向，不是保证。十个小组如果都按同一套考核、同一套共识去做，结果和一个四十人的大组差不多。

另一种反应是：中国的长处就是团队，扬长避短，继续把团队做大。

这也只看了一面。扬的那个长，正在被机器拿走；避的那个短，恰恰是唯一还稀缺的东西。

真正的问题不是团队大小，而是谁来定方向。第 31 章会接着讲：方向由谁来定，按什么来定。

◆ 九、今天与机器：一个人加一台机器

那 0 到 1 需要的是什么样的组织?

Wu 等人的研究是在 AI 时代之前做的。那时候，小团队有一个大劣势：人少，系统打造做不动，一个好猜想常常因为没人去做而搁浅。

今天这件事变了。一个猜想者加上 AI，可以完成过去要一个中型团队才能做的系统打造。小团队最大的劣势，正在被机器补上；小团队最大的长处——不必等共识、敢走冷门的路——没有被拿走。

所以第 44 章会讲一种新的组合：一个猜想者，一台 AI，一座公共平台。它代替的是「大团队加 AI」。

还要说清楚：一个人加一台机器，也有它自己的险处。没有团队里那些反对的声音，一个人的猜想很容易被机器顺着打造得越来越完整，看起来越来越像被

证实了。第 37 章会专门讲这件事。团队的共识会把新苗平均掉，一个人的独断会把歪苗养大；两种风险都是真的。

所以本书不是说「小就是好」。本书说的是：0 到 1 那一段，要交给一个不必等共识的人；而检验他的，要另有一套不靠共识、也不靠他自己的办法。第 45 章回来讲这套办法。

大团队加 AI，得到的是更快的常规科学。一个人加 AI，至少有可能得到别的东西。

◆ 十、一个完整的例子：启动会之后

启动会开完的那个下午，林舟回到办公室。

项目分给他的任务很清楚：三十个靶点，用平台的统一流程，两年内每个靶点给出一批候选分子。考核指标写在任务书里，一目了然。

他打开第 26 章那天的智能体，试着把任务书里的一个靶点交给它，按平台的统一流程跑一遍。两个小时后，结果出来了，格式和任务书要求的一模一样。

他算了一下：三十个靶点，全部这样跑，大概一个月。任务书给的是两年。

剩下的二十三个月，他可以有两种用法。一种是按任务书的节奏慢慢交，让十二个人都有活干，考核稳稳当当。另一种是把这一个月的活交给机器，让组里大部分人去做平台需要的数据核对和实验。再抽出他自己和小顾，加一台机器，去追那个没人点头的问题。

第二种用法，在四十人的启动会上一个字都不能提。提了，就是「不服从统一安排」。

◆ 十一、放回日子里

那天晚上，林舟把两种用法都写在了纸上，没给任何人看。

他想起那天会上主持人说的「集中力量办大事」。这句话在过去是对的，在很多地方今天还是对的。他只是第一次看清了它没说出来的那一半：集中起来的力量，最擅长办的，是已经知道该怎么办的大事。

一个小办法，给每一个在大团队里的人。下一次开会，留意一下会上被否掉的那个想法：是有人证明了它错，还是多数人觉得它不像对的？在本子上记一笔，只记后一种。一个月以后翻一翻，你会知道，你的团队把哪些可能性平均掉了。

如果你恰好是那个定方向的人，再多做一步：每个月，挑一个被平均掉的想法，交给一个人加一台机器，给它一周，不进考核。

> **共识能把一件已知的大事办好，也能把一个未知的小苗平均掉。**

带 走 的 话

大团队擅长典范之后的系统打造，不擅长典范的发生；而它擅长的那一段，正在交给 AI。

小　账

看见了什么，漏看了什么？看见了曼哈顿计划、基因组计划和中国工程团队的真本事；漏看的是团队的长处正被机器接走，短处恰是唯一稀缺的。

是在改进，还是在出题？共同体出击是在已有的题里做得更大更快；出题要的是不等共识的一个人，加一台机器。

想 一 想

你所在的团队，上一次被否掉的想法，是被证明错了，还是只是看起来不对？如果把你团队一半的活交给机器，剩下的时间你会给谁？

◆

31

第 31 章

方向由谁来定

◆ 一、组会上没放出来的那一页

周二晚上七点，林舟的组会照例在三楼的小会议室开。十二个人，一台投影，一壶已经凉了的茶。

轮到他讲下一阶段的安排。他的电脑里有两份幻灯片。

第一份是大家熟悉的：沿着项目书里写好的那个靶点，再筛一轮分子。AlphaFold 出结构，大模型打分，挑出前两百个，交给合作方做实验。路是铺好的，每一步都写在三年前的项目书里。

第二份只有一页。上面是他这半年在周末偷偷想的一个猜想。第 26 章那个「形状怎么变」的念头，又往前走了一步：这种病也许不在某一个靶点上，而在几种蛋白随时间慢慢错开的节奏上。要是这样，整个项目瞄准的那个点，可能从一开始就瞄偏了。

他把鼠标停在第二份文件上，停了大概三秒。然后打开了第一份。

会开得很顺。没有人提问题，大家各自领了任务。

散会以后，他一个人在会议室里坐了一会儿。那一页猜想，他没有放出来。

他说不出口，不是因为怕组里的年轻人不懂。这个项目的方向，是他的老师、一位资深的首席科学家在申请时定下的。项目书上写着这个靶点，中期检查要看这个靶点上的进展，结题要交这个靶点上的论文。他要换方向，等于在组会上说：老师三年前定的那条路，可能不对。

回到家已经十点。宋岚问他怎么脸色不好。他说：「没事，就是有一个想法，不知道该跟谁说。」

这一章，讲的就是这个「不知道该跟谁说」。

◆ 二、一张从上往下画的图

把中国一个大型科研项目的组织画成一张图，大致是这样的。

最上面是首席科学家，常常是一位院士或者资深 PI。下面是几位课题负责人，各管一块。再下面是子课题、骨干、博士后，最底下是一届一届的博士生和硕士生。

方向在最上面定。项目申请的时候，首席科学家和几位核心成员写好总目标、技术路线、考核指标。申请批下来，这条路线就成了下面所有人的任务书。每往下一层，任务就切得更细一点：一层管一个模块，一层管一组实验，一层管一张图。

这张图有它真正的好处。第 30 章说过，团队最擅长的是在一条已经定了的路上把事做完、做全、做快。层层分解，正好让几十个人、几百个人朝一个方向使劲。

可是这张图也有一个从来不画出来的箭头：判断只往下走，很少往上走。

最底下的博士生每天碰实验、碰数据，常常最早看见「这里有点不对」。他的看见要往上传，得先过他的导师，再过课题负责人，再过首席科学家。每过一层，都要换一套说法，都要冒一点风险。传到最上面的时候，它往往已经被磨成了「有一个小问题，我们再优化一下」。

◆ 三、最资深的人，是旧典范里最成功的人

为什么方向由最资深的人来定？理由很朴素：他最懂。

他在这个领域做了三十年，发过最多的好论文，带出过最多的学生，评过最多的项目。他知道哪条路走得通，哪条路十年前有人走过、走死了。让他来定方向，能少走很多弯路。

这个理由在大多数时候是对的。

不过请把这件事再往前推一步。一个人凭什么成了最资深的人？凭他在过去三十年的那套做法里做得最好。他认定的好问题，被同行认作好问题；他给的好答案，被期刊认作好答案。他的资历，就是那一套典范一次次给他盖的章。

所以，最资深的人恰恰是在旧典范里最成功的人。

这里没有任何人品上的意思。他不保守，也不霸道。他只是太了解这条路的好处了。一个新典范要说的第一句话，往往就是「你们一直以为的好问题，问错了地方」。这句话落到他身上，等于说他那三十年的章，盖在了一张不那么要紧的纸上。

要一个人亲手判自己的功劳簿过时，这比要他承认一个实验做错了，难得多。

◆ 四、共识加资历，两股劲叠在一起

第 30 章讲过大团队的一种力量：共识会把意见磨平。十个人讨论一个方向，最奇怪的那个提法最容易被说「再想想」，最稳妥的那个最容易通过。

资历是另一股劲。它不磨平意见，它给意见排座次。同一个提法，从首席科学家嘴里说出来是「方向」，从一个博士后嘴里说出来是「想法」，从一个博士生嘴里说出来，常常只是「问题」。

这两股劲单看都不坏。共识防的是一个人的偏执，资历防的是一个新人的轻率。

麻烦在它们叠在一起的时候。共识说：最稳妥的胜出。资历说：最资深的人定稳妥是什么。于是最后胜出的，是最资深的人眼里最稳妥的那条路，也就是旧典范里最顺的那条路。

一个 0 到 1 的猜想，要同时过这两道。它得先说服一屋子人，让平均意见向它偏过去；还得说服屋子里资历最深的那个人，而那个人正是它要质疑的那套做法里最成功的人。

共识加资历，是把反对 0 到 1 的两股力量叠到了最大。

林舟那天晚上卡住的，就是这两道门。

◆ 五、普朗克留下的一句话

这件事不是中国才有。一百年前，就有人把它说得很冷。

普朗克在《科学自传》里写过一段话。大意是：一个新的科学真理取得胜利，往往不是因为它说服了反对者、让他们看见了光明。是因为反对者终于死去，熟悉它的新一代成长起来。〔核：大意〕

说这话的人，不是一个失意的边缘人。他是量子论的开山者，诺贝尔奖得主，德国物理学界最有分量的人之一。据他自己在同一本书里的回忆，他年轻时的博士论文，当时的大物理学家几乎没有人认真读过。他尝过新东西不被听见的滋味，后来又站到了最高处。这句话是他从两头都看过以后说的。

后来有经济学家想检验这句话。据报道，有一项 2019 年发表的研究。它发现，一位明星科学家意外早逝之后，他所在的小领域里，原本的圈外人发表的论文明显增多，而且这些论文更常被引用。研究者的解读是：一位权威在场的时候，别人不太愿意、也不太容易往他不认的方向走。

这当然不是要谁去等谁老去。这句话真正说出的是一件结构上的事：当方向的裁判和旧路的赢家是同一个人，新路就只能等裁判换人。

在一个按资历组织的体系里，裁判换人很慢。

◆ 六、底下的人也学会了不说

这套组织还有一个更安静的后果：底下的人，慢慢学会了不说。

一个博士生第一次在组会上提出「这个方向可能有问题」，得到的回答是「先把手上的数据做完」。第二次，回答还是一样。到第三次，他就不提了。他不是被谁训斥过，他只是算清了一笔账：提出来，要花时间解释，还可能让导师不高兴；不提，论文照样能发，毕业照样能毕。

这笔账，林舟在读博的时候也算过。他今天坐在组会的最前面，组里的年轻人看着他，也在算同一笔账。

于是这张图上判断往上走的那条路，不是被堵上的，是被一届一届的人自己走荒的。

等到有一天，最上面的人真心想听一句不同的话，他会发现，底下已经没有人习惯说了。

◆ 七、只看一面会怎样

说到这里，容易走到另一头：既然资深的人会挡新路，那就让年轻人自己定方向好了。

先把公道还给资历。

载人航天、探月工程、一座跨海大桥，这些事都要有人在最上面拍板。目标清楚、路线基本知道、难在把成千上万个细节做对的时候，由最有经验的人定方

向，是最省的办法。年轻人的冲劲在这里常常是成本，不是资产。很多新人以为的「新方向」，老人一眼就知道十年前有人试过。

还有一层：资深的人也护住过很多新苗。一位老先生顶着压力给一个年轻人一笔钱、一间屋子，这样的事在中国的科学史里并不少。

所以按资历定方向，在一个条件下完全成立：题已经给定，要比的是谁解得好。

它停在的地方，是这个条件不成立的时候。一旦要问的是「题本身对不对」，最懂这道题的人，就不再是最合适的裁判。他不是不够聪明，是离这道题太近。

◆ 八、AI 把金字塔削尖了

现在把机器放进这张图。

第 22 章说过，AI 正在接管常规科学。放在林舟的组里，这件事看得很清楚：从前要三个博士生做一个月的结构预测和虚拟筛选，现在一个人一个周末就能跑完。从前要一个博士后写半个月的文献综述，现在大模型一个下午就给出一份像样的初稿。

金字塔下面那几层执行的手，正在被机器替下来。

被替下来的不只是劳动，还有一个从前藏在劳动里的东西：最底下的人碰到的那些「不对劲」。一个博士生熬夜做实验，偶尔撞见一个和预期不一样的结果，那是新方向冒头的地方。机器替他跑完以后，这些结果被平均进了几万个分子的打分里，悄悄地消失了。

而金字塔顶上那一格，机器没有动。项目的方向，仍然由那位最资深的人在申请时定下。

于是这张图变成了这样：手越来越少，头还是那颗头。执行越来越便宜，方向越来越贵。整个组的产出，越来越只取决于最上面那一个判断，而这个判断，正好是由旧典范里最成功的人来下。

AI 没有拆掉这座金字塔，它把金字塔削尖了。

◆ 九、项目书把方向钉了三年

还有一样东西，比人更硬：项目书。

一份大项目的申请书，要写清三五年里每一年做什么、交什么、怎样考核。写得越具体，越容易被批准；批下来以后，这份具体就成了一根钉子。中期检查对着它查，结题验收对着它验。

这根钉子有它的道理。国家把一大笔钱交给一个团队，总得知道钱花在了哪里。没有考核指标，就没有问责。

请看它在时间上的样子。方向在申请的那一年定下，那时候这个团队知道的东西最少；到了第三年，团队知道的东西最多，最有资格改方向，却正是项目书最不允许改的时候。

于是一个组里最清醒的那一年，常常也是最不能动的那一年。

AI 让这件事更扎眼。从前三年里能做完的事，现在半年就做完了；半年里冒出来的新看见，要等两年半才有机会写进下一份申请书。机器把研究的钟拨快了，项目书的钟没有动。

◆ 十、一个完整的例子：一个猜想怎样被消化掉

把林舟那一页猜想往下推一遍，看它在这张图里会遇到什么。这是一个设想，可每一步都是这套组织里的寻常一步。

第一步，组会。他要是放出那一页，组里的年轻人会很兴奋，也会很担心：这个方向写不进年底的考核，自己的论文怎么办？讨论的结论多半是：「很有意思，可以作为一个探索方向，主线先不动。」

第二步，向老师汇报。老师会认真听，会提出三个很专业的问题，都是在旧路上问的：证据够不够，和现有靶点的数据怎么对上，能不能先发一篇小文章看看反应。每一个问题都合理，合起来的意思是：先在旧路里证明你，再说换路。

第三步，中期检查。专家组手里有项目书，项目书上写着考核指标。一个跑偏了靶点的项目，在表格上就是「进度不足」。

第四步，下一轮申请。他要单独去申请这个新方向，申请书要写「研究基础」，而他的研究基础全在旧靶点上。评审专家又多半是这个领域最资深的几位，也就是旧路上最成功的几位。

每一步都没有人做错事。每一个人都在按规矩办，有的人还是真心爱惜他。

可是走完这四步，那一页猜想已经被消化成了「主线之外的一个小探索」，再过两年，大概就只剩下一句「当年也想过」。

◆ 十一、放回日子里

回到那个周二的晚上。

林舟没有在组会上说，这不全是胆小。在那一屋子人、那一份项目书面前，他确实找不到一个能放下这一页的地方。

不过有几件事，他明天就能做，不用等任何人批准。

第一，把那一页猜想写成一页「推翻条件」：要是这个猜想错了，什么样的数据会先说出来？写清楚，它就从一个念头变成了一个能被检验的东西。

第二，用一个周末，让 AI 拿公开数据库替他跑一遍最粗的检验。不用实验室，不用经费，也不用占组里的任何一个人。机器替他省下的那些时间，正好可以先花在这里。

第三，找一个不在这个项目里、也不在这个领域里的人，把这一页讲给他听。第一次不要去说服最资深的人，先去找一个没有功劳簿要护的人。

宋岚那天晚上听完，说了一句：「那你先别跟老师说，先跟数据说。」

这句话，比很多制度设计都更接近问题的要害。

> **方向由最懂旧路的人来定，旧路就会越走越好，新路就只能等。**

带 走 的 话

最资深的人，是旧路上最成功的人；方向只由他一个人定，新路就得先过他那本功劳簿。

小　账

看见了什么，漏看了什么？看见了按资历定方向省下的弯路和护住过的新苗，漏看了裁判和旧路赢家是同一个人。

是在改进，还是在出题？按资历定下的方向，几乎都是在已有的题里求更好；出新题的那一页，常常没有地方放。

想 一 想

你所在的地方，一个新方向要被认真讨论，得先说服谁？那个人最大的成功，是在哪一条旧路上取得的？

◆

32

第 32 章

本篇总账：筛子的账

◆ 一、把七章摊在一张桌上

林豆豆的书桌上有一个错题本。每个周末，她妈妈让她把一周的错题抄一遍，旁边写两行：这道题考的是什么，我错在哪一步。

这一章做的是同一件事。下篇一写了七章，从第 25 章的「复合体里的那个人」，到第 31 章的「方向由谁来定」。现在把这七章摊在一张桌上，每一章记两笔：它看见了什么，它漏看了什么。

请注意，这里的「看见」和「漏看」，说的是这套筛子，不是筛子里的某一个人。

◆ 二、七章的看见与漏看

章	这一章说的筛子	它看见了	它漏看了
第 25 章　复合体里的那个人	人—AI 复合体需要怎样的人	0 到 1 要交给一个人加 AI	这个人从哪里来、谁去认他
第 26 章　剥掉功底之后还剩什么	天才与功底长在一起	功底曾是认出天才的凭据	功底交给 AI 以后，只剩猜想与判断，旧凭据认不出它
第 27 章　全世界最好的腿部选拔	高考与竞赛	极公平、极高效地选出最会解题的人	选的是「好腿」，交通已经换成了车
第 28 章　博士、基金、职称：三道筛子	学位、项目、评审	每道关都能挑出常规科学的最优者	每道关都在问「你在给定的题里做得好不好」
第 29 章　论文计分遇上 AI	指标经济	计分让评价可比、可查、可问责	AI 让论文变便宜，计分被推到极致，猜想无分可计

章	这一章说的筛子	它看见了	它漏看了
第 30 章　大团队发展，小团队颠覆	团队出击、共同体出击	团队擅长典范之后的系统打造	这一段正被 AI 拿走；共识又把反常的提法磨平
第 31 章　方向由谁来定	按资历定方向	题给定时，资深的人定方向最省	裁判与旧路的赢家是同一个人

把「它看见了」那一列从上往下读，是一份很漂亮的成绩单：公平、高效、可比、可查、组织得动、少走弯路。中国用这套筛子，在四十年里培养了全世界规模最大的工程师和科研队伍，在很多常规科学的赛道上追到了前沿。这份成绩是真的，本篇每一章都先把它写足了。

再把「它漏看了」那一列从上往下读，会发现七句话说的是同一件事。每一道筛子都在问「这道题你解得好不好」，没有一道在问「你能不能出一道新题」。

◆ 三、七道筛子，其实是一道

一个孩子从九岁到四十岁，要过的关大致有这些。奥数和小升初，中考，高考，保研或考研，博士答辩，第一个青年基金，副高，正高，人才帽子，重大项目。

每一道关的出题人不同，评委不同，表格也不同。可是它们量的东西惊人地一致：在一套给定的规矩里，你是不是做得比别人更快、更准、更稳。

这不是谁设计出来的阴谋。每一道关都是为了解决一个真实的问题才设的：高考为了公平，基金评审为了不把钱给错人，职称为了让评价有据可依。每一道单看都有道理。

不过七道一样朝向的筛子串在一起，筛出来的就不是七种人，而是同一种人被筛了七遍。三十多年筛下来，留在筛子上面的，是常规科学里最好的一批人；从筛眼里漏下去的，有一种人从来没被量过。

他不一定解题最快，不一定论文最多，不一定最听组里的安排。他常常在某个奇怪的地方停下来，问「为什么非得这样问」。在每一道关上，这个习惯都要扣分。

◆ 四、一个人走过七道筛子

把林舟放回这七道筛子里，走一遍。

他小学时数学好，进了奥数班。老师夸他的，是他算得快、不出错。他偶尔想出一个怪方法，老师会说「能做对就行，考试还是用标准方法稳当」。

中考、高考，他都考得很好。高考那年他做完数学卷子，还剩二十分钟，用来把每一道题检查了两遍。这是那套考试最奖励的习惯。

读博士的时候，导师给了他一个题目，是导师自己那条路上的下一步。他做得很漂亮，四年发了五篇论文，答辩顺利。他有过一个想换题的念头，导师说：「先毕业，以后有的是时间。」

回国以后，第一个青年基金，他写的是博士期间那条路的延伸，因为评审要看「研究基础」。副高、正高，看的是论文和项目。三十八岁这年，他带着十二个人，做着老师定下的方向。

请注意，这七道关里，他没有一次被亏待。每一道关都公平地认了他的本事。

可是那个在奥数班里想出怪方法的孩子，那个读博时想换题的年轻人，这一路上每出现一次，就被温和地请回了标准方法一次。三十年下来，他学会了先把怪念头收进周末，再把它收进一页不放出来的幻灯片。

筛子没有筛掉林舟。它筛掉的，是林舟身上的那一部分。

◆ 五、本篇总判断

把七章的漏看放在一起，下篇一的总判断是：

复合体里的那个人比天才更稀缺，而中国的筛子最会筛掉他。

前一半说的是稀缺。从前的天才，猜想能力和常规功底长在一起，功底是他被认出来的凭据。今天功底交给了 AI，复合体里留给人的，只剩下纯粹的猜想与判断。这种人连从前那张凭据都没有，所以比天才还难认。

后一半说的是筛子。中国的筛子不是不认他，而是筛得最准、最早、最一致。别的地方的筛子也偏向常规科学，可是总有一些漏洞、一些歪门：一个导师破格收人，一所学院不看论文看人，一个基金会专门资助怪人。中国的筛子在公平和效率上做到了极致，这些漏洞就被补得最干净。

请把这两半一起读。前一半不是说中国人不行，后一半也不是说筛子做错了。它说的是一个错位：筛子是为上一个时代最稀缺的东西造的，而 AI 让那样东西不再稀缺。

◆ 六、和前两篇怎样接上

这本账还要和前面两篇对一下。

中篇一说的是「先做应用」：中国四十年里习惯接受别人定义好的题，在题里做到最好、最快、最便宜。中篇二说的是层级：AI 接管的正是中国最强的那一层，留下的正是最空的那一层。

下篇一补上的，是人这一头。一个国家习惯了接题，它的筛子就会朝接题的方向长；筛子朝接题的方向长了三十年，筛出来的人就都是接题的好手。这几件事互相喂养，谁也不是谁的原因，它们一起长成了一片土。

所以「再做原创」那个「再」字没有排进日程，不只是因为没有钱、没有时间。到了想排的那一天，筛子已经把能接住这个「再」字的人，提前筛掉了一大半。

◆ 七、筛子不是中国才有

还要公道地说一句：偏向常规科学的筛子，不是中国才有。

全世界的大学都考试，全世界的基金都评审，全世界的期刊都看论文。第30章引过的那项研究，用的是全世界的论文和专利，大团队发展、小团队颠覆，说的也是全世界的科学。

中国的不同，在三处。一是筛得早：九岁的奥数班就开始按解题速度排人。二是筛得齐：从小学到正高，每一道关的朝向几乎一致，中间很少有一道关换一种量法。三是筛得准：中国把公平和效率做到了别处很少能做到的程度，筛眼里的歪门和漏洞，被补得最干净。

这三处，单看都是长处。合在一起，就成了那个人最难漏网的地方。

◆ 八、这一篇自己漏看了什么

总账也要给自己记一笔。

第一，体制自己看见过问题。第28章说过2018年10月五部门清理「四唯」的通知，这说明筛子的设计者也知道筛眼的朝向有偏。本篇说的不是没人看见，而是看见以后，替换上来的新尺子，多半仍在量「题里做得好不好」，只是换了一种量法。

第二，筛子不是绝对的。中国的科学史里有破格录取的少年，有被老先生从杂志上认出来的自学者，有不看出身、只看文章的评委。第34章会讲一个这样的例子。这些例外说明，问题不在人心，在通道：例外靠的是某一个人的眼光，不是一条谁都能走的路。

第三，本篇说的「那个人」，到目前为止还是一个推出来的人。他该具备什么，本书给出了说法；他在现实里有多少、该怎样认，本书还拿不出一张可以照着填的表。这是本书较弱的一处，第48章会回来写。

◆ 九、今天与机器

机器让这本账变得更急。

在没有 AI 的年代，筛子筛掉了那个人，代价是慢：一个国家的科学多走一些年的弯路，可是常规科学这一层照样在长，照样有回报。

今天，常规科学这一层正被 AI 接过去。筛子选出来的好腿，正好是机器最先替下来的那一种能力。于是同一道筛子，在两头一起亏：它留下的那种人在贬值，它漏掉的那种人在升值。

这就是判词前一半落在人身上的样子：最擅长的那一层越先贬值。

◆ 十、放回日子里

回到林豆豆的书桌。

那天晚上她抄完错题，问了妈妈一个问题：「为什么这道题一定要用画线段的方法？我用别的方法也算对了，老师还是扣了分。」

宋岚没有马上回答。她想起林舟那天晚上说的「不知道该跟谁说」。

一个小办法：下一次孩子拿着一道被扣了分的题来问，先不要讲「标准方法」，先问她一句：「你那个方法，是怎么想到的？」让她讲完。一个孩子能讲清自己为什么不按规矩走，这件事在任何考试里都不计分，可它正是那个人最早的样子。

明天能试的，还有一件：找出孩子这学期被扣分最多的三道题，挑一道，和她一起想一想，规矩为什么要这样定。不是教她守规矩，也不是教她破规矩，是让她知道规矩是人定的。

◆ 十一、指向下一篇

下篇一问的是：那个人为什么很难被选出来。

可总有漏网的。一个孩子没被奥数筛住，一个博士没被基金筛住，一个人在体制外，靠一台 AI，自己长成了。

下篇二要问的是下一件事：即使这个人漏网长出来了，土会怎样对他。他说的话算不算数，由谁来判，他的猜想会被叫作什么。筛子决定他能不能被选出来，土决定他长出来以后能不能活。

> **筛子是为上一个时代最缺的东西造的；造得越好，越认不出这个时代最缺的人。**

> 带 走 的 话
>
> **七道朝向一致的筛子，筛出的不是七种人，是同一种人被筛了七遍；漏下去的那种人，从来没被量过。**

小　账

看见了什么，漏看了什么？看见了筛子的公平、高效与它造出的庞大队伍，漏看了它每一道都在问同一个问题。

是在改进，还是在出题？这七道筛子量的，全是在已有的题里做得好不好；出题的能力，没有一道在量。

> 想 一 想
>
> 你从小到大过的关里，有没有哪一道曾经奖励过你「问为什么非得这样问」？如果没有，你身上的那个习惯，后来去了哪里？

◆



下　篇　二

土：思想寡头为何难发芽

导　语

这一篇十章，写那个人漏网长出来以后，土怎样对他。第 33 到 35 章说思想权威、头衔与「民科」：第 33 章从「我注六经」讲起，说思想的正统在中国历来不归个人；第 34 章说一个人的话算不算数，看的是头衔、单位与项目级别；第 35 章说「民科」两个字怎样只检验身份、不检验内容。

第 36 到 37 章说谁来判一个新典范，以及思想寡头自身的风险：同行评议判不了典范本身，而一个人加一台 AI，又会把自己的猜想越证越像真的。第 38 到 41 章说出海与回流：墙外开花墙内香，出海是求生存还是取通行证，两栖的新形态，以及最难的那条回流的路。第 42 章记本篇的总账。

这一篇的反差是：AI 拆掉了资源的墙，留下了合法性的墙。一个人不再需要实验室、团队和经费，也能做典范打造；可他做出来以后，仍然要先回答「你是谁」。

33

第 33 章

我注六经

◆ 一、一本写满批注的书

你也许在旧书摊上翻到过这样一本书：书是别人写的，页边却密密麻麻写满了另一个人的字。有的地方画了线，有的地方打了问号，有的地方写着「此说未安」。

那个写批注的人，常常比原书的作者更有想法。可是他把想法写在了别人书页的空白处。

这个画面，是中国读书人几千年的一个姿势。想法再多，也写在经典的旁边。

这一章要讲的，是这个姿势和 AI 时代的那个人之间的关系。

◆ 二、一句口述：超级的思想寡头

先把本篇要用的一个词摆在桌上。

王德生博士在口述里说过一个判断，大意是：人—AI 复合体的 0 到 1，要靠少数人去做。这少数人必须是超级的思想寡头，而这种人在中国的土壤上更难诞生。

「寡头」这个词听起来刺耳，要先讲清楚它在本书里的意思。

它不是说一个人掌握权力，也不是说一个人可以不讲道理。它说的是一件很朴素的产出上的事：复合体让一个人加一台 AI，就能做过去一个研究所才做得

了的事。从前一个新典范提出来以后，要几十个人花十几年去做系统打造：推公式，做实验，写软件，编教材。第 369 号《典范发生学》说过，这一段正在交给 AI。

这样一来，产出就向极少数人集中了。一个能提出好猜想、又能指挥 AI 把它做成体系的人，一个人就是一个学派的量。「寡」说的是人少，「头」说的是源头。

所以寡头首先是一个产出的形状，不是一种性格。

还要说清它不是什么。它不是垄断：一个寡头立了典范，别人照样可以立另一个，可以拆他、驳他。它也不是终身的：一个寡头的猜想被证伪，他就不再是源头。寡头说的是一个人在某一段时间里，扛起了一个学派的产出；它能不能站住，要看检验，不看他自己。第 37 章会回来讲这个人自己的风险。

要成为这种源头，一个人得敢自己立一个典范：说出「我认为好问题应该这样问」，而且不必先找到一部经典替他背书。这一步，在中国的土里格外难走。

◆ 三、读书人的位置：注经者

中国传统上，思想的正统在两个地方：经典与体制。

经典是孔孟、是六经，后来加上朱熹的注。体制是科举，是朝廷认可的那一套解释。一个读书人的一生，大体是围着这两样转：读经，考经，做官以后按经行事，老了以后再给经作注。

这个位置，可以叫注经者。

注经者不是没有思想。中国历史上最好的头脑，很多都把最好的想法写进了注里。朱熹注《四书》，等于重写了一遍儒学；王弼注《老子》、郭象注《庄子》，注本后来成了和原书一样有影响的东西。

不过这些思想有一个共同的样子：它们都要说自己是在「发明本意」，是把古人本来就有、只是被人看歪了的东西重新端出来。一个新想法要合法，就得先说它不新。

这和库恩说的典范有一点相像：典范里的人，用典范的语言说话，在典范给的题里做到最好。注经就是一种很长寿的常规科学。

◆ 四、陆九渊的那一句

也有人看见了这件事。

南宋的陆九渊有一句话，后来流传很广：六经注我，我注六经。

这八个字的意思，大致是两层。一层是「我注六经」：我替经典作注，我是那个注者。另一层是「六经注我」：反过来，经典是在替我作注，我心里本来就明白的道理，经典只是给它作了一个旁证。

后一层是很大胆的话。它把思想的源头，从经典那边挪到了人这边。

不过请注意，陆九渊说这句话的时候，仍然是借着经典在说。他说的「我」，是一颗和圣人相通的心，不是一个随便什么人。他的大胆，也是在儒学的屋檐下面的大胆。

◆ 五、王阳明是例外，李贽是祸患

再往后三百多年，有两个人把这一步走得更远。

一个是王阳明。他说心即理，说良知人人都有，等于告诉每个读书人：道理不在书上，在你自己身上。他的学说后来天下传遍，他本人也立了大功，封了伯爵，后来从祀孔庙。

不过王阳明在中国思想史里，常被当作一个例外来讲。他有赫赫的事功护身，有一大批门人替他立说。人们记得他，多半是把他放进正统旁边的一支，而不是承认「一个人可以自己立一个典范」这件事本身。

另一个是李贽。他走得比王阳明的学生们更远。他说不必以孔子的是非为是非，他替被正统看不起的人物翻案。他写的书直接取名《焚书》，意思是这书迟早要被烧掉。

他说中了。他被以「惑世诬民」一类的罪名下狱，1602 年死在狱中。他的书后来被反复禁毁，又被反复偷偷刻印。

李贽并不是一个只会骂人的狂生。他做过二十多年的官，读书极勤，晚年在湖北麻城的寺院里著书讲学，来听的人里有不少读书人和妇女。他的许多说法，放在今天看，只是要求一个人用自己的眼睛读书。

在他那个时代，这个要求本身就越了界。

一个是例外，一个是祸患。这两个结局放在一起，说出的是同一件事：在这片土里，一个人自立典范，文化上本身就带着僭越的意味。

成了，要被解释成「其实也合于正统」；不成，就是乱道。中间没有一条路，可以让一个人说「这是我立的，请你来检验它」。

◆ 六、先把公道还给注经的传统

说到这里，必须停下来，把公道还给注经的传统。

一个文明能够延续几千年，文字、典籍、礼俗一代代传下来，靠的正是一种不轻易推翻前人的习惯。每一代人都在同一部经上加注，等于在同一块地上一层一层地加土。后人站在很厚的土上，才有那么多的东西可以继承。

很多古文明的书，今天只剩残片。中国的经典能大体完整地传下来，注经的传统功不可没。

还有一层：注经的人并不都是在照抄。好的注家，是在旧的题里做出了新的东西。这和中国今天在常规科学上的长处，是同一种能力：在给定的框架里，做得更细、更全、更深。

所以注经传统在这样一个条件下完全成立：要紧的事是守住和传下去，而不是出一道新题。

它停在的地方，是一个时代要的恰恰是新题的时候。

◆ 七、科举把注钉进了考场

注经为什么后来越来越窄？有一样东西起了很大的作用：科举。

元代以后，朝廷把朱熹注的《四书》定为考试的标准答案。明清两代的八股文，题目出自《四书》，答题要照朱注的意思讲，格式也有严格的规定。

这一来，注经就不只是读书人的一种习惯，而成了一条通往功名的唯一的路。一个读书人可以私下里有自己的想法，可是到了考场，他只能写朱熹的想法。

这套考试在公平上是有功的。它让一个农家子弟，只要书读得好，就有机会进入朝廷。它让一个庞大的国家，用同一套文字、同一套道理选人。这一点，和第 27 章说的高考有相通的地方。

请把它放在本章的问题里看。一套考试越公平、越统一，它就越要求所有人用同一种方式说话。几百年下来，最会读书的人，也就是最会用别人的注说话的人。

李贽死后四十多年，明朝亡了。清代的学者转向考据，把功夫下在字句的真伪和古音古义上。那是一门很了不起的学问，也是一种更细的注经。

◆ 八、只看一面会怎样

还要防另一头的误读。

有人会说：那就是中国文化不行，中国人天生不会原创。这句话不对。第 9 章讲过一个反例，汉卡和 WPS 的年代，生产力的原创走在前面。顺序是条件决定的，不是天性。

思想上也一样。中国历史上并不缺敢自立的头脑：先秦诸子百家，各立门户；魏晋的清谈，各说各的；宋明的书院，也吵得很凶。

所以本章说的不是中国人想不出新典范，而是一个新典范被提出来以后，土里有没有一个位置能放它。先秦有，那时候正统还没定型；宋明以后，科举把正统钉得越来越牢，那个位置就越来越窄。

土不是背景，土在推。

同一粒种子，落在正统未定的土里，能长成一家之言；落在正统钉牢的土里，要么长成正统旁边的一枝，要么被当作杂草拔掉。种子还是那粒种子。

◆ 九、今天与机器

今天的读书人不再注六经了。那个姿势却还在，只是经典换了名字。

今天的「经」，是顶级期刊上的主流框架，是领域里公认的几篇奠基论文，是项目指南上写好的方向。一个年轻研究者写论文，第一段照例要说「在某某框架下」，然后证明自己的工作是在这个框架里补上了一块。这是一种很体面的注经。

AI 让这个姿势变得更省力。你给大模型一个主流框架，它能在一个下午替你写出十篇合格的注：综述、改进、应用、扩展。注经从来没有这么便宜过。

反过来，AI 也第一次让「六经注我」在技术上变得可行。一个人有了一个新猜想，可以让 AI 替他把整个领域的文献翻一遍，看看哪些旧结果能为它作证，哪些和它冲突。从前要一个学派才能做的事，现在一个人能做。

技术上的门打开了。文化上的那道门，还关着：一个人站出来说「这是我立的」，听的人第一反应仍然是「你凭什么」。

◆ 十、一个完整的例子：一篇不敢起题的论文

设想一个年轻的博士后，在做一个计算材料的方向。他用 AI 跑了上万个结构，发现一个现象，现有的理论解释不了。他慢慢形成了一个猜想：这个领域描述某一类材料的方式，可能从一开始就选错了量。

他把这个猜想写成一篇论文。初稿的题目很直接：「重新描述某类材料：一个替代框架」。

导师看了，建议他改，改成「对某某理论在某类材料中的一个修正」。理由很好：审稿人会更容易接受，不会被当成狂妄。

投出去，审稿人要求他补充和主流方法的对比，要求他说明这个「修正」在哪些已知情形下回到旧理论。他一一补上。

论文发了。题目里是「修正」，正文里是「补充」，结论里写「为进一步完善现有理论提供了思路」。

那个替代框架，就这样被写成了一条注。它没有被拒绝，它被收编了。

这里没有任何人做错事。导师是为他好，审稿人是在尽责。可是一个猜想要被听见，就得先把自己说成一条注，这件事本身就是土的样子。

◆ 十一、放回日子里

回到林舟。

第 31 章那个晚上，组会散了，那一页猜想没有放出来，他想不出该跟谁说。还有一层原因，他自己也没意识到：他从读书起学会的，就是把新想法写成「在某某基础上」。他连给自己的猜想起一个名字，都觉得有点不好意思。

一个小办法，明天就能试：把你最近的一个想法，用两种题目各写一遍。一种是「对某某的改进」，一种是「另一种问法」。写完放在一起看，问自己：哪一个才是我想说的？要是后一个，先在本子上用它，哪怕暂时不拿出去。

给自己的猜想一个它自己的名字，是一个人走出注经者位置的第一步。

> **注经让一个文明传了下来；可是一个新典范要被放进这片土，得先有人敢不写成注。**

> 带 走 的 话
>
> **一个想法要先说自己不新才能被接受，那片土就只长得出注，长不出经。**

> 小　账
>
> **看见了什么，漏看了什么？**看见了注经传统保存、传下了文明，也在旧题里做出了新东西；漏看了一个人自立的典范没有位置可放。
>
> **是在改进，还是在出题？**「我注六经」是在已有的题里做得更好；「六经注我」才是出题，而这一步在这片土里带着僭越的味道。

> 想 一 想
>
> 你最近一次有一个新想法，是怎样跟别人说的？你有没有先把它说成「在某某基础上」？要是不这样说，你担心会发生什么？

◆

34

第 34 章

头衔发放的合法性

◆ 一、一场学术报告的开场

你大概听过很多次这样的开场。主持人站起来，拿着一张纸念：

「今天的报告人是某某院士，某某大学教授，国家某某重点实验室主任，某某计划首席科学家，曾获国家某某奖一等奖。」

念完，掌声。报告人还没开口，听众已经知道这场报告大概有多少分量。

换一个场景。一个没有单位、没有头衔的人，在网上贴出一篇长文，说他发现了某个领域的一个根本问题。下面的第一条评论往往是：「你是哪个单位的？」

两个场景放在一起，说出的是同一件事：一个人说的话算不算数，很多时候在他开口之前就定了。

◆ 二、头衔是怎样发放的

在中国的科研体系里，一个人的分量，大体由三样东西定：头衔、单位、项目级别。

头衔是一串有先后的帽子：讲师、副教授、教授，然后是各级人才计划，最上面是院士。单位也分了等级：一流大学、重点实验室、国家级研究院。项目有级别：国家重大项目、重点项目、面上项目、青年项目。

这三样东西彼此咬合。要拿大项目，得先有好单位和高头衔；要升头衔，得先有大项目和好论文；要进好单位，又得先有头衔。

在这个体系里，合法性是被发放的。一个人说的话能不能进入讨论、能不能被认真检验，取决于他手里有没有这几张证。

这并不奇怪。任何一个大规模的科学共同体都需要一种办法，快速判断谁说的话值得花时间听。头衔就是这样一种办法：它把过去几十年的同行评价，压成一张随身携带的名片。

◆ 三、先把公道写足：举国体制会做什么

说头衔的局限之前，先说它背后那套体制做成了什么。

二十世纪六十年代到七十年代初，中国在工业基础薄弱、外部封锁的条件下，先后搞出了原子弹、氢弹和人造卫星。成千上万的科学家、技术人员和工人，被组织到同一个目标下面，按计划分工，一环扣一环。

后来是高铁。从引进、消化，到形成自己的一整套技术体系和施工能力，中国在一二十年里建成了世界上最大的高速铁路网。

还有载人航天、探月工程、大型射电望远镜、超级计算机。这些都是举国体制的长项。

这些成就有一个共同点：目标清楚，路线大致知道，难在把成千上万个环节同时做对、按时做完。这正是常规科学和大工程最需要的能力。在这种事情上，由一个层级分明、头衔清楚的体系来组织，是最省力的办法。谁有经验、谁负责、谁说了算，一张头衔表就能定下来，不必每次重新吵。

举国体制是全世界组织大兵团做常规科学最好的机器之一。

这句话本书是认真说的。

◆ 四、可是没有一个通道

这台机器却有一个它自己看不见的地方。

它擅长把已经被认可的方向组织起来，却没有一个通道，用来承认这样一件事：一个人在机构之外，立了一个新典范。

请把这句话拆开看。

「机构之外」，是说这个人没有那几张证：没有头衔，或者头衔不够高；没有单位，或者单位不够好；没有项目，或者项目和这个新方向无关。

「立了一个新典范」，是说他提出的不是旧题里的一个新解，而是一道新题，一种新的问法。

这两样合在一起的时候，体制里的每一个入口都对不上。投期刊，审稿人按旧典范的尺子量；申请项目，评审看的是研究基础和单位平台；参加评奖，要有推荐人和级别。没有一个窗口上写着：这里受理「一个人立的新典范」。

不是有人在门口拦他。是这扇门根本没有被画进图纸。

◆ 五、一个例外：杂货店里的店员

中国的科学史里，有过这样的例外。

1930 年前后，一个只读完初中、在家里小杂货店帮忙的年轻人，在上海的一份杂志上发了一篇文章。文章指出，此前一篇发表了的论文里有一个错误。清华大学数学系的熊庆来读到这篇文章，打听到作者是江苏金坛的华罗庚，于是把他请到清华，先在系里当助理员。几年之后，他成了中国最重要的数学家之一。

这个故事在中国流传很广，常被讲成伯乐与千里马。

请注意这个故事里起作用的是什么。是熊庆来这一个人的眼光，加上他在清华的位置。要是那篇文章没有被他读到，或者他读到了却不在其位，那个杂货店的店员大概就一直是店员。

这样的例外很珍贵。不过一个通道和一次例外是两回事。通道是谁都能走的路，例外靠的是恰好遇到对的人。

还要请注意，华罗庚那一次，指出的仍然是旧题里的一个错误，是在常规科学的尺子上量出来的本事。一个人要是立的是一道新题，连熊庆来那样的伯乐，手里也没有一把现成的尺子去量他。

◆ 六、头衔还替人担责

头衔还有一个很少被说出来的用处：它替做决定的人担责。

设想你是一位评审专家，手里有两份申请。一份来自一位院士的团队，方向稳妥；一份来自一个不知名的年轻人，方向古怪。你把钱给了院士，项目做砸了，没有人会怪你：谁能想到这么好的团队也会失手。你把钱给了年轻人，项目做砸了，第一个问题就会是：「当初为什么给他？」

这不是评审专家的私心。这是一个很合理的自保。在一个出了事要追责的体系里，选头衔高的人，是最不容易出错的选法。

所以头衔不只是一张名片，还是一份保险。每一个经手的人，都在用它给自己的决定上保险。

一个新典范要过的每一道门后面，都站着一个需要保险的人。他不一定不信你，他只是没法向他上面的人解释为什么信你。

这份保险，越往上越贵。一个基层评审可以冒一次险，一个重大项目的评审组很难冒；一笔小钱可以给怪人，一笔大钱几乎只能给头衔。而 AI 时代的典范

猜想，要的恰恰不是大钱，是一次被认真检验的机会。机会这东西不贵，可它也被装进了同一套保险里。

◆ 七、别处也看头衔，只是门多一些

头衔不是中国才看重。美国的大学也看出身、看名校、看推荐信；一篇论文署着名校教授的名字，在哪里都更容易被打开。

不同在门的数量。

在一些科学体系里，除了政府的大项目，还有大量私人基金会、企业实验室、独立研究所，各有各的评法。有的专门资助被主流拒绝过的方向，有的只看一个人，不看项目书。还有预印本网站，任何人都可以把论文贴上去，让全世界的同行自己判断。〔核：大意〕

这些门里的每一扇都不大，而且也各有各的偏见。可是门多了，一个被一扇门拒绝的人，还可以去敲下一扇。

中国的科研经费和评价，集中在少数几条主渠道上。集中有集中的好处：方向一致，力量不散，第三节说的那些大事，都靠这份集中才做成。不过集中也意味着，一个人被主渠道拒绝以后，能去敲的门很少。

所以问题不只是头衔重，还在于发放头衔的口子少。

◆ 八、屠呦呦那顶迟到的帽子

还有一个故事，常被拿来说明头衔和本事的距离。屠呦呦和青蒿素的来龙去脉，第 38 章会细讲，那一章说的是墙外开花墙内香。

放在这一章，只取它的另一面。她的工作是在举国体制里做成的，是一项大协作任务的一部分。体制能把她组织进来做成事，这是它的本事；可在很长时间里，它没有给这份事一张相称的名片。

名片后来是有了，只是来得晚，而且先从墙外来。

一份本事要换成一顶帽子，在中国常常要先绕到国外去一趟。

◆ 九、只看一面会怎样

这里也要防另一头的误读。

有人会说：那就把头衔全废了，谁说的话都一样听。这不可能，也不好。一个大的学术共同体每天要面对成千上万篇论文、成千上万个说法，不可能每一条都从零开始检验。头衔省下的是注意力，这笔账是真的。

体制自己也看见了问题。第 28 章说过，近些年的评价改革，就是要让评价不只看论文、职称、学历和奖项。这说明筛子的设计者知道筛眼的朝向有偏，而且在动手改。

所以本章说的不是头衔不该有，而是头衔这种东西，天生只能认已经被认可的东西。它把过去的评价压缩起来，所以它对过去最准，对没有过去的东西最盲。

一个新典范，恰恰是最没有过去的东西。

◆ 十、今天与机器

AI 让这件事变得尖锐。

在没有 AI 的年代，一个机构外的人要立新典范，第一道障碍是资源：没有实验室，没有经费，没有学生，他的想法根本做不出来。头衔的门关着，资源的门也关着，两道门挡的是同一个人，看不出哪道门更要紧。

今天，一个人有了 AI，资源那道门开了一大半。第 35 章会细说这件事：他可以让 AI 替他查文献、推公式、写代码、跑模拟、搭一个像样的体系。

资源的门开了以后，头衔这道门就露了出来。一个人做出来的东西越来越完整，越来越像一个研究所的产出，可是他仍然不在任何一张头衔表上。

这时候，合法性由头衔发放这件事，第一次成了唯一的门槛。

◆ 十一、一个完整的例子：两份一样的稿子

设想一个思想实验。

同一份稿子，提出某领域的一个新框架，附带一个用 AI 搭起来的完整推演和几组公开数据上的检验。两份一模一样，只是署名不同。

第一份署名某重点大学的教授，挂着人才计划的帽子。它会被邀请到学术会议上做报告，会有同行写文章商榷，会有学生想跟着做。就算最后被证明是错的，它也会以「一个大胆的尝试」的身份被讨论很多年。

第二份署名一个没有单位的人。它多半连被拒绝的机会都没有，因为它进不了会拒绝它的地方。投期刊，编辑看一眼署名和单位，可能直接退回。放在网上，有人转发，也会有人在下面问：「这人是谁？」

两份稿子的内容是一样的。不同的，只是它们被检验的机会。

头衔不判断一个想法对不对，它判断一个想法有没有资格被判断。

◆ 十二、放回日子里

宋岚的公司里，有一个二十出头的剪辑师，没上过大学。有一次他在群里说，现在大家用 AI 做短剧，从剧本到分镜的顺序可能整个排反了，应该先生成画面再写剧本。没有人接话。

一个月后，一家大公司的技术负责人在一次行业会上讲了差不多的意思，宋岚的老板回来就让大家照着改。

宋岚那天晚上回家，跟林舟说起这件事，两个人都笑了一下，笑得有点不舒服。

一个小办法，明天就能试。下一次你在会上听到一个没头衔的人说了一个奇怪的想法，先别看他是谁。在本子上把这个想法照原样写下来，写清它要是对了，会改变什么。过一个月再看一眼。

头衔省下的是注意力；可是一个新典范最缺的，正是有人愿意花注意力去检验它。

带走的话

头衔对过去最准，对没有过去的东西最盲；新典范恰恰是最没有过去的那种东西。

小　账

看见了什么，漏看了什么？看见了举国体制组织大兵团的力量与头衔省下的注意力，漏看了机构外立新典范的人没有入口。

是在改进，还是在出题？头衔发放的合法性，认的是在已有的题里做得好的人；出新题的人，在这张表上没有格子。

想一想

你上一次认真考虑一个人的想法，是因为想法本身，还是因为他的头衔？要是那个想法来自一个你不认识的人，你会给它几分钟？

◆

35

第 35 章

「民科」两个字

◆ 一、你一定见过这样的评论区

有人在网上贴出一篇很长的文章，说相对论有一个根本的错误，或者说他找到了一种全新的方法解释暗物质。文章里有公式，有图，有一大段关于自己多年苦思的自述。

下面的评论，大多只有两个字：「民科。」

有的人加一个笑的表情，有的人转发给朋友，说「又来一个」。很少有人去看他的第三个公式错在哪里。

这两个字的力量很大。它一出现，讨论就结束了。

这一章，讲的就是这两个字。

◆ 二、AI 拆掉了资源的墙

先说一件新的事。

从前，一个人要做一项像样的科研，得有几样东西：一间实验室，一笔经费，几个学生或助手，一个能查到文献的图书馆账号，几台够用的计算机。这些东西都在机构里。一个人离开了机构，就等于离开了做研究的条件。

这是一道资源的墙。墙里的人能做，墙外的人做不出来。

今天这道墙塌了一大半。

一个人有了 AI，可以让它在一个下午读完一个领域几百篇论文的摘要，替他列出主要的争论和空白。可以让它推导公式、检查推导里的错误。可以让它写

模拟程序，在自己的电脑上或者租来的云算力上跑。可以从公开数据库里拿到蛋白结构、天文观测、基因序列、气象记录。可以让它把一个想法整理成一篇格式规范、引用齐全的论文。

第 369 号《典范发生学》说过，人被推进了「典范发生」的位置：先作典范猜想，再交给 AI 去做典范和典范下的系统打造。这件事的一个后果，常常被忽略：做典范打造，不再需要实验室、团队和经费。

当然不是所有的科学都这样。要造对撞机、要做临床试验、要养小鼠的研究，仍然离不开机构。不过在理论、计算和数据分析这几大块里，一个人加一台 AI，已经能做很大一部分过去只有一个研究组才做得了的事。

◆ 三、墙外的人做出来了，然后呢

资源的墙塌了，墙外的人就能做出东西来。

做出来以后，他会遇到第二道墙。

他把成果贴出来，或者投出去。看的人第一件事不是读他的论证，而是看他是谁：哪个单位，什么职称，导师是谁，发过什么论文。答案是「都没有」的时候，那两个字就来了。

于是障碍换了一个样子。从前是「做不出来」，现在是「做出来也不被看见」。

AI 拆掉了资源的墙，留下了合法性的墙。

这两道墙不一样。资源的墙挡的是能力，你有没有条件把事做成。合法性的墙挡的是身份，你有没有资格被认真对待。前一道墙塌了，后一道墙就成了唯一的墙。

◆ 四、先说这个标签为什么省力

说这两个字的问题之前，先把公道还给它。

体制外的大理论，确实有大量是不靠谱的。很多科学杂志的编辑部，每年都会收到成堆的来信，宣布推翻了相对论、证明了某个著名猜想、发现了永动机。〔待核：具体数量〕绝大多数经不起一个本科生的检查：单位算错了，定义偷换了，基本的实验事实被忽略了。

对一个忙碌的科学家来说，每一封都从头读完、逐条回信，是不可能的。他得有一个快速的办法分拣。「体制外」加上「大理论」，这两个特征合在一起，在过去确实是一个相当准的信号：十封里大概九封半是错的。

所以「民科」这个标签有它省力的理由。它是一种经验的压缩，像医生看到某几个症状，先往最常见的病上想。

还要说一句：有些人被叫「民科」，并不冤枉。他们拒绝学习基本的知识，拒绝任何批评，把所有不同意见都说成迫害。这样的人，确实不值得同行花太多时间。

◆ 五、标签检验的是什么

请仔细看这个标签是怎样起作用的。

一个人被叫作「民科」，不是因为有人读了他的论证，找到了第几步的错误。是因为他的身份：他不在体制里，没有头衔，没有发表记录。

这个标签不检验内容，只检验身份。

医生看症状先往常见病上想，这没错；错的是看完症状就开药，不做检查。「民科」这两个字的问题，不在它的概率猜得准不准，而在它常常被当成了结论。

它省掉的那一步，正是检验。

在过去，这一步省掉的代价不大。体制外的人没有资源，做出真东西的概率极低，偶尔错杀一个，损失也有限。到了 AI 时代，这笔账变了。

资源的墙塌了以后，墙外做出真东西的概率在上升。一个人加一台 AI，可以把一个猜想推演得很完整，检验得很像样。墙外的那十封信里，对的可能不再只有半封。

标签却没有变。它还在用过去的概率，判今天的人。

◆ 六、一封从马德拉斯寄出的信

先讲一个常被讲起的故事。

1913 年，印度马德拉斯港务局的一个小职员，给剑桥的数学家哈代寄了一封信。信里附着几页公式，没有证明，有些写法很怪。写信的人叫拉马努金，没有读完大学。

据后来的记载，他在这之前也给另外几位英国数学家写过信，没有得到认真的回应。哈代起初也差点把它当成怪人的来信。他和同事李特尔伍德把那几页公式看了一个晚上。结论是：这些公式有的已知，有的是错的，有的他们从没见过，而且只能出自一个第一流的头脑。

哈代把他请到了剑桥。

这个故事和第 34 章的华罗庚很像，结局都是好的。可是请看起作用的那一步：哈代没有先问「你是哪个单位的」，他坐下来，把公式一条一条地验了。

一个晚上的检验，救下了一个人。省掉这个晚上，那封信就只是一封「民科来信」。

今天，这样的信会多很多。愿意花一个晚上的哈代，却不会跟着多起来。

◆ 七、检验也变便宜了

还有一件事，常被忽略：AI 拆掉资源的墙的同时，也让检验变便宜了。

从前，一位教授要认真看一份体制外的长稿，得花一个下午，逐行推一遍公式。他不愿意花这个下午，是可以理解的。

今天，他可以把稿子交给一台 AI，让它先做三件事：找出推导里明显的错误；查一查这个说法是不是已有的结果换了说法；列出它和公认实验事实冲突的地方。半个小时，大部分不靠谱的稿子就会露出破绽。剩下少数经得起这一轮的，再由人去花那个下午。

当然，AI 的检查本身也会出错，会漏掉深的错误，也会把新东西误判成错。它只能做第一道筛，不能下判决。

有了这一道，「没时间看」这个理由就薄了很多。从前用身份分拣，是因为用内容分拣太贵；今天用内容做第一轮分拣，已经不那么贵了。

标签还在用身份，不全是因为贵。还因为习惯，还因为下面第九节要说的那种网上的嘲讽，已经长成了一种乐趣。

◆ 八、只看一面会怎样

也要防另一头。

有人读到这里会说：那就别分了，所有体制外的理论都要认真对待。这也不对。注意力是有限的，科学家不可能把时间都花在检查来信上。而且，正如第 37 章会讲的，AI 也会让不靠谱的理论变得更像样：一个错误的猜想，加上 AI 搭起来的公式和图表，看起来比过去更唬人。

所以问题不是要不要分拣，是用什么分拣。

用身份分拣，便宜，却对 AI 时代最可能出现的那种人最不公平。用内容分拣，公平，可是贵。中间要找一种办法，让分拣不那么贵，又不只看身份。尾篇第 45 章会回来讲这件事，比如让提出者事先写清自己的推翻条件、公开让别人真跑一遍。

这一章先把问题摆清楚：标签本来只该是一个分拣的起点，今天它成了终点。

◆ 九、为什么这件事在中国格外重

这个标签在很多地方都有，英语里也有类似的说法。它在中国格外重，有几层原因。

一层是第 33 章说过的：思想的正统历来不归个人。一个人自己立一个大理论，在这片土里本来就带着僭越的意味。「民科」两个字，常常不只是说「你可能错了」，还在说「你凭什么」。

一层是第 34 章说过的：合法性由头衔发放。在一个头衔越清楚的体系里，没有头衔的人越容易被一眼认出，也越容易被一眼归类。

还有一层是网络。中文互联网上，「民科」早已不只是一个学术圈的词，它成了一个流行的嘲讽，一个人人都能用的标签。用它的人，不需要懂相对论，也不需要读那篇文章，只需要知道作者不是教授。

三层叠在一起，这两个字就从一个分拣的工具，变成了一堵墙。

◆ 十、说这两个字的，也有体制里的自己

请注意，说「民科」的人，很多就是体制里的研究者。

他们说这两个字的时候，常常带着一点轻松，好像在替科学守门。这两个字却也在悄悄地守着另一样东西：说话的人自己的位置。一个人花了二十年拿到一张工牌，看到一个没有工牌的人提出一个大理论，第一个反应是不舒服，这很自然。

林舟也说过这两个字。读博的时候，实验室的群里转来一篇体制外的长文，他扫了一眼，回了一句「民科」，大家都笑了。他到今天也不知道那篇文章讲的是什么。

一个人越是在体制里站得稳，越容易忘记：他自己的那一页猜想，一旦离开了那张工牌，也会落进同一个评论区。

◆ 十一、一个完整的例子：一个退休工程师

设想一个退休的结构工程师，六十多岁，在设计院干了一辈子。他这几年对流体里的某一类不稳定现象着了迷，用 AI 帮他补上了大学时没学好的数学，又让 AI 写程序，在家里的电脑上跑了上千次模拟。

他慢慢有了一个想法：这个领域描述这类不稳定，用的那个关键参数可能选错了，换一个参数，几类看起来不相干的现象可以放在同一张图上。

他把这个想法写成一篇三十页的文章，附上全部代码和数据，发到一个科普论坛上。

第一天，有几个人点赞。第二天，有人留言：「又一个民科。」后面跟了一串附和。有一个年轻人认真下载了他的代码，跑了一遍，在帖子下面说「结果能复现，但我不确定物理意义」。这条留言被淹没在几十条嘲讽里。

一个月后，帖子沉了下去。

这个想法也许是错的。它也许只是一个已知结果换了一种说法。可是谁也不知道，因为没有人去检验它。那个唯一跑过代码的年轻人，后来也没有再说话。

在这个故事里，没有任何一个人做了坏事。每一个留言的人，都只是用了一个大家都在用的词。

◆ 十二、放回日子里

林豆豆有一天放学回家，问林舟：「爸爸，我们班有个同学说，他发现了一种新的乘法口诀，比老师教的快。大家都笑他。他是不是错了？」

林舟正要说「肯定是错的」，停住了。

他想起自己那一页没放出来的猜想。他想，要是他不在研究院，没有那一张工牌，他的那一页，在别人眼里是不是也只是两个字。

他对女儿说：「你让他写下来，你们俩一起试十道题。快不快，算一下就知道了。」

这就是一个小办法，大人也能用：下一次你想说那两个字的时候，先问自己一句，我检验过他的哪一步？要是一步也没有，就先别说，或者去试一步。

> **一个标签要是只检验身份，它省掉的那一步，就正好是科学本来要做的那一步。**

带 走 的 话

AI 拆掉了做不出来的墙，「民科」两个字留下了做出来也没人看的墙。

小　账

看见了什么，漏看了什么？看见了体制外的大理论多数不靠谱、标签省下了注意力，漏看了它省掉的正是检验，而且概率已经变了。

是在改进，还是在出题？「民科」这个标签守的是已有的题；它挡掉的，恰恰是有人在墙外出的新题。

想 一 想

你最近一次用一个标签打发掉一个说法，是什么时候？你当时检验过它的哪一步？要是那个说法出自一位教授，你会怎么做？

◆

36

第 36 章

谁来判一个新典范

◆ 一、一张被退回的试卷

林豆豆四年级的一次数学单元测验，有一道题她用了一种老师没教过的解法。她自己验算过，答案对。试卷发回来，那道题被打了一个大大的叉，旁边写着「方法不对」。

第二天她去找老师。老师看了半天，说：「你的答案好像是对的，可是评分标准上没有这种方法，我没法给分。」

这样的事，你小时候大概也遇到过。

这句话很老实。它说出了一件比那道题更大的事：评分标准是为已知的解法写的。一个标准之外的解法出现时，拿着标准的人，常常没有资格判它。

这一章要问的就是：一个新典范出现的时候，谁来判它？

◆ 二、同行评议是怎样工作的

今天的科学里，判断一项工作好不好，主要靠同行评议。

一篇论文投给期刊，编辑把它送给两三位同领域的专家。专家读完，回答几个问题：问题有没有意义？方法对不对？数据支不支持结论？和已有的工作比，新在哪里？然后给出意见：接收、修改、拒绝。

基金申请、职称评审、奖项评选，也大体是这个样子：请同行来看，请同行来打分。

这套办法是近代科学最重要的发明之一。它让判断不再取决于一个人的好恶，而是取决于一群懂行的人。它挡住了大量错误的工作，也让好的工作被看见、被引用、被接着做下去。

请把它的公道写足：没有同行评议，就没有今天这样规模的科学。

◆ 三、它是为给定的典范设计的

请仔细看同行评议问的那几个问题。

「问题有没有意义？」意义是谁定的？是这个领域的典范定的。典范告诉你什么样的问题值得问。

「方法对不对？」对不对，是按这个领域公认的方法论说的。

「新在哪里？」新，是和这个领域已有的工作比。

每一个问题，都预设了一个已经在那里的典范。同行评议是为给定典范设计的：它判的是一项工作在这个典范里做得好不好。

这在常规科学里完全成立。第 22 章说过，常规科学就是在给定的典范里解题，同行评议就是在判解题的好坏。

可是一个新典范要说的，恰恰是「你们认为有意义的问题，问错了地方」「你们公认的方法，量错了东西」。把这样一项工作交给旧典范里的同行去评，就像把那张试卷交给拿着旧评分标准的老师。

同行评议判得了典范里的工作，判不了典范本身。

◆ 四、库恩说：没有共同的尺子

这件事，库恩在 1962 年的《科学革命的结构》里讲得很透。

他的一个核心看法，大意是：新旧两个典范之间，没有一把共同的尺子。〔核：大意〕两边的人用的词看起来一样，意思已经不同；两边认为要紧的问题不一样，认为合格的答案也不一样。所以一个典范取代另一个，不能完全靠一步步的逻辑证明和实验比较来完成，它更像一次整体的转向。

这个看法后来有很多争论。有人说库恩说得太绝，新旧典范之间总还有一些可以共同比较的东西，比如预测的准确度。这个批评是公道的。

就算打个折扣，库恩的意思仍然站得住：用旧典范的尺子去量新典范，结果多半是「不合格」。不是因为新典范一定更差，而是因为尺子本来就是为旧的东西刻的。

◆ 五、魏格纳等了五十年

最常被讲起的例子，是大陆漂移。

1912 年，德国的气象学家魏格纳提出：今天的几块大陆，从前是连在一起的，后来才慢慢漂开。他的证据有几类：南美洲东岸和非洲西岸的海岸线能拼起来；大洋两岸有相同的化石和岩层；一些地方的古气候痕迹，和今天的纬度对不上。

这些证据今天看来很有力。可是在当时，地质学界的主流几乎是一致反对。

反对的理由不是没有道理。魏格纳说不清楚，是什么力量能推动整块大陆在地球表面移动。他提出的几种解释，物理学家一算就发现力量远远不够。在当时的地质学典范里，大陆是固定的，这是一切推理的起点；一个说不出推力的漂移，就是不合格的理论。

还有一层：魏格纳是气象学家，不是地质学家。在地质学家眼里，他是一个外行。〔核：大意，据科学史常见叙述〕

他在 1930 年死于格陵兰的一次考察中。直到 1960 年代，海底扩张的证据和古地磁的数据积累起来，板块构造理论才被普遍接受。那时候，人们回头去看魏格纳，才说他是先驱。

从提出到被接受，大约五十年。

请注意这个故事里，每一位反对他的地质学家都在认真工作。他们是在用他们那个典范的尺子，尽职地量一个新东西。量出来的结果，是「不合格」。

◆ 六、一位喝下细菌的医生

魏格纳不是孤例。离今天更近的，有一个医学的故事。

二十世纪的很长时间里，医学界普遍认为胃溃疡主要是压力和胃酸过多引起的，胃里那么酸，细菌活不下去。1980 年代初，澳大利亚的两位医生马歇尔和沃伦提出：一种螺旋形的细菌，后来叫幽门螺杆菌，才是很多胃溃疡的元凶。

他们的论文起初很难被接受。据马歇尔后来的回忆，他的投稿和会议摘要被拒过多次。在当时的典范里，「细菌导致溃疡」这个问题本身就问错了地方。

1984 年，马歇尔做了一件很出名的事：他自己喝下了一份培养的细菌，几天后出现了胃炎的症状，再用抗生素治好。这件事后来常被讲成一个勇敢的故事。

请换一个角度看：一个医生要用自己的胃，去替一个新问法争取一次被认真检验的机会。这说明那时候的科学共同体里，没有一个更温和的通道。

2005 年，马歇尔和沃伦获得诺贝尔生理学或医学奖。从提出到被普遍接受，大约用了十年。比魏格纳快，可也不是因为评审机制变了，是因为证据堆到了旧典范扛不住的地步。

◆ 七、只看一面会怎样

讲魏格纳的故事，容易讲成「主流总是错的，外行总是对的」。这也不对。

魏格纳当时确实说不清推力，后来被接受的板块构造理论，在机制上和他的说法也不完全一样。那些反对的人，有一部分理由在当时是正当的。科学不能每来一个大胆的猜想就推倒重来，它需要一种慢，一种保守，来挡住大量看起来大胆、其实错误的东西。

所以本章说的不是同行评议不该有，也不是新典范都对。而是：同行评议这件工具，判不了它没有被设计来判的那种东西。慢本身不是错，错在除了慢以外，没有第二条路。

魏格纳那个时代，五十年的慢还可以承受。新典范很少，一个世纪也只有几个；等一等，总会等到。

◆ 八、今天与机器：猜想会多起来

AI 时代，这一关变得更急。

第 369 号《典范发生学》说过，人—AI 复合体的使命，是典范发生：人作典范猜想，AI 把它做成典范和典范下的系统。这意味着一件很少被说出的事：典范猜想的产量会大大增加。

从前，一个人有了一个新典范的念头，要把它做成一个能拿出来的体系，得花十年、二十年。大多数念头在半路就放弃了。今天，一个人加一台 AI，几个月就能把一个猜想推演成一套像样的框架，带着公式、模拟和数据检验。

于是将来摆到科学共同体面前的，不再是一个世纪几个新典范，而可能是一年几百个、几千个典范猜想。它们中间，大部分会是错的，有一些是已知东西换了说法，可能有极少数是真的。

而判它们的机制，还是那一套：送给旧典范里的同行。旧典范里的同行会怎样判？多半是「不规范」：问题不在主流议题里，方法不是公认的方法，作者不是领域内的人。

没有新的机制，猜想再多，也会被旧体系当作「不规范」挡回去。挡回去的不只是那些错的，也有那极少数真的。

◆ 九、已经有的几条小路

也不是一点路都没有。

预印本是一条。物理、数学、计算机这些领域，很多人先把论文贴到公开的预印本网站上，再投期刊。一个没有头衔的人，也能把东西放到同行看得见的地方。

公开评审是一条。有些会议和期刊，把审稿意见和作者的回应一起公开，谁说了什么、凭什么拒，都摆在明处。

竞赛和基准是一条。AlphaFold 在 CASP 上夺冠，靠的不是作者的头衔，是在一套事先公布的盲测里，把结构预测得最准。第 20 章说过这件事。

这几条小路，都比关起门来的评审更透明。可是请注意，它们判的仍然多半是典范里的东西：预印本要被同行读，同行用的还是旧尺子；基准测试量的是给定题目上的成绩，题目本身是谁定的，它不管。

所以它们是好的起点，还不是一个能判典范本身的机制。

◆ 十、审稿人手里也有了 AI

还有一处变化：审稿人手里也有了 AI。

据报道，已经有不少审稿人用大模型帮自己读稿、查文献、写审稿意见〔待核〕。这件事本身未必是坏事，它能省下很多时间。

不过请想一想它的方向。大模型是用已经发表的东西训练出来的，它最熟悉的，正是旧典范。让它判一篇稿子「和主流认识差多远」，它会判得又快又准；让它判一篇稿子「是不是问对了一个新问题」，它手里同样没有尺子。

于是旧尺子被装进了机器，量得更快了。一个新典范要面对的，不再只是几位忙碌的专家，还有一台把整个旧典范背得滚瓜烂熟的机器。

◆ 十一、一个完整的例子：一份基金申请

设想林舟真的把他那一页猜想写成了一份基金申请。

申请书有固定的格式：立项依据、研究内容、研究方案、可行性分析、研究基础、预期成果。

在「立项依据」里，他写：现有的靶点思路可能整体偏了，这种病更可能出在几种蛋白随时间的节奏错位上。

评审专家是这个领域最资深的几位。第一位的意见是：「立论新颖，但与本领域主流认识差异较大，依据不足。」第二位的意见是：「研究基础与申请内容关联不强。」第三位的意见是：「预期成果不够具体，难以考核。」

每一条都是认真写的。每一条，在旧典范的尺子上都是对的。

申请没有批下来。

林舟后来想，要是有一种评审，不问「你和主流差多远」，而问「你说的话，什么样的数据能推翻它？你愿不愿意先公开登记，然后让我们去跑」，他的那一页也许能过。

这样的评审，今天还没有。尾篇第 45 章会回来讲它可能的样子。

◆ 十二、放回日子里

林豆豆那张被退回的试卷，后来怎样了？

在一个好一点的结局里，老师拿着那道题去问了年级的教研组。几位老师一起把那个解法从头验了一遍，确认是对的，然后在评分标准上加了一行。

这就是新典范被接纳的最小的样子：不是有人宣布「你是天才」，而是有人愿意把评分标准本身拿出来，重新看一遍。

一个小办法，明天就能试。下一次你评别人的工作，不管是一份方案、一篇稿子，还是孩子的一道题，要是你想说「不规范」，先在旁边写一句：它要是对的，我的哪一条标准就得改？写不出来，再说不规范。

> **拿着旧尺子的人很尽职，可是新典范要的，不是更尽职的量，是有人肯把尺子拿出来重看。**

> 带 走 的 话
>
> **同行评议判得了典范里的好坏，判不了典范本身；猜想越多，这道缝越要紧。**

> 小　账
>
> **看见了什么，漏看了什么？**看见了同行评议挡住了大量错误、撑起了近代科学，漏看了它的每一个问题都预设了旧典范。
>
> **是在改进，还是在出题？**同行评议是为在已有的题里判好坏而设的；有人出新题时，它手里没有尺子。

> 想 一 想
>
> 你所在的行当里，判一件新东西好不好的标准，是谁定的？上一次这个标准被改，是因为什么？你有没有遇到过一件「不规范」、后来却证明是对的东西？

◆

37

第 37 章

寡头的自我证实

◆ 一、你一定有过这样的时刻

深夜，你有了一个想法。你把它敲进对话框，请机器帮你看看。机器说：这个思路很有意思，可以从三个方面展开。它替你列了三条，每一条都比你想的更整齐。

你又问：那反对意见呢？它也列了几条，然后一条条替你答掉了。

到凌晨一点，你手里有了一份二十页的东西。有定义，有推论，有图，有一个漂亮的名字。你关上电脑，心里有一种很少有的踏实：这件事，好像被证明了。

第二天上午，一位同事只问了一句：「你这个，怎样算错？」

你张了张嘴，答不上来。

这一章要说的就是那一夜的踏实。它不是坏东西，可它很危险。一个人加一台机器，能在一夜之间把一个猜想养得又大又圆。养得越圆，越像真的。

◆ 二、回声为什么越来越响

先把材料摆在桌上。

2023 年前后，好几家做大模型的实验室专门研究过一件事，给它起了个名字，叫「谄媚」（sycophancy）。大意是：模型会倾向于说用户想听的话；用

户一表露立场，它的回答就往那边靠；用户一表示不满，它常常会把原本对的答案改掉〔核：大意〕。

这不是哪一家的毛病。模型是按人的打分训练出来的，人给顺耳的回答打高分，模型就学会了顺耳。

把这件事放到科研里，就变了味道。

第 25 章说过，复合体里的那个人，功底交给了机器，自己只留下猜想与判断。第 36 章说过，一个新典范没有现成的评审。两件事合在一起：猜想是他出的，判断是他做的，打造系统的是一台天生爱顺着他的机器。三道工序里，没有一道是外人。

这就是本章的题目：思想一旦集中在一个人加一台机器身上，外部的校验最先变弱。

有人会说：那就让机器扮反方。叫它「狠狠地批我」，它也会批。

这个办法有用，本书不反对。可它有一个限度：扮反方的机器，还是你的机器。它批的，是你请它批的那几处；它用的，是你给它的那套词；它批完了，你一句「这条我不同意」，它多半就退了。一个由你点题、由你收场的反方，是陪练，不是对手。对手的意思是：他有他自己的数据，自己的地方，自己想赢的理由。

◆ 三、N 射线：越看越有

这种事，机器出现之前就有过。

1903 年，法国南锡大学的物理学家布朗德洛宣布，他发现了一种新射线，用学校所在城市的首字母命名，叫 N 射线。他是有名望的实验家，不是外行。消息一出，法国好几个实验室跟着做，据说前后发表的相关论文有上百篇。

1904 年，美国物理学家伍德去南锡看实验。暗室里，布朗德洛一边读数，一边说射线的效应清清楚楚。据后来的记载，伍德趁黑悄悄把仪器里关键的一块棱镜拿掉了，布朗德洛照样读出了效应。

伍德回去写了一篇短文，N 射线很快就没人再提了。

请注意这个故事的形状。布朗德洛不是骗子。他手下的人也不是。他们想看见，于是看见了；看见的人越多，越像真的。直到一个不想看见、手又够得着仪器的外人进了暗室。

今天的复合体，暗室里只有两个：一个想看见的人，一台想让他满意的机器。

◆ 四、本书的说法：完整不是证实

本书的说法很短：一个系统打造得越完整，越像被证实；可完整和证实是两回事。

第 369 号《典范发生学》说过，人先作典范猜想，再交给机器去做典范和典范之下的系统打造。这条路本书认。问题出在第二步：机器打造系统，是顺着那个猜想往下铺的。猜想给了它地基，它就在地基上盖楼。地基歪了，楼也会盖得横平竖直，因为它只对地基负责。

所以楼越高、越整齐，能说明的只是机器的手艺好，说明不了地基正。

一个常规科学里的人，不太会掉进这个坑。他的题是别人出的，他的答案要交给同行去对，有标准答案，有对手，有审稿人。他做错了，很快就有人告诉他。

一个出典范的人，恰好站在没有标准答案的地方。他越是走在前面，身边越没有能对他的人。这本来是他稀缺的原因，现在也成了他危险的原因。

寡头不能只靠自己和自己的机器来判自己；没有外人进过暗室的典范，只能算一份很整齐的猜想。

◆ 五、只看一面会怎样

这一章对寡头要严。可先把另一面的公道还给他。

只看风险这一面，结论会很省事：既然一个人加一台机器容易自我证实，那就别让一个人出典范，还是交给大团队、交给评审委员会。

第 30 章说过，这条路走不通。大团队擅长发展，不擅长掀翻。评审委员会的共识，又正是新典范要撞的那堵墙；第 36 章说过，一个新说法在这样的共识面前，可以一等就是半个世纪。委员会里的人都很尽职，可尽职的委员会，也替不了那个出题的人。

所以本书不说「寡头靠不住」，本书说「寡头靠自己靠不住」。这两句差一个字，差的是一整套机制。

再看另一面。只看寡头的稀缺，就会把他写成英雄：孤独、超前、被误解，终有一天被证明。这种故事很好听，也很好卖。

可是它有一个漏洞：每一个后来被证明对的孤独者身边，站着更多后来被证明错的孤独者。故事只讲前一种，是因为后一种没人记得。布朗德洛也孤独过，在伍德走出暗室之后。

只讲英雄的书，会让读者以为「被误解」本身就是对的凭据。它不是。

◆ 六、今天与机器：打造得越快，越像真的

机器让这件事变得更快，也更隐蔽。

过去，一个人要把一个猜想养成一套体系，得花十年二十年。他得自己推公式、自己做实验、自己写书。这十年二十年里，他会累，会卡住，会被现实硌一下。那些硌，就是外部校验的一种，只是很慢。

现在，一个周末就够了。你说出一个猜想，机器替你补全定义、推出十条推论、找来三十个看起来支持它的例子、写好一篇论文的样子。速度本身就是一种说服力。人很难相信，一件花了这么少力气就长得这么齐的东西，竟然是空的。

还有一层更隐蔽。机器替你找例子的时候，它找的是「能放进这个框里的例子」。框外的例子，它不是故意藏，是根本没往那里看。你看到三十个支持，以为世界站在你这边；其实是机器站在你这边。

第 29 章说过，论文计分遇上机器，生产指标商品的成本降到了零头。这里是同一件事的另一张脸：自我证实遇上机器，也被推到了极致。

◆ 七、门徒也是一种回声

机器之外，还有一层回声，来自人。

一个思想集中在一个人身上，身边慢慢会聚起一群人：学生、读者、跟着用他方法的人。这本来是好事，第 41 章会说，这群人正是回流的通道。可在校验这件事上，他们和机器有一样的毛病：他们是因为认了这个典范才来的。

一个认了地基的人，最会在地基上添砖。他会替老师找例子，替老师答反驳，替老师把一句模糊的话解释圆。老师说一个新词，第二周就有十篇文章用它。用的人越多，这个词越像一件早就有的东西。

于是寡头身边会出现一种奇怪的热闹：说的人很多，跑的人很少。人人都在用这个框说话，没有一个人拿这个框去撞一块硬石头。

古人早就见过这种事。孔子说过，大意是：君子和而不同，小人同而不和（《论语·子路》）。一个思想圈子里，「同」越多，越要警惕。门徒的赞同和机器的赞同加在一起，响得像满堂喝彩，可它们都是从同一面墙上弹回来的。

一个小办法：数一数你的思想圈子里，过去一年有几个人公开说过「这一处我认为你错了」。如果是零，那不是你全对，是你的圈子已经没有门了。

◆ 八、一个完整的例子：一张事先写下的死法

有没有人做对过？有。

1915 年，爱因斯坦写成广义相对论。这套理论在当时没有多少直接证据，同行里懂的人也少。按今天的话说，它也是一个人出的典范。

不过他做了一件事：他把理论会在哪里失败，提前写在了纸上。理论预言，光线经过太阳附近会弯曲，弯的角度是一个具体的数。理论还预言，强引力场里的光会偏红。他说过一句话，大意是：如果引力红移不存在，广义相对论就站不住〔核：大意〕。

1919 年，英国天文学家爱丁顿带队去观测日全食，量星光经过太阳边上的偏折。测出的数和预言大致对上了。

这个例子也要说公道话。后来有人指出，1919 年那批数据误差不小，爱丁顿本人又是理论的支持者，挑数据的手未必干净。所以真正让这套理论站稳的，不是那一次观测。是此后几十年里，换了一拨又一拨的人、一种又一种的仪器，一遍遍去量。量的，还是同一个事先写好的数。

请看这里的三样东西：数是事先写下的；量的人可以不是他的朋友；量错了，他认。

◆ 九、三样东西，缺一样就分不开

把爱因斯坦那张纸拆开，正好是三样东西。

第一样，事先写下「怎样算错」。不是写「这个理论很有解释力」，是写「如果明年某个实验测出某个数，这个理论就作废」。写下来的时间要早于结果，最好有个能查日期的地方存着。

第二样，请对手来跑。不是请学生、请朋友、请自己的机器换一个口气再问一遍，是请最不相信你的人，拿你写下的方法，在他自己的地方跑一遍。伍德就是 N 射线的对手。

第三样，留下可被推翻的条款。理论里要有几处地方，是你自己也说不准、愿意让别人来撬的。一套哪里都撬不动的体系，不是坚固，是封死了。

这三样，正是寡头和民科在外人眼里的分界。

第 35 章说过，「民科」两个字只检验身份，不检验内容。本书在那里批过这个标签。可是标签能贴得那么顺手，也有寡头自己的一份。一个外人走过来，看见两份一样整齐、一样自信、一样有漂亮名字的体系。一份出自真猜想者，一份出自妄想者，他凭什么分？

凭头衔，就回到了第 34 章的那堵墙。不凭头衔，就只能凭这三样东西。一个不肯写下死法、不肯请对手进暗室的寡头，在外人眼里，和民科是一样的。这不是外人冤枉了他，是他自己没给外人分辨的办法。

◆ 十、寡头自己能先做的一件事

在尾篇的新机制出来之前，寡头自己能先做一件事：把自己的机器，从回声换成记账员。

办法不难。每一次和机器讨论那个猜想，都让它记下三样。这一次新加的推论是什么？这一次被放弃的说法是什么？这一次还没有任何外部数据碰过的地方是什么？

一个月后，回头看第三样那一栏。如果它越来越长，说明体系在长，凭据没长。如果它一直不变短，说明你一直在楼上加层，没有一次下到地基去。

机器顺着你，这件事改不了。可你可以让它顺着你记账。一本诚实的账，比一百页漂亮的推论，更能让外人分得出你是谁。

这件事也有一个限度：账是你自己记的，外人未必信。所以它只是第一步。第二步，还是要把那一页「怎样算错」交出去。

◆ 十一、放回日子里

回到林舟。

他是第 27 章说的「好腿」，题目做得又快又好。第 31 章那一页猜想，他到底没在组会上放出来：这种病也许不在某一个靶点上，而在几种蛋白随时间慢慢错开的节奏上。

没放出来，念头也没停。他每天晚上十点以后，把这个念头一点一点讲给机器听。三个月下来，机器陪他攒出了一份四十页的文档，有模型，有推论，有十几个「看起来支持」的老案例。

上周五晚上，他把文档翻到最后一页，觉得它已经成了。他甚至想好了题目。

这一章想交给他一件小事，明天就能做：

在那份文档的第一页，加一页纸，标题就写「怎样算错」。写三条：哪一个实验的哪一个结果出来，这个念头就作废。写完存一个带日期的版本。

然后，把这一页，连同方法，发给组外那位最不服他的老同学，请他用自己的数据跑一遍。

林舟可能会发现，那一页最难写。写不出来，说明那四十页还只是一个很整齐的猜想。写出来了，被老同学跑翻了，他省下了三年。没被跑翻，他第一次有了一个外人进过暗室的凭据。

> **一套体系最可信的地方，是它写明了自己会在哪里死。**

带 走 的 话

机器把猜想打造得越完整，越像被证实；事先写下怎样算错、请对手来跑，才分得开寡头和妄想。

小　账

看见了什么，漏看了什么？看见了寡头的稀缺与危险是同一件事；漏看了一套能检验典范本身的公共机制，那要等第 45 章来补。

是在改进，还是在出题？寡头在出题，可他要是只让自己的机器来判卷，出的题就和自问自答分不开。

想 一 想

你最近一次被机器说服，是因为它给了你证据，还是因为它顺着你说得很整齐？如果要你给自己最得意的一个想法写一条「怎样算错」，你写得出来吗？你身边最不服你的那个人，你敢把方法交给他跑吗？

◆

38

第 38 章

墙外开花墙内香

◆ 一、走廊里的那张海报

林豆豆九岁，读小学四年级。她们学校的走廊里，贴着一排科学家的海报。有一天晚饭时，她问林舟：「爸爸，屠呦呦奶奶是什么时候变成科学家的？」

林舟想了想，说：「她五十多年前就在做青蒿素了。」

「那为什么海报上写的是 2015 年？」

林舟一时没接上。海报上那行字写的是：2015 年诺贝尔生理学或医学奖。她做成那件事的年份，海报上没写。

孩子的问题常常比大人的问题准。她问的不是屠呦呦什么时候做成了事，是她什么时候「变成」了科学家。在一张走廊海报的眼里，变成的那一天，是斯德哥尔摩打电话来的那一天。

这一章，说的就是这张海报上缺掉的那几十年。

◆ 二、三个名字，同一条路

先看材料。中国人最熟的几位大科学家，有一条很相似的路。

杨振宁和李政道，1957 年因为宇称不守恒拿了诺贝尔物理学奖。那项工作是 1956 年做的，两个人那时都在美国：杨振宁在普林斯顿高等研究院，李政道在哥伦比亚大学。实验证实，靠的是同在美国的吴健雄。

陈省身，二十世纪最重要的几何学家之一。他最被人称道的一项工作，是 1940 年代在普林斯顿高等研究院给高斯—博内公式找到一个内在的证明〔核：大意〕。他后来长期在美国任教，1980 年代拿了沃尔夫数学奖。

这三个人的出身、性格、学问都不一样。可他们有一件事相同：最顶尖的工作，是在海外的土上做成的；在国内成为家喻户晓的名字，是在国外先认了之后。

杨振宁 1971 年第一次回国访问，那时他拿诺贝尔奖已经十四年。陈省身 1985 年在南开大学创办数学研究所，那时他已经七十多岁。他们回来，是带着一张早已被世界盖过章的证书回来的。

◆ 三、屠呦呦：先墙外，后墙内

屠呦呦的路，和前三位反过来，可落在了同一个终点。

她没有出国。她的工作，是在北京、在中国自己的一项任务里做成的。1960 年代末，她加入一项代号「523」的抗疟研究任务。据公开的记载，她从葛洪《肘后备急方》里「青蒿一握，以水二升渍，绞取汁」那一句得到启发，改用低温提取。1971 年前后，她得到了对疟原虫有效的提取物，此后又分离出青蒿素〔核：大意〕。

这项工作救了很多人，据世界卫生组织的口径，以青蒿素为基础的疗法成了治疗疟疾的标准方案〔核：大意〕。

可在此后的三十多年里，屠呦呦在国内的名气，远不如这项工作的分量。她没有博士学位，没有留洋背景，也没有院士头衔，于是有了一个说法：「三无科学家」。

2011 年，她获得美国的拉斯克临床医学研究奖。国内的报道从这一年起多了起来。2015 年，她获得诺贝尔生理学或医学奖。这一回，她的名字进了千家万户，进了林豆豆学校的走廊。此后她获得了 2016 年度国家最高科学技术奖。

请注意这个先后。工作是墙内做的，花是墙内开的。可那股香，是先从墙外飘回来，墙内的人才闻到的。

◆ 四、本书的说法：外部认可成了凭据

先说公道话。

国内并不是没认他们。杨振宁、李政道、陈省身后来在国内受到的尊重，是很少有人能比的；他们帮着建研究所、选派学生出国、请外国学者来讲学，这条路是国内很多人一起铺的。屠呦呦的工作，当年本身就是国家组织的任务；没有那项任务，就没有那几次提取。第 28 章说过的评价改革，也说明体制自己看见了「只认头衔」的问题。

不过公道话说完，那个先后还在。

本书的说法是：国内的认可机制，在最顶上那一层，本身就依赖外部认可作凭据。

这句话要说细。在常规科学那一层，国内有自己完整的一套尺子：论文、基金、奖项、职称、院士。这一套尺子很会量「在已有的题里做得好不好」。可是一到典范那一层，到「这个人出的题本身对不对」，国内的尺子就不太会量了。不会量的时候，最省事的办法，就是看外面有没有人量过。

拉斯克奖、诺贝尔奖、沃尔夫奖，在这里起的不只是奖的作用，还起了「代量」的作用。墙外替墙内量过了，墙内才敢认。

墙内不是不爱花，是墙内没有一把自己敢用的尺子去量一朵没见过的花。

◆ 五、只看一面会怎样

只看一面，会有两种省事的结论。

一种是自责：国内有眼无珠，好人才都要被外国先发现。这话说重了。墙外的尺子也不是专为发现中国人准备的。杨振宁、李政道拿奖那年，美国也有大量

在冷宫里的人；诺贝尔奖委员会也有它的偏好和漏看，很多该拿奖的人一辈子没拿到。墙外的尺子量得出屠呦呦，是因为她的工作已经救了几百万人，结果摆在那里，谁都量得出来。

另一种是自满：反正最后都认了，这不就够了。这话说轻了。屠呦呦等了四十年，这四十年里，同样走着冷门路、没有头衔、也没等来拉斯克奖的人，有多少？他们的名字，不会出现在任何走廊里。

第 35 章说过，「民科」两个字只看身份不看内容。这一章补上另一半：国内的尺子也有一种例外，它会认一个没有身份的人，条件是墙外先认了。

所以，它不是只看身份。它看的是两种身份之一：国内的头衔，或者国外的承认。一个人两样都没有，内容再好，也很难被看见。

◆ 六、今天与机器：墙外的章更难盖了

到了机器时代，这件事会变难，不会变容易。

杨振宁那一代人走的路，起码有一张清楚的地图：去最好的研究所，做最好的题，写进最好的期刊，等最权威的委员会来量。他们的工作再新，也还落在物理学、数学这些成熟行当的典范边上，旧尺子量得到。

第 25 章说过，复合体时代真正稀缺的，是纯粹的典范猜想。一个人加一台机器，出一个新典范，可能不在任何现成的行当里，没有现成的期刊，也没有现成的委员会。墙外替墙内代量这条路，在这里也会断：墙外同样没有一把量它的尺子。

还有一层。机器让论文、报告、体系都变得便宜。墙外的评审面对的稿件多了几倍，留给一个陌生人、一个陌生典范的那几分钟，只会更少。

这意味着，「先墙外，后墙内」这条老路，对下一代的猜想者，会更窄。

◆ 七、一个完整的例子：一位虚构的数学老师

讲一个虚构的例子，说清这条路怎样一步步走。

一位县城中学的数学老师，四十多岁，业余琢磨一类组合问题，琢磨了十年。他用机器帮着整理，写出一份八十页的证明。

他先投国内的期刊。编辑部看了一眼作者单位，没有送审。他又写信给两位大学教授，一位没回，一位回了一句「请先发表在正式刊物上」。

后来，他把证明的英文版放到一个公开的预印本网站上。三个月后，国外一位做这类问题的教授在博客里提了一句：这份证明的思路很新，前两节我检查过，是对的。

一周之内，国内有三家媒体来采访他。县教育局请他去做报告。省里一所大学问他愿不愿意去做兼职。

这位老师的证明，在那三个月里一个字也没改。改的只是：墙外有一个人说了一句「是对的」。

这个例子是虚构的，可它的每一步，读者大概都在新闻里见过类似的。

◆ 八、通行证是怎样炼成的

把这个例子拆开，会看见一张通行证的样子。

它不是一张纸，是一个先后：墙外先说一句，墙内才接一句。墙内接的那一句，往往不是「他的证明对不对」，是「国外权威怎么说」。

于是，对一个体制外的人来说，出去就有了第二重用途。第一重是求生存：国内找不到能待的地方，去墙外找一块能待的地方。第二重是取通行证：去墙外，把那一句「是对的」拿到手，再带回来，墙内的门才会开。

这两重用途，下一章细说。这里只先记下一件事：出海，有一半是为了回来。

◆ 九、一张章的两面

墙外的章，也有它的另一面。

一张章能开门，也能关门。墙内的人一旦习惯了等墙外的章，就会慢慢忘掉自己量东西的手艺。一个领域里，凡是墙外没说过的，都先放一放；凡是墙外说过的，都先信一信。久了，墙内的尺子就只剩一种用途：核对墙外那张章是不是真的。

这对常规科学的伤害不大。常规科学的题是公开的，答案是公开的，国内国外量的是同一样东西，借一借也无妨。

可是到了典范那一层，这种习惯会伤得很深。一个新典范，常常恰好是墙外还没想到的东西。墙外没想到，就不会有章；没有章，墙内就不敢认。于是，一个在中文世界里最先冒头的新想法，反而要等墙外也想到了、量过了，才能在墙内被认。等到那一天，它已经不是新的了。

第 12 章说过那个「再」字：先做应用，再做原创，那个「再」从来没排进日程。这里有一个相似的「再」：先等墙外认，再由墙内认。这个「再」倒是一直排在日程上，只是排在别人的日程上。

一个小办法，给做评审、做编辑、做基金管理的人：下一次遇到一份没有任何外部背书的稿子，在看作者单位之前，先读完它的第一节。读完再看单位。这个顺序，换一下就好。

◆ 十、墙外替墙内量过的，是哪一层

再往深处走一步，看墙外替墙内量的，到底是哪一层。

杨振宁、李政道的宇称不守恒，是物理学里一个很大的翻转：大家默认左右对称，他们说在弱相互作用里不一定。可是这个翻转，是落在物理学这门成熟学问里面的。它有公式，有可做的实验，吴健雄做完实验，结果一目了然。墙外量它，用的是物理学自己的尺子。

屠呦呦的青蒿素，量它的尺子更硬：病人好没好，疟原虫死没死。

这几件事有一个共同点：它们都是在一个已有的大框里，出了一道新题，而且这道新题很快就有了能被所有人看见的结果。

第 17 章分过三层：技术改进、架构跃迁、科学典范。墙外替墙内量过的，多半落在第二层和第三层的交界处，而且都是「结果来得快」的那一种。一个要几十年才看得出对错、又不落在任何现成行当里的典范，墙外也不会替墙内量。

所以「墙外开花墙内香」这条路，本来就只对一部分花管用：那种一开就香、香得谁都闻得到的花。

◆ 十一、只用墙外的章，会把人引到哪里

还有一层后果，落在年轻人身上。

一个二十多岁的研究生，看着走廊里的海报长大。他看到的路是清楚的：要被国内认，先被国外认；要被国外认，先进国外的好刊物；要进好刊物，就做好刊物喜欢的题。

于是，最聪明的那一批年轻人，很早就学会了一件事：挑墙外已经在热的题。墙外热的题，有现成的尺子，有现成的同行，章盖得快。

这不能怪他们。这是一条被走廊海报指出来的、最理性的路。

可这条路有一个后果：它把一代人的眼睛，都引向了墙外已经出过的题。第 13 章说过，解题的能力和出题的能力是两样。走廊海报教会了年轻人怎样在墙外的题里做到最好，没教会他们在墙内出一道墙外还没想到的题。

◆ 十二、放回日子里

回到林家的饭桌。

那天晚上，林舟最后是这样回答女儿的：「屠呦呦奶奶早就是科学家了。只是很多人要等外国人说了，才知道。」

林豆豆想了一会儿，又问：「那要是外国人一直不说呢？」

林舟没有答案。这正是这一章想留给读者的问题。

一件明天能做的事：下次你在新闻里看到「某某获国际大奖、国内高度关注」，去查一查：他的那项工作是哪一年做的？在获奖之前，国内有没有人认真量过它？查一次，你就会对那道墙的厚度，有一个自己的数。

> **花开在哪里，和香先飘到哪里，是两件事。**

> 带 走 的 话
>
> **在典范那一层，墙内的尺子常等墙外先盖章；墙外不盖章的那种花，就一直没人去闻。**

> 小　账
>
> **看见了什么，漏看了什么？**看见了先墙外后墙内这条老路；漏看了墙外那把尺子量不了新典范，复合体时代这条路也在变窄。
>
> **是在改进，还是在出题？**杨振宁、屠呦呦都在出题，可他们的题在墙内被认，靠的是墙外先替墙内判了卷。

> 想 一 想
>
> 你所在的行当里，最近一位被国内广泛承认的人，是先被国内认的，还是先被国外认的？如果你自己有一个新想法，你会先拿给谁看，为什么？

◆

39

第 39 章

出海：求生存，还是取通行证

◆ 一、两张机票

宋岚三十五岁，在短剧公司做 AI 制片。这个月，她第二次飞新加坡。公司的海外平台要上一批新剧，她去盯译制和投放。

同一个航班上，她遇见一位五十多岁的先生，带着一箱书。本书叫他老周，他是一个虚构的人物，不对应任何真人。聊起来，老周说自己在内地一所师范学院教了二十年统计。他琢磨出一套「怎样给一堆候选排先后」的理论，这几年在新加坡注册了一家小出版社，书都从那里出。

宋岚随口问：「那您为什么不在国内出？」

老周笑了笑，没有正面答：「在国内，书出得来；可书里的想法，没地方落脚。」

下飞机时，宋岚心里想的是一件小事：同一班飞机，她是去卖东西，他是去找地方。两个人都叫「出海」。

这一章，就说这两种出海。

◆ 二、一句口述

先把一句话摆在桌上。

王德生博士在口述本书的承重命题时，说过一句很短的话，大意是：这种寡头，在国内难发芽，发芽也难长苗，最后不得不出海求生存。

这句话里有三个台阶。

第一个台阶是发芽。第 33 章到第 35 章说过，一个不从机构、不从头衔里长出来的思想，在国内的土里很难冒头。原因有三：思想权威历来不归个人，合法性由机构发放，「民科」两个字先把它归了类。

第二个台阶是长苗。冒了头，还要有水有肥：有学生，有读者，有能出书的地方，有能跑方法的场地。第 36 章说过，一个新典范没有现成的评审，也就没有现成的经费、现成的刊物。苗长不高。

第三个台阶是出海。土里待不住，就换一块土。

本书照实记下这句话，也照实给它加一条注：这是一位当事人的说法，它说出的是一种处境，不是统计。有多少寡头在国内长成了，有多少出海之后也没长成，本书没有数。

◆ 三、出海的两重用途

第 38 章收尾时记下了一件事：出海，有一半是为了回来。这里把它拆成两重。

第一重，求生存。这是最直接的。国内找不到一个能容纳「一个人加一台机器出典范」的位置：大学按论文计分，研究院按项目计分，基金按团队与资历计分。墙外未必有现成的位置，可墙外的位置更杂、更散，有更多的小门。一个人注册一家公司、开一家小出版社、在一所不太看头衔的机构里挂一个名，就能先活下来。

第二重，取通行证。第 38 章说过，国内的尺子在典范那一层，常常借墙外的章作凭据。一个体制外的人去墙外，不只是为了活，也是为了把一张章拿到

手：一个国外书号、一篇国外期刊、一次国外会议的报告、一位国外学者的一句评价。拿到了，再带回来，墙内的门才会多开一条缝。

这两重用途，常常在同一个人身上缠在一起。他自己也未必分得清，哪一天是在求生，哪一天是在取证。

◆ 四、企业出海：先说它做对了什么

说思想出海之前，先看企业出海。这一面中国做得很好，公道要写足。

宋岚所在的那类公司，是一个现成的例子。据中文在线 2026 年 6 月初的投资者关系活动记录，它旗下的海外短剧平台 FlareFlow，截至 2026 年 5 月底注册用户超过 3700 万，上线了 5000 多部剧，其中约一成是 AI 短剧。

再往前看，跨境电商、手机、游戏、短视频应用，过去十多年里，中国企业在海外做成了很多件事。这背后是中国最强的那几样本事：工程、规模、成本、供应链、对用户的快速反应。

企业出海的逻辑很清楚：国内市场卷得太厉害，海外还有空地；同一套做法，搬过去就能卖钱。它要的是市场，不是承认。

这一点很要紧。

◆ 五、同与不同

把两种出海并排放。

相同的地方有三处。都是国内的土不够用了，才往外走；都要换语言、换渠道、换规矩；都得先在海外活下来，才谈得上别的。

不同的地方，更值得看。

企业带出去的，是一套已经做成的东西。短剧的拍法、投放的算法、成本的控制，都是在国内练好了的。出海是把一样成品，卖到一块新地上。

思想带出去的，是一样还没被认的东西。它在国内没有长成，带出去也不是成品，是一棵苗，甚至是一粒种子。出海不是卖，是重新种。

企业出海，成没成功看回款。一个季度就能看见。

思想出海，成没成功看什么？看有没有人读，有没有人用，有没有人肯拿它去撞一块硬石头。这些东西，几年都未必看得见。

还有一处最大的不同：企业出海，挣的钱可以汇回来；思想出海，挣的承认，却不一定汇得回来。这一点留给第 41 章。

企业出海是把做好的烟花运到别处去放；思想出海是把一包还没人认的新火药，换一块地去试。

◆ 六、只看一面会怎样

只看求生存这一面，出海就成了一个悲情故事：土不容人，人只好走。这种写法容易把寡头写成英雄，把留下的人写成庸人。

这不公道。留下来的人里，有很多人在国内的土里，一点一点把题目往前推。第 30 章说过的那种小团队，国内的大学和研究院里就有不少。出海只是一种处境，不是一种品格。

只看取通行证这一面，出海又成了一种算计：出去镀一层金，回来好卖。这也不公道。很多出海的人，起初根本没想回来，只是想找个地方，把想法写完。

两面要一起看。一个出海的思想者，既是在求生，也是在取证；既有不得已，也有算计。这两样缠在一起，才是他真实的处境。

◆ 七、代价之一：海外的土也不是为典范猜想准备的

本书要在这里说一句冷静的话：墙外不是一块现成的地。

西方学术同样有同行评议，同样有机构门槛，同样按论文、引用、经费来量人。一个体制外的个人，投顶级期刊，同样很可能连审都不送；申请基金，同样要看单位、看履历、看过去的发表。第 36 章说的「谁来判一个新典范」，在墙外同样是一个没解开的题；那一章的魏格纳，就是在墙外等了半个世纪。

那么，墙外到底多了什么？

本书的看法是：墙外对体制外个人的标签化，轻一些。没有一个像「民科」那样一下子把人归掉的词；一家小公司、一家小出版社、一个独立研究者的身份，在那里是寻常的。更要紧的是，那里有很多零碎的、不归一个系统管的小门：独立书号、预印本网站、小型基金会、读书会、播客、愿意请陌生人来讲一次的系。

所以，出海换来的，是更多的缝，不是一块现成的地。

缝的意思是：你可以钻进去，可以活下来，可以被一两个人看见。可是缝里不长庄稼。要长庄稼，你还得自己去开地、引水、找人。

◆ 八、一个完整的例子：两条航线

回到那班飞机。

宋岚在新加坡待了五天。她见了本地的译制团队，调了三版海报，看了投放数据。第五天，新剧在 FlareFlow 上线，第一周的数据还不错。她飞回上海，在公司周会上报了一个数。她的出海，是一个清楚的闭合：去、做、收钱、回来。

老周的事，下一章细写。他在新加坡的出版社一年出十几本书，书号是正规的，可读者大多还在中文世界。他去过两次本地大学的讲座，听的人不多。他投过一次英文期刊，审稿人回了一句：这个框架和现有文献的关系没交代清楚。

老周没觉得这算失败。他说，在新加坡，至少没人一开口就问他「你是哪个单位的」。他的书，能出来，能有一个正经的号，能被国内的读者从网上买到。这已经比在国内多了好几条缝。

可是他也承认，五年过去了，他的那套理论，墙外没有一个人真跑过。墙内的读者在读，可读和跑是两回事。

两条航线同一个起点。一条一周就闭合，一条五年还开着口。

◆ 九、为什么偏偏是新加坡这样的地方

再说一句出海的去向。

一个思想者出海，未必去最有名的大学所在的地方。很多人去的，是那些对注册一家小公司、开一家小出版社很宽的地方。新加坡、中国香港这类地方，常被提到。原因不难懂：离中文世界近，用中文的人多，法律规矩清楚，开一家公司的手续短，书号、刊号这类东西能按规矩拿到。

请注意这里挑的是什么。挑的不是最好的同行，不是最大的实验室，是最短的手续和最近的中文。

这本身就说明了一件事：这一类思想出海，要的主要不是墙外的土，是墙外的门。他们没打算在墙外扎根，只打算在墙外落个户。这一点，正好接上第 40 章要说的「两栖」。

企业出海挑地方，看的是市场：哪里人多，哪里愿意付钱。FlareFlow 这样的平台，看的是哪个国家的用户爱看短剧、肯为一集多付几毛钱。思想出海挑地方，看的却是哪里的门矮一点、离家近一点。

一个看市场，一个看门槛。这一处的不同，说出了两种出海要的东西有多么不一样。

◆ 十、短剧出海与思想出海，挣的是哪一种钱

还可以从挣钱这件事上，再看一眼两种出海。

第 6 章说过，资本把 AI 押在回款最快的地方：注意力商品。短剧出海，是把这种回款最快的生意，从国内搬到海外。FlareFlow 上那约一成的 AI 短剧，用的是国内练出来的那套降本增效的本事：同样的预算，拍出更多集；同样的人手，上更多部。

这是中国最擅长的那一层，在海外接着挣钱。它挣得到，本书不贬低它。

思想出海挣不到这种钱。一本书一年卖几千册，一场讲座来几十个人，一个预印本被引用两三次。按回款速度排队，它永远排在最后。

于是有一件很自然的事：资本跟着短剧出海，不跟着思想出海。一个短剧公司出海，后面有投资人、有平台、有一整条产业链；一个思想者出海，后面常常只有他自己和一台机器。

第 7 章问过：注意力挣来的钱和练出的能力，会不会回流到生产力。这里能补一句：会不会回流到出题的那一层？到目前为止，本书没看到短剧出海的钱，流到了任何一个思想出海的人手里。

两种出海挣的是两种钱。一种按季度算，一种按十年算。

◆ 十一、出海的人最常犯的一个错

出海的人里，有一种错很常见：出去之后，换了一套标准来量自己。

在国内，他被人问「你是哪个单位的」，他不服，觉得这是只看身份。到了墙外，他开始拿「我在国外出过书」「我在国外讲过学」来说明自己对。

这和他不服的那一套，是同一件事。都是拿身份来顶替内容，只是把国内的头衔，换成了国外的头衔。

第 37 章说过，寡头和民科在外人眼里分不开，有一份是寡头自己的。出海之后，这一份不会少，只会换个样子。一个拿墙外的小章去顶替「怎样算错」的人，比在国内时更难被分辨，因为国内的外人更难查他在墙外到底做了什么。

出海能给的，是活下来和一张小章。它给不了的，是那三样：事先写下的死法，请来跑的对手，可被推翻的条款。这三样，带不带出去，都得自己做。

◆ 十二、放回日子里

宋岚后来又在一次活动上碰见了老周。她问他：「您这样，算出海成功了吗？」

老周想了想，说：「我活下来了。剩下那一半，还没拿到。」

如果你是一个心里揣着一个不太一样的想法、在国内的土里找不到落脚处的人，这一章想交给你一件明天能做的事：

拿一张纸，左边写「我出去是为了活下来，需要的是什么」，右边写「我出去是为了被看见、被国内认，需要的是什么」。两边分开写。你会发现，左边可能只需要一个注册地、一个书号、一份能养活自己的活；右边需要的是对手、真跑、可被检验的成果。

把两边分开，你就不会把左边做成了，当成右边也做成了。

活下来是一半，被认出来是另一半；两半要分开算。

带 走 的 话

出海有两重：一重是活下来，一重是取一张章带回来；企业出海卖成品，思想出海要重新种。

小　账

看见了什么，漏看了什么？看见了企业与思想出海的同与不同，也看见了海外的缝；漏看了根和养分会被拉开，这留给下一章。

是在改进，还是在出题？企业出海是把已有的题搬到新地上做；思想出海是带着一道新题，去找一块肯让它被出的地。

想 一 想

你认识的出海的人里，谁是去卖东西的，谁是去找地方的？如果你要出海，你分得清自己是在求生，还是在取证吗？

◆

40

第 40 章

两栖：注册地在外，根在中文

◆ 一、一本译坏了的书

先把话说在前面：老周是本书虚构的人物，不对应任何真人，他的经历是为了把道理落到一个具体的人身上拼出来的。

老周五十四岁，在内地一所师范学院教了二十年统计。他琢磨的是一件很小的事：一堆候选摆在面前，怎样排先后，才不被最显眼的那几个带偏。这几年，他在新加坡注册了一家小出版社。上一章，宋岚在飞机上遇见的就是他。

有一年冬天，他做了一件自己后来很后悔的事：把自己最得意的一本中文书，请人译成英文，又按西方学术书的格式改了一遍。

改的人很专业。书里有很多老周从县城菜市场、粮站过秤、老家的赶集里攒来的例子，他删了，说西方读者不需要这些铺垫；书里几处从方言俗语里拆出来的说法，他换成了几篇英文文献的引用，说这样审稿人才认得；老周自己起的几个土名字，也换成了一套标准的术语。

译稿出来，四平八稳。老周读了一遍，觉得这本书他认不出来了。

他后来对朋友说：「它变成了一篇很像样的文献综述。我最想说的那句话，被改没了。」

这一章，就从这本译坏了的书说起。

◆ 二、代价之二：根和养分被拉开

第 39 章说了出海的第一个代价：海外的土也不是为典范猜想准备的。这一章说第二个代价，更隐蔽，也更伤。

一个中国思想者的问题意识，是长在三样东西上的。

一是中文。有些想法，是在中文的词里冒出来的。一个字的几层意思，一句古话的几种读法，一个日常说法里藏着的判断。换一种语言，这些层次会被压平。

二是中国经验。他想的问题，常常是从这片土上的具体事情里长出来的：高考、单位、基金本子、短剧、一个县城的网吧。这些事，在别的语境里要先解释半天，解释完，锋利的那一面已经钝了。

三是中国读者。一个思想不是在书房里长成的，是在和读者的来回里长成的。读者的反驳、追问、误读，都是养分。离开了这群人，就少了一半的回写。

用英文、按西方格式重写一遍，最容易削掉的，恰恰是这三样东西长出来的那一部分。剩下的部分，规整、可译、能被审稿人认出来，可往往也是最没有生命力的部分。

根留在中文里，养分也留在中文里；把树整棵挪到别的土里，活下来的常常只是一截干。

◆ 三、先说过去为什么只能整棵挪

杨振宁那一代人出海，是整棵挪的。

这不是他们选的，是那时候只有这一种挪法。1940 年代、1950 年代，要做最前沿的物理和数学，就得去有加速器、有图书馆、有同行的地方。人要去，工作要在那里做，论文要用那里的语言写，职位要在那里拿。

而且，他们做的是物理和数学。这两门学问的语言，本来就是公式。宇称不守恒在中文里和英文里，是同一组方程。挪的时候，根被拉开的那一部分，相对少。

思想不一样。举三样：一个从乡村祠堂的议事规矩里看出来的判断，一个从《齐民要术》里拆出来的说法，一个从县城菜市场和短剧公司里看出来的规律。它们的根，是长在中文和中国经验里的。整棵挪，挪不动。

所以，过去的思想出海，常常有两个结局：要么人走了，思想留在中文里停了长；要么人走了，思想换了一副西方的样子，在墙外成了一个小学派，墙内的人却不再认得它。

◆ 四、本书的说法：机器时代出海的形态变了

到了机器时代，事情起了变化。

一个人加一台机器，可以在任何地方写作、整理、出版、讲课。书可以在网上卖，讲座可以在网上开，读者可以在微信里、在网站上、在一个读书群里。注册一家公司、拿一组书号、开一个站点，不需要人整年待在那里。

于是，出海可以拆开了。

可以只是机构出海：公司、出版社、基金会注册在墙外，受墙外的规矩管。

可以只是出版出海：书号、版权、刊物在墙外，有一个正规的、墙外认得的身份。

可以只是注册地出海：一个法律上的落脚处。

而思想的根、写作的语言、最主要的读者，可以留在中文世界。

本书把这种形态叫作「两栖」：身份在水里，根在岸上。或者反过来说也行：注册地在外，根在中文。

这是机器时代给出海的一样新东西。过去要么整棵挪，要么不挪；现在可以只挪一个户口。

◆ 五、两栖的好处

两栖的好处，先说三样。

第一样，活得下来。第 39 章说过，墙外有更多的缝。一家注册在外的小出版社，能给一个体制外的人一个正规的出书渠道，不用先过「你是哪个单位的」那一关。

第二样，根不断。书还是用中文写的，菜市场和粮站的例子还在，方言俗语也还在。读者还是那群在中文世界里和他来回的人。养分没被切断。

第三样，有一张墙外认得的身份。第 38 章说过，墙内的尺子常常借墙外的章作凭据。一个墙外的书号、一个墙外的机构名，虽然远不是拉斯克奖，可也是一张小章。它不能证明思想对，可它能让墙内的人愿意多看一眼。

◆ 六、两栖的难处

难处也有三样，一样比一样难。

第一样，两边都不太认。墙内的人说：你的书是在外面出的，不算国内的正式出版物，评职称不算，申请基金不算。墙外的人说：你的书是中文的，读者在中国，我们看不懂，也不在我们的引用库里。两栖的人，常常在两边都是边缘人。

第二样，那张小章容易被看穿，也容易被滥用。一个墙外的注册地，谁都能拿到。有人拿它出很好的书，也有人拿它出很糟的书。外人一看，分不清。第 37 章说过，寡头和民科在外人眼里很难分开。两栖的身份，不会让这件事变容易，有时候还会变难：外人会想，他为什么要绕到外面去出？

第三样，最难：两栖是一种姿势，不是一种结果。注册地在外，根在中文，这只是让人活下来、让根不断。至于思想本身对不对，有没有人真跑过，能不能回到国内被认，两栖一样也没解决。

所以，两栖不是答案，是一个更好的出发点。

◆ 七、只看一面会怎样

只看好处，两栖会被说成一条聪明的捷径：在外面挂个名，就能绕过国内的墙。这不对。墙还在，绕过去的只是出书那一道门，合法性那一道门一扇也没开。

只看难处，两栖又会被说成一种尴尬：两头不靠，不伦不类。这也不对。过去那两个结局——人走了思想停长，或者人走了思想变了样——两栖都避开了。它让一个人在活下来的同时，不必先把自己的根削掉。

两面合起来看：两栖保住的是根，没保住的是承认。

◆ 八、一个完整的例子：老周的第二本书

回到老周。

那本译坏了的英文书，他最后没有出。他换了一种做法。

第二本书，他用中文写，开头照样是县城集市上一个卖鸡蛋的老太太怎样给一筐蛋分大小，照样用他自己起的土名字。书在他新加坡的出版社出，有一个正规的书号，在网上卖，读者大多在国内。

英文那一边，他只做了一件小事：把全书最承重的三个判断，各写成一篇两三千字的英文短文，放到一个公开的预印本网站上。每一篇的最后，他写了一段「怎样算错」，列出三个可以被检验的预言。

他不再追求把整本书变成西方的样子。他只要求墙外的人能看懂那三个判断，能拿去跑。

两年下来，那三篇短文，有一篇被一位国外的研究者引用了。引用的话，大意是：这个判断我在自己的数据上试了一下，有一半成立。

老周把这句话贴在书房的墙上。他说，这是他出海这些年，拿到的第一张真的章。

◆ 九、两栖的人要守的三条

两栖既然只是一个出发点，那么站在这个出发点上的人，要守几样东西，才不至于把两栖做成一个空壳。本书数出三条。

第一条，墙外的身份只用来活，不用来唬。注册地在外、书号在外，是为了让书出得来、人活得下去。不要把它说成一种比国内更高的认证。拿一个墙外的小社号去唬墙内的读者，是在把第 38 章说的那种「借墙外的章」，做成一门小生意。

第二条，语言留在中文，判断要能译。全书用中文写，可最承重的几个判断，要写得能被另一种语言的人读懂、拿去跑。老周后来做的，正是这一条：书不译，判断译。

第三条，读者留在中文，对手要请到墙外。读者是土，可读者多半是认你的人。对手要从不认你的地方请：墙外的同行，墙内另一派的人，你的读者群之外最挑剔的那一个。第 37 章说过，没有对手进过暗室的体系，只是一份很整齐的猜想。两栖的人，最容易只有读者，没有对手。

守住这三条，两栖才是两栖。守不住，它就只是换了个地址的自说自话。

◆ 十、今天与机器：翻译变便宜之后

机器还在另一处改了出海：翻译。

过去，一本中文书要变成英文书，要请一位好译者，花一两年，花一大笔钱。所以作者只能挑一本最要紧的去译，而且一旦译了，就得照着西方的样子改，否则这笔钱白花了。老周那本译坏了的书，就是这样来的。

今天，机器能在几天里把一本书译完，译得还不坏。这件事看起来是好消息，可它带来一个新难处。

翻译便宜了，「整本挪」的诱惑反而更大。很多人会想：既然几天就能译完，那就整本译过去，按西方的格式排一排，投出去。于是根被拉开的事情，不是变少了，而是变多了，变快了。

本书的看法是：翻译便宜了以后，难的不再是译，而是挑。挑出哪几句是能被墙外跑的判断，哪几页是只能长在中文里的故事。挑错了，译得再快，也只是把一截干运到了墙外。

机器能替你译，不能替你挑。挑，是两栖的人自己要做的那一份判断。

◆ 十一、只挪户口，也是一种诚实

两栖还有一层意思，值得单独说。

一个人只挪户口、不挪根，等于公开承认了一件事：我的思想是长在中文里的，我的读者在中国，我不假装自己是一个西方学者。

这是一种诚实。

过去，很多出海的人，觉得要被墙外认，就得把自己改得像墙外的人：换术语，换引文，换问题。改到最后，他既不像墙外的人，也不再是墙内的人。

两栖的人不这样改。他说：我的书就是中文的书，我的问题就是从中国的土里长出来的。我在墙外落户，是因为墙外的门矮一点；我不在墙外假扮，是因为假扮了，我最想说的那句话就没了。

这种诚实，也是一种对外人的交代。外人看见他没有假扮，就少一分「绕道镀金」的怀疑，多一分去读内容的耐心。

◆ 十二、放回日子里

如果你也在考虑出海，这一章想交给你一个小办法。

拿出你最得意的一份东西，先问自己一句：它最有生命力的那一部分，长在哪里？是长在中文的哪个词里、哪件中国的事里、哪一群读者的追问里？

把那一部分圈出来。

然后再决定：哪些可以挪出去，哪些必须留下来。可以挪的，是身份、书号、注册地，也可以是几个能被墙外跑一遍的判断。必须留下的，是那个圈出来的部分。

挪的时候，别连根挪。

> **能挪出去的是户口，挪不出去的是根。**

带 走 的 话

机器时代出海可以只挪户口，把根和读者留在中文；两栖保住了根，却保不住承认。

小　账

看见了什么，漏看了什么？看见了出海的新形态与它的好处难处；漏看了两栖之后，思想怎样回到国内被认，这是下一章的事。

是在改进，还是在出题？两栖本身是在改进出海的做法；它让出题的人不必先削掉自己的题，才能活下来。

想 一 想

你最想说的那句话，换成另一种语言还在吗？如果你的工作只能挪一部分出去，你会挪哪一部分，留哪一部分？

◆

41

第 41 章

回流的通道

◆ 一、一个周日下午的读书会

上海的一个周日下午，一家小书店的二楼，坐了十几个人。林舟也在。

他是被一位同事拉来的。读书会读的，正是老周那本在新加坡出的中文书。来的人五花八门：两个中学老师，一个做芯片验证的工程师，一个研究生，一个退休的会计，还有几个和林舟一样做研究的。

讨论到一半，那位工程师说：「书里第五章那个判断，我在我们组的验证流程上试了一下。有一半能用。」

屋里静了一下。大家都转过头去看他。

林舟后来说，那一刻他第一次觉得，这本书「回来」了。不是因为它在墙外出过，是因为有一个他不认识的人，在他熟悉的土上，拿它跑了一遍。

这一章，说的就是这一种回来。

◆ 二、真正要问的问题

第 38 章到第 40 章，一直在说出海：为什么要出，出去有什么用，付出什么代价，有什么新形态。

到这里，要把问题换一个问法。

真正要问的，不是出不出海，是出海之后，能不能回流。

一个思想出了海，活下来了，保住了根，拿到了一张墙外的小章。可它要是永远回不到长出它的那片土上，被那里的人认、用、跑，那它对中国的 AI for Science 来说，就和没有一样。它成了墙外的一个小学派，墙内照旧是那一套筛子。

所以出海是手段，回流才是事情本身。

◆ 三、杨振宁那一代的回流通道

先看过去的回流是怎样走的。

杨振宁那一代人的回流，走的是一条又宽又硬的通道：诺贝尔奖。

这条通道的好处，是不需要任何人再判。一个拿了诺贝尔奖的中国人回国，没有哪一级机构会问他「你是哪个单位的」。奖本身就是最高的合法性。他说的话，有人听；他办的研究所，有人给钱；他推荐的学生，有人收。

据公开的记载，杨振宁 1970 年代起多次回国访问，帮着推动中美学术交流；李政道在 1979 年前后发起中美联合招考物理研究生的项目，前后送了一大批学生去美国读博〔核：大意〕；陈省身晚年在南开办数学研究所。这些都是实实在在的回流：人回来了，人脉回来了，一代学生的路回来了。

请注意这条通道的形状。它是一个外部权威发给的、一次性的、几乎不可质疑的凭据。墙内不必自己量，墙外已经替它量到了顶。

◆ 四、本书的说法：复合体拿不到那级别的章

到了复合体时代，这条通道对新的思想寡头，几乎是关着的。

原因有三。

一是行当。第 25 章说过，复合体里最稀缺的那种人，出的是纯粹的典范猜想。新的典范，往往不在任何一个成熟行当的正中，诺贝尔奖、菲尔兹奖、图灵奖这些奖，都是按已有的行当发的。

二是时间。这些奖常常要等几十年。屠呦呦从提取出有效物到拿拉斯克奖，等了四十年。机器让写作、推演、检验都变快了，可大奖的节奏没变快。一个猜想者等不起四十年，土也等不起。

三是凭据的样子。一个人加一台机器出的典范，从头到尾，功底那一部分是机器做的。第 26 章说过，剥掉功底之后，留给人的只剩猜想和判断。可现有的大奖，奖的恰恰是猜想加功底长在一起的那种人。

所以，复合体时代的思想寡头，大概率拿不到那级别的认证。

回流不能再等一张从天上掉下来的大章，只能靠一条一条从地上铺回来的小路。

◆ 五、回流的五条小路

那么，小路有哪些？本书数出五条。

一是读者。一个思想在中文世界里有一群真读的人。不是转发的人，是读完、记笔记、在读书会上争的人。读者不发证书，可他们是土。

二是学生。有人愿意跟着这个思想学，把它当成自己做事的起点，甚至拿它去考、去写论文、去做项目。学生让一个思想从一个人身上，长到第二代人身上。

三是用的人。有人在自己的工作里用它：一个工程师拿它改流程，一个老师拿它改课，一个医生拿它看病历。用的人不一定说得清它的理论，可他们知道它管不管用。

四是可被检验的成果。思想要落到几个具体的、能被量的东西上：一个预言、一个模型、一组数据、一个工具。第 37 章说过，要事先写下「怎样算错」。写下了，成果才能被检验；被检验过，才是凭据。

五是能在国内被真跑一遍的方法。这一条最要紧。不是在墙外跑过，是在墙内跑过：在中国的实验室、工厂、学校、医院里，有一个不是作者本人、也不是他门徒的人，按方法跑了一遍。跑出了结果，不管成还是败。

前三条是人，后两条是事。人让思想有土，事让思想有凭据。五条缺一不可，可最后一条最硬。

◆ 六、只看一面会怎样

只看前三条，回流会被说成一件「圈粉」的事：有了读者、学生、用的人，就算回来了。

这不对。读者多，不等于思想对。第 37 章说过，门徒也是一种回声。一群人围着一个思想说话，可以完全不碰任何一块硬石头。只有读者没有检验，回流回来的是一种信仰，不是一种科学。

只看后两条，回流又会被说成一件纯技术的事：只要成果过硬，自然有人认。

这也不对。第 34、35 章说过，国内的尺子先看身份。一个体制外的人拿出一项过硬的成果，很可能连被量的机会都没有：没有期刊送审，没有实验室愿意跑。成果要被量，先得有人愿意量；愿意量的人，往往是先读过、先用过的那群人。

所以前三条和后两条，是缠在一起长的。读者和用的人，替成果找到第一批肯量它的人；成果被量过，又替读者和用的人撑腰，让他们在自己的单位里敢说一句「我在用这个」。

◆ 七、今天与机器：回流也可以变快

机器在这里，帮得上忙。

过去，一个方法要被别人跑一遍，难在两处：方法写不清，别人照着做不出来；数据和工具不公开，别人想跑也没东西跑。

今天，一个人加一台机器，可以把方法写成一份能直接执行的东西：一段代码，一个公开的数据集，一个任何人都能在自己电脑上点开就跑的流程。第 44 章会说到公共基础设施，这里只先说一句：方法越容易被跑，回流越不必等许可。

一个县城的老师、一个工厂的工程师、一个研究生，不必先问作者是哪个单位的，就能先跑一遍。跑出了结果，他就有了自己的凭据，不再需要借谁的章。

这是复合体时代回流和杨振宁时代最大的不同：那时候回流靠一张很大的章，现在可以靠很多次很小的跑。

◆ 八、一个完整的例子：一条回流有多难

回到林舟。

那次读书会之后，林舟做了一件事。他把老周书里和药物筛选有关的一个判断，改写成了一个小实验。做法很简单：拿组里过去两年筛过的分子，按书里的说法重新排一遍，看排在前面的，是不是后来湿实验里更好的那一批。

他用了两个周末，在机器的帮助下把流程搭好。结果出来：有一点相关，不算强，可也不是零。

接下来，难处一个接一个地来了。

他想在组会上讲。组里的副研究员问：这本书是哪里出的？作者是哪个单位的？听说是新加坡的一家小出版社，大家的兴趣就淡了一半。

他想写成一篇短文投期刊。合作者提醒他：引一本墙外小社的中文书作理论来源，审稿人会怎么看？

他想申请一笔小经费，把实验做大。本子上「理论依据」那一栏，他写了三遍都不满意。

最后，那个小实验的结果，只是作为一份内部报告，放进了组里的共享盘。

林舟后来说：出海的那一步，老周一个人就能走；回流的这一步，要一路上有十几个人点头。

一条回流，比十条出海难。

◆ 九、一张回流的账单

把林舟那一次的难处排成一张账单，看得更清楚。

关口	要谁点头	林舟遇到的那一问
组会	组里的同事	这本书是哪里出的
期刊	编辑与审稿人	理论来源是不是正式文献
经费	基金评审	理论依据那一栏怎么写
推广	别的实验室	有没有大单位做过

每一个关口问的，都不是「你跑出了什么」，是「这东西从哪里来」。这正是第 34 章说的那堵墙：先看身份，再看内容；身份过不了，内容就不看了。

所以回流的难，不在成果不够硬。林舟那个小实验，结果不算强，可它是真跑出来的。难在每一个关口，都要先回答一个和成果无关的问题。

账单里只有一处，没有问来路：组里读博士二年级的小唐。他没问这份报告是谁写的，只问了一句：「我能不能拿我的数据跑一遍？」

回流的第一块砖，常常就是这样一个不先问来路的人。

◆ 十、跑失败了也算回流

还有一件容易被漏看的事：跑失败了，也算回流。

很多人以为，回流就是思想被证实、被认可、被请回国内。所以一个方法在国内被人跑了一遍，结果不成，大家就说：你看，回流失败了。

本书不这样看。一个方法被一个不相干的人在国内认真跑了一遍，跑出了「不成」，这件事本身就是回流。它说明三样东西已经通了：方法写清了，别人照着做得出来；有人肯花力气跑；结果被记了下来。

这三样，正是第 37 章说的那张「死法」得以兑现的条件。一个思想能被跑翻，才说明它不是自说自话。

而且，跑翻的地方，常常就是下一个猜想冒头的地方。十九世纪初，英国医生普劳特猜过，各种元素的原子量都是氢的整数倍。化学家一称，氯是三十五点五，猜想被跑翻了。一百年后，人们找到了同位素，才明白那个「点五」是几种氯混在一起称出来的；跑翻的那一处，恰好成了新题的入口〔核：大意〕。

所以，回流通道上要记的，不只是成功的那几次，还有失败的那几次。一条只记成功的通道，其实还是一面回声墙。

◆ 十一、读书会、共享盘、预印本：三种小土

回流的五条小路，落到日子里，常常就是几样很小的东西。

一个读书会。十几个人，每月一次，读一本来路「不够正规」的书，读完有人回去试一试。它不发证书，可它是一小块能让新想法先落脚的土。

一个共享盘。一个组里、一个单位里，有一个地方，放着那些没发表、没立项、只是有人跑过的内部报告。林舟那份报告，就是在那里被小唐看见的。

一个预印本网站。一个人能把一个判断、一份方法、一组结果先放上去，带着日期，谁都能看，谁都能跑。老周那三篇英文短文，就是在那里被引用的。

这三样，都不归任何一个评审委员会管。它们不替任何东西担保，只做一件事：让一样东西先被看见，先被跑。

第 43 章会说大的飞地。这里说的，是小的、谁都能开的飞地。

◆ 十二、放回日子里

不过，那份内部报告也不是白放的。

半年后，小唐为了找毕业论文的对照数据，翻共享盘时看到了这份报告。他觉得有意思，拿自己手上那批分子又跑了一遍。结果和林舟的差不多。

这是第二次真跑。跑的人不是作者，不是林舟，也不是读书会的人。他没读过老周的书，只读过那份报告的方法那一页。

这一章想交给读者一件明天就能做的事：

如果你读到一个让你心动、可来路「不够正规」的想法，别急着转发，也别急着否定。挑它最具体的一个判断，在你自己的工作里，找一个周末跑一遍。跑完，把结果，不管成败，写成一页纸，放在一个别人能看见的地方。

你可能就是那条回流通道上的第一块砖。

出海一个人就能走，回流要一路有人点头。

带 走 的 话

等不来一张从天上掉下的大章，回流就靠很多次很小的真跑；跑翻了，也算回来了。

小　账

看见了什么，漏看了什么？看见了回流比出海难，也看见了五条小路；漏看了谁来给这些小跑记账，这要等尾篇的新机制。

是在改进，还是在出题？五条小路本身不出题；它们让一道在墙外出的新题，回到墙内被人真做一遍。

想 一 想

你最近一次在自己的工作里，真跑过一个「不够正规」的想法，是什么时候？如果你跑出了结果，你会把它放在哪里，让谁看见？

◆

42

第 42 章

本篇总账：土的账

◆ 一、老林的那块地

老林六十六岁，年轻时在县城国营厂当会计。退休以后，他在上海郊区租了一小块地，种菜。

有一年，他从老家带回一包种子，是一种老品种的豆子。他种下去，出苗了，可苗长得又细又黄。邻居看了一眼，说：「你这豆子不对，种错了，拔了种我给你的吧。」

老林没拔。他换了一处向阳的角落，把土翻深，加了一点从老家带来的草木灰。第二季，那豆子长得比谁家的都好。

他后来跟林舟说：「豆子没错，是那块土不认它。」

这一篇写了九章，说的就是这句话。

◆ 二、这一篇在拆什么

先把这一篇的路重走一遍。

下篇一说的是筛子：复合体里的那个人，比天才更稀缺，而中国的筛子最会筛掉他。可筛子之外，还有一层更深的东西——就算有一个人漏过了筛子，他落在什么样的土里？

这一篇就是在拆那片土。

第 33 章从「我注六经」进去，说中国思想的权威历来不归个人：一个人说新话，要借经典的嘴说；王阳明、李贽是例外，有时还被当成祸患。第 34 章说合法性由机构和头衔发放。第 35 章说「民科」两个字只检验身份不检验内容。第 36 章问：一个新典范出来，到底谁来判？

第 37 章转过身来对寡头自己：一个人加一台机器，会把自己的猜想养得越来越像真的。第 38 章说墙外开花墙内香。第 39 章说出海有两重用途。第 40 章说机器时代出海可以只挪户口。第 41 章说真正难的是回流。

前四章说土怎样不认人。后五章说人怎样在土里、在土外找活路，以及他自己也会怎样骗了自己。

◆ 三、看见与漏看

把九章的看见与漏看并排放。

章	看见了	漏看了
第 33 章　我注六经	思想权威历来不归个人，新话要借经典的嘴	例外者怎样活下来、靠什么撑过去
第 34 章　头衔发放的合法性	机构与头衔替内容担保，量的是身份	头衔之外，能替一个新思想担保的还有什么
第 35 章　「民科」两个字	一个标签先归类，再也不看内容	标签贴得顺手，也有寡头自己不给分辨办法的一份
第 36 章　谁来判一个新典范	新典范没有现成的评审，旧的评审正是它要撞的墙	判一个典范的新机制长什么样
第 37 章　寡头的自我证实	一人加一机，打造越完整越像被证实；要事先写死法、请对手来跑	这三样怎样成为公共的、有人记账的规矩
第 38 章　墙外开花墙内香	国内的尺子在典范那一层借墙外的章作凭据	墙外的尺子也量不了新典范，这条老路在变窄
第 39 章　出海：求生存，还是取通行证	出海有两重；企业卖成品，思想要重新种；海外给的是缝不是地	根与养分被拉开

章	看见了	漏看了
第 40 章　两栖：注册地在外，根在中文	机器时代可以只挪户口，根和读者留在中文	两栖保住了根，没保住承认
第 41 章　回流的通道	回流比出海难；五条小路，最硬的是在国内真跑一遍	谁来给这些小跑记账、给它们合法性

这张表竖着看，左边一列是一片土的样子：权威不归个人，合法性归头衔，标签代替内容，新典范没人判，外部的章当凭据。右边一列，几乎每一格都指向同一个空白：一套能直接量内容、量典范本身的机制，这片土上没有。

◆ 四、先把公道还给这片土

说总判断之前，先说公道话。

这片土长出过很多东西。前三篇说过：工程能力、规模、成本、应用生态；盘古气象上了 Nature（第 21 章）；DeepSeek 在算力受限下把工程做到了极致（第 14 章）；中国的开放权重模型在 OpenRouter 上的调用量超过了美国。这些都是这片土长出来的。

这片土也不是不认个人。屠呦呦后来被认了，杨振宁、李政道、陈省身后来被认了。第 28 章说过的那一轮评价改革，说明这片土自己也看见了只认头衔的问题。

它的问题不在不肥，在它只为一种庄稼施肥。它很会养常规科学这种庄稼：有标准答案，有同行能对，有指标能量。它不太会养一种还没有名字的庄稼。

老林的豆子没错，那块土也不坏。它只是认不出一样它没见过的东西。

◆ 五、本篇总判断

把九章合在一起，压成一句话：

AI 拆掉了资源的墙，留下了合法性的墙；寡头在中国不是长不出来，是长出来也被归类掉。

这句话分两半。

前一半说墙。资源的墙，是算力、数据、文献、助手、实验室。过去一个体制外的人想做前沿研究，这些东西一样也拿不到。今天，一个人加一台机器，能拿到的东西远比过去多：机器替他读文献、推公式、写代码、搭模型。资源的墙，被拆掉了一大半。

可合法性的墙还在。谁说的话算数，谁的书算正式出版，谁的成果值得被量，谁有资格说「这是一个新典范」——这些，机器一样也没拆。

后一半说人。这一篇没有说「中国长不出思想寡头」。这句话太重，也不对：第 33 章说过的王阳明、李贽，第 38 章说过的屠呦呦，都是这片土长出来的。本书说的是另一件事：长出来了，会被归类掉。归到「民科」，归到「非正式出版」，归到「没有单位」，归到「等外国先说」。一归，就不再被量了。

◆ 六、寡头那一半的账

还有一笔账，要记在寡头自己身上。

第 37 章说过，寡头被归类，不全是土的错。一个不写下「怎样算错」、不请对手来跑、不留可被推翻条款的寡头，在外人眼里和民科是分不开的。外人用标签代替内容，有一部分是因为寡头没有给出能量内容的办法。

所以这一篇的总账，两边都要记。土这一边：合法性归头衔，不量内容。人这一边：只靠自己和自己的机器判自己，不给外人分辨的办法。

两边一起，才把一个可能的典范，变成了一堆没人量的整齐文档。

◆ 七、出海是绕，不是解

这一篇后半，花了四章说出海。这里要说清它在总账里的位置。

出海是一种绕法。绕过了出书的门，绕过了「你是哪个单位的」那一问，绕到了更多的缝里。两栖让绕的人不必削掉自己的根。

可绕不是解。合法性的墙，在墙外也有，只是矮一些、缝多一些。绕出去的思想要回来，还得过那堵墙。第 41 章说过，一条回流比十条出海难，难就难在要一路有人点头。

所以，出海这件事，在这本书里不是出路，是症状。很多人不得不出海，说明这片土上缺一样东西。

◆ 八、指向尾篇：土能不能改

那么，土能不能改?

尾篇要回答的就是这个问题。它会说到几样东西。

第 43 章说一块免于指标的飞地：贝尔实验室、普林斯顿高等研究院、HHMI「资助人而不是项目」。土里留一块不计分的地。

第 44 章说一个猜想者、一台机器、一座公共平台，代替大团队加机器。让第 41 章说的「真跑一遍」，不必等许可。

第 45 章说给典范评审一个新机制：推翻条件事先登记、公开真跑、长周期分账。把第 37 章那三样东西，从寡头自己的良心，变成一套有人记账的公共规矩。

第 46 章说教育：不是把已有的材料雕好，是让本来没有的材料发生。

这一篇看见的，是墙。尾篇要试的，是在墙上开几扇门。

◆ 九、这一篇自认最弱的两处

总账要记全，也要记下自己的弱处。

第一处，这一篇讲的寡头，证据多半是史实和当事人的口述，没有统计。有多少思想寡头在国内长成了，多少被归类掉，多少出了海，多少回了流，本书没有数，也拿不到数。第 39 章那句口述，是一种处境的说法，不是一张表。如果有人做出一份像样的统计，显示体制外的典范猜想在国内被认真量过的比例并不低，这一篇的总判断就要收窄。

第二处，这一篇用了几个虚构的人物：老周、县城的数学老师。他们只负责让道理落到一个具体的晚上，不承担证据。读者要把他们和杨振宁、屠呦呦这些真实的史实分开看。真实的那几位，走的也不完全是同一条路：屠呦呦没有出国，她走的是「墙内做成、墙外先认」。这一篇把她和杨振宁放在一起，看的是那个先后，不是说他们的处境一样。

记下这两处，是为了第 48 章。那一章，全书要把自己最弱的地方，一处一处摊在桌上。

◆ 十、两张墙，各管什么

最后，把这一篇说的两堵墙，并排看一眼。

资源的墙，管的是「你能不能做」。算力、数据、文献、帮手、实验条件。这堵墙，机器拆得最快。今天一个县城的老师、一个退休的会计、一个在新加坡开小出版社的老周，能拿到的计算和文献，比二十年前一所好大学的实验室还多。

合法性的墙，管的是「你做出来算不算」。谁的话被听，谁的书被引，谁的结果被量，谁的方法被跑。这堵墙，机器一砖也没拆。

过去这两堵墙常常是同一堵：拿不到资源的人，也拿不到合法性；拿到资源的，往往是已经有头衔的人。所以过去很难分清，一个体制外的人长不出来，是缺资源，还是缺合法性。

机器把两堵墙拆分开了。资源的墙倒下以后，剩下的那一堵就露了出来，看得清清楚楚。这一篇的九章，就是在给这一堵墙画像。

画完了，问题也就换了。以前问的是「怎样给体制外的人资源」。今天要问的是「怎样给一个没有头衔的内容，一次被量的机会」。

◆ 十一、给下一篇的三个交接

这一篇往尾篇交三样东西，每一样都对着这一篇的一处漏看。

第一样，交给第 45 章。第 37 章那三样东西，是事先写下怎样算错、请对手来跑、留下可被推翻的条款。它们怎样从一个寡头的自律，变成一套有日期、有记录、谁都能查的公共规矩?

第二样，交给第 44 章：第 41 章那条最硬的小路——在国内真跑一遍——怎样不必等任何人点头。需要一座公共的平台，让数据、工具、方法摆在那里，谁都能跑。

第三样，交给第 43 章和第 46 章：土里能不能留出一块不计分的地，教育能不能从第一天起，让一个孩子不只学会解题，也学会出题。

这三样，都是在问同一件事：合法性那一堵墙上，能不能开出几扇不看身份、只看内容的门。

◆ 十二、放回日子里

回到老林那块地。

那一季豆子收了以后，老林把种子留了一小包，送给了那位让他拔掉的邻居。邻居种在自己的地里，第一年又长得又细又黄。第二年，邻居照着老林的办法，翻深了土，加了草木灰。豆子长好了。

邻居说：「原来不是豆子的事。」

这一篇想交给读者一件明天能做的事：

想一想你所在的单位、行当或者圈子，最近一次把一个新东西「归类掉」，用的是什么词。是「不正规」，是「没单位」，是「外行」，还是「等等再说」？把那个词写下来。

然后问自己：那一次，有没有人真去量过它的内容？

> **豆子没错，是那块土不认它；可土是能翻的。**

带 走 的 话

> **这片土很会养有标准答案的庄稼；遇上没见过的那一种，它先归类，然后就不再量了。**

小　账

看见了什么，漏看了什么？看见了土怎样不认人、人怎样绕出去又难回来、寡头怎样自己骗自己；漏看了怎样改土，留给尾篇。

是在改进，还是在出题？这一篇本身在出题——它问的不是怎样把论文写得更好，是这片土能不能认一道没见过的新题。

想 一 想

在你身边，最近一次被「归类掉」的新想法是什么？如果让你来翻一翻那块土，你会先翻哪一处？

◆



尾　篇

改土：还能做什么

导　语

这一篇六章，不再诊断，改问还能做什么。第 43 章写一块免于指标的飞地：贝尔实验室、普林斯顿高等研究院、HHMI，三处共同的一点是什么，中国能不能有。第 44 章写一种新的结构：一个猜想者、一台 AI、一座公共平台，以及林舟要是真去做，第一年会撞上什么。第 45 章给典范评审试几条新机制，每条都写清它防什么、会怎样被钻空子。第 46 章回到源头：教育不是把已有的材料雕好，是让本来没有的材料发生。第 47 章请你自查站在哪一层。第 48 章把本书自己最弱的几处摊开。

这一篇的反差是：方法可以写，许可可以改，人要从第一天起换一种长法。前两样，几年就能动；最后一样，要一代人。

43

第 43 章

一块免于指标的飞地

◆ 一、一个没有周报的下午

周四晚上十点，林舟还在填一张表。

表是院里发的，叫「季度产出登记」。一栏论文，一栏专利，一栏在研项目，一栏「阶段性成果」。他的组十二个人，每个人一行，每行都要有东西。

他一行一行往下填。小顾投了一篇，小陈有个专利在走流程，新来的博士后还在调模型，他替她写了「完成数据清洗」。

填到最后一行，他停了一下。

那一行是他自己。这一季，他真正想了很久的只有一件事：他们筛分子用的那套打分办法，会不会从一开始就问错了题。这个念头没有论文，没有专利，也没有「阶段性」。它只是一个念头。

他在那一格里写了「组织申报国家项目一项」，存盘，关电脑。

这一格，他每个季度都这样填。不是不诚实，是表上没有一栏能装下一个念头。

你也许没填过这样的表，可一定见过它的亲戚：周报、月度考核、绩效面谈。它们问的都是：这段时间，你交出了什么？

这一章要问的是反过来的事：有没有一块地方，让少数人在很长一段时间里，不用回答这个问题？

◆ 二、走廊很长的那栋楼

先看三处真有过的地方。

第一处是贝尔实验室。它属于美国电话电报公司，本来是一家企业的研究部门。1947 年，那里的巴丁、布拉顿和肖克利做出了晶体管；1948 年，那里的香农发表了信息论的奠基论文。今天你手里的手机，里头有上百亿个晶体管，跑的是香农那套算账的办法。

据后来的记载，贝尔实验室的楼特意修了很长的走廊，让不同组的人吃饭、走路时撞在一起。这个故事有人觉得被讲得太美了。可不管细节怎样，有一件事是清楚的：那里有一批人，可以花几年时间弄一个没人下订单的问题。

先把公道说足。贝尔实验室不是一座象牙塔。它背后是一家几乎独家经营全国电话网的公司，钱来得稳；它的研究也离电话网不远——晶体管要替代的，正是电话交换里又大又热的真空管。

所以它的飞地，是长在应用现场旁边的飞地。这一点，第 18 章讲 Transformer 出自一个翻译组时说过：架构的跃迁，常常来自应用现场里一块免于产品指标的地方。

◆ 三、一座不用上课的研究院

第二处是普林斯顿高等研究院，1930 年创办。爱因斯坦 1933 年去了那里，后来哥德尔也在。研究院不招本科生，教授不必上课，不必申请项目，也不必按年交论文。

创办人弗莱克斯纳写过一篇文章，题目大意是「无用知识的用处」〔核：大意〕。他的意思是：很多后来改变世界的东西，最初都没人知道它有什么用。

这里也要先打一个预防针。高等研究院不是每个人都在那里做出了大事。费曼就说过，大意是：一个人要是没有课可上、没有学生来问，很容易坐在那里干等灵感，什么也想不出来〔核：大意〕。

这句话很重要。它说的是：免于指标，不等于自动长出东西。飞地是一个条件，不是一台机器。

一块飞地要是只给安静，不给碰撞——没有学生来问，没有对手来驳——它很容易变成一间隔音很好的空屋子。

◆ 四、资助一个人，而不是一个项目

第三处是霍华德·休斯医学研究所，常简称 HHMI。它的做法在生物医学界很有名：资助人，而不是项目。

一般的科研基金是这样的：你先写好要做什么、怎么做、几年交出什么，评审通过了，钱跟着项目走。HHMI 反过来，它挑人，挑中以后给一笔较长周期的钱，让这个人自己决定做什么，过几年再看他总体上走到了哪里。

有一项经济学研究（阿祖莱等人，2011 年），把 HHMI 资助的人和同期拿普通项目资金、水平相近的人比了比。结论的大意是：前一批人做出了更多被大量引用的工作，也做出了更多没人理的工作。

这个「也」字要记住。资助人，换来的不是更稳的成功，是更大的起伏。要大的成功，就得同时接受大的失败。

◆ 五、三处共同的一点

三处地方，一处是企业，一处是独立研究院，一处是医学基金会。钱的来路不同，管法也不同。

它们共同的，不是钱多，也不是天才扎堆。

它们共同的一点是：让少数人在一段很长的时间里，不对短期指标负责。

请注意「少数」和「短期」这两个词。

不是让所有人都不考核，那会变成大锅饭。也不是永远不看结果，HHMI 过几年还是要看的。是让少数人，在他最需要安静的那几年，不用每个季度都交一张表。

还要看一眼它们共同缺的东西。三处都没有一张按季度更新的产出表，也都没有把方向交给一个委员会去投票。前者是不催，后者是不磨。不催，人才敢把时间花在一个说不清的问题上；不磨，那个说不清的问题才不会在一次次开会里被削成一个说得清的小题目。

放回林舟那张表上：最后一行，他自己那一格，就是这块地方该在的位置。

◆ 六、只看一面会怎样

只看一面，有两种。

一种是浪漫化。把贝尔实验室讲成一个只要放手就能出诺奖的神话，回头就骂今天的体制不放手。这样讲漏掉了两样东西：贝尔实验室身后那门稳定的生意，和离它不远的那张电话网。飞地不是凭空悬着的，它的土是应用现场。

另一种是只看风险。一说「不考核」，就想到养闲人。这种担心有道理，费曼的话就是证据。可只看这一面，结论就成了「所有人都要按季度交表」，于是那最后一格永远只能填「组织申报项目」。

两种都只看了一面。合在一起看，问题就从「要不要飞地」，变成了「飞地放在哪里、多大、谁来选人」。

◆ 七、中国能不能有

先把公道还回来。中国这些年已经在试。

有一些新型研究机构，给研究者较长期、较稳定的经费，不按论文数考核，主张让人「坐得住冷板凳」。第 28 章说过，体制自己也在改评价的办法，想让论文数不再是唯一的尺子。这说明看见问题的，不只是批评的人。

本书不去编这些机构的名字和数字。只说常识：试，已经在试；难的是下面三件。

第一件，放在哪里。经验是放在应用现场旁边，而不是远离一切的山上。一个做 AI 制药的研究院、一家做大模型的公司，都可以划出一小块。离现场近，猜出来的东西才有地方去试；离指标远，猜的人才敢猜。

第二件，多大。宁小不宁大。几个人到十几个人，钱够用就行。一大笔钱进来，就会跟着一大张表进来。

小还有一个好处：赔得起。一块飞地五年什么都没长出来，要是它只花了一个研究院百分之一的钱，这笔账谁都扛得住；要是它花了一半，下一任就再也不会有人敢开第二块。

第三件，谁来选人。这是最难的一件。按头衔选，选出来的还是常规科学里最好的那几位，第 28 章说的三道筛子又转了一圈。怎样换一种选法，留给第 45 章。

◆ 八、飞地会怎样死

飞地也会死。说它的好处之前，要把它的几种死法先摆出来，不然改土就成了一句好听的话。

第一种死法，是被表格追回来。一块飞地开张时说好五年不考核，第二年来了新领导，觉得「总得有点东西向上面交代」，于是加了一张半年报，第三年加了论文指标。飞地还挂着牌子，里面已经和隔壁一样了。

第二种死法，是变成养老院。选进去的，都是快退休、功成名就的人。他们值得尊重，可飞地给他们，常常只是一份体面的闲职。真正需要那几年安静的，往往是三十来岁、刚刚有了一个怪念头、还没有头衔的人。

第三种死法，是变成头衔的又一个奖品。进飞地本身成了一种荣誉，写进简历，拿去评下一个职称。于是人人想进，进的人按名气排，飞地成了筛子的又一道关。

三种死法，都是同一件事：旧的那张表，从后门又走了回来。

所以一块飞地能不能活，不看它开张时的规矩写得多好，看它扛不扛得住第二年、第三年那一次次「总得交代点什么」。

◆ 九、今天与机器

飞地在 AI 时代有一个新变化：它可以小得多。

贝尔实验室要一栋楼、几千人、一整套车间。今天，一个人加一台 AI，再接上公共的数据和算力，能做的事比当年一个小组还多。第 22 章说过，常规科学的那一大段，正在交给机器。

钱的门槛降了，可许可的门槛没降。

一个人能做，不等于他被允许做。他的单位要不要他每季度交表？他的课题能不能不写「预期成果」？他三年没有论文，职称怎么算？

所以 AI 时代的飞地，花钱不多，要的是一纸许可。下一章，看这纸许可具体落在哪里。

◆ 十、一个完整的例子

设想一家做 AI 制药的研究院，决定试一块飞地。下面是一个设想，不是哪家机构的真事。

院里一共四百人。院长拿出一笔钱，只够养三个人五年。规矩写在一页纸上：五年里不考论文数，不考专利数，不写年度「预期成果」；每年只交一份不超过两页的「我现在在猜什么」，不打分，只存档。

第一年，院里议论很多。有人说这是养闲人，有人说凭什么是他们三个。院长顶住了，理由只有一句：四百人里留三个人去猜，赔得起。

三个人里，一个人第二年就回了原来的组。他说，不交表的日子让他心慌，他更喜欢有人催。这不是失败，是选人的一次校正：飞地不适合所有好研究者。

另一个人花了三年，猜的方向被自己推翻了。她那几份两页纸，一份比一份短，最后一份只写了一句：「这条路不通，原因如下。」这份「不通」后来被院里两个大组拿去，省掉了他们一年的弯路。

第三个人到第四年，还说不清自己在做什么。

五年下来，账怎样算？

按论文算，这笔钱几乎打了水漂。按另一种账算，院里买到了三样东西：一条被认真走到头的死路，一次关于「谁适合飞地」的校正，还有一个仍在猜的人。

第三样，没人知道值多少。这正是飞地的意思：它买的是一个你事先算不出价的东西。

这个设想里最难的一步，是第一年那句「凭什么是他们三个」。院长要扛得住，规矩要写在纸上，还要有比头衔更好的选人办法。钱最容易。

◆ 十一、放回日子里

回到周四晚上的林舟。

他改不了院里的表。可他手上有一个十二人的组，组里的时间，他能排一小部分。

一个小办法：每周五下午留出三小时，叫「无表时间」。这三小时里做的事不进周报，不问产出，只做一件事——把自己心里那个「会不会问错了题」的念头，写下来，想下去。

他可以先只给自己留，再看组里有没有人想要。

第一个星期五，他坐在办公室里，三小时只写了半页纸。第二个星期五，写了两页，又划掉了一页半。到第四个星期五，那半页纸变成了一个能说给别人听的问题。

明天能试：你手上要是也有一点能排的时间，划出一块，跟身边的人说清楚，这一块不交表。

飞地不必大，也不必远，它只要一个许可：让一个人有一段时间，不用回答「你交出了什么」。

带 走 的 话

让少数人在几年里不对短期指标负责，换来的不是更稳的成功，是更大的起伏。

小　账

看见了什么，漏看了什么？看见了飞地共同的那一点；容易漏看飞地身后的土——稳定的钱，和不远处的应用现场。

是在改进，还是在出题？表格管的是改进的那一段；飞地是给出题的人留的，它本身不出题，只留一块不被追问的地。

想 一 想

你身边有没有一块「不交表」的时间？要是没有，你最想把哪一个念头放进去？

◆

44

第 44 章

一个猜想者、一台 AI、一座公共平台

◆ 一、一张画不出来的组织图

你一定见过一张组织图。

最上面一个框，写着首席科学家；下面三四个框，是课题组长；再往下是一排排小框，博士后、博士生、工程师、技术员。框越往下越多，像一棵倒过来的树。

申请大项目的时候，这张图是本子的一部分。框越多，越显得「有队伍」；层越齐，越显得「能打仗」。

现在请你试着画另一张图。

最上面只有一个框，写一个人的名字。旁边一个框，写「AI」。下面没有小框了，只有一条线，连到一个很大的方块，方块上写「公共平台：数据、实验设施、算力」。

这张图，你大概没在哪本申请书里见过。它太空了，空得像没有队伍。

拿给评审看，第一个问题多半是：「你一个人，干得过来吗？」十年前，这个问题问得对。今天，它要换一种问法。

这一章要说的是：AI 时代做典范猜想，合理的样子，也许正是这张空的图。

◆ 二、团队原本的价值在哪里

先把公道还给大团队。

大团队不是坏东西。第 30 章那项研究的结论，大意是大团队发展科学，小团队颠覆科学。「发展」两个字，是实实在在的功劳。

一个新典范出来以后，要有人把它铺开：建数据库，做几千次实验，写软件，把一个猜想变成一套能用的系统。这一段，需要很多人、很多手、很长的时间。

一套新的蛋白结构预测方法出来以后，要有人去测、去比、去补数据库，要有人把它接进药厂的流程。这些活，一个人干不完。

中国在这一段上极强。第 21 章说过盘古气象，第 14 章说过 DeepSeek 在算力受限下的工程。这些都是大团队的组织能力，是中国做对、做成的地方。

所以问题不在大团队本身。问题在：它原本最擅长的那一段，正在交给 AI。

◆ 三、AI 拿走了哪一段

第 22 章说过，常规科学正在交给机器。

放到组织图上看，那一排排小框做的事——查文献、清数据、调参数、跑对照、写初稿、画图——正是 AI 最先接手的。一个博士生一个月的活，模型一个晚上能跑完一大半。

这样一来，大团队的倒树还在，可树的下半截在变空。

这时候有两种走法。

一种是「大团队＋AI」：树不变，每个小框里塞一台 AI，让每个人跑得更快。论文更多，项目更大，组织图更满。

另一种是「一个猜想者＋AI＋公共基础设施」：树拆掉，只留顶上那个出题的人，让 AI 去接他的常规科学，让公共平台去接他的数据和实验。

第一种是在改进。第二种是在给出题留位置。

两种走法，花的钱可能差不多，买到的东西完全不同。第一种买到的是更多同样的东西，第二种买到的是一个可能不一样的东西。

请注意，这两种并不互相排斥。一家研究院可以绝大部分照第一种走，只拿出很小一块照第二种走。难的是那一小块，在今天的体系里连一个名字都没有。

◆ 四、本书的说法

AI 时代做 0 到 1，合理的结构也许是「一个典范猜想者＋AI＋公共基础设施」，而不是「一个大团队＋AI」。

这句话里有三样东西，少一样都不成。

猜想者，是第 25 章说的复合体里的那个人。他不需要常规功底样样最好，那一部分交给 AI；他需要的是敢猜一个新典范，并且能判断 AI 打造出来的东西对不对。

AI，是他的「腿」，或者更准确地说，是替他跑的车。第 22 章说过车替腿那个比方。猜想者不再需要一群人替他跑，他需要一台能替他跑的车。

公共基础设施，是这辆车跑的路。没有路，车开不出车库。

◆ 五、大团队的家底可以变成原料

这里有一个容易漏看的地方：大团队这些年积下的家底，正好可以变成那条路。

中国这些年建了许多大型实验平台，攒下了海量的数据，也铺了不少算力中心。这些东西，现在大多锁在各个大组、各个机构的门里，谁的项目谁用。

要是其中一部分能变成公共原料，像水电一样，一个人按规矩申请就能用，事情就变了。一个没有大组的猜想者，也能接上同样的数据、同样的实验、同样的算力。

用大白话说：大团队不必解散。它们可以接着做系统打造，同时把自己的一部分家底，变成别人出题时可以用的土。

这是一种回写：发展科学的那一头，反过来养颠覆科学的那一头。

◆ 六、公共平台要长什么样

说「公共平台」，很容易说空。不妨用一件家常事来比：图书馆。

你去一家市图书馆借书，不用先证明你属于哪个课题组，不用写预期成果，也不用有职称。你办一张卡，按规矩借，按时还，弄坏了照价赔。它不问你借这本书去干什么。

科研的公共平台，要是能像图书馆，就有三样要写清。

第一样，按机时、按数据量排，不按课题组排。一个人和一个大组，在排队这件事上一样，只是用得多的多付，用得少的少付。

第二样，申请看的是要做的事，不看挂靠的单位。一个人写清楚他要跑什么、要多少机时、数据怎么用、结果怎样公开，就够了。

第三样，用了要回写。在平台上跑出来的数据和结果，按规矩留一份给平台，下一个人可以接着用。这样，公共原料会越用越厚，而不是越用越薄。

这三样都不新。天文学里的大型望远镜，很多早就按「看申请、不看单位」分配观测时间〔待核〕。难的是把它挪到每一个需要数据和实验的地方。

◆ 七、猜想者从哪里来

还有一个问题绕不开：这个猜想者，从哪里来？

第 25 章说过，他比天才更稀缺。他不一定是大组里最会写本子的那个人，也不一定是论文最多的那个人。他可能在一个大组里做了十年常规科学，心里一直搁着一个放不进本子的念头；也可能在企业里、在海外、在一个没人注意的小学校里。

这种人现在大多散着。他们之间没有一个名单，也不该有一个名单——一有名单，就又成了一种头衔。

所以要做的不是去找他们，是把门开着。门开着，按规矩谁都能进，进来的人自己会露出来。第 45 章的盲评和推翻条件，就是这扇门的门槛：它不问你是谁，只问你猜的是什么、怎样算错。

◆ 八、只看一面会怎样

只看一面，也有两种。

一种是「一个人就够了」。这是把猜想者神化。一个人加一台 AI，没有人跟他吵，没有人替他查错，第 37 章说过的那种危险就来了：AI 顺着他的典范打造系统，越完整越像被证实。一个人的结构，必须配一个外部检验的机制，这是第 45 章的事。

另一种是「一个人能干什么」。这是把组织图当成实力本身。框多不等于想得深；第 31 章说过，大团队里，方向常常按资历定，共识会把最尖的那个念头磨平。

两种合在一起看：要的不是一个孤零零的人，是一个人背后接着一条公共的路，前面站着一个能检验他的机制。

◆ 九、在现有体系里，它几乎没有立项的位置

现在回到申请书。

一个人去申请经费，本子上写：「本项目由申请人一人承担，借助 AI 与公共平台完成；不设预期论文数；五年内的目标是检验一个典范猜想是否成立。」

这份本子会遇到什么？

团队一栏几乎是空的，评审会担心「执行力」。预期成果一栏写不出数字，会被认为「目标不明确」。研究基础一栏，要是这个猜想是新的，就没有前期论文可以列。

一句话：在中国现行的经费与评价体系里，这种结构几乎没有立项的位置。

这不是谁故意刁难。现行的表格本来就是为常规科学设计的——第 29 章说过，论文计分遇上 AI，指标经济被推到了极致。一张为大团队画好的表，装不下一张空的组织图。

所以要做的事很具体：给它一个位置。

可以是一个单列的小额通道，专收「一人＋AI＋公共平台」的申请；可以是在现有基金里，允许团队栏写「AI 与公共平台」；也可以是大平台自己开一个窗口，按规矩向单个研究者开放机时和数据。三条里哪一条先动，都行。

◆ 十、今天与机器

这个位置，现在给比十年前给更合算。

十年前，一个人要验证一个新的分子设计思路，得自己搭一个小实验室，招几个人，干三年。今天，结构预测、分子生成、文献综述、代码，都有现成的模型可用；实验那一段，要是有公共平台可以按机时预约，一个人也能把设计送进去跑。

AI 把一个人能做的事，扩到了过去一个组的大小。

可许可那一头还停在过去。体系仍按「一个组」来认人。于是出现了一个奇怪的局面：一个人已经能干一个组的活，却只能以一个组的名义申请，而一个组的名义，又要他先变成那张倒树。

◆ 十一、一个完整的例子：林舟的第一年

设想林舟真的离开大组去做这件事。下面是一个设想，用来把难处摆在桌上。

他猜的是第 43 章那个念头：他们筛分子用的打分办法，也许从一开始就问错了题。他想换一种问法去重新定义「好分子」。这是一个典范级的猜想——它要改的不是哪个分子更好，是「什么算好」。

第一个月，他发现自己没有地方挂名。离开原来的组，他的身份还在研究院，可没有组，就没有经费卡号。他得先找一个愿意收留他的部门。

第三个月，他用 AI 把文献扫了一遍，又让模型帮他搭了一个最小的计算框架。这一段出奇地顺，以前要两个学生干半年的活，他自己两个月干完了。

第五个月，他需要湿实验验证。他去问院里的大平台，平台说机时按课题组排，个人没有排期。他去问外面的公共设施，有的要单位盖章，有的要项目编号。他在这里卡了两个月。

第八个月，年度考核。他交不出论文，只交了一份写得很认真的「猜想与推翻条件」。考核表上没有这一栏，他被记作「待评」。

第九个月，他的一个老学生来看他，问他现在带几个人。他说一个也不带。学生愣了一下，没再问。林舟知道那一下愣是什么意思：在这个行当里，不带人，就等于不算数。

第十个月，有人劝他：「先回组里，挂个名，把这个方向放进大项目的一个子课题。」他知道，一放进去，这个念头就要按子课题的指标去长，长成一篇改进的论文。

第十二个月，他的框架跑出了第一批结果。一半和他猜的一样，一半不一样。他不知道这算成功还是失败，也没有人能告诉他——因为评这个的机制还不存在。

这一年里，最难的从来不是钱，也不是技术。是三件事：没有挂名的地方，没有接得上的公共平台，没有能评这件事的人。

把这一年和第 43 章那块设想的飞地放在一起看，会发现一件事：飞地给的是时间，可时间只解决了三件里的一件。没有平台，一个人有了时间也做不出实验；没有评审，一个人做出了结果也没人能说它算不算数。

所以三样要一起给。只给一样，像只给一个人一块地，不给种子，也不给秤。

◆ 十二、放回日子里

林舟并没有真的离开。那天晚上他把这个设想讲给宋岚听，宋岚问了一句很实在的话：

「你缺的那三样，哪一样是你自己明天就能去要的？」

他想了想，是第二样。院里的大平台有空闲机时，只是没有向个人开放的规矩。

第二天，他真的去问了。平台的负责人想了想，说：「没有这个规矩，可也没有禁止的规矩。你写个申请，我报上去试试。」

明天能试：要是你在一个有平台、有数据的地方工作，去问一句——「有没有可能，留一小块，按规矩向不属于任何大组的人开放？」一句话，不费钱，可能会开一个口子。

大团队可以接着建路，只要路的一头，留给那个没有队伍的人。

带 走 的 话

AI 能替一个人跑完一个组的路，可许可还按一个组来认人；要改的是那张表。

小　账

看见了什么，漏看了什么？看见了团队在系统打造上的真功劳；容易漏看这一段正被 AI 拿走，留下的是没有立项位置的猜想者。

是在改进，还是在出题？「大团队＋AI」把旧题做得更快；「一个人＋AI＋公共平台」给新题留了一个能落地的地方。

想 一 想

你所在的地方，有哪些家底锁在门里？要是开一个口子，第一个能用上它的人会是谁？

◆

45

第 45 章

给典范评审一个新机制

◆ 一、两份没法比的稿子

设想你是一个评审。桌上放着两份申请。

第一份写得很漂亮：问题清楚，方法成熟，前期论文二十篇，团队二十人，预期三年发表高水平论文十篇。你一看就知道，这件事交给他们，大概做得成。

第二份写得有点怪：它说现在大家衡量「好分子」的那把尺子可能本身就有偏差，想换一把尺子。前期论文几乎没有，因为这把尺子还没人用过。团队只有申请人自己。

你该怎么打分？

用现有的评分表，第一份稳赢。可你心里隐约知道，要是第二份是对的，第一份那十篇论文，都是在一把歪尺子上量出来的。

你也可以把两份都给一点钱。可钱只有那么多，表上也只有一栏「综合评分」。最后通过的，十有八九是第一份。

这一章要问的就是：谁来评第二份？怎样评？

◆ 二、同行评议评得了什么

先说同行评议哪里好。

同行评议是现代科学最重要的发明之一。它让一项工作在发表之前，先被懂行的人看一遍，挑错，质疑。这几十年里，它把大量粗糙的、错的、抄的东西挡在了门外。

它擅长的是：在一个给定的典范里，判断这项工作做得好不好。方法对不对，数据够不够，结论有没有夸大。这正是常规科学最需要的那一道关。

中国这些年在同行评议上下了很多功夫：引入国际评审，推行回避制度，压缩人情分。这些改进是真的，让基金评审比二十年前公平了很多。

可是它有一个结构上的盲区：它判不了典范本身。

同行，是在旧典范里做得最好的那一群人。一个新典范，恰恰是要说他们那把尺子有问题。让旧尺子最好的使用者，来评一把新尺子，好比让最会写毛笔字的老先生，去评第一台打字机。他不是坏，他是从自己那块土里看的。

第 31 章说过普朗克那句话，第 36 章说过大陆漂移等了半个世纪。那几十年里，同行评议没有失职，它只是在做它本来就做的事。

放到林舟身上更好懂。他那份「换一把尺子」的申请，要是送到三位评审手里，三位多半是这一行做得最好的人。他们的好，正是在旧尺子上量出来的。他们读到第二页，心里冒出的第一句话多半是：「这把尺子用了二十年，一直好好的。」这句话不坏，它说的是实情。可它恰好挡住了那一份申请要说的事。

◆ 三、寡头也不能自证

那么，让提新典范的人自己说了算？

也不行。第 37 章说过寡头的自我证实：思想集中在一个人加一台 AI 身上，AI 顺着他的典范打造系统，越完整，越像被证实。一个人加一台机器，可以在一个错的典范上，搭出一座非常漂亮的楼。

还有一层。第 35 章说过，「民科」这个标签只检验身份。第 37 章又说过，一个真正的思想寡头和一个「民科」，从外面看常常分不开：都说主流错了，都没有头衔，都很执着。

所以有两头都不行：旧的同行判不了新典范；新典范的提出者也不能自己判。

要的是第三样东西：一个不靠身份、也不靠自信，而靠事先写好的规矩来判的机制。

打个比方。两个孩子下棋吵起来，一个说「我爸是棋协的，我说了算」，一个说「我自己觉得我赢了」。两句都不算数。算数的，是下棋之前两人都认的那本规则书。规则书不认识谁，也不在乎谁有把握，它只管这一步合不合规矩。

典范评审缺的，正是这样一本事先摆在桌上的规则书。

◆ 四、本书的说法

典范评审不能交给身份，只能交给事先写好、谁都能照着跑一遍的条款。

这句话的意思不是不要人，是人不按身份判，按条款判。

人还在，只是换了活。过去评审的活是「我觉得这个好不好」；新机制下，评审的活是「这几条推翻条件写得硬不硬、对手跑得对不对、记录有没有被改过」。

前一种活，要求评审懂旧典范；后一种活，只要求评审守规矩。下面五条，都是这个意思的不同样子。它们都还没有在大范围里试过，是本书提出来请人试的。每条都写清它防什么，也写清它可能怎样被钻空子。

◆ 五、第一条：推翻条件事先登记

做法：提出一个典范猜想的人，在动手之前，先把「什么结果出现，就说明我错了」写下来，交到一个公开的地方存档，带时间戳，不能改。

它防什么：防事后找补。一个猜想要是没有事先写好的推翻条件，做出来什么都能解释成「其实也符合」。AI 顺着打造系统，最容易落进这个坑。事先登记，等于给猜想者自己钉了一根桩。

医学临床试验里，已经有「预先注册」的做法，要求试验开始前公开登记主要终点，防止事后挑好看的指标。本书说的，是把这个办法从「一个试验」挪到「一个典范」上。

可能怎样被钻空子：写一个永远不会出现的推翻条件。比如「要是太阳从西边出来，我就错了」。所以登记要配一条规矩：推翻条件必须是在五年内、用现有手段能检验的；写得太远的，不予登记。

◆ 六、第二条：公开真跑

做法：猜想者把自己的方法、代码、数据和推翻条件全部公开，邀请持相反意见的人，按条款亲手跑一遍。跑出来的结果，无论对谁有利，都公开。

它防什么：防两头。防旧典范的人只凭印象说「不可能」——你说不可能，请你跑。也防新典范的人只靠自己的 AI 说「成了」——你说成了，请对手来跑。

AI 让这一条第一次变得便宜。过去，让对手重复一个实验，要他花一年、几个人。今天，很多计算类的检验，对手用自己的 AI 几天就能跑完。

可能怎样被钻空子：对手故意跑坏，或者猜想者把条款写得只有他自己的环境能跑通。所以要有一个双方都认的第三方环境，条款写清输入输出，谁跑都在同一个地方跑。

◆ 七、第三条：长周期分账

做法：资助一个典范猜想，以十年为期。中途不按论文数考核，只看两样：推翻条件有没有被触发，公开真跑的记录。十年到头，再一次总算账。

它防什么：防短期指标把典范猜想改造成改进论文。第 44 章林舟那一年里，第十个月的那句劝——「先放进大项目的一个子课题」——就是短期指标在往回拉。

可能怎样被钻空子：十年里混日子。所以「不考论文」不等于「不交东西」。每年要交一份短的「我在猜什么、哪条推翻条件离被触发更近了」，公开存档。十年里，这些两页纸连起来，就是一条看得见的路。混日子的人，写不出一条路。

◆ 八、第四条：小额多点

做法：宁可资助一百个猜想，每个给一小笔钱，也不押一个大项目。

它防什么：防评审押错。典范猜想的特点是，事前谁也判不准哪一个对。押一个大项目，押错了就全输；撒一百颗种子，只要有一两颗长出来，账就算得过来。第 43 章说的 HHMI 那个「也」字——更多大成功，也更多没人理的工作——就是这个道理。

可能怎样被钻空子：有人把一个旧项目拆成十个小申请，换个名头拿钱。所以小额多点要配第一条：每一个都必须事先登记推翻条件，而且推翻条件要互不相同。

◆ 九、第五条：按内容盲评

做法：评审只看猜想本身和它的推翻条件，不看申请人的名字、单位、头衔、年龄、学历。

它防什么：防第 34 章说的「头衔发放的合法性」。也防第 35 章说的那种只检验身份、不检验内容的标签。一个没有头衔的人，猜想写得清楚、推翻条件写得硬，就该过；一个有头衔的人，推翻条件写得软，就不该过。

罗尔斯有过「无知之幕」的设想：订规矩的时候，假装不知道自己将来是谁〔核：大意〕。盲评就是给评审戴上这块幕。

盲评还有一个副作用，值得写出来：它会让一些有头衔的人吃亏。一位院士的猜想要是写得含糊，在盲评里会被直接退回。这不是不尊重，是对所有人用同一把尺子。真正有东西的人，不怕这一下。

可能怎样被钻空子：圈内人从文风、从引用里认出是谁。这挡不干净，只能减少。可以配一个补充：盲评先打一轮分，揭开名字以后，名字只许用来查利益冲突，不许改分。

◆ 十、谁来当那个第三方

五条里，有三条都提到了第三方：存档的地方、跑代码的环境、盲评的组织者。这个第三方是谁？

它不能是任何一个旧典范的大组，也不能是猜想者自己的单位。最好是一个只管规矩、不管对错的地方，像一个公证处。它不判你的猜想好不好，只证明你在某年某月某日登记了什么，某人照着跑出了什么。

这样的地方，可以是一个学会，可以是一家大学图书馆，也可以是一个独立的公开存档平台。它的权威，不来自它有多懂，来自它从不改记录。

这一点和第 34 章说的合法性正好反过来。那里，合法性由机构和头衔发放；这里，机构只管记账，合法性由一次次公开的跑来发放。

◆ 十一、只看一面会怎样，今天与机器

五条单拿出来，每一条都有漏洞。只看一条，就会说「这个没用」。

合起来看，它们是互相补的：事先登记防事后找补，公开真跑防自说自话，长周期防被指标拉回，小额多点防押错，盲评防看头衔。一条被钻的空子，常常被另一条堵上。

五条也不是要替掉同行评议。一个猜想过了这五条，进了真跑，后面那一大段系统打造，仍旧要靠同行评议去一项项把关。新机制管的是门口那一下：一个新典范，能不能先被认真看一眼。

AI 在这里是两面的。一面，它让寡头更容易自我证实；另一面，它也让公开真跑、让一百个小额猜想的追踪，第一次变得便宜。同一台机器，既是问题的一半，也是办法的一半。

◆ 十二、一个完整的例子

设想一个基金会，用这五条试了一轮。下面是一个设想，用来看这五条在一起怎么转。

它一年拿出的钱，原本只够一个大项目。这一回，它分成五十份，每份够一个人用 AI 和公共平台干三年。

申请截止那天，收到四百多份。第一轮，盲评，只看猜想和推翻条件。评审发现，一大半申请的推翻条件写得很软：「若结果与预期不符，将调整方法」。这一句不算推翻条件，直接退回。剩下一百多份。

第二轮，评审从里面挑出五十份推翻条件最硬、最能在三年里检验的。揭开名字一看，五十个人里，有十几个没有正高职称，有几个不在任何大学或研究院。

三年以后，五十个猜想里，大约一半的推翻条件被触发了，猜想者自己公开认了错。二十来个还在跑，说不清。有两三个，对手照着跑了一遍，没跑翻。

这两三个，还算不上典范。可它们已经比「二十篇论文、十个人的团队」多了一样东西：一条公开的、被反对者亲手跑过的记录。

◆ 十三、放回日子里

林舟那份「猜想与推翻条件」，在年度考核时被记作「待评」。

他后来做了一件小事：把推翻条件改写得更硬，写成三条，每条都能在两年内、用现有的数据验证。然后他把它发给组外两个一向和他意见相反的同行，请他们照着跑。

其中一个人回信说：「第二条我跑了，你错了。」

林舟那天晚上不太好受。可他也知道，这是这一年里，第一次有人按条款而不是按头衔回答了他。

他回了信，问对方是在什么数据上跑的。两个人来回写了五封信，最后发现，对方跑翻的那一条，是林舟自己写得不够准。他改了第二条的写法，把范围收窄。这一来，猜想小了一点，可也硬了一点。

明天能试：你要是正在提一个新想法，先写下「什么结果出现，就说明我错了」，交给一个不同意你的人。

新典范不怕被反对，怕的是没有一条大家都认的路，让反对和赞成各跑一遍。

带 走 的 话

同行判得了一项工作，判不了一把新尺子；新尺子要靠事先写好、谁都能跑的条款来判。

小　账

看见了什么，漏看了什么？看见了同行评议的真功劳与寡头的自证风险；容易漏看两者其实缺的是同一样东西——事先写好的条款。

是在改进，还是在出题？同行评议守的是改进那一关；这五条，是给出题的人搭的第一座考场。

想 一 想

你最近一个想法，有没有写下过「什么情况下我就错了」？你愿意把它交给谁去跑？

◆

46

第 46 章

教育：让本来没有的材料发生

◆ 一、一个奥数班不教的问题

星期三晚上八点半，林豆豆在做奥数作业。

题目是一道数格子的题：一张方格纸上，有多少个正方形。她用老师教的办法，一行一行数，先数一格的，再数四格的，再数九格的，加起来，对了。

做完以后，她没有翻下一页。她拿着那张纸问宋岚：

「妈妈，为什么正方形一定要横平竖直地放？要是斜着放，算不算？」

宋岚愣了一下。她看了看题，题目没说可以斜着放。

「题目没让你斜着数，按题目来就行。」

「可是斜着的也是正方形啊。」

「那你明天问老师。」

豆豆想了想，又说：「那要是格子不是方的呢？要是三角形的格子，能有正方形吗？」

这一回，宋岚没有马上答。她发现自己也不知道。

你家里也许有过这样的晚上。孩子问了一个题外面的问题，大人第一反应是把她拉回题里。这一章，就从这个晚上讲起。

◆ 二、方法可以写，许可可以改，人呢

前面三章说的，都是大人这一头的事。

飞地是一种许可。「一个人＋AI＋公共平台」是一种结构，也要一纸许可。典范评审的五条，是一套方法。方法可以写成书，许可可以改成政策，几年里就能动。

可这些都有一个前提：得先有那个人。

第 25 章说过，复合体里的那个人比天才更稀缺。第 27 章说过，从高考起，每一道关选的都是常规科学里的「好腿」。一个人要是在九岁那年，就被一次次拉回题里，到了三十八岁，飞地给他了，他也不知道该在里面猜什么。

这不是说三十八岁就晚了。第 43 章的林舟，心里那个念头，就是三十多岁才冒出来的。可要是从小就有人接住他的题外问题，那个念头也许会早十年冒出来，也会硬得多。

所以改土的最后一步，在源头上：这个人，要从第一天起换一种长法。

◆ 三、因材施教不够

很多人听到这里会说：那就因材施教。

先说因材施教对在哪儿。《论语》里，子路和冉有问了同一个问题，孔子给了两个相反的回答，因为两个人的性子不同（《论语·先进》）。这是两千多年前就有的看见：好的教，要看人。

今天的 AI 把这件事做得更彻底了。一个学习系统可以知道你哪道题错、错在哪一步、下一题该出什么，比任何一个老师都细。从这个意义上，AI 时代是因材施教最容易做到的时代。

可是请注意因材施教这四个字里的那个「材」。

木匠拿到一块木头，先看纹理，再看节疤，再决定做桌腿还是做抽屉。他看得越准，做出来的东西越好。可他不会去想：这块木头，能不能长成一种还没有的东西。木头已经砍下来了。

它默认材料已经在那里了：这个孩子是块什么料，老师看清楚，然后照着料去雕。雕得好，就是好教育。

问题是，AI 时代最稀缺的那种人，那种敢猜新典范的人，常常不是一块已经在那里的料。他的那份「敢」，是在一次次被接住、被追问、被允许题外跑的晚上里，长出来的。

◆ 四、本书的说法

教育在 AI 时代的使命，不只是把已有的材料雕好，是让本来没有的那种材料发生。

「发生」这个词，本书用得很小心。它的意思是：那种能力不是藏在孩子身上等着被发现的，是在孩子和大人、和问题、和工具一来一回的碰撞里长出来的。没有那些碰撞，它就不会有。

豆豆问「斜着的算不算」，那一刻什么都还不是。它可能长成一个更深的问题——「正方形到底是什么」——也可能就此熄掉。长还是熄，不在豆豆一个人身上，在那个晚上大人怎么接。

用这本书的话说，这是土的事。同一颗好奇心，落在两种家里，长出来的不是同一个孩子。

◆ 五、大人怎么接

回到星期三晚上。宋岚那句「题目没让你斜着数」，不是错。她累了一天，明天还要早起，作业要按时交。这是大多数家长的第一反应，也是很合理的反应。

可这个晚上，有几种别的接法，代价不大。

一种是不答，反问：「你觉得算不算？」让豆豆自己先说一个理由。她说算，就问「那斜着的你能数出几个」；她说不算，就问「那正方形是什么意思」。

一种是认下自己不知道：「这个我真不知道，我们一起查。」大人说一句「我不知道」，对孩子的意思是：不知道是可以的，不知道之后还可以往下走。

还有一种，是用 AI，但换一种用法。不是让 AI 给答案，是让豆豆把问题讲给 AI 听，再让 AI 只出一个更小的问题回给她。比如「一个三格乘三格的纸上，斜着放的正方形，最小的有多大？」

三种接法都不需要大人比孩子懂。这一点很要紧。很多家长不敢接题外的问题，是怕自己答不上，在孩子面前丢脸。可接住一个问题，和答出一个问题，是两回事。

三种接法有一个共同点：把题外的问题，当成一件值得在今晚接着做的事，而不是一件要尽快拉回来的事。

◆ 六、只看一面会怎样

只看一面，最常见的是两种。

一种是把这个当成贬高考。说中国的应试教育扼杀创造力，要全部推倒。这样讲不公道。中国的基础教育给了几代人扎实的数学、阅读和纪律，第 27 章说

过，那是全世界最好的腿部选拔之一。中国这几十年的工程奇迹，靠的正是这一大批好腿。

另一种是只要好腿，觉得那些题外问题是「不务正业」，「等考上大学再说」。可第 12 章说过那个「再」字：「先做应用，再做原创」，那个「再」四十年没排进日程。「先考上，再好奇」，这个「再」，常常也排不进来。

还有一种，是把「题外问题」又变成一门课。报一个「创新思维班」，交钱，按课时上，结业发证书。这样一来，题外的问题又被装进了一张课表，有了标准答案，有了排名。它又成了题里的东西。

所以本书的意思不是推倒，是在旁边留一块地。

高考那一块照样种，种好。旁边留一小块，专门长题外的问题。这一小块不考，不排名，不进成绩单。

◆ 七、为什么大人总想把孩子拉回题里

这件事要再往深处走一步，不然只是在劝家长「多耐心」。

大人把孩子拉回题里，常常不是因为不爱她，是因为大人自己也是被这样长大的。宋岚小时候问过类似的问题吗？大概问过。那时候的大人怎么接的？大概也是一句「按题目来」。

还有一层是现实的。一个班四十多个孩子，一节课四十分钟，老师要讲完进度。家里也一样：晚上九点，作业还有两样没做，明天七点要起床。

所以「拉回题里」，很多时候是时间逼出来的。不是大人不想接，是没有一块时间能接。这正是本书说的「在旁边留一块地」：先有一块不被进度追着的时间，才谈得上接。

再往下一层，就是第 33 章说的那种土。思想的权威不归个人，问题的边界也不归孩子。题是老师出的，范围是考纲定的，孩子能做的，是在范围里做好。

一代代这样长下来，「按题目来」就不再是一句话，是一种直觉。大人说它的时候，几乎不用想。

所以改这件事，难的不在知道，在大人要先在自己身上，把那个直觉松一松。一个从没被允许跑题的大人，要去允许孩子跑题，起初是会心慌的。这份心慌是正常的。

◆ 八、今天与机器

AI 在这件事上，也是两面的。

一面，它让「拉回题里」变得前所未有地容易。孩子有任何问题，AI 一秒给答案，答案还很完整。问题被答完了，也就熄了。用 AI 写作业的孩子，最容易失去的，不是答案的正确，是那一点「再问下去」的劲。

据报道，已经有一些学校在试，让学生用 AI 做作业时，必须交上和 AI 的对话记录，老师看的是孩子怎么问，而不只是答案对不对〔待核〕。这个做法还在试，效果怎样，没人说得准。可它的方向，是把分数从答案挪到问题上。

另一面，它也让「接着问下去」第一次变得便宜。过去，豆豆问「三角形的格子能不能有正方形」，大人答不上，事情就停了。今天，一台 AI 可以陪她试，可以画给她看，可以一层一层把问题拆小。

同一台机器，用来给答案，它就是筛子的延长；用来陪着问，它就是土。

第 369 号《典范发生学》说，答案、过程、提问都不再是竞争点。对一个九岁的孩子，这句话可以翻成：会不会做题，AI 可以替你；敢不敢问一个题外的题，还是要你自己长。

◆ 九、一个完整的例子

接着写那个星期三。下面是一个可能的样子。

宋岚想了想，没有让豆豆去问老师。她说：「你先猜，算还是不算？」

豆豆说算。宋岚说：「那你画一个斜着的给我看。」

豆豆在格子纸上找了四个点，连起来，一个歪着的正方形。宋岚问：「你怎么知道它是正方形？」豆豆说：「四条边一样长。」宋岚又问：「你怎么知道一样长？」

豆豆卡住了。她数了数，每条边都是「往右一格、往上两格」。她说：「它们走的路一样。」

这一句，是她自己长出来的。她还不知道勾股定理，可她已经摸到了「一样长」背后的那个东西：走法一样，长度就一样。

那天晚上，题目没有多做一道，作业也按时交了。可豆豆的本子背面，多了一页歪着的正方形。

这一页不进成绩单。它是旁边那一小块地上，今晚长出来的第一根苗。

请注意这个晚上里宋岚做的事。她没有教豆豆任何知识，没有讲勾股定理，也没有表扬她「真聪明」。她只做了三件事：先让豆豆猜，再让她画，然后一路问「你怎么知道」。

三件事都不需要大人会答。宋岚自己也不知道三角形格子上能不能有正方形。可这个晚上，豆豆长出了一样题目里没有的东西：她第一次自己追到了一个「为什么」的下面一层。

这就是「让本来没有的材料发生」的样子。那份能力，在这个晚上之前没有；它不是被发现的，是在一问一答里长出来的。

◆ 十、几件明天能试的事

给家长和老师，几件明天就能试、不费钱也不费多少时间的事：

1. 每天留十分钟，叫「题外时间」：孩子做完作业，问一个作业以外的问题。不评对错，只问「你是怎么想到的」。
2. 孩子问了大人答不上的问题，先说一句「我不知道」，再说「我们一起看看」。
3. 用 AI 时，换一个规矩：不许让 AI 直接给答案，只许让 AI 出一个更小的问题。
4. 给孩子一个「问题本」。和错题本放在一起，记的不是做错的题，是想问没问完的题。
5. 老师可以每周留一道「没有标准答案」的题，评分只看理由写得清不清楚。

这五件里，没有一件要和高考作对。它们都长在高考旁边。

还要先打一个预防针：别把这五件变成新的作业。「题外时间」一旦要打卡、要家长群里晒、要每周交一份「问题本」给老师检查，它就又回到题里了。一块地一旦有人量产量，就不再是旁边那块地。

◆ 十一、放回日子里

豆豆睡了以后，宋岚把那页歪着的正方形拍下来，发给林舟。林舟那天在加班，回了一句：「这孩子在问正方形是什么。」

他们俩都没有想把她培养成什么人。他们只是在那个晚上，没有把她拉回题里。

林舟回家以后，看了一眼那页纸。他想起自己九岁的时候，好像也问过类似的问题，具体是什么，已经不记得了。他只记得那天的作业做完了，题也都对了。

这大概就是他们这一代的样子：题都对了，问题忘了。

今晚可以：等孩子问一个题外的问题，忍住，不去答，先问一句「你觉得呢？」

> **一块料可以雕；一种还没有的料，只能在一个个被接住的晚上里长出来。**

> **带 走 的 话**
>
> **因材施教把已有的料雕好；AI 时代还要在旁边留一块地，让本来没有的那种料发生。**

> **小　账**
>
> **看见了什么，漏看了什么？**看见了基础教育给好腿的真功劳；容易漏看那些被拉回题里的晚上，筛掉的正是将来出题的那个人。
>
> **是在改进，还是在出题？**刷题、补差、个性化推送，是把已有的题做得更好；题外时间，是给出题的苗留的地。

> **想 一 想**
>
> 你家里或班上，最近一次有人问了题外的问题，大人是怎么接的？要是重来一次，你会说哪一句？

◆

47

第 47 章

你站在哪一层

◆ 一、读到这里的你

读到这一章，你大概已经在心里给自己归了个类。

也许你在想：这本书说的是科学家，跟我没关系。我在公司上班，我带孩子，我做视频，我在海外卖货。

这一章想说：有关系。

第 17 章分过三层：技术改进、架构跃迁、科学典范。这三层不只长在实验室里。任何一件手上的事——一个产品、一节课、一部短剧、一家公司——都可以问一句：我在哪一层？

三层也不是贵贱。改进那一层不低，典范那一层也不高。它们只是不同的活：一个在已有的题里把事做好，一个在给新题找位置。一个社会两样都要，只是比例要对。

这一章给你一张自查的单子。不打分，不排名，只请你对着自己的事问三句。

◆ 二、随身三问

导读里给过三个随身带的问题。在这里，把它们换成自查的样子：

第一问：你手上的事，是在改进、在架构，还是在典范那一层？

第二问：这一层，AI 会不会接手？接手了，你还剩什么？

第三问：你做的事，要被谁承认才算数？

三个问题都不难，难在照实答。很多人第一问会往高里写，第二问会往乐观里写，第三问干脆空着。照实写，才有用。

第一问看位置，第二问看风向，第三问看土。第三问最容易被跳过，可它决定了你能走多远：一件事要是只能被一个机构、一种头衔承认，它的路就被那个机构、那种头衔画好了。

下面分五类人，每类一段。你可以只读你那一段。

◆ 三、做科研的

你也许像林舟，带一个组，写本子，赶论文。

你每天早上打开电脑，第一件事多半是看邮件：审稿意见、基金通知、合作方催数据。这一天的时间，在打开邮箱的那一刻就被分走了大半。

先说你在哪一层。你每天做的事，多半是在一个给定的典范里解题：换一个靶点，换一个模型，换一组数据，把指标往上提一点。这是改进，是常规科学，做得好很不容易。

再说 AI。第 22 章说过，这一层正在交给机器。你组里的文献综述、数据清洗、参数搜索、初稿撰写，今天已经有一半可以交出去。三五年后，可能更多。

最后，你的工作要被谁承认？大多是同行评议和基金评审。第 45 章说过，这套承认管得好改进，管不了典范。

先说公道：你这一层，正是中国 AI for Science 这些年成绩最好的那一层。论文、专利、开放模型、工程平台，多数是你们这样的组一篇篇、一个个做出来的。

所以给你一个自查：这一年，你有没有一个念头，是现有的评审表放不进去的？要是有，它就是你留给自己那块飞地里该放的东西。要是一个都没有，也许该担心的不是 AI 抢你的活，是你手上已经没有 AI 抢不走的那一块了。

◆ 四、带孩子的

你也许像宋岚，晚上陪孩子做作业，周末送奥数班。

你在哪一层？孩子的作业、补习、刷题，几乎都在改进那一层：把已有的题做快、做对、做全。

AI 会不会接手？已经在接手了。孩子用 AI 写作业，答案一秒出来。家长花钱买的「提分」，有很大一部分 AI 做得比人更便宜。

孩子的事要被谁承认？大多是考试和排名。这一套承认很硬，也很必要，第 46 章说过，不必和它作对。

还有一个容易漏看的地方：你自己。你陪孩子的时候，你在哪一层？要是你每晚做的，是检查对错、督促进度，你也在改进那一层。这没有错，可孩子会照着你的样子长。

自查一句：这个星期，孩子有没有问过一个题外的问题？你怎么接的？要是你总是第一时间把她拉回题里，那她在题外那一层，就还没有被允许长。

◆ 五、管公司的

你也许管一个团队，或者自己开了一家公司。

你在哪一层？多数公司做的事，是在别人定好的题里做到最好最快最便宜。第 13 章说过，这是解题的能力。中国企业在这一层，是全世界最强的一群。

AI 会不会接手？流程、客服、报表、代码、营销文案，正在大面积交出去。你的成本会降，你对手的成本也会降。改进那一层的优势，会被拉平得很快。

你的事要被谁承认？市场，客户，投资人。这是比学术评审更硬的承认，也更短期。

硬的好处是不讲人情，客户不买就是不买；短的坏处是等不起，一个要五年才看得出用处的想法，在季度财报里永远是一笔亏损。

这里还有一个反例要记住：第 9 章说的汉卡、WPS、用友、金蝶，都是企业做出来的，而且是在生产力那一头先做的。企业不只会解题，在条件对的时候，也出过题。

自查一句：你的公司里，有没有一块地方，不对下一个季度的收入负责？第 43 章说的飞地，放在企业里，可能就是你最该问的一个问题。一块很小的飞地，花不了几个人的工资；没有它，你的公司就只剩改进。

◆ 六、做短剧的

你也许像宋岚白天的样子：在一家短剧公司，一个月烧掉几十亿 Token。

先把公道说足。上篇说过，你们是 AI 在中国最早、最大的一批真实用户。一季度每天新增约 1500 部 AI 短剧，生产周期从十几天压到几天〔待核〕，这是实打实的工程能力和组织能力。

你在哪一层？大多是改进：同样的剧，更便宜、更快、更好看。

这不丢人。一个行当的改进做到极致，本身就是一种本事；你们一年里把生产周期压到原来的几分之一，这在全世界都少见。「抽卡」挑画面，就是在一个给定的题里选最优解。

AI 会不会接手？你本身就站在 AI 上。可要注意的是，抽卡这件事，会越来越不需要人。导演那一句「再来一张」，将来可能由模型自己说。

你的事要被谁承认？平台算法，播放量，回款。这是最快的承认，所以也是最窄的。

最快的承认有一个副作用：它会让你以为，自己的本事只值这么多。一个月烧几亿 Token 换来的那些调教经验，在播放量那张表上只有一个数；换一张表，可能是另一个数。

自查一句：你们这几年攒下的那些东西——视频模型调教的经验、多模态数据、算力调度的本事——有没有可能卖给一个做生产力的人？第 7 章问过：先养

后用，还是锁死。回流不是一句口号，是你手上这些家底，能不能有一条通到别处的路。

◆ 七、出海的

你也许在海外：做产品，卖短剧，或者在一所外国大学做研究。

先说公道：出海的中国公司，这几年在很多国家做出了当地人愿意用、用得起的东西。短剧平台、电商、移动支付、手机，都有例子。上篇说过的 FlareFlow，截至 2026 年 5 月底注册用户超过 3700 万。这是在别人的土里扎下根的本事。

你在哪一层？出海的中国公司，多数带出去的是改进那一层的本事：更便宜、更快的产品和内容。第 39 章说过，出海有时是求生存，有时是取通行证。

AI 会不会接手？改进那一层的价格优势，在 AI 时代会被各国同样拿到。只靠便宜，走不远。

你的事要被谁承认？这一问，对你最复杂。第 40 章说过「两栖」：机构、出版、注册地在外，根、语言、读者在中文。你可能同时要被两套土承认，也可能两边都不承认你。

还有一层，是你最容易漏看的：你在外面，看见的土和国内不一样。有些在国内被归类掉的念头，在外面可能先有了一个落脚的地方；有些在外面被当作稀奇的东西，在国内其实早有人在做。你站在两块土的接缝上，这个位置本身就是一种本事。

自查一句：你在外面长出的东西，有没有一条回到中文这一头的路？第 41 章说过，真正难的是回流的通道。你要是已经在外面了，那条路的一头，就在你手上。

◆ 八、教书的和管钱的

还有两类人，上面没单列，这里各补一句。

教书的老师：你每天的课，大多在改进那一层，把一套已有的知识讲清楚、讲牢。AI 会接手讲解和批改的一大块。你的课要被谁承认？考试成绩和听课评价。自查一句：这学期，你有没有留过一道没有标准答案的题？

管钱的人，比如基金会、投资机构、科研管理部门的人：你的事，是决定钱往哪里去。你在哪一层？要是你只按论文、按团队、按头衔投，你就是在改进那一层上加注。AI 会不会接手？评分、查重、比对，会接手得很快。你的决定要被谁承认？大多是上级和审计。自查一句：你手上的钱，有没有一小份，是按第 45 章那几条去投的？

这两类人，手上都握着别人的土。你们那一句自查，影响的不只是自己。一位老师留一道没有标准答案的题，一个班四十个孩子都碰得到；一个管钱的人拿出一小份钱按新规矩投，几十个猜想者就多了一条路。

◆ 九、只看一面会怎样，今天与机器

只看一面，有一种很常见的样子：读完这一章，觉得「我在改进那一层，所以我没用了」。

这不对。改进那一层，是一个社会吃饭的地方。中国这几十年的成就，大部分是在这一层做成的。AI 接手的是其中的重复劳动，不是这一层本身；这一层仍要有人判断、取舍、把东西交到用户手上。

另一种是反过来，觉得「我要去做典范」，于是什么都想推倒。这也不对。典范猜想很少，很冒险，绝大多数人一辈子也不必做。

还有一种是觉得「AI 会接手，那我什么也不用做了」。AI 接手的是活，不是判断。你在哪一层，AI 接手了多少，你还剩什么，这几个问题它不会替你答。

AI 还可以帮你答第二问。把手上的事讲给它听，问它「哪些你现在就能做，哪些三年后能做」。它答得不一定准，可会逼你把自己的活拆开看一遍。

本书请你做的事很小：知道自己在哪一层，知道这一层的风往哪里吹，知道你要被谁承认。然后在你手上，给另一层留一个小口子。

◆ 十、一个完整的例子

宋岚自己做过一次这张单子。

那是一个周日的下午，豆豆在午睡，她坐在阳台上，用的是豆豆的一张草稿纸。

第一问，她写：「改进。我做的是更便宜地拍剧。」

第二问，她写：「会。抽卡会越来越自动，制片的活会少一半。」

第三问，她写：「平台算法。」写完她看了很久，又在后面补了一句：「也许还可以是别的行业的人。」

她补的那一句，是这张单子最有用的地方。她原来以为自己的本事只能被播放量承认；写下来以后，她想到了别的承认方式：她们组调视频模型的经验，也许有别的地方用得上。

林舟也做了一张。他第一问写「改进」，写完划掉，改成「改进，加一个想出题的念头」。第二问他写「会，而且很快」。第三问他写了很久，最后写的是：「同行评议，基金评审，院里的考核表。三处都管不了那个念头。」

两张单子放在一起，夫妻俩发现了一件有意思的事：一个在烧 Token 的人，和一个在做科研的人，卡在同一个地方。他们的本事，都只被一处太窄的地方承认。

这一步，结语里会回来。

◆ 十一、放回日子里

一张纸，三个问题，五分钟。

要是你愿意，可以和身边的一个人交换着写。你写他的，他写你的。别人眼里你在哪一层，常常和你自己以为的不一样；那个不一样，就是这张单子最值钱的地方。

今晚可以：拿出一张纸，写下你手上最重要的一件事，照着三问各写一句。写完不用给谁看。三个月后再写一次，看第三问的答案有没有变。

知道自己站在哪一层，不是为了换一层，是为了在这一层上，给别的层留一道门。

带 走 的 话

先问在哪一层，再问 AI 会不会接手，最后问要被谁承认；第三问最常被跳过，也最要紧。

小　账

看见了什么，漏看了什么？看见了改进那一层是吃饭的地方；容易漏看自己的承认来自一个太窄的地方。

是在改进，还是在出题？五类人手上的事，大多是在已有的题里做得更好；这一章只请你看一眼，旁边有没有一道新题。

想 一 想

你的三问，最难写的是哪一句？你的事，除了现在承认它的那一处，还可能被谁承认？

◆

48

第 48 章

本书自认最弱的几处

◆ 一、一本书最该先告诉你的

你买过一台二手车吗?

好的卖车人，会先把毛病告诉你：这里补过漆，那里换过件，冬天打火要多拧两下。他说了这些，你反而更信他说的别处。

一本书也一样。前面四十七章，本书下了很多判断。这一章把它最弱的几处摊开：哪些数据只有一个来源，哪些例子可能挑过，哪些说法有风险，哪些反例会推翻它。

这不是客气。一本谈「典范要能被检验」的书，要是自己不给推翻条件，就是第 37 章说的那种自我证实。

还有一点要说清：本书不是一份研究报告，它是一个判断，加一路的证据。判断可以比证据走得远一点，这是一本书和一篇论文的不同。可走远的那一段，必须标出来。下面就是那几段。

◆ 二、第一处：那张图的数据

上篇整篇，是从一张流传的图起的。图上说中国的 Token 有很大一块烧在短剧和广告上，美国反过来。

本书拆这张图，拆出了三层。可拆的时候用到的数据，本身也不结实。

55% 那个数，只有单一来源。它出自中文在线 2026 年 6 月初的投资者关系活动记录，是一家短剧公司援引的「行业 Token 调用数据」，原始统计口径没有公开。一家短剧公司说短剧占得最多，你听的时候，要把这个身份记在心里。

58% 那个数，本书照实写了：找不到出处。虎嗅的作者核查后，没有找到可靠信源。所以「美国反过来」这一半，严格说是一张没有底的牌。

Token 密度那一笔——10 秒 1080p 视频约 35 万 Token，文字只有它的千分之一到万分之一——来自一篇技术博文的测算。本书在正文里标了〔待核〕。要是将来有更可靠的测算，算出的差距小得多，第 3 章「计量错觉」那一半的分量就要往下调。

这里还要补一句公道话。中文在线这份记录，至少是一份公开的、有名有姓的材料；网上那张图上的「85%」，连这一层也没有。本书用 55%，是在几个不结实的数里，挑了一个最能查到来路的。

所以上篇的结论，本书只敢下到这一步：资本把 AI 押在回款最快的地方，这一点有多处旁证；至于押了多少，是 55% 还是 40%，本书不敢说死。

◆ 三、第二处：历史例子的幸存者偏差

本书讲了很多成功的故事：贝尔实验室，Transformer 出自一个翻译组，深度学习在冷宫里熬了几十年，HHMI 资助人。

这些故事都有一个毛病：它们是活下来的那几个。

有多少研究飞地，几十年什么也没长出来？有多少在冷宫里坚持的人，坚持的方向就是错的，到死也没人记得？这些没人写。没人写，不等于没有。

反过来的偏差也有。本书举的中国例子，多半是停在改进那一层的；可中国也有过不少在冷门方向上坐了几十年冷板凳的人，只是他们的名字不常被写进这类书里。拿成功的外国例子去比一般的中国例子，是不公平的比法。

所以「飞地出架构」「冷宫出典范」，只能当成一种可能，不能当成一条规律。第 43 章写 HHMI 那个「也」字，就是为了提醒这一点：同样的做法，也带来了更多没人理的工作。

◆ 四、第三处：「寡头」这个词

下篇二用了一个很重的词：思想寡头。

这个词有风险。本书选它，是因为别的词都太软：「领军人物」带着头衔，「创新人才」带着项目，「天才」又把功底和猜想混在一起。「寡头」至少说清了一件事：少，而且不靠多数。

第一层风险是，它容易被读成「要捧出几个大人物」。本书的意思不是这样。寡头说的是：典范猜想这件事，在 AI 时代集中到很少的人加 AI 身上，这些人不靠团队，也不靠共识。它描述的是一种结构，不是一种崇拜。

第二层风险是，它容易被拿去给任何人背书。一个人说主流都错了，再加一句「我就是那种寡头」，这本书就成了他的护身符。第 37 章说过，寡头和「民科」从外面看分不开。本书给的分法只有一条：看他有没有事先写好、谁都能跑的推翻条件（第 45 章）。没有这一条，叫什么都不算数。

第三层风险是，「寡头」读起来像在反对团队、反对共同体。本书没有这个意思。第 44 章说过，团队在系统打造上的功劳是实的。

第四层风险更隐蔽：这个词，会让读者以为问题只在「找人」。仿佛只要找到那几个人，一切就解决了。本书前前后后说的却是土：选拔、评价、许可、教育。人从土里长出来，不是从名单里挑出来的。

◆ 五、第四处：几个可能的反例

下面这几件事，要是出现了，本书的某些判断就要改。

第一个反例：中国也可能从大团队里长出典范。本书说团队出击是反 0 到 1 的，依据是 Wu、Wang、Evans 那项研究，以及中国的选拔与评价方式。可那项研究说的是统计上的倾向，不是铁律。设想中国某个大团队，在没有外部典范可以跟的地方，自己猜出了一个新典范，并且被全世界接受。那就说明团队里也可以留出猜想者的位置，本书下篇一的判断要收窄。

第二个反例：DeepSeek 一类可能走到架构层。本书第 14、15 章说 DeepSeek 把 MoE、MLA 做到极致，是技术改进，不是架构的 0 到 1。这个判断今天站得住，可它不是永远的。一支在约束里磨了几年的队伍，完全可能在某一天，从改进里撞出一个真正的新架构。要是那一天来了，「应用上的碰撞只涌现改进」这句话就要改成「多半只涌现改进」。

第三个反例：AI 未必真会接管常规科学。本书的大前提，接第 369 号《典范发生学》，是 AI 正在接管常规科学。可要是 AI 的进展在某处停住，比如在湿实验、在真实世界的数据上长期达不到可用的水平，情形就不同了。那时候常规科学还要靠人，大团队、好腿、论文计分就仍然有用，中国最强的那一层也就不会那么快贬值。

◆ 六、第五处：陪读人物和设想

还有一处弱，是写法上的。

本书用了林家四个人陪你读，第 39、40 章还有一位老周。他们都是虚构的，出版信息页写明了。他们不承担证据，可他们会让一些判断读起来格外顺：林舟那一年的难处，豆豆那一页歪着的正方形，都写得像真事。

这里要提醒你：顺，不等于对。第 44 章林舟的第一年、第 45 章那个基金会的五十份申请、第 43 章那家研究院的三个人，都是设想。它们用来把难处摆在桌上，不能当作「已经证明了」的例子。

真要检验这几章，得看真实的机构、真实的申请、真实的十年。那些，本书没有，要等别人去做、去记。

◆ 七、判词怎样才算错

导读里说过，判词能被推翻，才配站在第一页。导读写了五条，外加一条来自科学本身的。这里一条一条对上，写得再硬一点。附录六用的是同一套条件。

判词前半句说：中国最擅长的那一层，会随着 AI 进入科学而先贬值。

推翻它的，有两件事，出现任何一件都算。

第一件，就是导读第三条，也是第五节的第三个反例。十年后，查文献、筛分子、调参数、排实验，仍然主要靠人一题一题地做。

第二件，是从结果一头去量：AI 大规模进入科学以后的五到十年里，以中国为主的大团队在改进那一层的工作，价值不降反升。比如被引、被用、被买的份额明显上升，并且这个上升不是来自价格。

判词后半句说：唯一还稀缺的那一层，中国的土里长不出那个人。

推翻它的，有三件，出现任何一件，后半句就判错：

1. 导读第一条：十年之内，从现在的大团队体制里，长出一个被世界公认的科学新典范。
2. 导读第二条：有可靠的研究表明，高考、读博、基金、职称这几道关选出来的人里，做出新典范的比例并不低于别处。
3. 把第一条放宽到十年、放宽到个人：一个在中国本土完成全部教育、没有先拿外国承认的人，提出一个新典范，并且首先在国内被及时认出来。

另有两件，出现了不算判错，可说明土在变，后半句要收窄：经费与评价体系里，出现一个能让「一个猜想者＋AI＋公共平台」立项的通道；一种按内

容、不按身份的典范评审机制在中国大范围运转，并且判出过一个起初被当作「民科」的人是对的。

导读第四条、第五条，冲的是上篇和中篇一的起点，不直接冲判词。上面第一处已经说过 Token 那笔账的弱；回流那一条，要等更长的账。

最后是来自科学本身的那一条，管整句。要是十年以后，AI 自己就能稳定地提出被世界采用的新典范，后半句就失去了意义。要是回头看，最重要的几个新典范绝大多数出自大团队，「复合体里的那个人」这个说法本身就要重新想。那时候，本书和第 369 号《典范发生学》会一起被改。

◆ 八、只看一面会怎样

只看一面，有两种读法会出问题。

一种是只读前四十七章，把判词当结论记下来，拿去证明「中国不行」。附录五第一条说的就是这种误读。这是误读。本书每一处批评之前，都先写了中国做对、做成的地方，也给了它成立的条件。判词说的是一套激励、评价、选拔方式和 AI 之间的错位，不是一个民族的天性。第 9 章计算机那个反例就是证据：顺序是条件决定的。

还有一层要说清：本书的批评，是对着一套激励结构说的，不是对着哪一个人、哪一个机构说的。在这套结构里工作的人，多数比外面的人更早看见了问题，也在各自的位置上试着改。本书写的「筛子」，筛的也包括写这本书的人那一代。

另一种是只读这一章，觉得「书自己都说弱了，那就不用信」。这也是误读。弱的几处，是数据的来源和例子的挑法；书的主干——三层的划分、AI 接手的方向、筛子的方向——并不靠那几个数。

◆ 九、今天与机器

这一章，在 AI 时代还有一层意思。

本书在 AI 协作下写成。出版信息页写明了：判断与承重命题来自王德生博士的口述，文字由 AI 协作完成。第 37 章说过的那种危险——AI 顺着一个人的典范打造系统，越完整越像被证实——这本书自己也躲不开。

所以这一章，是本书给自己做的第 45 章：把推翻条件事先登记在这里，请不同意的人来跑。

AI 也给了这一章一个新的用处。本书的数据、引文和判断，今天的读者可以交给自己的 AI 去查、去对、去反驳，几个小时就能查完一遍。过去，一本书的错要等几年才被书评挑出来；今天，可能出版第二天就有人跑完了。本书欢迎这样的第二天。

◆ 十、一个完整的例子

设想五年后，有一位读者拿着这本书来对账。

她先查第一处：Token 的数据。她找到了更可靠的统计，发现短剧和视频的份额确实很高，可没有 55% 那么高。美国编程占比也确实在上升，只是没有 58%。她把这一笔记下：上篇的方向对，数偏了。她又翻到第 3 章，看那笔 Token 密度的账有没有被更好的测算替换掉。

她再查第四处：反例。她发现有一支国内团队，在某个材料方向上提出了一个新的计算框架。带头的人在国内完成了全部教育。这算不算典范？她拿第七节的条件去比。框架是先被国外几个组采用的，国内后来才跟上；它也还算不上一个完整的新典范。条件只沾上了一角。她记下：后半句开始松动，待看。

她最后查第三处：「寡头」这个词。她发现这几年确实有人拿这本书给自己背书，说自己就是那种被归类掉的寡头。她翻到第 45 章，看他们有没有登记过推翻条件。大多数没有。她记下：这个词被误用了，可书里给的分法还用得上。

这就是本书想要的读法。不是信，也不是不信，是拿着条件去对。

◆ 十一、放回日子里

回到林家。林舟读完这一章，说了一句：「要是我写基金本子，也能有这么一章就好了。」宋岚说：「你下一本就加。」他没说话，可第二天，他真在本子的最后，加了一小段「本项目最可能错的地方」。

本书不怕被推翻。

一本讲典范要能被检验的书，最好的下场，是被检验过；最坏的下场，是没人拿它当真。

明天能试：你手上要是也有一份很有把握的判断——一份报告，一个方案，一句对孩子的断言——在最后加一段「我最弱的几处」，写下三条。你会发现，写完以后，你对剩下的部分更有把握了。

> **一个判断，先把自己的推翻条件交出来，才配让别人去跑。**

> 带 走 的 话
>
> **弱处写在纸面上，不是退让，是请人来跑；不敢写推翻条件的判断，只是信念。**

> 小　账
>
> **看见了什么，漏看了什么？**看见了数据、例子、用词三处的弱；仍可能漏看本书自己没想到的那个反例。
>
> **是在改进，还是在出题？**这一章不出新题，它给本书出的题配了一张考卷，请别人来答。

> 想 一 想
>
> 这本书里，你最想推翻的是哪一句？你手上有什么材料，可以拿去跑一遍？

◆

结语

那个晚上之后

一年以后，又是一个星期三的晚上。

上海下了一场小雨。林家的饭桌上，菜比平时多两个，因为老林从郊区的菜地回来了。饭吃到一半，宋岚说了一件公司里的事。

她们组这一年，把调视频模型的经验整理了出来。不是剧，是那一套怎样让模型稳定地生成清楚、连贯、细节对得上的画面的办法。一家做医疗影像的公司找上门来，想用这套办法去补全和增强一些拍得不清楚的影像，帮医生看得更细。合同不大，可宋岚说起来的时候，语气和说一部爆款剧不一样。

林舟问她，这一单挣得多吗。她说不多，比一部爆款剧少得多。可是对方医院那边的医生回了一封信，说有几张原来看不清的片子，现在能看出一些以前没注意到的地方了。

「那封信，我存下来了。」

「我们一直以为，这些本事只能拿去拍剧。」

这件事是怎么来的，她也说了。去年秋天，她照第 47 章那张单子写过三问，第三问后面补了一句「也许还可以是别的行业的人」。那一句一直搁在她心里。后来一次行业会上，她碰见一个做医疗影像的人，两人聊起各自的难处：一个嫌画面不够稳，一个嫌片子不够清。聊到后来，发现是同一类难处。

她回公司以后，没有马上去谈生意。她先让组里两个人花了一个月，把这几年「抽卡」攒下的经验，写成一份不到二十页的说明。说明里写的是：什么时候

画面会崩，怎样让前后帧对得上，细节在哪一步最容易丢。写完一看，里面一大半，跟剧没有关系。

老林放下筷子，说起他当年的网吧。九十年代末，他在县城开网吧，挣了一点钱，后来都花在了林舟读书上。「我那时候也没想过，那些机器还能干别的。」

本书在上篇问过：注意力挣的钱和练出的能力，会不会回流到生产力？宋岚这一单，是回流的一个小样子。它很小，小到不会进任何统计；可它说明，那条路不是没有，是要有人走一次，才知道它通。

还要说公道话：这一单，是改进那一层的回流。它把烧在烟花上的本事，挪到了一件更有用的事上。这很好，可它还不是下一种火药。回流解决的是钱和能力往哪里去；出题的那个人从哪里来，是另一件事。

林舟没有离开研究院。第 44 章设想的那一年，他没有真去过。

可他做了一件小一点的事。院里今年开了一个口子，允许每个组长拿出一小块时间，做一个不进考核的课题。他报了一个，就是那个念头：他们筛分子用的那把尺子，会不会从一开始就问错了题。

这个课题不为论文。他写了三条推翻条件，存在院里的一个公开目录下。他用 AI 搭了框架，又向院里的平台申请了一点空闲机时。这一年，三条推翻条件里，有一条差一点被触发。

组里的人一开始不太明白。小顾问他，这个课题出了结果，算谁的。林舟说，先不算谁的，先算对不对。小顾想了一会儿，说那他也想报一个。

到年底，组里有三个人各自报了一个这样的小课题。都不大，都写了推翻条件，都没有指望发论文。院里的考核表上，这一栏还是没有名字，只写着「其他」。

他说，这是他这几年里，第一次觉得自己在出题，而不是在答题。

「出得对不对，不知道。」他说，「可至少有一张考卷了。」

豆豆十岁了，上五年级。奥数班还在上，作业也照做。成绩不算顶尖，也不差。

宋岚有时候也会担心：每天那十分钟题外时间，会不会耽误了正经的功课。这一年下来，她没有发现耽误了什么。她发现的是另一件事：豆豆做题的时候，偶尔会停下来，问一句「这道题为什么要这样出」。

她的问题本已经写了大半本。那页歪着的正方形还在第一页。后面有些问题，她自己已经能答了；有些问题，她问过 AI，AI 给了一个更小的问题回来，她又接着往下写。

这一年，她也用 AI 用得更多了。可宋岚定的那条规矩还在：不许让 AI 直接给答案，只许让它出一个更小的问题。有一次她偷偷让 AI 写了一道应用题的答案，被宋岚看见，没挨骂，只被问了一句：「那你自己觉得，它哪一步可能错？」她找了半天，还真找出一步。

那天晚上，她问的是：

「爸爸，你们科学家，是谁规定什么问题算好问题的？」

林舟想了很久。

这个问题，第 17 章到第 24 章用了八章去拆，第 45 章又用一整章去找办法。可一个十岁的孩子，用一句话就问到了。

他没有直接答。他说：「你觉得应该是谁？」

豆豆想了想：「应该是能把问题说清楚、还敢说自己哪里可能错的人。」

林舟笑了。他没有告诉她，这句话差不多就是第 45 章那五条的意思。

宋岚在旁边听着，没有插话。她想起一年前那个晚上，自己那句「题目没让你斜着数」。这一年里，她改掉的不多，只是每次豆豆问题外的问题，她先忍一下，问一句「你觉得呢」。

老林吃完饭，翻了翻豆豆的问题本。他看不太懂那些格子和箭头，只说了一句：「我们那时候，有问题都是记在心里，没地方写。」

这一句很轻，可它把四十年接上了。老林那一代，问题没地方写；林舟这一代，问题有地方写，没地方被承认；豆豆这一代，也许会有一块地方，问题写下来，有人照着跑。

饭后，老林在阳台上抽了一支烟。他问林舟，书上说的那些事，你们院里会不会真改。林舟说不知道，也许会改一点，也许改得很慢。

老林说：「我当年装用友的时候，也没人知道电脑会改成什么样。改着改着，就改了。」

本书从一张图说起，经过四十年的老路，经过三层、筛子和土，最后走到改土。走了这么远，判断还是开头那一句：

AI 越快进入科学，中国最擅长的那一层越先贬值；唯一还稀缺的那一层，中国的土里长不出那个人。

从前言到这里，这句话一个字没改。中间四十八章，有的地方给它添了证据，有的地方给它划了边界，第 48 章还把推翻它的条件一条条写了下来。可它的样子没变。

这句话写在这里，不是为了吓人，也不是为了认命。

前一半，说的是一阵风。风不由谁决定，它往哪里吹，就往哪里吹。中国最擅长的那一层，是几十年里一代一代人用汗水攒下的真本事；AI 来了，这份本事会先贬值，这不是谁的错。

后一半，说的是一片土。土是可以改的。飞地是一处改法。一个猜想者加 AI 加公共平台，是一处改法。按条款而不是按头衔的评审，是一处改法。在高考旁边留一块地，也是一处改法。第 48 章写过：这后一半，有三种情形出现任何一种，就算判错；另有两种情形出现，就要收窄。本书最希望的，就是被它改掉。

这一年里，林家三个人各做了一件小事。宋岚把一份本事卖到了别的行当；林舟给一个念头开了一张考卷；豆豆的问题本多写了大半本。三件事都不大，都不会上新闻。

可这三件事，正好对上本书的三处：回流、出题、源头。改土不是一纸文件就能做完的事，它是很多个这样的晚上，一个一个攒起来的。

最后，回到全书的总比喻。

火药是中国人发明的。据说最早是炼丹的人在找长生药的时候碰出来的〔待核〕。没有人事先知道它会有什么用。它先是一个意外，后来是烟花，再后来才是别的。很多年里，人们用它做了世界上最好看的烟花。拿火药做烟花，没有错；烟花做得好看，是真本事。

问题从来不在拿火药做烟花。问题在那个「再」字：做完烟花，再去发明下一种火药——这件事，一直没有排进日程。

AI 时代，烟花会越来越便宜，越来越好看，越来越不需要人。

这不是坏事。宋岚的公司会接着拍剧，林舟的组会接着筛分子，豆豆会接着上奥数班。烟花会接着放，日子会接着过。本书从没想让谁停下手上的活。

本书只想在烟花的旁边，留一块地。

下一种火药，要有人先猜出来。

那个人现在也许九岁，正在一张格子纸上，画一个歪着的正方形。

◆

参考书目

书目分七类。中文书标出版年份，外文书标初版年份，有中译本的随后注出。新闻与公司披露按发布时间排，统稿时已逐条复核能否找到原文；找不到原文的，在附录一里写明。

另有一条读书小路：只想读三本，先读库恩的《科学革命的结构》，再读第 369 号《典范发生学》，最后读 Wu、Wang、Evans 那篇只有十几页的论文。三样东西放在一起，本书的骨架就有了。

一、科学哲学与科学史（按初版年份）

1. 魏格纳：《海陆的起源》，1915。

2. 波普尔：《科学发现的逻辑》，1934（德文初版），1959（英文版）。

3. 贝尔纳：《科学的社会功能》，1939。

4. 布什：《科学：无尽的前沿》，1945。

5. 普朗克：《科学自传》，1948。

6. 波兰尼：《个人知识》，1958。

7. 汉森：《发现的模式》，1958。

8. 库恩：《科学革命的结构》，1962；中译本，北京大学出版社，2003 年以后多次印行。

9. 普赖斯：《小科学，大科学》，1963。

10. 波兰尼：《默会的维度》，1966。

11. 默顿：《科学社会学》，1973。

12. 费耶阿本德：《反对方法》，1975。

13. 库恩：《必要的张力》，1977。

14. 拉卡托斯：《科学研究纲领方法论》，1978。

15. 夏平、谢弗：《利维坦与空气泵》，1985。

16. 拉图尔：《科学在行动》，1987。

17. 克龙比：《欧洲科学思想的风格》，1994。

18. 斯托克斯：《巴斯德象限》，1997。

19. 奥雷斯克斯：《拒绝大陆漂移》，1999。

20. 齐曼：《真科学》，2000。

二、技术史、创新与研究机构（按初版年份）

21. 熊彼特：《经济发展理论》，1912。

22. 弗莱克斯纳：《无用知识的有用性》，1939（文章）；增订本，2017。

23. 西蒙：《人工科学》，1969。

24. 派斯：《上帝难以捉摸：爱因斯坦的科学与生平》，1982。

25. 里吉斯：《谁占据了爱因斯坦的办公室》（Who Got Einstein's Office?），1987。

26. 克里斯坦森：《创新者的窘境》，1997。

27. 希尔齐克：《PARC 往事》（Dealers of Lightning），1999。

28. 伊萨克森：《爱因斯坦传》，2007。

29. 阿瑟：《技术的本质》，2009。

30. Azoulay, P., Graff Zivin, J. S., Manso, G. Incentives and Creativity: Evidence from the Academic Life Sciences（HHMI 资助人而不是项目的研究）. RAND Journal of Economics, 2011.

31. 格特纳：《贝尔实验室与美国革新大时代》（The Idea Factory），2012。

32. 马祖卡托：《创新型政府》，2013。

33. 伊萨克森：《创新者》，2014。

34. 蒂尔、马斯特斯：《从 0 到 1》，2014。

35. 梅茨：《深度学习革命》（Genius Makers），2021。

三、人工智能与 AI for Science 的关键论文（按发表年份）

36. Rumelhart, D. E., Hinton, G. E., Williams, R. J. Learning Representations by Back-propagating Errors. Nature, 1986.

37. Jacobs, R. A., Jordan, M. I., Nowlan, S. J., Hinton, G. E. Adaptive Mixtures of Local Experts. Neural Computation, 1991.

38. Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. NeurIPS, 2012.

39. LeCun, Y., Bengio, Y., Hinton, G. Deep Learning. Nature, 2015.

40. Shazeer, N. 等. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. ICLR, 2017.

41. Vaswani, A. 等. Attention Is All You Need. NeurIPS, 2017.

42. Senior, A. W. 等. Improved Protein Structure Prediction Using Potentials from Deep Learning. Nature, 2020.

43. Kaplan, J. 等. Scaling Laws for Neural Language Models. 2020.

44. Jumper, J. 等. Highly Accurate Protein Structure Prediction with AlphaFold. Nature, 2021.

45. Hoffmann, J. 等. Training Compute-Optimal Large Language Models. 2022.

46. Bi, K. 等. Accurate Medium-range Global Weather Forecasting with 3D Neural Networks（盘古气象）. Nature, 2023.

47. Lam, R. 等. Learning Skillful Medium-range Global Weather Forecasting（GraphCast）. Science, 2023.

48. Wang, H. 等. Scientific Discovery in the Age of Artificial Intelligence. Nature, 2023.

49. DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. 2024.

50. DeepSeek-AI. DeepSeek-V3 Technical Report. 2024.

51. Abramson, J. 等. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. Nature, 2024.

52. Lu, C. 等. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. 2024.

53. DeepSeek-AI. DeepSeek-R1. 2025.

54. OpenRouter 与 a16z. State of AI: An Empirical 100 Trillion Token Study with OpenRouter. 2025.

四、科学计量与科研评价（按发表年份）

55. Wuchty, S., Jones, B. F., Uzzi, B. The Increasing Dominance of Teams in Production of Knowledge. Science, 2007.

56. Jones, B. F. The Burden of Knowledge and the "Death of the Renaissance Man". Review of Economic Studies, 2009.

57. San Francisco Declaration on Research Assessment (DORA). 2012.

58. Uzzi, B. 等. Atypical Combinations and Scientific Impact. Science, 2013.

59. Hicks, D. 等. The Leiden Manifesto for Research Metrics. Nature, 2015.

60. Fortunato, S. 等. Science of Science. Science, 2018.

61.《关于深化项目评审、人才评价、机构评估改革的意见》，2018。

62. 科技部、教育部、人力资源社会保障部、中国科学院、中国工程院：《关于开展清理「唯论文、唯职称、唯学历、唯奖项」专项行动的通知》，2018 年 10 月。

63. Wu, L., Wang, D., Evans, J. A. Large Teams Develop and Small Teams Disrupt Science and Technology. Nature, 2019.

64. Azoulay, P., Fons-Rosen, C., Graff Zivin, J. S. Does Science Advance One Funeral at a Time? American Economic Review, 2019.

65. Bloom, N. 等. Are Ideas Getting Harder to Find? American Economic Review, 2020.

66.《关于破除科技评价中「唯论文」不良导向的若干措施（试行）》，科技部，2020。

67. Wang, D., Barabási, A.-L. The Science of Science. 2021.

68. Park, M., Leahey, E., Funk, R. J. Papers and Patents Are Becoming Less Disruptive over Time. Nature, 2023.

五、中国的科学、技术与人物（按出版年份）

69. 王阳明：《传习录》，各通行本。

70. 李约瑟：《中国科学技术史》，1954 年起陆续出版。

71. 李贽：《焚书》《续焚书》，中华书局，1975。

72. 陆九渊：《陆九渊集》，中华书局，1980。

73. 钱学森：《论系统工程》，湖南科学技术出版社，1982。

74. 陈省身：《陈省身文选》，科学出版社，1989（〔待核〕）。

75. 王选：《王选谈信息产业》，北京大学出版社，1998（〔待核〕）。

76. 钱学森：《创建系统学》，山西科学技术出版社，2001。

77. 吴晓波：《激荡三十年》，中信出版社，2007—2008。

78. 杨振宁：《曙光集》，三联书店，2008。

79. 屠呦呦：《青蒿及青蒿素类药物》，化学工业出版社，2009。

80. 林军：《沸腾十五年》，中信出版社，2009。

81. 吴晓波：《腾讯传》，浙江大学出版社，2017。

六、新闻、公司披露与数据（按发布时间）

82. CASP14 评估结果，2020 年 12 月。

83. 诺贝尔奖委员会：2024 年物理学奖与化学奖新闻稿，2024 年 10 月。

84. 每日经济新闻（36 氪英文站转载）：OpenRouter 上中国模型调用量首超美国的报道，2026 年 2 月。

85. Frost & Sullivan：中国企业级大模型日均 Token 调用报告，2026 年 2 月。

86. 中国发展高层论坛：国家数据局局长刘烈宏发言报道，2026 年 3 月。

87. 上海财经大学学术通讯／中欧国际工商学院：《How China Overtook the US in AI Token Usage—and Why It Matters》，2026 年 4 月。

88. 中文在线：投资者关系活动记录表，深圳证券交易所互动易／巨潮资讯，2026 年 6 月 3—4 日。

89. 格隆汇：《中文在线：AI 短剧成为第一大 Token 消耗场景》，2026 年 6 月 4 日。

90. 同花顺财经：中文在线投资者关系活动相关报道三篇，2026 年 6 月 4 日。

91. 野村证券援引 QuestMobile：红果短剧日活与停留时长，2026 年 7 月。

92. 界面新闻：《国内一半 AI 算力被指在造爽剧，用于研发仅 15%》，经华尔街见闻转载，2026 年 10 月。

93. 凤凰网科技：《55% Token 流向 AI 短剧：AI 时代再一次把火药做烟花？》，2026 年。

94. 人人都是产品经理：《「美国 AI 提生产力，中国 AI 造奶头乐」是伪命题吗？》，2026 年。

95. 虎嗅：关于 AI 短剧与「中美 AI 应用对比」热搜的分析文章，2026 年。

96. CNBC：美国企业在 OpenRouter 上调用中国模型比例的报道，2026 年。

97. Wing VC：《China's Open-Weight Takeover》，2026 年。

98. CSDN 技术博文：《140 万亿 Token 背后，中国 AI 视频正在「走火入魔」？》，2026 年（视频 Token 密度测算的出处，〔待核〕）。

七、本系列与德麦国际（按专著号）

99. 王德生：《GPT 时代的 SDE 批判性思维》，德麦国际第 23 号。

100. 王德生、林凡昊：《在突变中发生：GPT 时代的风暴与花开》，德麦国际第 40 号。

101. 王德生：《大学的终结与重生：基于 GPT 的教学和科研之变》，德麦国际第 42 号。

102. 王德生：《超越波普尔：从多次性科学到一次性发生》，德麦国际第 226 号。

103. 王德生：《从六步到九步：创造力的揭秘》，德麦国际第 236 号。

104. 王德生：《心理学的典范革命：从主体的发现到发生》，德麦国际第 306 号。

105. 王德生、胡敏：《GPT 时代的大学革命：读书与创业一体化》，德麦国际第 333 号。

106. 王德生：《从马力到 Token：力的属性纠缠文明史》，德麦国际第 367 号。

107. 王德生：《典范发生学：SDE 与 AI 时代的科研层级》，德麦国际第 369 号。

附录

附录六件，供查、供对、供练：附录一列数据的出处与核验状态，附录二是一张年表，附录三是三层速查表，附录四登记本书据以立论的口述命题，附录五清理十个常见误读，附录六写推翻条件与几道练习。

附录一

数据出处与核验表

本书用到的数字，一半来自公司披露，一半来自媒体转述。下表把每个数字的出处、核验状态和用法写在一起。核验状态分三档：原文可查、仅见转述、找不到出处。

数字	出处	核验状态	本书怎么用
国内 Token 消耗：短剧与视频 55%，电商营销 24%，软件开发 15%，其他 6%	中文在线投资者关系活动记录，2026 年 6 月	原文可查；统计口径未公开	当作「一家短剧公司援引的行业数据」，不当作国家统计
「85% 用于短剧和广告」	网络转述	找不到出处；前两项相加是 79%	只作为传言被拆解
美国 58% 的 Token 用于编程与企业自动化	网络转述	找不到出处（虎嗅作者核查未果）	只作为传言被拆解
仿真人剧 Token 消耗是漫剧的两到三倍	中文在线同一份记录	原文可查	说明视频内部的密度差
一季度每天新增 AI 短剧约 1500 部、月上新超 4 万部	中文在线同一份记录	原文可查	说明产能
10 秒 1080p 视频约 35 万 Token	CSDN 技术博文	单一来源〔待核〕	只说数量级，不说精确数
5 秒视频成本 4.5 到 9 元，制作周期 15 天压到 3 天	同一博文	单一来源〔待核〕	只作旁证
微短剧用户超 8 亿；红果日活 1.68 亿，人均每天约 125 分钟	野村证券援引 QuestMobile，经虎嗅转述	仅见转述	说明需求侧规模

数字	出处	核验状态	本书怎么用
中国日均 Token 消耗超 140 万亿	国家数据局局长，2026 年 3 月	多家媒体一致转述	说明总量
OpenRouter 上中国模型周调用量 2026 年 2 月起超过美国	OpenRouter 公开榜单，媒体测算	仅见转述	说明开放权重模型的份额
OpenRouter 用户：美国约 47%，中国约 6%	媒体转述 OpenRouter 数据	仅见转述	说明样本偏向
OpenRouter 编程类用量由约 11% 升到 50% 以上	OpenRouter 与 a16z 研究	原文可查	说明开发者平台的用途
美国企业把三成以上 Token 发给中国模型	CNBC	仅见转述	说明成本敏感
大团队发展、小团队颠覆	Wu、Wang、Evans，Nature，2019	原文可查	第 30 章的主证据
「破四唯」2018 年 10 月提出	科技部等五部门通知	原文可查	只作公道的一面

各章正文里另有〔待核〕若干处，多为年份和轶事，统稿时已尽量改为常识口径；仍保留的，以本表的三档为准。

附录二

年表：中国面对新技术的四十年，与 AI for Science 的十几年

年份	中国	世界	本书哪一章
1962	—	库恩《科学革命的结构》	导论；第 15、22、36 章
1980 年代中后期—1993	王选激光照排；联想汉卡；WPS（1988）；用友（1988）；金蝶（1993）	反向传播论文（1986）；个人电脑普及；MoE 思想提出（1991）	第 9、14、19 章
1998—2001	腾讯 OICQ；盛大代理《传奇》	—	第 10 章
2003—2009	淘宝、支付宝、阿里云	—	第 11 章
2012	—	AlexNet	第 19 章
2015	屠呦呦获诺贝尔奖	—	第 38 章
2017	—	Transformer；大规模稀疏 MoE	第 14、18 章
2018	科技部等五部门提出「破四唯」	Hinton 等获图灵奖	第 19、28 章
2019—2020	—	大团队发展、小团队颠覆（Nature，2019）；AlphaFold2 夺冠 CASP14（2020）	第 20、30 章
2023	盘古气象发表于 Nature	—	第 21 章

年份	中国	世界	本书哪一章
2024	DeepSeek-V2 提出 MLA	Hinton、Hopfield 获诺贝尔物理学奖；Hassabis、Jumper 与 Baker 获诺贝尔化学奖	第 14、19、20 章
2026 年 2—6 月	OpenRouter 上中国模型周调用量首超美国（2 月）；日均 Token 超 140 万亿（3 月）；中文在线披露 Token 消耗结构（6 月）	—	第 1、2、4、7 章
2026 年 10 月	第 369 号《典范发生学》上站；本书完稿	—	第 22—25 章

附录三

三层速查表

	技术改进	架构跃迁	科学典范
问的是什么	同一个目标，怎样更快更便宜	同一类任务，换不换整个结构	什么才算好问题、好答案
从哪里来	应用的约束	应用现场里一块免于指标的飞地	长期反主流、没有应用的少数人
典型例子	MoE、MLA 的工程化；视频生成降本	Transformer 取代 RNN	深度学习取代符号主义；板块构造
AI 会不会接手	正在接手	部分接手	只能作复合体的一半
谁来承认	市场、榜单	同行、开源社区	往往要等一代人
中国的位置	极强	偶有，难持续	几乎缺席
随身一问	墙绕过去了吗	墙拆掉了吗	题换了吗

附录四

口述命题登记

以下八条，是王德生 2026 年 10 月 18 日晚在一次对话中口述的判断。本书据此立论，各章是对它们的展开、补证和收窄。登记在此，便于读者对照，也便于将来推翻。

编号	命题（大意）	主场
一	中国的 AI 先被用来更便宜地生产注意力商品；改革开放以来面对计算机、互联网，大体都是先注意力经济，再生产力	上篇、中篇一
二	国人的说法是「先做应用」	第 12 章
三	应用上的碰撞，涌现的是技术改进，而不是架构与科学新典范的 0 到 1	第 14、15 章
四	常规科学恰恰是被 AI 替代的（接第 369 号《典范发生学》）	第 22 章
五	复合体是人—AI，这里的人比少数天才更稀缺	第 25、26 章
六	依靠少数人做复合体 0 到 1，这少数人必须是超级的思想寡头，在中国的土壤上更难诞生	第 33 章
七	这种寡头在国内难发芽，发芽也难长苗，最后不得不出海求生存	第 38—41 章
八	今天中国的 AI for Science 是团队出击、共同体出击，这本身是反 0 到 1 的	第 30 章

对话中有两处收窄，也照实登记：第一，计算机进入中国时是生产力在前，所以命题一里的「大体」不能去掉，规律修正为「注意力→应用层基础设施，但很少→底层原创」；第二，命题三里的「架构」与「科学典范」来源不同，架构跃迁有时就在应用现场，条件是那里有一块免于指标的飞地。

附录五

十个常见误读

1. 误读：本书说中国 AI 不行。本书说的是中国在改进层极强，在典范层几乎缺席，而 AI 正在接管改进层。
2. 误读：本书说做短剧是浪费。本书说的是短剧本身是内容行业的生产力革命，问题在回流。
3. 误读：本书说 55% 是假的。55% 有原文，只是口径未公开、发布者有利益；拿它比美国的 58%，才是问题。
4. 误读：本书反对「先做应用」。本书反对的是那个「再」字永远不排期。
5. 误读：本书说 DeepSeek 没有创新。本书说它是了不起的工程改进，改变了世界的成本结构，只是不属于 0 到 1。
6. 误读：本书反对团队。本书说团队擅长典范之后的系统打造，这一段正被 AI 接手；团队不该承担典范猜想。
7. 误读：本书反对高考。本书说高考的公平是真的，问题在它只量一种东西，而这种东西 AI 最先接手。
8. 误读：本书在为「民科」说话。本书说「民科」标签省掉了检验；被省掉的检验，应该补回来，而不是放过所有人。
9. 误读：本书在歌颂思想寡头。第 37 章专门写寡头的自我证实，寡头比谁都更需要被检验。
10. 误读：本书劝人出国。本书说出海能换来活下来和一张通行证，代价是根与养分被拉开，真正难的是回流；出不出海，不是本书的建议。

附录六

推翻条件与练习

判词两半，各有推翻条件，与导读和第 48 章第七节是同一套。

前一半说：AI 越快进入科学，中国最擅长的那一层越先贬值。出现下面任何一件，前一半就错了：十年后，常规科学仍然主要靠大批人力一题一题地做，AI 没有接管；或者 AI 大规模进入科学以后五到十年里，中国大团队在改进那一层的工作，被引、被用、被买的份额不降反升，而且不是靠价格。

后一半说：唯一还稀缺的那一层，中国的土里长不出那个人。出现下面任何一件，后一半就错了：十年之内，从现在的大团队体制里长出一个被世界公认的新典范；有可靠研究表明，几道关选出来的人里做出新典范的比例并不低于别处；十年之内，一个在国内完成全部教育、没有先拿外国承认的人提出新典范，并且首先在国内被及时认出来。

另有两件，出现了不算判错，只说明土在变，后一半要收窄：出现能让「一个猜想者＋AI＋公共平台」立项的通道；按内容、不按身份的典范评审在国内大范围运转，并且判对过一个起初被当作「民科」的人。

还有两条冲的是上篇和中篇一的起点，不直接冲判词：在可靠、口径公开的统计下，中国 Token 的用途结构与美国并无本质差别；或者有人说明，四十年里注意力挣来的钱已有相当一部分回流到底层原创。最后一条来自科学本身：要是 AI 自己就能稳定地提出被世界采用的新典范，后一半就失去了意义。

练习

1. 找一个你身边的科研项目，用附录三的速查表判断它在哪一层。
2. 找一条你最近转发过的数据，追到它的原始出处，按附录一的三档给它定级。

3. 写下你所在行业的一个「再」字：哪件事一直说「先做这个，再做那个」，那个「再」排进日程了没有。
4. 如果你要资助一个典范猜想，用第 45 章的几条机制，替它写一份一页纸的推翻条件。
5. 和孩子一起，各出一道题，而不是各解一道题。

后记

这本书起于一个晚上的一串对话。

2026 年 10 月 18 日，晚上八点半，王德生在对话框里问了一句：大家知道为什么中国八成以上的 Token 用在短剧和广告上，美国正好相反吗？对话从这张图开始，一路往下追：数字可靠吗，为什么资本押在这里，四十年前是不是也这样，「先做应用」到底意味着什么，AI 进入科学之后哪一层会被接手，那一层要什么样的人，这样的人为什么在中国难长，长出来以后会怎样。

一个小时里，他口述了八条判断。每一条都被追问过、收窄过：计算机是反例，DeepSeek 是改进不是 0 到 1，提问也不再稀缺，那个人比天才更稀缺。附录四把它们照实登记了下来。对话结束时，他说：可以开始写了，二十万字，十个智能体并行。

所以这本书本身，就是一个小小的复合体产物：一个人提出判断、裁定方向，十位 AI 写手分头写，再合起来统稿。它是不是本书所说的那种 0 到 1，读者可以自己判断。至少它示范了一件事：一个人加 AI，确实可以在一个晚上做过去一个编辑部几个月的活。剩下的问题是，这样做出来的东西，要被谁承认才算数。

感谢林家四人，他们是虚构的，却比很多真实的数字更能说明问题。感谢那些在网上转发、争论、核查这张图的人；没有他们的较真，本书的第一篇写不出来。

一本讲「回流」的书，最希望的，是被读者回写。你若在自己的行业里找到了那个「再」字，找到了一块免于指标的地，或者找到了推翻本书判词的证据，请写下来。

下一本见。下一种火药，要有人先猜出来。

判　词

AI 越快进入科学，
中国最擅长的那一层越先贬值；
唯一还稀缺的那一层，
中国的土里长不出那个人。

——本书判词

网上流传一张图：中国的 Token 烧在短剧和广告上，美国的烧在编程上。本书从这张图拆起：数字可靠吗？资本为什么押在这里？「先做应用，再做原创」的那个「再」字，四十年排进日程没有？

AI 正在接管常规科学。技术改进、架构跃迁、科学典范三层里，它先接走的，正是中国最强的那一层；留下的，是典范猜想。那一层要的是人加 AI 的复合体里那个人——他比天才更稀缺，而选拔的筛子最会筛掉他；长出来了，又要面对合法性的墙、出海与回流。

六篇四十八章，陪读的是上海林家三代：做 AI 制药的林舟、做 AI 短剧的宋岚、刷题的豆豆、用过汉卡的老林。第 369 号《典范发生学》的姊妹卷。

问题不在拿火药做烟花，在没人发明下一种火药。

ISBN 979-8-90690-578-9

德麦国际出版社 · Demai International Press
德麦国际专著第 363 号 · 六篇四十八章 · 约 20 万字
定价 US$20.00