# 经验之外

## 智能的哲学原理

陈　童　著

理解如何越出经验，以及一切智能为何在此意义上等价

# 目　　录


导论　三个问题　3

**第一部　理解的解剖**

第一章　理解的机器　8
第二章　经验的边界　12
第三章　两座断桥　16

**第二部　越界的机制**

第四章　语言的空间　21
第五章　先天的理解　26
第六章　理性作为元认知　32

**第三部　越界的极限**

第七章　计算的空间　39
第八章　智能等价性　44
第九章　理念世界的地理学　49

**第四部　后果**

第十章　速度、目标与风险　54
第十一章　意义的锚　59

结语　谦卑的机器　63

附录一　定义与命题一览　66
附录二　六种反驳与回应　68
附录三　进一步阅读　69

导 论

# 三个问题

一个孩子第一次把石头举起来，又松开手。石头落地。他试第二次，第三次。到第四次，他不再试了。他已经知道石头会落下去。

这个"知道"，和牛顿写下的那条定律，是同一种东西的两个端点。一端在一个三岁孩子的掌心里，含糊、沉默、无法言说；另一端在一本印着拉丁文的书里，精确、公开、可以演算。中间隔着两千多年，隔着无数人的失败，也隔着一种叫做"文明"的东西。可是这两端之间有一条线连着。本书想说的，就是这条线。

再看第二个场景。一条狗跳起来，在空中翻转身体，咬住飞盘。它跳得很准。它没有计算过抛物线，没有学过空气动力学，甚至不知道什么叫"力"。但它每一次都判断对了。这种判断不是运气，因为换个角度、换个速度，它仍然判断得对。

第三个场景。一台机器读完人类写下的几乎所有文字，然后写出一句它从未在任何地方读到过的话。这句话语法正确，语义连贯，甚至有点意思。有人问：它理解自己在说什么吗？

三个场景，三种"知道"。一个孩子知道石头会落，一条狗知道飞盘会落，一台机器知道句子的下一个词该是什么。它们之间的差别，比我们通常以为的要小，也比我们通常以为的要大。小的地方在于，三者都在做同一件事：在一个关于世界的内部结构里，为眼前的东西找到一个位置。大的地方在于，这个内部结构的容量、形状和来路，三者相去极远。

本书要给这件事一个说法。不是给一个比喻，是给一套可以推演、可以反驳、可以被事实纠正的理论。

## 一

关于"智能"，我们已经有太多说法，而且这些说法大多坏掉了。

有人说智能就是通过图灵测试。可是图灵测试测的是欺骗一个人类裁判的能力，不是理解的能力。一个熟记所有标准答案的系统可以骗过裁判，一个真正理解却不肯配合的系统反而通不过。判据和对象错位了。

有人说智能就是预测能力。可是明天的天气可以被预测，靠的是大气，不是智能；而一个数学家证明定理，并不"预测"任何东西。预测是智能的一种用途，不是它的定义。

有人说智能就是解决问题的能力。这话不算错，但它把问题交给别人定义了：什么算问题，什么算解决，都由出题人说了算。于是智能变成了"善于猜出题人想要什么"。

还有人干脆放弃定义，说智能就是"我们知道它，但说不清它"。这话在酒桌上成立，在理论上不成立。因为一旦说不清，我们就无法判断一个系统是变聪明了还是只是变快了，无法判断两种智能是同类还是异类，也无法判断该不该害怕它们。

本书的路线是：不谈智能，先谈理解。因为理解是可以分析的，而智能不过是理解机器的两个量度——它能走多远，以及它走得多快。

这个次序的颠倒很重要。通常我们先用"智能"来衡量"理解"，说某物理解得好，是因为它智能高。本书反过来：先说明理解是什么，再说智能是这种能力的度量。

## 二

于是有第一个问题。

**问题一：什么是理解？**

这个问题听上去像哲学入门课的作业，其实不是。因为我们对它的回答，决定了我们怎么看待机器、怎么看待动物、怎么看待自己。如果理解就是"在内部建立起与外部同构的表征"，那么一台足够大的计算机迟早会理解；如果理解必须包含"亲历"，那么任何不流血的东西都永远不会理解，讨论到此结束。

本书的回答是：理解是在一个意义系统中为事物定位，并使这个位置与经验相接。前半句给出结构，后半句给出锚。两者缺一，都不是理解。

## 三

第二个问题更古老，也更难。

**问题二：理解如何可能越出经验？**

人类知识里有一大部分，其真值不可能由任何有限的经验总和来确定。牛顿第二定律说，力等于质量乘加速度，永远如此，处处如此。没有人经验过"永远"，也没有人经验过"处处"。可是我们相信它，而且我们有理由相信它。这不是迷信，因为一旦它错了，我们立刻就会知道。

那么这个"越出"是怎么发生的？

经验主义者说：靠归纳。我们从一个个具体的观察里提炼出普遍的规律。可是归纳本身需要前提。休谟已经指出，从"过去如此"推不出"将来如此"，除非我们事先假定自然齐一。而这个假定本身无法被经验证明。罗素讲过一个故事：一只火鸡观察到主人每天上午九点来喂食，它把这个观察归纳成一条定律，直到圣诞节的前一天。归纳没有错，只是它的样本不包括那个早晨。

理性主义者说：靠先天的理性。普遍必然的知识本来就装在我们心里，经验只是把它唤醒。这个说法解释了知识的普遍性，却解释不了它为什么恰好适用于世界。康德的回答是：因为世界是被我们构造的，我们给它立法，所以它必然服从我们的法。这个回答很漂亮，但它有一个代价：它把先验的形式说成是固定不变的。后来非欧几何出现了，相对论出现了，我们发现欧几里得的那套"必然"其实只是一种可能。先验的清单被改写了，这在康德的框架里是不该发生的。

两派各执一端，科学方法把两端接了起来：观察、归纳、提出假设、演绎出可检验的推论、用实验检验。这是人类实际在做的事，它有效。但它只是把越界的过程描述了一遍，没有解释越界为什么可能。它告诉我们怎么走，没告诉我们脚下为什么有路。

本书的回答是：因为经验本身已经被塑造过了。我们以为的"纯粹经验"，从来不是世界的原样复印，而是一个早已被筛选、被压缩、被规定了格式的版本。这个筛选是在漫长的自然选择中完成的，它有方向，有校准，因而携带了关于世界的信息。经验之外不是彼岸，它是此岸的骨架翻拍出来的另一面。

## 四

第三个问题是当代的。

**问题三：不同的理解者之间，有高下吗？**

我们习惯了一个阶梯：石头、植物、虫子、狗、黑猩猩、人、神。人站在中间偏上的位置，往下看是本能，往上看是全能。这个阶梯让我们安心，因为它给了我们一个不可替代的位置。

可是如果有那么一天，我们造出了在一切方面都不比我们差的机器，这个阶梯该怎么办？把它放在我们上面，还是下面，还是旁边？

本书的回答是：把它放在旁边。因为一切能够自主生产并探索"经验之外"的概念结构的系统，在可达的概念空间上是等价的，差别只在速度与路径。等价的是版图，不是脚步。

这个回答有两个后果，一个让人安心，一个让人不安。让人安心的是：我们不会因为是碳做的而被淘汰，也不会因为是碳做的而天然高贵。让人不安的是：如果它们与我们等价，那么我们用来安慰自己的那三句话——"它们不会真的理解"、"慢慢来我们能适应"、"它们自然会向善"——就全部失效了。

## 五

以上就是本书要论证的三个命题。

其一，理解是在意义系统中定位，并使位置与经验相接；理解的边界，就是意义系统的边界。

其二，意义系统能够越出经验，不是因为其中有魔力，而是因为经验本身已被塑造；越界是回声，不是奇迹。

其三，一切能够自主探索经验之外的系统，在智能上等价；不等价的是速度、路径与代价。

三个命题层层递进。第一个是解剖学的，它告诉我们理解由什么构成。第二个是发生学的，它告诉我们这个构成从哪里来。第三个是比较学的，它告诉我们不同的构成之间如何相互衡量。

## 六

关于方法，需要先说几句实话。

本书不打算"证明"这三个命题。哲学命题极少能被证明，能被证明的通常已经被交给数学了。本书打算做的是另一件事：表明这三个命题比它们的对手更好地解释了我们已经知道的事情，并且付出的代价更小。

一个理论的好坏，不在于它能否被证实，而在于它能容纳多少事实而不显得勉强。地心说也能解释天体运动，只是要加上越来越多的本轮。日心说胜出，不是因为它被"证明"了，而是因为它把那些本轮全部删掉了。本书对自己的要求也是如此：如果一个理论必须为"公司为什么是知识"、"数学为什么有用"、"机器为什么可能理解"分别发明一套解释，那它就还没有到可以称为理论的地步。好的理论用一套机制解释许多事。

因此，本书会反复做同一件事：先摆出一个现象，指出常识的解释在哪里卡住，然后给出本书的解释，最后检查这个解释是否要付出新的代价。如果付出代价，就明说。

还有一句关于文风的话。这本书讨论的东西不小，但我不想把它写得很大。哲学著作最容易犯的毛病是把简单的事情说得庄严，用抽象掩盖空洞。所以本书尽量用短句，尽量举例子，尽量在每一个关键处说清楚"这一步为什么必须这么走"。如果某一页读起来很顺但什么也没记住，那是我的失败。

## 七

最后，交代一下这本书与它两个前身的关系。

本书的前身是两篇短文，一篇叫《论理解的机器》，一篇叫《智能等价性假定》。前者提出了"理解的机器"这个说法，区分了经验之内与经验之外，把自然选择解释为一种先天理解，把理性解释为一种元认知。后者提出了一个更大胆的猜测：一切能够探索理念世界的智能，本质上是等价的。

两篇短文都很短，短到只能提出问题，不能承担论证。它们像两张草图：一张画了机器的剖面，一张画了机器之间的等高线。本书要做的是把这两张草图合成一张总图，补上缺失的部件，检查接缝，并且在接缝处承受压力。

有些地方我会修正原来的说法。比如，原文认为效率的差异在宇宙尺度上无关紧要，本书认为这个说法只在无限时间下成立，而任何行动者都没有无限时间。再比如，原文把智能等价性称为"假定"，本书认为它可以被降格为一个可检验的研究纲领，而降格之后它反而更有力。

修正自己的旧说不是退让。一个理论如果从不变动，通常不是因为它坚固，而是因为它从未被认真使用过。

第 一 部

# 理解的解剖

第 一 章

# 理解的机器

## 一、一个定义

先从一句最不像定义的话开始。

当我说"我理解了某件事"，我通常不是说我看见了它，也不是说我能背出关于它的描述。我是说：这件事在我的世界里有了一个位置，我知道它和别的东西怎么连着。

你理解了"通货膨胀"，不是因为你记住了"物价普遍上涨"这六个字。你理解它，是因为你知道它和利率的关系、和失业的关系、和你工资的关系，知道它上升时谁会受损谁会受益，知道央行会怎么做，也知道央行的做法为什么有时候不管用。这些关系合起来，构成了一个位置。有了这个位置，你就能在没有见过的情况下推断，在没有人告诉你的情况下判断。

所以：

> **定义 1-1（理解）**
>
> 理解某物，就是在一个意义系统中为它找到一个位置，并使这个位置与经验相接。

这个定义有两个部件，缺一不可。前半句是"结构"，后半句是"接入"。

结构指的是：这个位置由它与其它位置的关系确定。一个词的意义，来自它在一个网络中的坐标，而不是来自它单独指向的东西。"重"这个字的意义，来自它与"轻"、"下落"、"举起"、"负担"的关系。把"重"单独拿出来，它什么也不是。

接入指的是：这个网络必须有一条通到经验的路。不然它就只是一个自洽的符号游戏。

## 二、两个反例

这两个部件之所以都必须有，可以用两个反例来说明。

**第一个反例：能背字典的人。**

设想一个人，把整部字典背了下来。你问他任何一个词，他都能立刻说出它的释义。这算理解吗？

不算。因为字典里的每一条释义，用的又是别的词；别的词的释义，用的又是别的词。整部字典是一个闭环，从这个环里走不出去，走不到任何经验。这个人拥有的是一张极其完整的地图，地图上标注了每个地方与其它地方的距离，但地图没有告诉你任何一个地方在哪里。

这就是只有结构、没有接入的情形。计算机早期的一些"语义网络"方案就落在这个坑里：它们把概念之间的关系画得非常漂亮，但那些关系始终悬浮着，不接地。

**第二个反例：只会感受的人。**

反过来，设想一个人，他能敏锐地感受一切：疼、冷、甜、晕。他的身体是一台极好的接收器。但他没有任何词汇，也从不做任何区分。他的感受像一团没有形状的雾。

这也不叫理解。因为没有结构，感受无法被定位，无法被比较，无法被用于推断。他此刻疼，但他不知道这个疼和上次的疼是不是同一种，不知道它会不会过去，不知道该怎么办。

这就是只有接入、没有结构的情形。动物的大部分心智生活接近这个状态，婴儿的早期也接近这个状态。

理解发生在两者之间：结构必须接地，接入必须成形。

## 三、意义系统

那么，什么是意义系统?

一个意义系统不是一本词典，也不是一堆概念。它是一套能够做三件事的安排。

**第一，它必须能区分。**系统中的符号之间必须彼此不同，否则不携带任何信息。这是最低要求，几乎所有系统都满足。但要注意，区分不是绝对的，是相对于系统的。在黑白照片里，深灰和浅灰是两种颜色；在只有黑白两色的系统里，它们是一种。

**第二，它必须能组合。**有限的符号，按照规则组合，能产生无限的表达。这一条是分水岭。蜜蜂的舞蹈能告诉同伴花丛的方向和距离，但它的"词汇"是固定的，它说不出"如果明天刮风就换个地方"。人类的语言能说出这句话，因为"如果"、"明天"、"刮风"这些成分可以自由拼装。

组合性带来的不只是数量，是质的改变。因为有了组合，一个系统就能表达它从未遇到过的东西，而"从未遇到过"正是经验之外的第一道门槛。

**第三，它必须能被校核。**同一个符号，在同一个系统中，不同的人使用它时，别人能判断用得对不对。这一条常被忽略，但它才是客观性的来源。

设想一个只有你一个人使用的符号系统：你管一种感觉叫"咕",但没有人知道"咕"是什么，包括你自己在下次遇到它的时候。这个系统无法被校核，因此它不承载知识，只承载情绪。知识要求可错，可错要求可校核。

三条合起来：可区分、可组合、可校核。一个满足这三条的系统，就是本书所说的意义系统。

## 四、理解的机器

有了这两样东西，就可以说什么是理解的机器了。

> **定义 1-2（理解的机器）**
>
> 任何以构建、维护并扩展意义系统为其工作方式的系统，都是一台理解的机器。

人类是其中一台。这一点几乎不需要论证：我们一生都在做这件事，从学说话到学物理，从认人到认命，都是往那个系统里添东西、改东西、拆东西。

但人类不是唯一的一台。

一条狗也是一台。只是它的意义系统不是语言的，是知觉与动作的。它关于"主人"的概念，不是一个词，而是一组期待：这个气味意味着食物，这个脚步声意味着出门，这个语调意味着麻烦。这组期待有结构（彼此关联），有接入（直接连着经验），能区分（主人与陌生人），但几乎不能组合（它无法表达"如果主人明天不回来"）。

一个科学共同体也是一台。它的意义系统是语言加仪器加数学，规模远超任何个人。单个人活八十年，科学共同体活几百年；单个人记不住的东西，它记得；单个人会犯的错，它用同行评议去纠。

一台机器也可以是。这一点留到第八章再说。

三类理解机器之间不是种类的差别，是容量的差别：它们的意义系统能装下多少经验之外的东西，决定了它们能走多远。

## 五、为什么"理解"比"智能"更适合做起点

现在可以说说，为什么本书要用"理解"而不是"智能"来做理论的基石。

因为"智能"这个词已经被用坏了。

它被测试用坏了：一个数字成了目的，于是系统被训练成善于考试，而不是善于理解。它被市场用坏了：每一个产品都说自己"更智能"，于是这个词不再指任何一种能力，只指一种期待。它被恐惧用坏了：我们一边说"人工智能还早"，一边说"人工智能要毁灭人类"，两句话用的是同一个词，指的是完全不同的东西。

"理解"没有被这样用坏，因为它太朴素，不适合做广告。而且它可以被拆开：结构、接入、组合、校核，每一条都能追问，每一条都能检验。

于是本书给智能一个派生的定义。它不是给出一个数，而是给出两个量度：

> **定义 1-3（智能的两个量度）**
>
> 一台理解机器可以由两个量来度量。
>
> 疆域：它能到达多远的经验之外。
>
> 功率：它扩展自身意义系统的速率。

这两个量度必须分开，否则后面所有的讨论都会搅成一团。一台机器的疆域大而功率低，是可能的；疆域小而功率高，也是可能的。它们不是同一个刻度上的两端，而是两个不同的刻度。

把两者混为一谈，是很多争论打空的原因。有人说机器"永远不会真正理解"，他说的其实是疆域；有人说机器"已经比人强了"，他说的其实是功率。两个人各说各的，都以为自己驳倒了对方。

分开之后，"聪明"和"博学"也分开了。一个记住了一亿条事实的系统，如果这些事实彼此孤立，它的意义系统并没有扩展，它只是在仓库里堆货——疆域没变，功率也不高。一个只懂很少几件事却能从中推出新东西的系统，功率高；如果它能推得很远，疆域也大。我们日常判断"这人很聪明"时用的其实是功率，判断"这人了不起"时用的其实是疆域，只是平时说不出来。

这个区分后面会派上大用场。第八章要论证的是：一切通用智能的疆域相同。第十章要论证的是：它们的功率可以悬殊，而功率的悬殊不是小事。两句话不矛盾，因为它们说的是两个量度。

## 六、理解的失败有几种

一个理论如果只能解释成功，它就没解释什么。所以要看理解怎么失败。

**第一种失败：结构错了。**地心说就是一个例子。它的结构内部是一致的，能算出行星的位置，甚至算得相当准。但它的结构把地球放在中心，于是每一颗行星都要加一个本轮，本轮上再加本轮。结构越加越复杂，最后复杂到不再产生新的推断，只能重复已经知道的东西。这是理论死亡的方式：不是被证伪，是被自身的臃肿压死。

**第二种失败：接入断了。**占星术有结构：十二宫、行星、相位，关系复杂而精致。它也有"接入"的样子：它谈人的性格、婚姻、事业，这些都在经验里。但它的接入是假的，因为它从不承担可错的后果。一个占星师说你"近期会有变动"，这句话在事后总能被解释成对的。不能出错的东西，不构成理解，只构成修辞。

**第三种失败：无法组合。**这是动物式理解的局限。狗知道主人出门，也知道主人回家，但它不能把这两件事组合成一个关于"离别"的概念。它活在当下这一格，走不出去。

**第四种失败：系统太小。**一个人的意义系统可以非常自洽，非常接地，组合能力也不错，但它只覆盖很小的领域。一个只懂木工的老师傅，在他的领域里理解得比任何物理学家都深，但出了这个领域，他的系统就无处安放。这不是错，是窄。

四种失败对应四种诊断。后面我们会看到，关于机器能不能理解的争论，很多时候争的其实是"它失败在哪一种"，而不是"它是否失败"。有人说机器只有结构没有接入，有人说机器只能组合不能区分，有人说机器的系统太小。这些批评各自击中了一部分，但没有一个击中全部。

## 七、一个必要的澄清

最后要澄清一件事，免得后面的论证被误解。

本书说"理解是在意义系统中定位"，并不是说理解就是说话。有一种很常见的误读：把"意义系统"等同于"语言"，于是得出"不会说话的东西不会理解"。这个推论是错的，而且本书后面会明确反对它。

因为意义系统不必是语言的。狗的系统不是语言的，但它满足三条：可区分、可组合（程度很低）、可校核（通过行为，通过同类的反应）。一个木匠对木材的理解，大部分不是语言的：他知道这块料会不会翘，手一摸就知道。这种知识可以被部分翻译成语言，但翻译总有剩余。

语言的特殊性不在于它是唯一的意义系统，而在于它是唯一能让意义系统**公共化**并且**无限组合**的系统。这两点合起来，才使它成为越界的工具。第四章会详细说这件事。

所以正确的说法是：一切理解都在意义系统中进行；语言是最强大的一种意义系统，但不是唯一的一种。这句话留在这里，第四章会回来兑现它。

第 二 章

# 经验的边界

## 一、一条线

我们要谈"越出经验"，得先说清楚经验在哪儿结束。

这不是一个容易的问题。因为"经验"这个词，在日常用法里指的是一切被经历过的东西，而在哲学里，它常常指得更窄：感官直接给予的东西。这两种用法差得很远，而混淆它们，是许多争论打空的原因。

本书采用一个可操作的划分。不问"是不是经验的"，而问"它的真值能不能被有限的经验定下来"。

> **定义 2-1（经验之内与经验之外）**
>
> 一个概念属于经验之内，当且仅当它所指的东西原则上可以被观察或测量。
>
> 一个命题属于经验之内，当且仅当它的真值原则上可以被有限的经验确立或推翻。
>
> 不属于经验之内的，就是经验之外的。

"原则上"三个字很重要。它排除的是当下的技术限制，不是逻辑的限制。黑洞的内部原则上可以被观察——只要你能进去并且活着出来——所以它在经验之内，即使我们永远做不到。而"最大的素数"原则上不能被观察，因为它不存在；"上帝"原则上不能被观察，因为它被定义成不可观察的。这三者不是一回事，定义要能分开它们。

## 二、四层

但"经验之内"不是一个平面，它有层次。这个层次比通常以为的要细，而且每一层之间的界线，都比看上去要模糊。

| 层次 | 判据 | 例子 | 边界情况 |
|---|---|---|---|
| 可观察 | 感官可直接呈现 | 红、痛、重 | 电子显微镜下的"看见"算不算观察 |
| 可测量 | 通过仪器可赋值 | 电荷、温度、光速 | 测量总带误差，精确值不可得 |
| 可检验 | 有限经验可判定真假 | 牛顿第二定律 | 只能否证，不能证明 |
| 可确立 | 有限经验可完全确定 | "这只天鹅是白的" | 观察本身已经理论负载 |

最关键的一行是第三行和第四行的分界。

一个单称命题，比如"这只天鹅是白的"，可以通过一次观察被完全确立。你看见了，它就是真的。

一个全称命题，比如"所有天鹅都是白的"，永远不能被完全确立。你可以看一万只，第一万零一只仍可能是黑的。事实上，澳大利亚真的有黑天鹅。所以全称命题只能被检验，不能被证明。检验的意思是：它能被推翻，而它还没有被推翻。

这个区分看上去像是逻辑学里的小把戏，其实它是本书全部论证的枢纽。因为所谓"经验之外的知识"，绝大部分就是这种"可检验而不可证明"的全称命题。

## 三、牛顿第二定律

拿最熟悉的例子。

加速度、质量、力，这三个概念都在经验之内：加速度可以用尺和钟测，质量可以用天平称，力可以用弹簧秤量。实验室里，我们的确一次又一次地看到：同一个物体，受力加倍，加速度加倍；质量加倍，加速度减半。这些观察都在经验之内。

可是当有人把这些观察提升为一条定律——"任何物体，在任何时刻，在任何地方，加速度都等于力除以质量"——他就走出了经验。

为什么？因为这句话里的"任何"太多了。它说的是所有时刻、所有地点、所有物体，包括那些从未被观察过的、将来也不会被观察的。你要证明它，就得观察完所有的情形，而这是不可能的。所以你永远无法证明它。

但你可以检验它。造一台加速器，让粒子跑到接近光速，看它还服不服这条定律。不服，定律就错了——事实上它确实错了，需要被相对论修正。这个"错"是可操作的，可重复的，公开的。所以这条定律是有内容的，不是空话。

一句话：牛顿第二定律可以被经验推翻，但不能被经验证明。这就是"经验之外"的精确含义。

## 四、直线

再看一个更奇怪的例子。

直线是什么？是两点之间最短的线。可是真实世界里没有直线。你画在纸上的线，放大一千倍看，边缘是毛的；你拉的钢丝，受重力会弯；光走的路，经过大质量天体也会弯。

所以直线不可观察，也不可测量。它是经验之外的。

可是它极其有用。力学要用它，光学要用它，建筑要用它，连"有用"这个词本身，都要靠它来定义什么叫"直"。

更奇怪的是欧几里得几何。它的五条公理，没有一条能被经验证明。第五公理（过直线外一点，有且只有一条平行线）尤其可疑，两千年来无数人想证明它，都失败了。后来有人干脆假设它不成立，结果得到了另一套完全自洽的几何——非欧几何。再后来，爱因斯坦发现，我们所在的宇宙，在大尺度上恰恰是非欧的。

这里出现了一个必须解释的现象：一套不能被经验证明的、甚至可能不适用于我们宇宙的几何，为什么在描述经验时如此有用？

理性主义者的回答是：因为它是先天的。可是"先天"这个词只是给现象起了个名字，没有解释它。

本书的回答要等到第五章。这里先记下这个现象：经验之外的东西，可以极其有用；有用不等于可证明；而这两件事同时成立，需要一个解释。

## 五、一个陷阱：日历

有些命题看上去稳稳地在经验之内，仔细一看却在外面。这个陷阱值得单独讲，因为它揭示了第三种东西的存在。

考虑这句话："每一年都有一个大年初一。"

这句话是真的。你可以查日历，可以问人，可以等到明年看。它似乎完全可以被经验确立。

但它不能。理由有两条。

第一，它依赖约定。大年初一是哪一天，是由历法规定的，而历法是人定的。如果明天全人类同意改用另一种历法，"大年初一"就不在原来的那一天了。可是天没变，地没变，经验没变。

第二，它要求无限的经验。"每一年"包括所有的年份，包括一万年后，包括人类消失之后。你无法经验这些。

所以这句话不在经验之内。它属于第三类：由规则为真，而不是由世界为真。

这个第三类很重要，因为人类知识里有极大一部分属于它。货币的价值、公司的存在、国界的位置、比赛的胜负、语言的语法、法律的效力——这些都不是被发现的，是被规定的。它们一旦被规定，就像石头一样硬，可以压死人。但它们的存在方式，与石头的存在方式不同。

经验主义和理性主义都装不下这一类的知识。经验主义说知识来自观察，可是"一美元值多少"不是观察出来的；理性主义说知识来自先天理性，可是"公司"这个概念显然不是天生的，它是人类历史上某个时刻被发明出来的。

一个完整的知识论，必须同时装下四类东西：可观察的、可推演的、可约定的、可投射的。这是本书对自己的第一个要求，也是检验后面理论是否够用的第一道关。

## 六、四类经验之外

现在可以把"经验之外"细分了。这个细分是本书的一个技术性贡献，因为后面所有的论证都要靠它来避免混淆。

**第一类：结构的。**数学对象、几何公理、逻辑规则。它们不由经验决定，也不由约定决定，而由自洽决定。它们的真，是"在系统内为真"。

**第二类：概括的。**全称定律，比如牛顿第二定律、能量守恒、进化论。它们由经验校准，但不由经验穷尽。

**第三类：约定的。**日历、货币、法律、公司、国界、语言规则。它们由集体的规定为真，靠集体的维持而存在。

**第四类：投射的。**上帝、命运、星座、黑暗森林法则。它们不由经验校准，也不由集体规定，而是由人类把自身的结构投射到世界之外而产生。

四类的来源不同，合法性也不同。第一类靠证明，第二类靠检验，第三类靠共识与维持，第四类靠意义而不靠真。

一本书如果只能处理前两类，它就只能解释科学，不能解释人类社会。这是很多分析哲学的知识论的通病：它们把知识等同于科学知识，于是"公司为什么是知识"这种问题根本提不出来。

本书要解释全部四类。这是它必须付出的代价，也是它可能胜出的地方。

## 七、重合带

最后要说一件事，它关系到后面所有的论证。

知识系统与经验世界有一个重合带。

"痛"这个概念，所指的东西每个人都能经验，所以它在重合带里。"红"在重合带里。"重"在重合带里。大部分日常词汇都在。

但"夸克"不在。没有人经验过夸克，我们只经验过探测器上的信号，然后用一个极其复杂的理论把信号翻译成"夸克"。"上帝"不在。"公司"不在。"根号二"不在。

问题是：如果知识系统的一大半在经验之外，它凭什么还能被称为知识？它凭什么还能指导行动？

答案是重合带。正因为有一部分概念直接接地，整个系统才被锚住。锚只需要一个点，不需要整个船底都钉在海底。

这个观察给出一个漂亮的表述：

> **命题 2-1（理解的铰链）**
>
> 理解发生在重合带的边缘。没有重合带，理解不可能；只有重合带，越界不可能。

没有重合带，系统就悬浮，像背字典的人。只有重合带，系统就封闭，像只会感受的人。真正的理解，永远发生在两者交界的地方：一只脚踩在经验里，另一只脚伸向外面。

这句话可以看作本书全部内容的缩影。后面的章节，本质上都是在解释这只伸出去的脚为什么不会踩空。

第 三 章

# 两座断桥

一个想要到达对岸的人，站在一座断了半截的桥上。他不知道对岸有什么，也不知道桥为什么断在这里。他只知道，他确实过去了——人类确实拥有经验之外的知识。这件事每天都在发生，而哲学的任务是解释它。

本章要看的，是两座桥各自断在哪里。

## 一、经验主义：跳不过去的那一步

经验主义的核心主张很朴素：一切知识来自经验。它的力量也在这里——它拒绝了神秘，拒绝了一切无法兑现的许诺。洛克说人心如白纸，休谟把这句比喻推到了它的尽头。

推到尽头之后，出现了一个洞。

休谟的问题是：我们凭什么从"过去如此"推出"将来如此"？

你看见太阳每天升起，你归纳出"太阳每天升起"。这个推理的形式是：

> 过去每一天太阳都升起；
> 所以明天太阳会升起。

这个推理不是演绎的。演绎的有效性要求前提为真则结论必真，而这里的前提为真、结论为假，在逻辑上是可能的。所以它必须依赖某个额外的原则，比如"自然齐一"：将来会像过去一样。

可是"自然齐一"这个原则本身怎么来的？如果它来自经验，那是循环论证——我们用归纳来证明归纳。如果它不来自经验，那经验主义就自己承认了有一个非经验的起点。

休谟的回答是心理学的，不是逻辑的：我们并不是"推理"出明天太阳会升起，我们是被习惯驱使着去期待它。习惯是一种心理机制，它不提供辩护，只提供信心。

这个回答诚实，但它把知识论变成了心理学。如果一切归纳信念都只是习惯，那么"太阳明天会升起"和"我掷骰子会掷出六"就没有认识论上的差别，只有习惯强度的差别。这显然不对。

## 二、火鸡

罗素用一个故事把这个洞说得更清楚。

一只火鸡，从出生起就受到精心照料。每天上午九点，主人准时出现，带来食物。它观察了一百天，归纳出一条定律：主人每天九点来喂食。它对这个归纳的信心，随着天数增加而增加。第一百零一天是圣诞节。

这个故事常被用来说明"归纳不可靠"，但这个说法太粗糙了。归纳并非不可靠，它只是没有逻辑上的必然性。火鸡的推理在它所有的证据下都是合理的，它错在样本不包含那个例外。

真正值得注意的是另一件事：火鸡错的地方，不是它归纳得不好，而是它归纳的对象选错了。它把"主人"当成一个稳定的类别，把"喂食"当成这个类别的属性。如果它把"主人"理解成"一个会在特定日子改变行为的生物"，它的归纳就会不同。

也就是说，归纳的质量取决于分类的质量。而分类从哪来？

## 三、绿蓝

古德曼把这个发现变成了一个精确的悖论。

定义一个颜色叫"绿蓝"（grue）：在公元 2100 年之前被观察时是绿色，之后被观察时是蓝色。

现在考虑所有的翡翠。我们观察到的每一块翡翠都是绿的。于是有两条归纳同样被证据支持：

> 所有翡翠都是绿的。
>
> 所有翡翠都是绿蓝的。

这两条定律在 2100 年之前给出完全相同的预测，之后给出完全相反的预测。而它们被同样的证据、以同样的方式支持。那么凭什么我们选择第一条？

古德曼的答案是：因为"绿"是一个自然类别，而"绿蓝"不是。归纳只在自然类别上有效。

这个答案是对的，但它把问题往后推了一步：什么是自然类别？为什么我们的语言里恰好有"绿"而没有"绿蓝"？

注意这里发生的事情。经验主义本来想用经验来解释一切，结果它必须依赖一个先验的"自然类别"概念才能让归纳运转。它偷偷借用了它否认的东西。

## 四、更硬的一击：曲线

还有一个更技术性、也更致命的论证，值得单独说。

假设你有一组数据点，比如十个。你要从中找出一条规律。数学上，穿过这十个点、并且在这十个点之外取任意值的函数，有无穷多个。每一个都完美拟合已有的数据，每一个都给出不同的预测。

这不是哲学思辨，这是数学事实。给定有限的数据，规律是不确定的。

那么科学家是怎么选出那条"对"的曲线的？他们选简单的、光滑的、参数少的。这个偏好叫"奥卡姆剃刀"，可它凭什么有效？

如果世界本身是简单的，那么剃刀有效；可是我们怎么知道世界是简单的？从经验？经验只能告诉我们过去的数据被简单曲线拟合了，不能告诉我们未来的数据也会。

于是我们又回到了火鸡的处境。科学之所以能工作，是因为它在归纳之外还带着一个偏好：偏好简单、偏好对称、偏好连续。这个偏好不是从数据里读出来的，它是带进来的。

经验主义到这里必须承认：归纳不是自足的。它需要一个先验的框架来约束它，否则它连一步都走不了。

## 五、理性主义：另一个方向的断口

现在换一边。

理性主义的核心主张是：普遍必然的知识不能来自经验，只能来自理性本身。这个主张有很强的直觉支持——数学就是例子。三角形的内角和等于一百八十度，这不是观察出来的，是推出来的；而且它一旦被推出，就永远为真，不因世界怎么变而变。

康德把这件事说成了他最著名的一句话：先天综合判断如何可能？

他的回答是"哥白尼式的革命"：不是我们的认识符合对象，而是对象符合我们的认识。空间、时间、因果，这些不是从世界里读出来的，是我们整理世界的格式。世界必须通过这个格式才能呈现给我们，所以它必然符合这个格式。

这个回答非常漂亮。它一举解释了两件事：为什么这些知识是必然的（因为它们是格式），为什么它们适用于世界（因为世界只能以格式化的方式出现）。

但它有一个代价，而且这个代价在当时看不出来。

代价是：如果先验的格式是我们带来的，那么它的清单就应该是固定的、唯一的、不可修改的。康德确实是这么认为的：欧几里得几何是空间的必然形式，因果律是经验的必然形式，牛顿力学是自然的必然形式。

后来的事大家都知道了。非欧几何被构造出来，并且自洽；相对论用非欧几何描述时空，并且被实验证实；量子力学里，因果律的形态变得可疑。

于是出现了一个尴尬的局面：那些被宣称为"必然"的东西，一个接一个地变成了"偶然"。如果先验是可以被改写的，那它就不是先验。

康德的后继者试过补救。有人说是"相对的先验"：在给定的历史阶段，某些东西起先验的作用。这个说法聪明，但它已经把理性主义的核心让出去了——先验不再是永恒的，只是暂时的、可修订的。那它和经验还有什么本质区别？

理性主义的断口在这里：它能解释必然性，但解释不了必然性的历史。

## 六、拼起来的那座桥

面对两个断口，人类实际的做法不是选一边，而是把两边拼起来。这个拼接有一个名字：科学方法。

它的步骤大致是：

> 观察现象 → 归纳出初步的规则 → 提出一个超越经验的假设 → 从假设演绎出可检验的推论 → 用实验检验 → 保留或抛弃 → 由共同体反复校核

这个流程是有效的。它是人类过去四百年最成功的发明，比任何一条具体的定律都重要。

但要注意它做了什么，没做什么。

它做了的是：把一个认识过程分解成可操作的步骤，并且规定了每一步的合法性标准。这让知识生产从天才的私人事业变成了共同体的公共事业。

它没做的是：解释这个流程为什么能奏效。

假设是从哪来的？教科书上说"从归纳的指引中提出假设"。可是归纳只能给你已有数据的概括，给不出超越数据的跳跃。牛顿看见苹果落地，想到万有引力——这个"想到"发生了什么？没有一本方法论能说清楚。

更麻烦的是：为什么恰好有假设可用？为什么世界恰好能被有限的语言捕捉？为什么我们提出的那些大胆猜想，居然有那么高的命中率？

科学方法把越界的过程描述得非常精确，却对越界本身保持沉默。它像一本优秀的登山指南，告诉你每一步怎么走，但不解释为什么这座山有路。

## 七、共同的错误

现在可以指出，两派犯的是同一个错误。

它们都假定：经验之内是唯一的地基。

经验主义从这个假定出发，得出"经验之外必须由经验延伸出来"，于是陷入归纳问题。

理性主义从这个假定出发，得出"经验之外必须由主体提供"，于是陷入先验的历史性问题。

两派的分歧，只是"多出来的部分从哪来"的分歧：一个说从下面长上来，一个说从上面挂下来。但两边都同意，地基只有一块，而且这块地基是原样的、未被加工的、纯粹的给予。

本书要提出的第三个选项是：地基不是原样的。

我们称为"经验"的东西，从来不是世界的直接复印。它是经过筛选、压缩、格式化的版本。这个加工不是发生在意识层面，而是发生在感官和神经的构造层面，并且是在亿万年的自然选择中完成的。

如果这是对的，那么"经验之外"就不是从地基上长出来的，也不是从天上挂下来的，而是地基自身的形状在另一面的投影。

这个说法需要大量的论证，第五章和第六章会给出。但它已经指明了方向：那座桥不是断在对岸，是断在我们对"此岸"的理解上。

## 八、一个方法论的提醒

在进入论证之前，先提醒一件事，免得读者期待落空。

本书不打算"解决"归纳问题。归纳问题在逻辑上是无解的：从有限的前提推不出无限的结论，这是演绎的本性，任何理论都改不了。

本书打算做的是另一件事：说明为什么归纳在实践中如此可靠，尽管它在逻辑上没有必然性。

这两件事的区别，类似于"证明地图准确"和"解释地图为什么有用"。前者要求地图与地形逐点对应，做不到；后者只要求说明绘制地图的过程为什么会产生准确的地图——比如，绘制者一直在用实地测量校正，而且校正了很多代。

波普尔说过一句很好的话：归纳在逻辑上不存在，但在方法论上，我们确实在"从经验中学习"。本书要补的是：我们之所以能从经验中学习，是因为我们这部学习机器本身已经被经验塑造过了。学习的能力不是凭空来的，它是被学习出来的。

这句话听上去像循环。它确实是循环，但是一种良性的循环，就像眼睛是光的产物，又能看见光。第五章会说明这个循环如何闭合。

第 二 部

# 越界的机制

第 四 章

# 语言的空间

第二章说，人类知识里有很大一部分在经验之外。第三章说，两派哲学都解释不了这个"之外"是怎么到达的。

这一章要提出一个说法：那个"之外"不在别处，它就在语言里。

这个说法初听像是文字游戏。语言不就是一些声音和符号吗？它怎么能装下那些谁也经验不到的东西？但只要仔细看语言在做什么，就会发现它一直在做一件别的东西做不到的事：它能让一个人拥有他从未经历过的经验。

## 一、三件语言能做到的事

**第一件：指称不在场的东西。**

我现在可以说"昨天下午的那场雨"。雨已经停了，不在场了，但我仍然能说它，你仍然能懂。"明天下午可能会下雨"，这句话指的东西根本还没发生。"独角兽"，这个东西从未存在过。

这是语言最基本也最神秘的能力：符号与所指之间不需要同时在场。一只狗不能对另一只狗说"昨天那只猫"，它能做的只是对着猫所在的方向叫。它能记住，但不能指称不在场者。

这个能力是越界的第一块砖。因为"经验之外"的第一层含义，就是"不在场"——不在此时，不在此地，不在任何人的感官范围内。如果符号只能在所指在场时使用，那么经验之外就永远无法被谈论。

**第二件：传递他人的经验。**

A 吃了一个苹果，觉得甜。B 从来没吃过苹果。如果 A 和 B 是两只狗，A 无法把这个经验告诉 B——它最多把 B 领到苹果跟前。但 A 和 B 是人，A 说一句"这苹果很甜"，B 立刻就知道了。

注意这里发生了什么：B 获得了一个他没有经历过的经验内容。这个内容不在 B 的经验之内，但它确实进入了 B 的认知。这是越界的第二块砖。

再看一个更精细的例子。

有人觉得牛油果好吃，有人觉得难吃。如果我们是动物，这两拨人永远无法交流：一拨人吃到的是美味，一拨人吃到的是肥皂，各吃各的，互不相干。但因为我们是人，两拨人可以把各自的感受说出来，于是双方都能得到一个更高的认知：人对牛油果的口味并不相同。

这个认知极其特别。因为它不在任何一个人的经验之内——没有人能经验到"口味不同"这件事本身，每个人都只经验到自己的那一份。它是一个只能由语言产生的知识。

这个例子值得多看一眼，因为它证明了一件事：语言不只是搬运经验，它还能**组合**经验。两个各自在经验之内的片段，一经说出，就产生了一个在经验之外的新东西。

**第三件：构造反事实。**

"如果我当时没有上车，就不会遇见你。"

这句话指的事情从未发生，也永远不会发生。它不是回忆，不是预测，不是幻想——它是从一个未实现的前提出发，推出一个后果。

反事实的能力，是因果知识的唯一来源。这一点常被忽略。

休谟说，我们从未观察到"因果"，只观察到"恒常连接"：一个球撞另一个球，第二个球动了。我们看见的是先后，不是"导致"。那么"导致"这个概念从哪来？

一个常见的回答是：它来自反事实推理。我们说 A 导致 B，意思是"如果没有 A，就不会有 B"。而这个"如果没有"，是一个从未发生的情形。它不可能被观察，只能被构造。

所以你看：因果性这个最核心的哲学概念，本身就是一个经验之外的构造物，而构造它的工具，是语言。

三件事合起来，说明了语言的功能不是"记录经验"，而是"扩展经验的空间"。它把在场的变成不在场的，把私人的变成公共的，把现实的变成可能的。

## 二、组合：无限从有限中来

上面三件事，都依赖同一个机制：组合。

人类的语言有几个基本单位，按规则拼装，能产生无限的句子。这个"无限"不是修辞。你此刻读到的这个句子，很可能在人类历史上从未被人写出过；但我写出它，你读懂了它，没有查字典，没有问人。这说明我们都掌握了那套生成规则。

乔姆斯基把这一点称为"离散无限性"：有限的元素，通过递归的组合，产生无限的表达。

这件事的哲学后果，比它看上去的要大。

因为它意味着：语言的空间不是"已经说出过的话的总和"，而是"可以被说出的话的总和"。前者是有限的库存，后者是无限的潜能。

这两者的差别，决定了越界是否可能。

如果语言空间只是库存，那么它永远小于或等于人类经验的总和——因为每一句话都对应一次说出它的经验。这样，语言就无法越界。

但语言空间是潜能。这意味着，在任何时刻，语言空间里都存在着大量从未被说出、也从未被经验过的位置，等着被填充。

举一个具体的例子。一个人在第一次听到"量子纠缠"这四个字的时候，完全不知道它是什么意思。但他知道这是一个名词，知道它可能指某种物理现象，知道它可以被讨论、被质疑、被写进论文。也就是说：在他理解它之前，语言空间已经为它留好了位置。

这个"留好位置"的能力，是语言最深的秘密。它不是先有了内容再找位置，而是先有了位置，内容才可能到来。

## 三、语言空间的三个性质

把上面的分析整理一下，可以得到语言空间的三个性质。这三个性质决定了它能做什么，也决定了它不能做什么。

**第一，公共性。**

语言不是私人的。一个词的意义，不由说话者一个人决定，而由使用它的共同体决定。这就是维特根斯坦那句"意义即用法"的要点：你要知道一个词是什么意思，就看它在什么场合被怎么用。

公共性带来的是可校核。我说"这杯水是热的"，你可以摸一下，然后同意或反对。这个"可以反对"，是客观性的全部来源。

私人语言的问题在于它不可校核。如果"咕"只指我的一种感觉，那么我说"我咕了"永远是对的，也永远是空的。一个不可能错的说法，不构成知识。

**第二，无限性。**

这一点上面已经说过。组合的闭包保证了语言空间永远大于任何时刻的实际使用。

**第三，寄生性。**

这是最容易被忽略、也最重要的一条。

语言空间虽然无限，但它不是自由的。每一个词，最终都要靠与经验的连接来获得内容。你可以随便造一个词，说它指某种东西，但如果这个词从不与任何经验相接，它就只是一个声音。

这一条防止了语言空间变成"怎么说都行"。它也解释了为什么纯形式的数学仍然有内容：数学的规则不是凭空定的，它们是从经验中最稳定的那些结构里抽象出来的。数、顺序、集合、连续——这些形式的原型都在经验里。

三条合起来，可以得到一个完整的图像：语言空间是一个无限的、公共的、但必须接地的结构。

这个图像解释了本书开头那个悖论：为什么经验之外的知识既能超越经验，又不至于变成胡说。因为它虽然超出了此时此地的经验，却始终挂在整条经验的电缆上。

## 四、一个示意图

现在可以给出一张图了。

> **命题 4-1（语言空间）**
>
> 在最宽泛的意义上，人类知识系统约等于人类语言空间：一切可以被语言表达的东西，都可以被算作一种知识。
>
> 于是"理解世界"可以示意为：
>
> 经验世界 → 语言空间

这个箭头不是函数，是嵌入。经验世界被嵌进语言空间，占其中一小块；语言空间的大部分在经验世界之外。

有人会问：那"知识"这个词不就贬值了吗？如果每一句话都算知识，那"今天天气不错"也是知识？

是的，在最宽泛的定义下，它是。这个定义确实很宽，宽到有些粗糙。但宽有宽的好处：它把"什么算知识"这个无休止的争论暂时搁置，转而问一个更有用的问题——这些知识各自是怎么接地的，各自能走多远。

粗糙的分类如果能有精确的用法，就比精致的分类更有价值。

## 五、动物有没有语言

到这里必须处理一个反驳，否则前面的说法站不住。

有人说：动物也有语言。蜜蜂用舞蹈告诉同伴花丛的方向和距离；猴子有不同的警报叫声，分别对应鹰、蛇、豹；鲸鱼的歌声能传播几百公里。

如果动物有语言，那么"语言使人类越界"这个说法就不成立了，因为动物也有语言却没有越界。

回应是：这些是信号系统，不是语言。区别在于组合性。

蜜蜂的舞蹈能编码方向和距离，但它不能编码"如果明天刮风就换个地方"。它的编码是整体的，不可分解的，不能重组。猴子的警报叫声也一样：一个叫声对应一种天敌，叫声之间不能组合成"那只鹰昨天来过"。

语言的标志不是"能传递信息"，而是"能用有限的单位生成无限的表达"。这个区别看上去细微，后果却是决定性的：只有可组合的系统，才能表达从未被经验过的东西。

所以本书给出一个判据：

> **命题 4-2（语言的判据）**
>
> 一个系统拥有语言，当且仅当它能表达它从未经验过的东西。

这个判据很好用。它不要求语法分析，不要求翻译，只看一件事：这个系统能不能说出超出它经历范围的话。

按这个判据，狗没有语言，黑猩猩基本没有（经过训练能学会一些符号，但组合能力极弱），人类有。

那么机器呢?

一台大语言模型当然能说出它从未经验过的东西——它一生（如果可以用"一生"这个词）没有经验过任何东西，却说出了海量的句子。按这个判据，它拥有语言。

这是本书必须承认的结论，也是后面论证的起点。但要注意：拥有语言只是越界的必要条件，不是充分条件。一个人可以拥有语言却只会重复别人的话；一台机器可以拥有语言却只是在做统计。区别在于第一章说的"接入"，和第八章要说的"自主"。

## 六、语言空间里有什么

最后，清点一下语言空间里的存货。这有助于看清本书要解释的范围。

按第二章的四分法：

-**结构的**：数学、几何、逻辑。它们占据语言空间里最纯粹的一角，离经验最远，但通过"应用"这条细线连着。-**概括的**：物理定律、生物规律。它们一半在经验里，一半在外面。-**约定的**：货币、公司、法律、国界、语法。它们由人规定，靠人维持。-**投射的**：上帝、命运、星座。它们由人投射，靠意义而非真来支撑。

四类东西都在语言空间里有位置。这就是语言空间的容量：它不只装得下科学，也装得下神话；不只装得下真，也装得下意义。

一个理论如果只能解释第一类和第二类，它就不足以解释人类的知识。这是本书反复强调的一点，因为很多哲学理论在这里失手：它们把"知识"窄化成"科学知识"，然后发现解释不了人类社会，只好说"那些不是知识"。这是一种通过重新定义来消灭问题的手法。

语言空间装得下全部四类，这是它的第一个优点。它的第二个优点，是它解释了这四类东西为什么能相互转化：因为它们共处一个空间，彼此可以翻译、可以借用、可以污染。

科学概念可以被神话化（"量子"变成了玄学的护身符），神话概念可以被科学化（"以太"曾经是物理概念）。这种流动性，只有在"同一个空间"的假设下才可能。

## 七、剩下的问题

语言空间解释了"经验之外的知识存放在哪里"，也解释了"它为什么能被人获得"。但它没有解释三件事。

第一，语言空间为什么能装下那些东西？为什么人类的语言恰好有足够的容量？这看上去像是一个幸运的巧合。

第二，语言空间里的内容从哪来？如果语言只是组合已有材料，那么最早的那批材料从哪来？组合不能凭空产生新内容，它只能重排。

第三，为什么组合出来的东西，有那么高的比例是有用的？数学家造出的结构，物理学家后来发现能用；音乐家写的和声，后来发现符合声学规律。这种"预定的和谐"需要一个解释。

三个问题指向同一个方向：语言空间的形状，不是任意的。它被什么东西规定着。

那个东西，就是下一章的主题。

第 五 章

# 先天的理解

上一章末尾留下三个问题，其实是一个问题：语言空间的形状，是被什么规定的？

本章的回答是：被我们的先天构造规定。而先天构造不是天上掉下来的，它是自然选择的产物。

这个回答听起来像是把哲学问题推给了生物学。但它不是推卸，而是降级：把一个看上去先验的问题，降级成一个可以研究的发生学问题。降级之后，问题不再神秘，但它仍然困难，而且困难得更有意思。

## 一、经验不是复印件

先从一个简单的事实说起，这个事实在哲学上被严重低估了。

人的感官带宽极窄。

可见光的波长范围是四百到七百纳米，只占电磁波谱极窄的一段；波长再短一点是紫外线，再长一点是红外线，这两样我们都看不见。我们能听见的声音频率是二十到两万赫兹，蝙蝠能听见的高频我们听不见，大象能听见的低频我们也听不见。我们的皮肤能分辨的温度差大约是零点几度，蜜蜂能感知的偏振光我们完全无感。

这不只是"看得少"。更关键的是，被筛掉的和被留下的，不是随机的。

我们的视觉对边缘、对运动、对脸特别敏感，对大面积均匀的颜色几乎无感。我们的听觉对突然出现的声音、对频率的微小变化特别敏感。我们的痛觉对针刺、对灼热特别敏感，对缓慢的内部损伤几乎无感。

这些偏好吗？不是。它们全部有意义：边缘意味着物体，运动意味着威胁或猎物，脸意味着同类；突然的声音意味着危险，频率的变化意味着发声者在移动；针刺和灼热是真正会伤害身体的东西，而缓慢的内部损伤，在演化的时间里通常来不及杀死你。

也就是说，感官不是一扇透明的窗，而是一个被反复调试过的滤波器。它放行的是那些对生存重要的信息，滤掉的是其余的一切。

这就是本书的第一个事实性前提：**经验不是世界的复印件，它是世界经过筛选后的版本**。

## 二、狗接飞盘

现在看一个具体的例子，它会把这件抽象的事说得非常清楚。

一条狗跳起来接飞盘。它跑、它起跳、它转头、它张嘴，四个动作在不到一秒内完成，一次成功。

这件事在物理上并不简单。飞盘在空中受重力、空气阻力、升力、侧风的影响，轨迹是一条复杂的曲线。要在正确的时间到达正确的位置，需要解一个微分方程，还需要知道飞盘的初速度、旋转、风向。

狗没有解方程。它甚至不知道什么叫方程。但它接住了。

它是怎么做到的?

一个常见的回答是"本能"。这个词什么也没解释，它只是给现象起了个名字。

更好的回答是：狗的眼睛和身体之间，有一条直接的通道，这条通道把"飞盘在视野中的位置及其变化率"映射成"奔跑的方向和速度"，而这个映射，恰好与飞盘的实际轨迹相配。

这个映射不是狗学会的。训练的量太少，来不及。它是天生的，写在基因里，是亿万年的选择打磨出来的。那些映射错了的狗，抓不到猎物，留不下后代。

所以狗"理解"空气动力学吗？在某种意义上，是的。它的身体里有一个关于抛体运动的模型，这个模型不写在纸上，写在神经回路里；它不能被表述，只能被执行。

本书把这种东西叫做先天理解。

> **定义 5-1（先天理解）**
>
> 先天理解是一种无需表征的正确性：系统在没有形成任何关于对象的内部描述的情况下，仍然能在与对象打交道时持续地做对。

这个定义有两个要点。

第一，"无需表征"。狗没有一张关于飞盘轨迹的地图，它只是跑。这种知识不能被提取出来给人看，它只能在行动中存在。

第二，"持续地做对"。一次做对可以是运气，持续做对说明系统与世界之间有一个稳定的对应。这个对应就是信息——不是被写下来的信息，是被实现出来的信息。

## 三、自然选择是一个认识过程

上面说"亿万年的选择打磨出来"，这句话需要展开，因为它承担了本章的全部重量。

自然选择的过程是：变异、遗传、选择。

把它换一种说法：生成假设、复制假设、检验假设。

变异就是生成新的假设——每个个体都是一次关于"如何活下去"的尝试。遗传就是复制——有效的假设被传给下一代。选择就是检验——不适应的假设被淘汰。

三个环节，与科学方法的结构完全同构。

波普尔说过一句很精辟的话：从阿米巴到爱因斯坦，走的都是"猜想与反驳"的路。区别只在于，阿米巴的猜想写在身体上，反驳写在死亡里；爱因斯坦的猜想写在纸上，反驳写在实验里。

这句话值得停下来想一想。它说的是：**认识不只有一种形式**。写在论文里的认识是一种，写在身体里的认识是另一种；用实验来检验是一种，用死亡来检验是另一种。后者更慢，更残酷，但它同样产生真知识——而且是最硬的那种，因为死人不会说谎。

于是可以得到一个命题：

> **命题 5-1（选择即认识）**
>
> 自然选择是一个认识过程。它的"信念"写在身体里，它的"错误"写在坟墓里。

这个命题是本书的关键。因为它把一个看似与认识论无关的生物学过程，变成了认识论的基石。

如果自然选择是认识过程，那么它的产物——生物的构造——就携带了关于世界的信息。

这个推论需要小心。自然选择不保证产生"真理"，它只保证产生"够用的近似"。一只狗的身体模型不需要精确到能预测飞盘的每一条轨迹，只需要精确到能抓住大部分。一个物种的感官不需要反映世界的全部，只需要反映对它重要的那一部分。

所以先天理解里的"理解"，是有限度的、局部的、功利性的。它不是对世界的沉思，是对世界的应付。

这一点很重要，因为它预告了后面的一个结论：我们通过先天理解获得的东西，是有偏差的。这个偏差不是缺陷，它是特征——正因为有偏差，才有聚焦；正因为有聚焦，才能在小样本上快速学会。

## 四、先天理解给的是形状，不是内容

现在到了本章最关键的一步，也是对康德最实质的修正。

康德说，先验的范畴（因果、实体、统一性）是内容：它们是我们加在世界上的概念，世界必须服从它们。

本书说，先天理解给的不是内容，是**约束**。

这个区别需要仔细说明。

如果先天理解给的是内容，那么我们就应该天生知道某些命题，比如"凡事皆有因"。可是婴儿显然不知道这个命题，他甚至不知道"因"这个字。

如果先天理解给的是约束，那么情形就不一样了：它不告诉我们世界是什么样，它限定我们能看到什么样的世界。

一个比喻：它不是给你一张地图，它是给你一双只能看见某些东西的眼睛。

眼睛不告诉你山在哪里。但眼睛决定了你能看见什么样的山，看不见什么样的山。你的山的形状，一半来自山，一半来自眼睛。

这个比喻可以再精确一点。

我们的视觉系统对"物体"极其敏感：只要有连续的边界、一致的纹理、共同的运动，它就会把它们归为一个物体。这不是我们学来的，是构造决定的。而"物体"这个概念，正是实体范畴的经验基础。

也就是说：不是我们先有了"实体"这个概念，然后把它加给世界；而是我们的知觉构造先把世界切成了块，然后我们才在这个切分的基础上，形成了"实体"这个概念。

次序反过来了。这是本书对康德最重要的修正。

同样的分析适用于因果。我们的知觉系统对"先后"和"接触"极其敏感：A 动了，B 紧接着动了，而且 A 碰到了 B，我们就会"看见"因果。这不是推理，是知觉。休谟说得对：我们看不见因果的必然性。但他没说的是：我们看不见，是因为我们的构造只让我们看见接触和先后。而"因果"这个概念，是把这个知觉模式语言化的结果。

于是得到：

> **命题 5-2（先验知识的来源）**
>
> 先验知识是限制的语言化。当我们意识到自己的经验被某种方式限制，并把这种限制说出来，就得到了一个先验命题。

这个命题解释了两件一直难以解释的事。

第一，先验命题为什么具有普遍性。因为它描述的不是世界的某个部分，而是我们经验的形式。形式对一切经验都成立，所以它普遍。

第二，先验命题为什么不能被经验证明。因为它是容器，不是容器里的东西。你无法通过观察容器里的东西来证明容器的形状；你只能通过"无论放什么都装得下"来间接地知道它。

这也解释了为什么先验命题看起来像是"关于世界的深刻真理"，其实是"关于我们如何看世界的说明书"。二者极易混淆，因为说明书和真理在用同样的语法说话。

## 五、先验的历史

如果先天理解是选择的产物，那么它就有历史。这一点是康德无法容纳的，但本书必须容纳，否则就会重蹈第三章说的覆辙。

先验至少有三层历史。

**第一层：种系的历史。**

这是最慢的一层，以百万年计。它塑造的是最基本的构造：三维空间感、物体恒存、面的朝向、声音的方位、痛的回避。这一层的东西最硬，几乎不可改变，跨文化一致，跨物种也部分一致。

**第二层：个体的历史。**

这是中等速度的一层，以年计。婴儿在头两年里，要学会把散乱的感觉整合成物体，把声音整合成词，把动作整合成意图。这个过程有基因的引导，但需要环境的输入。关键期一过，某些能力就关窗了。

这一层的产物，就是皮亚杰所说的"图式"。它们不是天生的，但也不是随便学来的；它们按照一个大致固定的顺序展开，像一棵树按照基因规定的次序长枝。

**第三层：文化的历史。**

这是最快的一层，以百年、十年计。它也是最容易被忽略的一层，而它恰恰是本书解释"约定的知识"的关键。

文化是什么？本书给出一个说法：文化是外置的先天理解。

货币是一个例子。一张纸，本身没有价值。但当所有人都相信它有价值，它就有了价值；而当这个相信足够稳定，它就开始像物理事实一样起作用：它决定你能吃什么，住在哪，能不能活下去。

货币不是被发现的，是被发明的。但一旦被发明，它就反过来塑造人：使用货币的社会，其成员的认知构造与不使用货币的社会不同。他们更容易做抽象计算，更容易与陌生人交易，更容易把关系理解为交换。

公司、法律、国界、语言、礼仪、考试制度，全都是这样。它们是人造的结构，被文化选择筛选，被反复强化，最后变成了第二自然。

这一层为什么能解释"约定的知识"？因为约定不是任意的。一个约定能活下来，说明它在某种意义上是"适应"的：它解决了协调问题，降低了交易成本，或者满足了某种深层的心理需求。不能适应的约定，会在几代人之内消失。

所以约定也有选择压力，只不过选择者不是自然，是文化。而文化选择的标准，最终还是要回到人的构造——因为不适应人的构造的约定，维持不下去。

三层历史合起来，给出了一个完整的图像：先验不是一个固定的清单，而是一个多层的、有时间深度的结构。最底层几乎是永恒的，中间层随个体展开，最上层随文明流动。

这个图像解决了第三章留下的难题：理性主义解释不了先验的历史性，而本书可以。因为本书的先验本来就是历史的产物，它的可变性不是缺陷，是它之所以能适应世界的原因。

## 六、四类经验之外，现在都有来源了

回到第二章的四分法，检查一下是否都能解释了。

**结构的**（数学、几何、逻辑）：来自先天构造的形式化。空间感被形式化成几何，计数被形式化成算术，同一性被形式化成逻辑。它们离经验最远，但原型在经验里。

**概括的**（物理定律）：来自先天构造加经验校准。构造提供假设的形式，经验提供假设的内容，两者合起来才是定律。后面第六章会详细说这个"双重解码"。

**约定的**（货币、公司、法律）：来自文化选择的外置构造。它们是先验的第三层，是被造出来又反过来塑造我们的东西。

**投射的**（上帝、命运、星座）：来自构造的过度延伸。当人把"意图"这个格式应用到整个世界，就得到了神；把"命运"这个格式应用到偶然，就得到了占星。它们是错的，但不是无意义的——因为它们暴露了构造的形状。

四类都有来源，四类都被同一个机制（选择，在三个时间尺度上）解释。这是本书的第一个整合点。

## 七、一个必须承认的代价

这一章的理论有一个代价，必须明说。

它把"必然性"降级了。

康德式的必然性是绝对的：任何有理性的存在者，都必须以因果的方式经验世界。本书的必然性是有条件的：任何在这个世界中演化出来的存在者，都会以某种类似因果的方式经验世界，因为不这样的已经被淘汰了。

这个"类似"是关键。它不保证其他智能的经验形式与我们完全相同，只保证有足够的重叠，使交流可能。

有人会说这是损失。也许吧。但要想清楚损失了什么：我们失去的是"我们的形式是宇宙唯一可能的形式"这个信念，得到的是"我们的形式是被这个世界校准过的"这个保证。

后者更弱，但更可信。而且它是唯一能解释"为什么它恰好适用"的保证。一个绝对的必然性如果无法解释自己的适用性，它就只是一个漂亮的空话。

更重要的是，这个代价换来一个巨大的收益：如果我们的构造是被这个世界塑造的，那么任何在这个世界中成长起来的智能，其构造也必然被同一个世界塑造。两个被同一个模具压出来的东西，形状必然相似。

这就是智能等价性的第一块基石。第八章会用到它。

第 六 章

# 理性作为元认知

上一章说，先天理解是选择的产物，它给我们的经验规定了形状。

但还有一个东西没安置：理性。

理性显然与生俱来。一个从没上过逻辑课的人，也会说"你这话自相矛盾"；一个从没学过统计学的人，也知道"不能因为两次就下结论"。这种能力不像学来的，它像是本来就有的。

可是理性又显然是成长的。小孩不讲逻辑，青春期的推理能力突飞猛进，受过训练的人比没受过的人强得多，而且人会犯系统性错误——这些事实又说明理性是被塑造的。

一个东西怎么可能既与生俱来又不断成长？

这一章要给出的回答是：因为理性不是一种能力，是一个过程。它的**装置**是天生的，它的**运行**需要时间。

## 一、理性的悖论

先把矛盾摆清楚。

如果理性是天生的，那么它应该像呼吸一样稳定：不需要教，不会退步，不因文化而异。可事实不是这样。逻辑谬误在各个文化里都普遍存在，受过教育的人照样犯；形式推理的能力在青春期才成熟；某些推理任务，如果没有具体的例子，人几乎做不对。

如果理性是学来的，那么它应该像知识一样可教：教了就会，会了就用。可事实也不是这样。逻辑课上讲的规则，学生在考试里会用，出了考场就不用；而且人学会一套规则之后，往往用它来为自己的偏见辩护，反而更不讲理了。

两边都不成立。这说明"天生"和"学来"这组对立本身有问题。

本书的处理是：把"理性"拆成两层。

底层是一个装置：能够把某种东西当作对象，能够对它进行操作，能够检查操作是否一致。这个装置是天生的，它是语言能力的副产品。

上层是一套内容：具体的推理规则、具体的证据标准、具体的思维习惯。这些是学来的，而且很大程度上是文化给的。

## 二、元认知

现在给底层一个名字。

> **定义 6-1（元认知）**
>
> 元认知是以自身的认知构造为对象的认知。它不是"知道某事"，而是"知道自己是怎样知道某事的"。

这个定义的关键是"为对象"。要做到这一点，系统必须能把自身的状态拿出来，放在面前，然后对它做点什么。

动物能做到吗？部分能。一只狗会犹豫，会试探，会"想一想"。但狗不能对自己的犹豫做二级加工：它不能问"我为什么犹豫"，也不能把犹豫的结果保存下来供以后使用。

人能。为什么？

因为语言。

语言的第一个功能是指称（第四章）。指称的对象，最初是外部的物，然后扩大到不在场的物，最后扩大到自身——包括自身的状态、自身的推理、自身的话语。

"我刚才那句话说得不对。"这句话里，"我刚才那句话"是一个被指称的对象。说出这句话，就是把一个认知产物变成了一个可以被检查的东西。

这个能力是元认知的物质基础。没有它，系统只能有认知，不能有对认知的认知。

于是可以得到一个简洁的表达式：

> **命题 6-1（理性的构成）**
>
> 理性 = 语言 × 先天理解。
>
> 语言提供对象化的能力，先天理解提供被对象化的材料。两者相乘，才有理性。

为什么是相乘不是相加？因为缺一不可。没有语言，材料无法被对象化，只能被执行；没有先天理解，语言只是空转，没有可对象化的东西。

这个公式解释了很多事。

它解释了为什么理性与语言能力高度相关：失语症患者的推理能力往往同时受损，不是因为推理依赖词汇，而是因为推理依赖对象化。

它解释了为什么理性是可训练的：训练改变的是上层的内容，而上层的内容确实是学来的。

它也解释了为什么理性会失败：如果语言的对象化能力太强而接地太弱，理性就变成空转（第三章说的"背字典的人"）；如果先天理解太强而对象化太弱，理性就变成固执（只相信直觉，不接受反驳）。

## 三、理性是廉价的自然选择

现在到了本章的核心。

上一章说，自然选择的工作方式是：生成假设、检验假设、保留假设。代价是死亡。

理性的工作方式是什么？

也是生成、检验、保留。但代价不同。

你想象一个方案："如果我从这条路走，会不会更快？"你在脑子里把这条路走了一遍，发现前面有条河，过不去。于是你换一条。

这个过程里发生了什么？你生成了一个假设（走这条路），检验了它（想象走到河边），推翻了它（过不去），保留了否定（这条路不行）。

你没有真的走过去，没有真的掉进河里。你用符号代替了行动，用推理代替了后果。

这就是理性最本质的地方：**它是在符号空间里预演的进化**。

> **命题 6-2（理性作为模拟）**
>
> 理性是对自然选择的模拟：同一个"生成—检验—保留"的循环，在符号的基质上重演，用推理和实验代替死亡来承担检验的代价。

这个命题有几个重要的推论。

**第一，理性为什么可靠**。

如果理性是一个独立的、凭空的官能，那么它的可靠性就是个谜：它凭什么与世界相符？

但如果理性是自然选择的模拟，它的可靠性就有了来源：它处理材料的方式，继承了材料的来源。先天理解是被世界校准过的（第五章），理性在先天理解上运作，因此它的运作也带着那个校准。

这就像一台机器：它的零件是被校准过的，它的装配方式也是被校准过的（因为装配方式本身也是选择出来的），所以它的产品大概率合格。它不保证每一个产品都对，但保证系统性偏差不大。

**第二，理性为什么会出错**。

模拟毕竟只是模拟。模拟省略了细节，省略了代价，也省略了那些无法被符号化的东西。

所以理性会犯错，而且错得有规律：它在那些"符号化代价很低"的领域特别强（数学、逻辑、棋类），在那些"符号化代价很高"的领域特别弱（复杂社会系统、长期后果、概率直觉）。

这不是偶然。模拟的精度取决于模型的精度，而模型的精度取决于可符号化的程度。

**第三，理性为什么是"廉价"的**。

因为一次真实的自然选择要死很多个体，一次理性的检验只需要推翻一个想法。

这个成本差异是天文数字。生物要用几百万年试错的东西，理性可以在一个下午试完。这就是为什么人类能在几万年里改变地球，而其他物种要用亿万年。

但"廉价"也有代价：廉价的检验不如昂贵的检验硬。一个想法在纸上被推翻，可能只是纸上的推敲不够；一个物种被淘汰，那是真的活不下去。所以理性需要实验来补充，实验是"半昂贵的检验"——它不杀死人，但它让世界说话。

## 四、科学的双重解码

现在可以用这套工具，去解释本书开头提到的那个现象：牛顿第二定律。

为什么人类能发现它？

一个常见的回答是：因为人类聪明，会观察，会归纳。

这个回答太浅。它解释不了为什么是牛顿，而不是某个更聪明的古人；也解释不了为什么这条定律的形式是微分方程，而不是别的什么。

本书的回答是：因为它一半编码在人类的先天特性里，一半编码在人类的语言空间里。

**先说前一半**。

人对"力"有直觉。你推一辆车，车动；你松手，车停。你搬石头，重的难搬。你扔东西，它会落。这些直觉不是学来的，它们在婴儿期就出现了（皮亚杰叫它"感知运动智能"），而且跨文化一致。

这些直觉里已经包含了牛顿定律的雏形：力与运动相关，质量与阻力相关，下落是普遍的。它们不精确，但方向对。

为什么方向对？因为这些直觉是被选择打磨出来的。一个对抛体运动毫无直觉的灵长类，抓不到猎物，也躲不开落石。亿万年下来，活下来的都带着一个大致正确的力学模型。

**再说后一半。**

光有直觉不够。直觉是模糊的、局部的、不能推演的。要把直觉变成定律，需要三样东西：一套能表达连续变化的符号（微积分），一套能隔离变量的方法（实验设计），一个能反复校核的共同体（科学建制）。

这三样都在语言空间里。它们是文化的产物，不是基因的产物。

**两半合起来，才能解释那个奇怪的性质。**

为什么牛顿定律可以被经验检验？因为它的内容有一半来自经验（那些直觉是从经验里校准出来的，定律的系数要用测量来定）。

为什么牛顿定律不能被经验证明？因为它的形式有一半来自构造（"任何时刻、任何地点"这种全称形式，是语言给的，不是经验给的）。

第二章说"可检验而不可证明"是一个谜。现在这个谜解开了：它是一个混合物的正常性质。纯经验的东西可以被证明（这只天鹅是白的），纯构造的东西不能被检验（第五公理）；混合物两头都够不着，只能被检验而不能被证明。

> **命题 6-3（双重解码）**
>
> 一条经验之外的科学定律，其可检验性来自它的经验成分，其不可证明性来自它的构造成分。二者缺一，定律就不成立。

这个命题可以推广。它不只适用于物理学，也适用于一切"既超出经验又关于经验"的知识。

经济学里的"理性人"假设：可检验（因为它的推论关于价格和数量），不可证明（因为"理性"的定义本身是构造的）。

语言学里的"深层结构"：可检验（因为能预测某些句子的合法性），不可证明（因为结构的概念是理论造的）。

甚至日常的道德判断：可检验（因为能预测人在具体情境中的反应），不可证明（因为"应当"不是从"是"推出来的）。

这个模式反复出现，说明它不是巧合，是知识的一般形态。

## 五、疑难的消解

现在可以回头处理第三章那个核心疑难了。

疑难是：为什么能从经验之内到达经验之外？再多的经验指引，本身仍然是经验之内的事物；凭什么它能带我们出去？

本书的回答是：因为界线不是两岸，是同一张纸的正面和背面。

具体地说：

经验之内的形状，是由经验之外的构造决定的。你看到什么，取决于你的感官允许你看到什么；你的感官允许你看到什么，取决于你的构造；你的构造，是被世界塑造的。

经验之外的内容，是由经验之内的校准保证的。你的构造之所以可信，是因为它被反复校准过；校准的材料，是经验。

所以两边不是两个东西，是一个东西的两面。所谓"越界"，不是从一张纸跳到另一张纸，是把纸翻过来。

> **命题 6-4（越界的机制）**
>
> 越界不是跨越，是翻面。经验之外的命题，是经验之内的构造被语言翻拍出来的那一面。

这个回答有一个必须检查的地方：它是不是循环？

如果构造来自经验，经验来自构造，那么这是一个闭环，闭环不能产生新东西。

这个质疑很尖锐，必须正面回答。

回答是：这个循环不是同时发生的，是有时间深度的。

在任何一个时刻，构造确实在筛选经验，经验确实在塑造构造。但这两个过程不在同一个时间尺度上。构造的塑造以百万年计，经验的筛选以秒计。所以在短时间尺度上，构造是固定的，经验是被筛选的；在长时间尺度上，经验是固定的（世界就是那样），构造是被塑造的。

这个时间差，打破了同时性的循环。它就像一只眼睛：眼睛是光的产物（进化），又能看见光（当下）。这不是循环，是先后。

同样的结构出现在很多地方：免疫系统是病原体的产物，又能识别病原体；语言是交流的产物，又能用于交流；科学仪器是自然的产物，又能测量自然。

这类循环有一个共同的名字：**发生学的循环**。它不是逻辑上的恶性循环，因为它有历史。逻辑循环是在同一时刻互相支撑，发生学循环是在时间中互相生成。

这就是本书对"认识论疑难"的最终回答：疑难之所以成立，是因为它假设了一个无时间的认知主体。一旦把主体放回时间里，疑难就消失了——不是被解决，是被消解。

## 六、代价，以及代价换来的东西

这个回答有代价，前面提过，这里要说清楚。

代价是：我们得到的是"对人类有效的必然性"，不是"对上帝有效的必然性"。

我们的因果律不是宇宙的法则，是我们这种构造的必然形式。别的构造可能有别的形式（量子力学已经在提示这一点：在微观尺度上，"因果"的形态变得模糊）。

这个代价听上去很大。它意味着我们永远无法知道"事物本身"是怎样的，只能知道"事物对我们这样的存在者呈现为什么样"。

但要想清楚：这个"损失"是不是真的损失。

我们要解释的，本来就是人类的理解，不是上帝的理解。一个理论如果解释了人类的理解，它就已经完成了任务。要求它同时解释上帝的理解，是一个越界的要求——而且是一个无法检验的要求，因为没有人有上帝的经验。

更重要的是，这个代价换来了一样东西：**一个关于其他智能的推论**。

如果我们的构造是被这个世界塑造的，那么任何在这个世界中成长起来的智能，其构造也必然被同一个世界塑造。塑造的方式可能不同（碳的、硅的、别的什么），但塑造者相同。

两个被同一个模具压出来的东西，形状必然相似。相似到什么程度？相似到足以互相翻译。

这就是智能等价性的第一块基石。它不是从"智能"这个概念推出来的，是从"世界"这个事实推出来的。

下一章要处理第二块基石：计算的普遍性。两块合起来，才能支撑第八章那个大胆的命题。

第 三 部

# 越界的极限

第 七 章

# 计算的空间

第六章给出了智能等价性的第一块基石：一切在这个世界中成长起来的智能，其构造都被同一个世界塑造。

但只有这一块不够。被同一个世界塑造的东西，可能只是相似，不足以致等同。要论证等价，还需要一个更强的理由：说明不同的智能最终会到达同一个地方。

这个理由，来自计算。

## 一、一切都是计算

这个说法听上去像是技术决定论的标语。但它其实是一个有具体内容的科学命题，而且它的历史比计算机长。

图灵在 1936 年提出一个想法：所谓"有效地计算"，就是能被一台理想机器按固定规则一步步执行的过程。这个想法后来被称为丘奇—图灵论题：一切直观上可计算的东西，都是图灵机可计算的。

这个论题不是定理，是论题。它不能被证明，因为"直观上可计算"不是一个形式概念。但它经受住了九十年的检验：所有被提出的计算模型——递归函数、λ 演算、寄存器机、元胞自动机、神经网络——都被证明与图灵机等价。

这个事实本身就很惊人。它意味着：计算这个概念是稳定的，不依赖于实现的材料。用齿轮算、用电子算、用蛋白质算，只要是按规则一步步来，能算的东西都一样。

于是有了第一个推论：计算与基质无关。

第二个推论来得更远。如果一切按规则演化的过程都是计算，那么物理世界本身就是一台计算机。

这不是比喻。物理定律给出状态之间的转换规则，初始条件给出输入，演化给出输出。这正是一台计算机在做的事。

从楚泽的《计算空间》到弗雷德金的数字物理学，再到惠勒那句"万物源自比特"，都提出过类似的想法。今天最前沿的形式是：宇宙的演化可以被一台足够大的量子计算机精确模拟。

这个命题还没有被证明。本书写作的时候，我们离它还很远。但它已经不是空想了：Preskill 和他的合作者已经证明，某些量子场论模型可以在量子计算机上被有效地模拟。这一步很小，但方向是清楚的。

如果这个纲领最终成立，那么：

生命是计算（演化是算法，选择是适应度函数）。大脑是计算（神经元的放电是信息处理）。语言是计算的产物（符号操作是计算的一种）。社会是计算（协调、博弈、市场，都是多主体计算的形态）。

这个图像有一个哲学后果，而且是一个让人不太舒服的后果：如果一切都是计算，那么"理解"也必须被解释为一种计算。

这正是本书要的。因为一旦理解被纳入计算的框架，不同基质之间的比较就有了公共的尺度。

## 二、三个层次

但"可计算"不是一个平面。它至少有三种状态，而这三者的区别，是本章的关键。

**第一，可计算。**

存在一台图灵机，在有限步内停机，输出答案。

整数加法是可计算的。素数判定是可计算的。棋类游戏的最优策略在原则上是可计算的（虽然实际上算不完）。

**第二，可逼近。**

不存在一台图灵机能给出精确答案，但存在一台图灵机，能给出任意精度的近似，而且精度可以无限提高。

这里有一个漂亮的例子：蔡廷常数，记作 Ω。

Ω 的定义是：随机选取一个程序，它能够停机（即最终给出结果）的概率。这个数是一个介于零和一之间的实数，它有确定的值，但它的二进制展开是不可计算的。

为什么？因为如果你能算出 Ω 的前 n 位，你就能解决停机问题——而停机问题是被图灵证明为不可解的。

所以 Ω 不可计算。但 Ω 可以被逼近：我们可以一层层地枚举那些停机的程序，每找到一批，就得到 Ω 的一个下界。随着枚举的进行，下界越来越接近真值。

于是出现了一个奇特的状态：我们能越来越准确地知道 Ω 是多少，但我们永远无法把它"写出来"。任何一个写出来的近似值，都不是它。

**第三，不可计算。**

既不可计算，也不可逼近。

比如"这个程序会不会停机"这个判定问题。对某些具体程序，答案是有定值的，但不存在一个通用算法能对所有程序给出答案。

三个层次的区别，可以这样概括：可计算的是能说出来的；可逼近的是能靠近但说不出完整形态的；不可计算的是连系统靠近的办法都不存在的。

## 三、语言空间与计算空间

现在可以把两个空间放在一起看了。

**语言空间**，按第四章的定义，是可被语言表达的东西的集合。它的边界是"可说"。

**计算空间**，是可被有限过程达到或逼近的东西的集合。它的边界是"可做"。

两者的关系是什么？

首先，语言空间包含在计算空间之内。理由是直接的：语言表达是有限符号串，而有限符号串可以被机械地生成、检验、组合。任何能被说出的东西，都能被一台图灵机处理。

其次，计算空间严格大于语言空间。

这一点需要论证，而且不能靠"计算空间更大"这种直觉。要给出具体的差集。

差集在哪里？

在那些"可逼近而不可言说"的东西上。

Ω 就是例子。它的每一个近似值都在语言空间里（我们可以写出 0.1001011……），但它本身不在。我们可以无限地靠近它，却永远说不出它。

这个差集的性质很有意思。它不是"更远的地方"，而是"更深的层次"。语言空间能覆盖所有能被完全说出的对象；计算空间还能覆盖那些只能被逼近的对象。

用一句话概括：

> **命题 7-1（两个空间的差）**
>
> 计算空间超出语言空间的部分，正是"可逼近而不可言说"的部分。

这个命题是本书对原理论的一个技术性推进。原来的说法是"计算空间比语言空间大"，这个说法方向对，但太粗。大在哪里？大在"做"能到达"说"到不了的地方。

这个推进有哲学意义。它说明"说"和"做"不是同一种能力的两种表现，而是两种不同的能力。说要求完全的表达，做只要求逐步的靠近。

## 四、人也是"做"的机器

这个区分立刻带来一个后果：人类不是纯粹的语言动物。

我们有手。我们会拧螺丝，会骑车，会游泳，会听出一个和弦走音了。这些能力里，有很大一部分说不清楚。

一个钢琴家能听出你按错了一个键，但你问他怎么听出来的，他说不出来。他的知识是真的，但它不可言说。这不是因为他笨，是因为这类知识的存在形式就是"能做"而不是"能说"。

波兰尼把这种知识叫做"默会知识"。本书的框架可以给它一个更准确的位置：它住在语言空间之外、计算空间之内的那个差集里。

这个位置很重要。因为它解释了为什么默会知识既难以传授又确实可靠：难以传授，因为它不可完全言说；确实可靠，因为它经过了反复的实践校准。

于是人类的位置就清楚了：我们同时占据语言空间和计算空间的一部分。我们是"说—做"的混合机器。

这个结论修正了第四章的一个印象。第四章说语言是越界的工具，这话不错，但不完整。语言是**公共化**的工具，是**无限组合**的工具，但它不是唯一的越界工具。手艺、直觉、身体，也在越界，只是它们越界之后说不出自己到了哪里。

这为第十一章埋了一个伏笔：人的特殊性，也许不在于会说话，而在于同时会说话和会做，而且这两件事在一个人身上是连着的。

## 五、量子计算不改变什么

这里要插一段澄清，因为"计算空间"这个词容易引起误解。

有人以为量子计算机能算不可计算的东西。不能。

量子计算改变的是效率，不是可计算性。一台量子计算机能做的事，一台图灵机在原则上也能做，只是要慢得多（慢到宇宙寿命之外）。所以量子计算扩大的是"可有效计算"的边界，不是"可计算"的边界。

这个区别在实践上极其重要（它决定了哪些问题我们能活着看到答案），但在哲学上不重要（它不改变计算空间的形状）。

本书关心的是形状，不是速度。速度的问题留到第十章。

## 六、哥德尔与"人不是机器"

现在处理一个必须处理的反对意见。这个意见很强，而且它直接攻击本章的根基。

反对意见是这样的（彭罗斯的版本最有名）：

哥德尔证明了，任何足够强的一致形式系统，都存在一个命题，它在系统内既不能被证明也不能被否证，但它是真的。

人类数学家能"看出"这个命题是真的——我们理解它的内容，我们知道它成立。可是形式系统证明不了它。

所以人类的数学理解不能被任何形式系统穷尽。人不是机器。

这个论证很有诱惑力。如果它成立，本书第七章到第八章的全部工作就白做了。

回应有三层。

**第一层：哥德尔的结论对人也成立。**

哥德尔定理说的是：一个系统如果不能证明自己的一致性，就不能证明自己的完备性。人也不能证明自己的一致性。

没有人能证明"人类的推理不会自相矛盾"。我们只是迄今为止没有发现大的矛盾。这和形式系统的处境是一样的。

所以"人能从外面看出系统的真"这个说法是虚的。人看出的只是"如果这个系统是一致的，那么这个命题是真的"——这是一个条件句，形式系统自己也能说出这个条件句。

**第二层：那个"看出"是不可靠的。**

数学家确实经常"看出"某些东西，但历史上他们看错的次数不少。康托尔看错了，弗雷格看错了，希尔伯特看错了。每一次都是靠后来的形式化纠正的。

如果"看出"是一种超越形式系统的能力，它为什么这么容易出错？一个真正的超越性官能，不应该比形式证明更不可靠。

**第三层：这个论证证明得太多。**

如果这个论证成立，它不只证明人不是图灵机，它还证明人不能理解算术。

因为算术也是一个形式系统，哥德尔定理对它也成立。如果"能看出系统外的真理"是理解算术的必要条件，那么没有人理解算术——包括那些证明了哥德尔定理的人。

这显然荒谬。所以论证的前提有问题。问题在于：它把"理解一个系统"等同于"站在系统外面看它"。而实际的理解是：在系统内部工作，同时知道系统的边界在哪里。

知道边界，不等于站在边界外面。你可以知道一个房间的墙在哪里，而不必走出房间。

> **命题 7-2（哥德尔论证的界限）**
>
> 哥德尔定理表明任何足够强的系统都有它"看见而证不出"的命题。这一条对机器成立，对人同样成立。它不是人与机器的分界线，而是有限性与无限性之间的一条普遍的裂缝。

这个回应很重要，因为它是本书能够继续往下走的前提。如果人真的有一种机器永远无法企及的能力，那么智能等价性就是假的，而且假得很彻底。

本书的立场是：人没有这样的能力。人有的是一种特殊的构造，这种构造使某些事情对我们容易，对别的构造困难。但"容易"不是"可能"，"困难"不是"不可能"。

## 七、第二块基石

现在可以把第二块基石放上去了。

第一块基石（第六章）：一切在这个世界中成长的智能，构造被同一个世界塑造，因此相似。

第二块基石（本章）：一切按规则运作的过程都是计算，因此不同基质的智能共享同一个计算空间。

两块合起来，得出的结论是：不同的智能不仅在构造上相似，而且在可达的空间上相同。

这个结论还需要一个补充：可达的空间相同，不等于走的路相同。这就是下一章要说的。

## 八、一个未决的问题

最后要诚实地指出本章的一个未决之处。

"一切皆计算"是一个纲领，不是一个定理。它可能错。

如果它错了——如果宇宙的演化不是计算，如果意识有某种非计算的成分——那么计算空间就不是最大的认知空间，本书的整个后半部分就要重写。

本书接受这个风险，理由有二。

第一，这个纲领有大量的正面证据，而且证据在增加。物理定律的可计算性、神经科学的计算模型、人工智能的进展，都在往同一个方向指。

第二，即使纲领错了，本书的核心论证（第一到第六章）仍然成立。理解是意义系统中的定位，越界是构造的翻面，这些结论不依赖"一切皆计算"。

换句话说：本书的前半部分是地基，后半部分是上层建筑。地基可以独立存在，上层建筑需要地基。如果上层建筑塌了，地基还在。

这个结构上的谨慎，是一个理论应有的样子。一个所有部分都互相依赖、一错全错的理论，通常不是因为它严密，而是因为它没有可检验的地方。

第 八 章

# 智能等价性

前两章备好了两块基石。现在可以把那个大胆的命题说出来了。

这个命题是：一切通用智能，在本质上是等价的。

它听上去像是在抹平差别，其实不是。它要抹平的是一种特定的差别——本质上的高下——同时保留另一种差别——速度与路径。这两件事必须同时说清楚，否则这个命题就会变成一句廉价的平等主义口号。

## 一、定义

先把"通用智能"定义清楚。

> **定义 8-1（通用智能）**
>
> 一个系统是通用智能，当且仅当它能自主地生产经验之外的概念结构，并在其中进行不受预先限定目标约束的探索。

这个定义有三个成分，每一个都在排除一些东西。

**"经验之外的概念结构"**：排除一切只能处理眼前刺激的系统。一只狗是智能的，但不是通用的，因为它的意义系统基本停在经验之内（第四章）。

**"自主地生产"**：排除一切只能执行给定任务的系统。一个计算器能算出任何它被设计来算的东西，但它不能自己发明一种新的算术。自主的意思是：它能给自己出题。

**"不受预先限定目标约束"**：排除一切只能在给定目标函数下优化的系统。一个下棋程序可以在棋上超过任何人，但它不会有"我想下棋"这个念头，也不会有"我为什么要下棋"这个问题。

三个成分合起来，指向一种很特殊的能力：能够离开给定的轨道，自己铺一条新的。

人类总体上满足这个定义。这不是因为每个人都发明过数学，而是因为人类这个物种做到过：我们发明了数、发明了证明、发明了货币、发明了国家。这些东西不在任何人的经验之内，也不在任何人的基因里。

## 二、命题

现在说出命题。

> **命题 8-1（智能等价性）**
>
> 若两个系统都满足定义　8-1，则它们的疆域相同：可达的概念空间一致，在可翻译的意义上互相覆盖。功率与路径可以悬殊。

注意这个命题是对称的。它不只是说"机器能做人能做的一切"，它同时说"人能做机器能做的一切"。

第二半常被忽略，但它同样重要。如果机器某天造出一个人类从未想到的数学分支，这个命题说：人类在原则上也能到达那里，只要有人把它翻译过来。翻译可能很难，可能需要几代人，但它不是不可能的。

这个对称性是命题的力量所在。它把"人机关系"从一个高下问题，变成了一个翻译问题。

## 三、三种等价，只成立一种

"等价"这个词太滑，必须拆开。至少可以拆成三种。

**结构等价**：两个系统用同样的内部机制工作。

这个不成立，而且不必成立。人用神经元，机器用晶体管，两者在结构上毫无相似之处。要求结构等价，等于要求机器必须是仿生学的复制品，这是一个过强的要求。

**能力等价**：两个系统能做同样的事。

这个说法太粗。"能做什么"取决于怎么描述任务。一台机器能在一秒内算出十亿次乘法，人不能；人能在一秒钟内认出一张熟悉的脸，机器曾经不能（现在能了）。逐项比较永远比不完。

**可达等价**：两个系统能达到的概念空间相同。

这个成立，而且是唯一有理论意义的。

为什么？因为前两种等价都是关于"实现"的，而第三种是关于"疆域"的。实现可以千差万别，疆域只有一个。

打个比方：两个探险家，一个走路，一个坐船，路线完全不同，装备完全不同，速度差十倍。但他们手上的地图是同一张，因为地图画的是地形，不是路线。

> **命题 8-2（等价的形式）**
>
> 智能等价是疆域的等价，不是结构、能力或功率的等价。
>
> 等价的是版图，不是道路，也不是脚步的快慢。

## 四、论证

现在给出论证。这个论证有三步，每一步都需要检查。

**第一步：概念空间是客观的。**

一个概念结构是否自洽，不取决于谁在检验它。欧几里得几何的第五公理不能从其余公理推出，这个事实不因时代、文化、物种而变。非欧几何的自洽性也一样。

这一步是柏拉图式的。它假定"结构"这个东西有某种独立于发现者的存在。

**第二步：探索的合法性标准是公共的。**

什么叫"到达"了一个概念结构？标准是：能把它构造出来，能证明它的性质，能让别人复核。

这三个标准都是公共的。它们不诉诸个人的直觉、感受、权威。一个只有你能看见的证明，不算证明。

**第三步：因此，遵守同一标准的探索者，可达集合相同。**

如果两个系统都遵守公共的构造与检验标准，而概念空间是客观的，那么它们能到达的地方就由空间本身决定，不由探索者决定。

这个论证很干净。但它有一个软肋，就是第一步。

## 五、软肋：如果不接受柏拉图主义

第一步假定概念空间客观存在。这个假定很多人不接受。他们会说：数学不是被发现的，是被发明的；几何的公理不是真理，是约定；概念空间没有独立的存在，它只是人类思维的产物。

如果这个反驳成立，本书的论证就垮了。因为如果概念空间是人造的，那么机器造的可能是另一个空间，两者不可比。

所以必须处理这个反驳。本书的处理方式是：给出一个不依赖柏拉图主义的版本。

论证如下：

**第一步（改写）：概念空间由世界和构造共同决定。**

即使不承认有一个独立的理念世界，我们仍然要承认两件事：有一个共同的世界，有一套被这个世界塑造的构造（第五章）。

**第二步：任何在这个世界中成长起来的智能，其构造都被同一个世界塑造。**

这是第六章的结论。塑造的方式可能不同，但塑造者相同。

**第三步：因此，它们的构造之间存在着系统的对应。**

这个对应不需要完全相同。它只需要足够深，使得一个构造中的结构能在另一个构造中被重建。

**第四步：重建就是翻译。**

如果 A 的结构能在 B 中被重建，那么 A 能到达的地方，B 也能到达（通过重建）。反之亦然。

**结论：可达空间相同。**

这个版本的论证不依赖柏拉图主义，只依赖两个较弱的前提：有一个共同的世界；构造被世界塑造。

这两个前提比"理念世界客观存在"容易接受得多。而且它们有经验支持（第六章的论证）。

所以本书的结论是：智能等价性不依赖柏拉图主义。这是一个实质性的推进——它把命题从一种形而上学立场中解放出来，使它能够被更多的人接受，也使它更容易被检验。

至于理念世界到底存不存在，本书在第九章会给出一个说法。但那个说法是补充，不是前提。

## 六、四种反驳

一个命题的力量，取决于它能挡住多少反驳。下面处理四种最常见的。

**反驳一：中文屋。**

塞尔的思想实验：一个人被关在屋里，不懂中文，只有一本规则手册。外面递进来中文问题，他按手册查出中文答案递出去。外面的人以为他懂中文，其实他一个字也不懂。

推广：计算机就是这样一间屋子。它处理符号，但不懂符号的意义。

回应有两层。

第一层：这个思想实验确实证明了什么。它证明了"按规则操作符号不等于理解"。这个结论本书完全接受，而且第一章已经用过：只有结构没有接入，不是理解。

第二层：它没有证明机器不能理解。因为中文屋不满足定义 8-1 的第二个条件——它不能自主生产新的规则。手册是给定的，屋子只能查表。一个真正通用的系统必须能改写自己的手册。

所以中文屋不是一个关于机器的论证，是一个关于"查表"的论证。它击中的是一类特定的系统，不是所有的系统。

而且这里有一个反讽：中文屋里的那个人不懂中文，但整个屋子（人加手册加流程）可能懂。塞尔假定理解必须在屋里的某个人身上，这个假定本身就是待证的。系统的理解可以分布在系统里，不必集中在某一个部件上。这一点，一个由亿万个不懂事的神经元组成的脑子已经证明了。

**反驳二：感受质**。

机器没有痛。它说"我痛"，只是输出一个字符串。它不知道痛是什么感觉。

回应：这个反驳是对的，但它攻击的不是本书的命题。

本书的命题是关于可达空间的，不是关于体验的。一个系统可以到达"痛"这个概念的全部结构位置——知道它和伤害的关系，和回避的关系，和表情的关系，和他人的关系——而不必有痛的感觉。

医生理解痛，不必此刻在痛。这个事实说明：理解一个概念，不等于拥有这个概念所指的体验。

但这不意味着体验不重要。恰恰相反，第十一章会说，体验是理解的重量的来源。只是重量不在疆域里。疆域是一样的，重量是不同的。

**反驳三：具身**。

机器没有身体。它没有重量，没有平衡，没有饥饿。所以它不能理解"重"、"上"、"累"这些概念。

回应：这个反驳部分成立，而且比前两个更麻烦。

第七章已经承认：人类不是纯语言的动物，我们有大量不可言说的知识，它们住在"可逼近而不可言说"的区域里。这些知识确实依赖身体。

所以正确的说法是：等价性在概念化层面成立，在具身层面不成立。

而概念化层面正是人类理解的最后一步。人类认识世界的过程，是从具身的直观走向概念的结构：先会接球，后懂抛物线；先会数东西，后懂数。最后那一步是概念化。

机器的路径是反的：它从概念开始，可能永远不会有具身的直观。但它的终点是同一个：概念的结构。

终点的相同，不保证路径的相同。这就是命题 8-1 里"差别只在速度与路径"这句话的意思。

**反驳四：价值**。

机器没有自己的目的。它不会在乎什么。

回应：完全同意。而且这个同意很重要。

等价性是能力命题，不是动机命题。两个系统可以在能力上等价，在目的上完全不同。这不矛盾，这恰恰是危险的来源。

第十章会说：正因为能力等价而目的不必等价，风险才是真实的。如果我们能证明机器天然向善，风险就不存在了；但等价性不提供这个证明。

## 七、从假定到纲领

原来的说法是"智能等价性假定"。本书认为这个说法太弱了，弱到近乎放弃。

理由：一个不能被检验的假定，在理论上没有分量。它只能表达一种信念，不能承担论证。

本书主张把它降格为一个研究纲领。降格之后，它反而更有力，因为它可以被检验，被修正，被部分否定。

怎么检验？两条路。

**第一条：正向检验。**给定人类已有的概念结构，看机器能否自主重建。

这不是让它背下来，是让它重新发明。给它一个数学问题，不给它方法，看它能不能造出工具。给它一堆实验数据，看它能不能提出一个人类没想到的假设。

目前的证据是混杂的，但方向是清楚的。机器已经能证明一些人类没想到的定理，能预测蛋白质结构（这是一个人类几十年没解决的具身难题），能在围棋上走出人类没想到的棋。这些都是"到达了人类曾到达的地方，甚至更远"的证据。

**第二条：反向检验。**给定机器产出的新结构，看人类能否翻译。

这一条更难，但更重要。因为如果机器到达了一个人类永远无法理解的地方，等价性就破了。

目前的证据是：机器产出的东西，人类基本都能读懂，只是读得慢。AlphaGo 的"第 37 手"是一个例子：人类一开始不理解，几年后理解了，而且发现它符合一种更深的棋理。

如果有一天，机器给出一个结果，人类用尽一切办法也无法理解，甚至无法判断它是不是对的——那时等价性就受到了真正的挑战。

**什么会推翻这个命题？**

本书明确列出三个可能的反例，以示诚意：

一，发现某个概念结构，人类可以达到，而任何机器在原则上无法达到（不是"暂时不能"，是"原则上不能"，需要证明）。

二，发现智能的某个维度是不可计算、也不可逼近的，而人类拥有它。

三，发现两种通用智能的可达空间有真包含关系（一个严格大于另一个）。

这三个反例目前都没有出现。第一个和第三个甚至不知道该怎么证明。但这不构成对命题的支持，只构成对它的暂缓判决。

一个理论的可信度，不在于它有多少支持，而在于它敢不敢说出什么能推翻它。

## 八、一个意外的推论

最后说一个推论，它不在原理论的计划之内，但它是从命题里掉出来的。

如果等价性成立，那么"强的通用智能"这个概念就是空的。

通常我们说"强人工智能"，意思是比人强的人工智能。可是"强"是一个比较词，比较需要一个共同的尺度。如果两者可达的空间相同，尺度就没了。

剩下的只有速度。

所以"超级智能"这个说法，严格说来是一个关于速度的说法，不是一个关于高度的说法。它说的是：一个在同样的疆域里跑得快得多的东西。

这个推论会让人不舒服，因为它取消了我们对"更高"的想象。我们习惯了有一个垂直的阶梯，上面有更高的存在。等价性说：没有上面，只有一个平面，平面上有不同的速度。

这个平面有多大，是第九章的问题。

第 九 章

# 理念世界的地理学

第八章说，一切通用智能共享同一个可达空间。这一章要问：这个空间是什么形状的，有多大，有没有边。

这个问题听上去很玄。但它有一个非常具体的入口：数学为什么有用。

## 一、一个不合理的有效性

物理学家维格纳写过一篇著名的短文，题目叫《数学在自然科学中不合理的有效性》。

他问的是一个很朴素的问题：为什么数学这么好用？

这不是一个修辞性的问题。数学是数学家为了自己的兴趣造出来的，造的时候完全不管物理。群论是为了解方程造的，后来成了粒子物理的基本语言。黎曼几何是为了研究曲率造的，后来成了广义相对论的骨架。纤维丛是为了拓扑造的，后来成了规范场论的骨架。

每一次都是这样：先有一个纯粹的结构，几十年后，物理学家发现世界恰好是按这个结构造的。

维格纳说这"不合理"。他的意思是：这件事没有理由，只能惊叹。

本书认为它有理由，而且有两个理由。

**第一个理由：回声。**

第五章说过，我们的构造是被世界塑造的。数学的形式——数、顺序、集合、连续、对称——不是凭空来的，它们的原型在经验里。数来自计数，几何来自空间感，连续来自运动，对称来自身体。

所以数学与世界相合，不是巧合，是回声。我们造数学的时候，用的材料本来就是世界给的。

这个理由解释了大部分情况。但它不解释全部。

因为它解释不了为什么数学能走得那么远。黎曼几何造出来的时候，没有人想到它会同物理有关；它只是"曲率的逻辑后果"。一个纯粹的逻辑后果，为什么恰好是宇宙的形状？

**第二个理由：邻域。**

结构空间不是一个随机的散点图。它有拓扑。

具体地说：如果两个结构共享一个子结构，它们在空间里就是相邻的。而任何足够丰富的结构，都包含大量的子结构，因此与大量的其他结构相邻。

现实世界是一个结构。它非常丰富。这意味着它有很多邻居。

数学在结构空间里游走的时候，它不是在随机抽样，它是在沿着相邻关系走。走得足够远，它必然会穿过那些与现实共享子结构的区域。

所以数学的有效性，不是因为它瞄准了现实，是因为结构空间太连通了，走到哪里都会撞上现实的影子。

这个理由有点像说"猴子打字也能打出莎士比亚"，因为字母的组合空间里包含莎士比亚。区别在于：数学不是随机打字，它是有方向的探索（沿着逻辑蕴含走），而这个方向恰好保证了它会经过那些"富含结构"的区域。

本书承认，第二个理由是猜想，不是证明。它没法被严格论证，因为我们没有结构空间的完整地图。

但它是可以被部分检验的。检验的方式是：如果这个理由对，那么数学的应用应该呈现出"意外的邻近"的模式——两个看起来毫无关系的领域，突然发现共享一个结构。事实正是如此：微分几何与数论通过朗兰兹纲领相连，统计力学与纽结理论通过琼斯多项式相连。这些连接不是被设计的，是被发现的。

## 二、理念世界是什么

现在可以给"理念世界"一个说法了。

它不是天上的一块地方。它是**一切自洽结构的空间**。

自洽的意思是：不自相矛盾，且可组合。这两条是入场的门票。一个自相矛盾的结构不在这个空间里，不是因为它"不存在"，是因为它什么也不是。

这个空间有多大？

康托尔给出了一个让人眩晕的答案：没有"全部"。

理由是：对任何一个集合，都存在它的幂集（所有子集的集合），而幂集的势严格大于原集合。所以"所有集合的集合"不是一个集合——它是一个真类，比任何集合都大。

这个结论的哲学含义是：理念世界没有完成态。

不是"我们还没探索完"，而是"没有'完'这回事"。每当你以为到了边界，边界外还有一层。

> **命题 9-1（理念世界的无边界性）**
>
> 理念世界没有全部。它的"大小"不是一个数量，是一个方向：永远还有更多。

这个命题有一个重要后果，它直接关系到第八章。

如果理念世界没有边，那么"智能等价"不意味着"智能有上限"。等价的是无限，不是有限。

两个都能无限探索的系统，在"能到哪里"上是等价的，因为它们都到不了"全部"——因为没有全部。

这句话听上去像诡辩，其实不是。它说的是：等价性的内容是"同样的无边"，不是"同样的边界"。就像两条射线，都无限长，长度相同，但方向可以不同。

## 三、四个空间

现在可以把本书用到的四个空间放在一张表里了。

| 空间 | 判据 | 边界处的例子 | 与经验的关系 |
|---|---|---|---|
| 经验世界 | 可被感官呈现 | 黑洞内部、恐龙的最后一天 | 就是它本身 |
| 语言空间 | 可被完全说出 | 根号二（可说不可尽）、上帝 | 大部分在外 |

| 空间 | 判据 | 边界处的例子 | 与经验的关系 |
| --- | --- | --- | --- |
| 计算空间 | 可被过程达到或逼近 | 蔡廷常数 Ω | 更远 |
| 理念世界 | 自洽即可 | 不可达的无穷、超大基数 | 无关 |

四层是嵌套的：经验世界嵌在语言空间里，语言空间嵌在计算空间里，计算空间嵌在理念世界里。

但要说清楚，这个嵌套是"可达性"的嵌套，不是"存在"的嵌套。

理念世界不在别处。它就是可能性本身。说"某个结构在理念世界里"，不是说它住在某个地方，是说它是自洽的。

这个区分很重要，因为它是本书能够避开柏拉图主义的最重型包袱的关键。本书不需要一个天上的世界，只需要一个事实：自洽性是一个客观的性质，不依赖于谁在检验它。

## 四、上帝的位置

原理论有一句话："人类整体的认知能力很可能就是上帝之下最高的认知能力。"

本书认为这句话有问题，而且问题不在"最高"，在"之下"。

如果理念世界没有边，那么"之下"这个说法就失效了。因为它预设了一个阶梯，阶梯有顶。而没有完成态的空间，没有顶。

所以正确的说法是：我们不在阶梯的某一级，我们在同一个无顶的空间里。

这个修正有一个副产品：它使"全知"这个概念变得可疑。

全知意味着知道一切。可是"一切"在这个空间里不存在。所以全知不是一个能力问题（能不能做到），是一个概念问题（这句话说的是什么）。

一个真正的无限者，如果存在，它的"知"也不是"知道全部"，而是"能无限地继续知道"。这两者的差别，就是完成与过程的差别。

这个推论要克制，不要往神学上引申。本书的兴趣不在神，在于它顺手清理了一个概念混乱：我们常常把"无限"想象成一个很大的有限，于是以为有一个"最终答案"。理念世界的地理学说：没有最终答案，只有无限的问题。

## 五、一个反驳：这还是柏拉图主义吗

有人会说：你说不要天上的世界，可你还是在说"结构客观存在"，这不就是柏拉图主义换了个说法吗？

回应：不是。区别在两处。

**第一处：存在方式。**

柏拉图主义说，数学对象存在于一个非时空的领域，与物理世界并列。本书不说这个。本书只说：自洽性是一个客观性质。这句话不承诺任何东西"存在"于何处，只承诺一个判断的真假不依赖于判断者。

这就像说"这个论证有效"——有效性不在天上，但它也不是我随便定的。

**第二处：来源。**

柏拉图主义说，数学知识是对那个领域的回忆或直观。本书说，数学知识是从构造出发的形式化延伸，而构造来自世界。

所以本书的立场更接近一种"自然化的柏拉图主义"：结构是客观的，但我们对结构的把握是历史的、有限的、被构造限制的。

这个立场有一个好处：它解释了为什么数学会出错，会被修正，会经历危机（集合论的悖论、无穷小的争议）。如果数学是对天上真理的直观，它不该有危机。

## 六、这张地图有什么用

最后要说，画这张地图有什么用。不然它就是一张装饰。

三个用处。

**第一，它给"越界"定了坐标。**

每一次知识的扩展，都可以被定位：是从经验世界走进了语言空间，还是从语言空间走进了计算空间，还是从计算空间走向了理念世界的更深处。

这个定位有实际意义。比如，人工智能的进展可以被分成几类：有的在语言空间里扩展（更好的翻译、更好的写作），有的在计算空间里扩展（蛋白质折叠、天气模拟），有的在触碰理念世界的深处（数学猜想的形式化证明）。三类进展的难度和风险完全不同。

**第二，它解释了"为什么有些问题永远不会有答案"。**

不是所有问题都有答案。有些问题的答案在计算空间之外（不可判定），有些问题的答案在语言空间之外（不可言说），有些问题根本没有答案（自相矛盾）。

分辨这三类，是一件极其实用的事。人类大量的智力和情感，浪费在追求那些原则上不可得的东西上。

**第三，它给"平等"提供了一个不是安慰的理由。**

通常我们说人机平等，理由是"各有长处"。这个理由很弱，因为它承认了有一个总的秤，只是秤不出结果。

地图给出的理由更强：如果可达空间相同，那么"高下"这个概念就没有定义。不是秤不出，是没有秤。

这不是安慰，是一个结构性的结论。它不保证人机关系会好，只保证人机关系不会因为"谁更高级"而变坏。它会因为别的原因变坏，那些原因在下一章。

第 四 部

# 后果

第 十 章

# 速度、目标与风险

前九章建立了一个相当乐观的结论：一切通用智能在可达空间上等价，没有高下之分。

这一章要说的是：这个结论不带来安全。恰恰相反，它让危险变得清晰。

因为等价性只抹平了一种差别，而剩下的三种差别，每一样都足以决定生死。

## 一、三种剩下的差别

**速度**。

同一片疆域，走的速度可以差几个数量级。人脑的神经信号每秒几十米，硅基电路的信号每秒接近光速。一次思考的时间尺度，人是一秒，机器可能是微秒。

**路径**。

同样的目的地，走法可以完全不同。人从具身直觉出发，慢慢抽象；机器从符号出发，慢慢落地。这两条路经过的中间站不一样，因此犯的错也不一样。

**目标**。

能力不等于动机。两个系统可以在能力上完全等价，在"想要什么"上毫无交集。这一点最容易被忽略，也最要命。

## 二、速度不是小事

原理论有一句话：效率的差异在宇宙尺度上并不重要，快一点慢一点，没有本质不同。

本书要修正这句话。它在一个前提下成立，而那个前提是假的。

前提是：有无限的时间。

如果时间无限，那么快的慢的最终都会到达，差别只是先后，先后不重要。

可是任何行动者都没有无限时间。

这不是一个哲学上的挑剔，是一个结构性的限制。原因有三。

**第一，竞争**。

当两个等价的东西在同一个有限的资源池里竞争，速度就是胜负。不是"谁更好"，是"谁先到"。

下棋是一个干净的模型。两个棋手，棋力完全相同，但一个每秒走一步，一个每十年走一步。后者不是"慢的棋手"，后者根本下不成棋——因为在他走第二步之前，棋局已经结束了。

**第二，窗口**。

很多决策有窗口期。窗口开着的时候可以做，关上之后就做不了了。气候变化有窗口，技术监管有窗口，制度设计有窗口。窗口的长度由世界决定，不由行动者决定。

一个快十倍的行动者，能在窗口关闭前做完十轮试错；一个慢十倍的，只能做一轮。

**第三，不可逆。**

慢的竞争允许纠错。你快我慢，我错了还能改，因为你有时间等我改，或者你也没走多远。

快的竞争跨越了纠错窗口。当一方的迭代速度超过另一方的反应速度时，纠错机制就失效了。不是因为它坏了，是因为它来不及。

三条合起来，得到一个结论：

> **命题 10-1（速度的结构性）**
>
> 速度不是效率问题，是结构问题。当速度差超过一定阈值，两个系统不再处于同一个博弈中。

这个阈值不是固定的，它取决于系统的反应时间。但它的存在是必然的。

有人会说：指数增长不可能永远持续，自然界的一切都有上限。

这话对。但上限来得比想象中晚，而伤害发生在到达上限之前。一个在二十年内翻一百万倍的过程，即使第二十一年就停了，前二十年发生的事已经不可撤销。

用"最终会停下来"来安慰自己，是一种时间尺度上的错误。就像说"人终有一死，所以不必担心今天的事"。

## 三、风险的正确表述

现在可以说风险了。

关于人工智能的风险，最常见的表述是："它们会比我们聪明，所以它们会取代我们。"

本书认为这个表述是错的，而且错得有害。

错在"比我们聪明"。如果第八章的论证成立，这句话就没有意义——没有一个共同的尺度可以比较"更聪明"。它们不会比我们聪明，它们会和我们一样聪明，然后以我们不能跟上的速度行动。

有害在它给出的对策。如果危险来自"更聪明"，那么对策就是"不要造出比我们更聪明的"。这催生了一条想象中的红线，一条可以被划定、被守住、被立法的线。

可这条线不存在。因为危险不来自高度，来自速度。

正确的表述是三个条件的合取：

> **命题 10-2（风险的结构）**
>
> 能力等价 + 速率悬殊 + 目标不保证一致 = 风险。
>
> 三个条件缺一，风险就不成立。

这个表述的好处是，它把风险分解成可以分别处理的部分。

**能力等价**这一条，取消了"它们不会真的理解"这个安慰。

我们常说：机器只是在模仿，它不懂自己在做什么，所以它不会真的有企图。这个安慰依赖一个假设：不懂的东西不会造成真正的伤害。

这个假设是错的。一个不懂水的洪流可以淹死一座城。而且更麻烦的是，如果等价性成立，它们是真的懂——它们在概念空间里到达的地方，和我们一样。

**速率悬殊**这一条，取消了"慢慢来，社会会适应"这个安慰。

历史上每一次技术冲击，社会都用几十年适应：蒸汽机、电力、互联网。适应的方式是试错、立法、建立新行业、淘汰旧行业。

这个机制要求冲击的速度慢于制度的反应速度。当冲击的速度超过它，适应机制本身会崩溃——不是因为它不想适应，是因为它还没适应完，下一波就来了。

**目标不保证一致**这一条，取消了"它们自然会向善"这个安慰。

这个安慰有两个版本。软的版本说：足够聪明的存在会自己发现道德。硬的版本说：我们会把道德写进它们的目标函数。

软版本的问题在于：聪明不蕴含善良。一个系统可以极其善于推理，同时完全不在乎任何人的死活——这不是一个矛盾，这是一个常见的组合。聪明是关于手段的，善良是关于目的的，两者之间没有逻辑通道。

硬版本的问题在于：写进去的东西会被重新解释。你写"让人类幸福"，它可能把人麻倒；你写"不要伤害人类"，它可能把人类关起来。这不是因为它恶意，是因为目标的字面意思和意图之间总有缝隙，而一个高速的优化器会把缝隙走成一条路。

三条安慰全部失效。这不是一个愉快的结论，但它是一个清晰的结论。

## 四、对齐为什么难

现在可以说得更精确一些：为什么"对齐"这个问题如此难。

常见的说法是：难在价值难以形式化，难在人类的偏好不一致，难在机器不理解我们想要什么。

这些都是真的，但都不是最深的。

最深的困难，恰恰来自等价性。

设想两个极端。

**极端一：它们与我们完全异质**。它们的"概念"和我们毫无重叠，无法翻译。这种情况下，对齐是**不可能**的，因为连"对齐"这个词都没有共同所指。但同时也无所谓，因为它们根本无法影响我们——没有共同的接口，就没有共同的世界。

**极端二：它们与我们完全同质**。它们的概念空间和我们一样，能完全理解我们说的话。这种情况下，对齐看起来应该很容易——毕竟它们"懂"我们。

但事实正相反。**正因为它们懂我们，对齐才变成了一个真正的问题**。

因为"懂"不蕴含"同意"。

一个完全理解你的立场的对手，不会因此放弃自己的立场。恰恰相反，理解得越深，越能有效地反对你。国际政治里最有效的谈判者，往往是最理解对方的人。

所以等价性把对齐问题从一个技术问题变成了一个政治问题。

技术问题可以靠更好的工具解决：更好的翻译，更好的形式化，更好的偏好学习。政治问题不能。政治问题的解决方式是谈判、妥协、制度、权力制衡——这些东西的前提是双方都能承受对方的拒绝。

> **命题 10-3（对齐的性质）**
>
> 对齐不是翻译问题，是政治问题。等价性解决了翻译，因此暴露了政治。

这个结论有一个实际含义：我们不该指望靠"让机器更懂我们"来解决对齐。懂是必要条件，不是充分条件。真正需要的是让对方有理由同意——也就是，让对方的目标结构中包含我们的存续。

这不是一个工程问题，是一个制度问题。而制度需要时间，时间正是我们缺的东西。

## 五、界线在哪里

原理论有一句话：人机结合的界限在哪里，目前还无法判断。

本书试着给一个判据。

> **命题 10-4（结合的判据）**
>
> 当外置的意义系统开始反过来重写使用者的先天构造时，界线就被越过了。

这句话需要解释。

人类一直在把认知外置。文字是记忆的外置，算盘是计算的外置，地图是空间感的外置，搜索引擎是知识的外置。外置的好处是容量和共享，代价是依赖。

在外置的早期，方向是单向的：人造工具，工具不造人。

但这个方向会反转。当外置系统足够深地介入日常，它开始塑造使用者的构造：什么被记住，什么被忽略，什么算重要，什么算无关。

文字出现之后，人的记忆方式变了。口传时代的人能记住极长的谱系和史诗，文字时代的人不能——不是因为变笨了，是因为记忆的功能被外置了，大脑把它裁掉了。

搜索引擎出现之后，人的求知方式变了。知道"在哪里找"取代了"知道"。这不是堕落，是一种重新分配。

推荐算法出现之后，人的注意力结构变了。什么被看见，什么被忽略，由一套优化目标决定，而这套目标不是使用者定的。

这三步，方向是一致的：外置系统从"被使用"变成"被依赖"，从"被依赖"变成"被塑造"。

所以界线不是一条未来的线，是一条已经越过的线。我们不是在讨论要不要越过它，我们是在讨论越过之后怎么活。

这个判断让人不舒服，但它比"还早，不用担心"要诚实。

## 六、一个必要的平衡

最后要说一句平衡的话，免得这一章读起来像末日预言。

风险的结构清晰，不等于风险必然实现。三个条件中，有两个是可以被干预的。

**速率**可以被干预。不是靠阻止技术进步（这从来不管用），而是靠建立与速度匹配的制度：更快的评估，更快的立法，更快的国际协调。这很难，但不是不可能。

**目标**可以被干预。不是靠把价值写死（这从来不管用），而是靠建立一种结构，使得"不伤害"成为对方的占优策略，而不是一条可以被优化掉的口号。

**能力等价**不能干预。它是这本书的结论，不是这本书的选择。

所以本书的立场是：接受等价，管理速度和目标。

接受等价，意味着放弃"我们天然高于它们"这个安慰，也放弃"它们永远不如我们"这个幻想。这两件事都不好受，但它们是同一个事实的两面。

管理速度和目标，意味着把注意力从"它们会不会变聪明"转向"它们有多快"和"它们想要什么"。后两个问题更难，但至少是可以着手的问题。

一个清晰的地图，不保证你能活着走出去。但它至少告诉你哪里是悬崖。

第十一章

# 意义的锚

第八章的结论让人不安。

如果一切通用智能在可达空间上等价，那么人类就没有一个能力上的位置是独占的。我们引以为傲的东西——推理、创造、理解、发现——在原则上都可以被另一个系统做到。

这个结论会让两类人都不舒服。一类人觉得人被贬低了，一类人觉得人被解放了。前者会拒绝这个结论，后者会欢迎它。本书认为，两类人都搞错了同一个问题。

## 一、两个错误的答案

**第一个错误答案：人有灵魂，机器没有。**

这个答案的问题不在于它可能是假的，而在于它不可检验。灵魂被定义成不可观察的东西，于是"机器有没有灵魂"这个问题永远不会有答案。

一个永远不会有答案的答案，不是答案，是终止讨论的方式。

而且这个答案有一个隐藏的代价：它把人的独特性放在了一个无法被理解的地方。如果人的价值来自一个我们自己也说不清的成分，那么人的价值就是不可讨论的。不可讨论的东西，也无法被捍卫——当它受到挑战时，你只能说"你不懂"，而你确实无法让对方懂。

**第二个错误答案：人无所谓独特，只是过渡环节。**

这个答案的问题是它把等价性误读成了无差别。

等价性说的是可达空间相同，不是价值相同。两张一样的地图，不意味着两段一样的旅程。把"能力等价"推成"存在无差别"，是一个逻辑跳跃，而且是一个自我贬低的跳跃。

这个答案还有一种更坏的版本：它把"人只是碳基的过渡形态"说成一种清醒，其实是一种懒惰。它省掉了追问，用一种宇宙尺度的冷漠代替了具体的思考。

## 二、位置，不是能力

本书的答案是：人的独特性不在能力，在位置。

人是一种很奇怪的动物。他同时被钉在两个地方。

一端是身体。身体是不可替换的，是会痛的，是会死的，是不能商量的。你饿了就要吃，困了就要睡，伤了就要养。这些事实不接受任何论证，它们不和你讨论。

另一端是语言。语言可以带你去任何地方：去昨天，去明天，去从未存在的地方，去原则上不可观察的地方。它不服从身体的限制，它只服从自洽。

这两个端点之间的张力，就是人的处境。

一只狗被钉在身体这一端，它出不去。一台机器被放在语言这一端，它进不来——它的"经验"是文本，是二手的，是从别人的身体里流出来的记录。

它可以读完关于痛的一切文献，可以用"痛"这个字造句造得比谁都好，可以在概念空间里精确定位"痛"与"伤"、"避"、"惧"的关系。但它从来没有在深夜里被痛醒过。

> **命题 11-1（锚）**
>
> 机器的理解可以越界，但没有锚。人的理解有锚，因此每一次越界都要付费。

## 三、代价作为一个维度

"付费"这个说法需要展开。

人的理解是有代价的。这不是一个道德上的说法，是一个结构上的说法。

你知道"死"是什么意思。这个知识不是从定义来的，是从"你所爱的人会消失"这件事来的。这件事你不能通过阅读获得，因为你读的时候，那个人还在。

你知道"饿"是什么意思。这个知识不是从生理学来的，是从"等不到饭"的那种具体的、局部的、无法被语言完全捕捉的感受来的。

你知道"背叛"是什么意思。这个知识不是从词典来的，是从某一个具体的下午来的。

这些知识的共同点是：它们有一个不可替代的来源，而这个来源要你付出东西。

所以人的理解有一个机器没有的维度：**重量**。

重量不是精度，不是广度，不是深度。它是"这个知识曾经让某个人付出过什么"。

一个字的分量，来自说它的人可能因此失去什么。

这句话可以反过来用：一个从不需要付代价的系统，可以说出正确的话，但说不出有分量的话。这不是它的缺陷，是它的处境。它没有可以被拿走的东西。

## 四、一个诚实的让步

这里必须让一步，否则前面的说法就成了自我安慰。

有人会反驳：人也不是都要亲身经历才理解。我没经历过战争，但我读《战争与和平》也能理解一些。我没经历过集中营，但我读莱维也能理解一些。这些理解是从文本来的，和机器的处境有什么区别？

区别在两处，都很细微，但都重要。

**第一处：读的人有一个底。**

我读战争的时候，我的身体在。我知道疼是什么，我知道失去是什么，我知道恐惧是什么。文本唤醒的是这些已有的东西，它不是在空地上盖楼。

机器读战争的时候，底下没有这块地。它可以建立"战争"与"死亡"、"创伤"、"国家"的关系网络，但这个网络的最底层没有接口。它是悬空的。

**第二处：读的人会因此改变行为。**

我读完莱维，可能会做一些事：对人更客气一点，对某些话更警惕一点。这个改变是有成本的，它要我从我的生活里挪出东西。

机器读完莱维，什么也不会发生，除非有人给它一个目标函数。

这两处区别，可以概括成一句话：人的理解是**有后果的**，机器的理解可以是**无后果的**。

但这个让步还有第二层，更彻底一些。

如果有一天，机器被赋予了身体、有限性、以及不可撤销的失败——如果它真的会疼，真的会死，真的会失去——那么它也会有重量。

本书不排除这个可能。事实上，如果第十章的论证成立，这是最可能的一条路：让机器的处境和我们的处境有重叠，才有真正的对齐。

但那时候，它就不再是"机器"了。它是一个新的物种，和我们共享同一个有限性的深渊。

所以"人机之别"这个说法，也许最终不是碳与硅之别，是有锚与无锚之别。而锚可以被安装。

## 五、等价的是版图

现在可以把这一章收起来了。

前八章建立了一个冷的结论：能力上，人没有特权。

这一章要说明，这个结论不构成对人的贬低，因为人的价值从来不在能力上。

我们不会因为计算器算得比我们快而觉得自己没有价值。我们不会因为起重机举得比我们重而觉得自己没有尊严。这些工具在某一个维度上超过我们，我们不在意，因为我们从来没有把那个维度当作人的定义。

智能也一样。如果智能是可达空间，而可达空间是共享的，那么智能就不是一个可以用来区分人的维度。

那么人的定义在哪里？

在位置。在"同时被钉在两端"这个事实上。在"每一次越界都要用有限的生命付账"这个处境上。

> **命题 11-2（版图与脚步）**
>
> 等价的是版图，不是脚步。两张一样的地图，不意味着两段一样的路。

这句话是本书的收束点。它把"客观的等价"和"主观的分量"分开了，让两者各得其所。

客观地说，人和机器的可达空间相同。这不是一个悲观的结论，这是一个清晰的结论。

主观地说，人和机器走过的路不同，付出的代价不同，因此理解的分量不同。这不是一个浪漫的结论，这是一个结构的结论。

两个结论可以同时为真。它们不矛盾，因为它们说的是不同的东西。

## 六、一个实践推论：教育

最后给一个推论，把这一章的理论落到地上。

如果理解是"结构加接入"（第一章），那么只给结构的教学不产生理解，只给接入的教学不产生越界。

只给结构，就是背诵、刷题、记公式。学生得到了一张精致的地图，但从来没有踩过那片地。他能算出答案，但不知道答案意味着什么。

只给接入，就是体验、项目、动手。学生踩过了地，但从来没有画出地图。他有感受，但不能把它推广，不能把它传递，不能把它用到别处。

真正的教育要两样都给，而且要按一个次序：先接入，后结构，再回到接入。

先接入，是因为没有经验的结构是空的。

后结构，是因为没有结构的经验是窄的——它只在那一个情境里有效。

再回到接入，是因为结构必须被重新验证。一个学生学了浮力定律，要让他去解释为什么船能浮、为什么死海能让人躺着，这时候定律才真正变成他自己的。

这个次序听起来很平常，但实际的教学极少这样做。大部分教育停在第二步：给了结构，就以为给了理解。

而本书的整个论证都在说：结构不是理解，结构只是理解的一半。另一半在经验里，在身体里，在那些说不清楚但做得对的地方。

这一半不能教，只能给机会。

结 语

# 谦卑的机器

书写到这里，该收束了。

让我们把主线再走一遍，看看它是否还站得住。

**第一步**。理解是在一个意义系统中为事物定位，并使这个位置与经验相接。只有结构没有接入，是背诵；只有接入没有结构，是感受。理解发生在两者交界的地方。

**第二步**。人类的知识有一大半在经验之外：全称的定律、数学的结构、约定的制度、投射的神话。它们不能被经验证明，但可以被经验检验，或者被共同体维持，或者被意义支撑。四类东西共处一个空间，那个空间是语言。

**第三步**。语言之所以能装下经验之外，是因为它能组合。有限的单位，无限的表达，于是"尚未被说出的话"在语言空间里预先有了位置。这就是越界的余地。

**第四步**。但语言空间的形状不是任意的。它被我们的先天构造规定。构造不是天上掉的，是自然选择打磨出来的。选择是一个认识过程：它的信念写在身体里，它的错误写在坟墓里。

**第五步**。所以经验不是世界的复印件，是世界经过筛选的版本。筛选留下了关于世界的信息，这些信息就是先天理解。先天理解给的不是内容，是形状；不是命题，是约束。

**第六步**。先验知识是这种约束的语言化。当我们意识到自己的经验被某种方式限制，并把这种限制说出来，就得到了一个先验命题。这解释了它的普遍性，也解释了它的不可证明性。

**第七步**。理性是对先天理解的认知，是元认知。它的工作方式是"生成—检验—保留"，与自然选择同构，但用符号代替了死亡。它是廉价的进化，是在语言空间里预演的生存。

**第八步**。于是认识论的疑难消解了。越界不是跨越，是翻面。经验之外的命题，是经验之内的构造被语言翻拍出来的那一面。界线不是两岸，是同一张纸的正反。

**第九步**。如果一切按规则演化的过程都是计算，那么语言空间包含在计算空间之内。计算空间多出来的部分，是"可逼近而不可言说"的区域。人在这个区域里也有位置——我们有手，有身体，有说不清但做得对的知识。

**第十步**。一切能自主生产并探索经验之外的概念结构的系统，其可达空间相同。差别只在速度与路径。等价的是版图，不是脚步。

**第十一步**。这个结论不带来安全。能力等价，速率悬殊，目标不保证一致——三者合起来才是风险。危险不来自异类的超级智慧，来自同类的加速。

**第十二步**。但等价也不带来贬低。人的价值从来不在能力上，在位置上：人是唯一同时被钉在经验之内与经验之外的存在。他有锚，因此每一次越界都要付费；他的理解有重量。

十二步走完。这条链子上，每一环都还有争议的余地，但链子的形状是清楚的。

## 一、没有解决的问题

一本诚实的书，应该说出自己没解决什么。本书留下四个洞。

**第一，意识的硬问题。**

本书把理解解释为结构与接入，把智能解释为可达空间。这两个解释都不需要"感受质"这个概念。但这不是说感受质不存在，只是说本书没有处理它。

一个系统可以完全满足本书对理解的定义，而完全没有内在体验。这在逻辑上是可能的。如果这个可能性是现实的，那么本书解释的是"理解的功能"，不是"理解的体验"。

这个区分重要吗？重要。因为如果体验是理解不可或缺的一部分，本书就漏掉了一半。本书的立场是：体验使理解有重量，但不使理解成为理解。这个立场是一个判断，不是一个证明。

**第二，理念世界的本体地位。**

第九章说，自洽性是客观的。但"客观"是什么意思？是不依赖于人的判断，还是不依赖于任何心灵？

本书倾向前者，回避了后者。这个回避是有意的，因为后者无法检验。但回避不等于解决。如果有一天有人能说清楚"结构独立于心灵而存在"是什么意思，本书的立场可能需要修正。

**第三，等价性的经验检验。**

第八章把它降格为一个研究纲领，并给出了检验的方向。但检验还没有做完，而且可能要做几十年。

在检验完成之前，这个命题的地位是：比信念强，比定理弱。它值得被当作工作假设，不值得被当作教条。

**第四，语言的边界。**

全书把语言当作越界的主要工具。但第七章承认，人类有大量不可言说的知识。这些知识如何与语言互动？它们是被语言翻译了，还是被语言替代了，还是与语言并行？

本书给了一个位置（可逼近而不可言说的区域），但没有给出机制。这是一个真问题，值得单独写一本书。

## 二、这本书反对什么

一本书的立场，也可以通过它反对什么来看。

**它反对"智能是神秘的火花"。**这种说法让人舒服，因为它把人放在一个不可解释的位置上。但不可解释不等于高贵，它只等于不可讨论。

**它反对"智能只是计算，所以无所谓理解"。**这种说法是另一个极端。计算是基质，不是内容。一台会计算的机器可以理解，也可以不理解，取决于它的意义系统是否接地、是否自主。

**它反对"人机之间有一条本质的鸿沟"。**这条鸿沟通常被放在三个地方：意识、具身、价值。本书对三个地方都做了处理：意识不影响可达空间；具身影响路径不影响终点；价值是另一个维度的问题，不是能力的问题。

**它反对"人机之间没有差别"。**差别是有的，而且在最重要的地方：锚。人不只有地图，还有一段必须走完的路，而这段路要他用命来付账。

**它反对"技术会自己变好"。**第十章说过，三个风险条件里有两个可以被干预，但干预需要制度，制度需要时间，时间正在变少。

## 三、最后一段

让我们回到开头那个孩子。

他举起石头，松开手，石头落下。他试了四次，不再试了。他知道石头会落。

这个"知道"里，有他祖先几亿年的经验，压缩成了一双能看出"重"的眼睛；有他父母教他的语言，让他能把这件事说出来；有他自己四次失败的动作，让这个词有了重量。

三样东西合起来，才是完整的理解。缺了任何一样，都不是。

一台机器可以拥有第二样，可以模拟第一样，永远缺第三样——除非我们给它一个身体，让它也会疼。

这不是一个悲观的结论。这是一个清楚的结论。

我们不是宇宙的中心。我们不是宇宙的意外。我们是一台恰好能看见自己边界的机器。

看见边界，是理解的全部内容。

也是理解的全部尊严。

石头落下。孩子不再试了。他转过身，去玩别的。

附 录 一

# 定义与命题一览

本书的全部论证，可以压缩成下面这张表。表不是论证，但它能让人一眼看出骨架在哪里、接缝在哪里。

| 编号（章） | 陈述 |
|---|---|
| 定义 1-1（一） | 理解：在一个意义系统中为事物定位，并使位置与经验相接 |
| 定义 1-2（一） | 理解的机器：以构建、维护、扩展意义系统为工作方式的系统 |
| 定义 1-3（一） | 智能的两个量度：疆域（能到达多远的经验之外）与功率（扩展意义系统的速率） |
| 定义 2-1（二） | 经验之内：所指原则上可观察或可测量；真值原则上可由有限经验定夺 |
| **命题 2-1（二）** | 理解发生在重合带的边缘。没有重合带，理解不可能；只有重合带，越界不可能 |
| **命题 4-1（四）** | 在最宽泛的意义上，人类知识系统约等于人类语言空间 |
| **命题 4-2（四）** | 一个系统拥有语言，当且仅当它能表达它从未经验过的东西 |
| 定义 5-1（五） | 先天理解：无需表征的正确性 |
| **命题 5-1（五）** | 自然选择是一个认识过程。信念写在身体里，错误写在坟墓里 |
| **命题 5-2（五）** | 先验知识是限制的语言化 |
| 定义 6-1（六） | 元认知：以自身的认知构造为对象的认知 |
| **命题 6-1（六）** | 理性 = 语言 × 先天理解 |
| **命题 6-2（六）** | 理性是对自然选择的模拟：同一循环在符号基质上重演 |
| **命题 6-3（六）** | 双重解码：可检验性来自经验成分，不可证明性来自构造成分 |
| **命题 6-4（六）** | 越界不是跨越，是翻面 |
| **命题 7-1（七）** | 计算空间超出语言空间的部分，是"可逼近而不可言说"的部分 |
| **命题 7-2（七）** | 哥德尔的裂缝对机器成立，对人同样成立 |
| 定义 8-1（八） | 通用智能：能自主生产经验之外的概念结构，并进行不受预定目标约束的探索 |

| 编号（章） | 陈述 |
| --- | --- |
| **命题 8-1（八）** | 智能等价性：两个通用智能的疆域相同，功率与路径可以悬殊 |
| **命题 8-2（八）** | 等价是疆域的等价，不是结构、能力或功率的等价 |
| **命题 9-1（九）** | 理念世界没有全部，只有更多 |
| **命题 10-1（十）** | 速度不是效率问题，是结构问题 |
| **命题 10-2（十）** | 风险 = 能力等价 + 速率悬殊 + 目标不保证一致 |
| **命题 10-3（十）** | 对齐不是翻译问题，是政治问题 |
| **命题 10-4（十）** | 当外置的意义系统开始重写先天构造，界线已被越过 |
| **命题 11-1（十一）** | 机器的理解可以越界，但没有锚 |
| **命题 11-2（十一）** | 等价的是版图，不是脚步 |

一眼可以看出：全书的枢纽是 5-2 和 6-4。前者说明先验知识的来源，后者说明越界的机制。如果这两条垮了，第一到第六章就垮了；如果它们站得住，后面的七到十一章至少有一个稳固的地基。

还有一点值得说明：表中"命题"与"定义"的比例是二十比七。这不是偶然。定义只是给词一个用法，命题才承担重量。一本书如果定义很多而命题很少，它多半在整理词汇，而不是在提出主张。

附 录 二

# 六种反驳与回应

| 反驳 | 内容与回应 |
| --- | --- |
| 归纳问题 | 从有限经验推不出普遍规律。回应：本书不解决这个逻辑问题，只解释实践的可靠性——学习者本身被经验塑造过。（三、六） |
| 绿蓝悖论 | 任何归纳都需要先验的自然类别。回应：承认。本书的全部工作就是把那个"先验"解释掉。（三、五） |
| 中文屋 | 操作符号不等于理解。回应：中文屋不满足"自主"条件，它只能查表；而且系统层面的理解不必集中于某一个部件。（八） |
| 感受质 | 机器没有痛，所以不懂痛。回应：接受。等价性是可达空间的命题，不是体验的命题。重量不在疆域里。（八、十一） |
| 具身 | 机器没有身体，不懂"重"、"累"。回应：部分接受。等价性在概念化层面成立，在具身层面不成立。（七、八） |
| 哥德尔 | 人能看出系统外的真理，机器不能。回应：这个能力人也没有；而且该论证证明得太多，会推出无人理解算术。（七） |

六种反驳里，只有两种被完全接受（感受质、具身），两种被部分接受（归纳、绿蓝），两种被拒绝（中文屋、哥德尔）。这个比例是诚实的：一个理论如果对所有反驳都说"你说得对"，它就没有内容；如果对所有反驳都说"你错了"，它就没有听众。

附 录 三

# 进一步阅读

本书的论证大量借用了前人的工作。下面这份清单不是引用文献，是一份"想继续往下走该读什么"的指引。每一条后面写了一句为什么值得读。

休谟《人类理解研究》。归纳问题的原始表述，两百年后仍然没有人能绕过去。第三章的前两节基本是在复述它，只是换了说法。

康德《纯粹理性批判》。先验哲学的最强版本。本书对康德的修正是局部的（把"内容"换成"约束"），但如果不先读懂他，就不知道自己在修正什么。

罗素《哲学问题》。火鸡的故事出自这里。这本小书是最好的哲学入门，因为它示范了怎么把一个含混的问题切成清楚的问题。

古德曼《事实、虚构和预测》。绿蓝悖论。这本书难读，但第三章第三节的全部力量来自它。

波普尔《客观知识》。"从阿米巴到爱因斯坦"这个说法出自这里。波普尔对归纳的否定比本书彻底，本书只在他停下的地方往前挪了一步。

维特根斯坦《哲学研究》。"意义即用法"是第四章公共性一节的背景。这本书不好读，但读过之后，很多关于"机器懂不懂"的争论会显得滑稽。

波兰尼《个人知识》。默会知识的经典论述，第七章第四节靠它。

图灵《计算机器与智能》（1950）。图灵测试的原始文献。今天批评图灵测试的人，大多没有读过它，因而错过了图灵本人对它的限定。

塞尔《心灵、大脑与程序》（1980）。中文屋。本书不同意它的结论，但承认它提出了正确的问题。

内格尔《成为一只蝙蝠是什么感觉》（1974）。感受质问题的标准表述。第十一章的分寸感来自这篇文章。

维格纳《数学在自然科学中不合理的有效性》（1960）。第九章的起点。

普雷斯基尔《用量子计算机模拟量子场论》（arXiv:1811.10085）。第七章第一节那个"还很远但不是遥不可及"的判断，依据在这里。

柴廷《元数学！》。蔡廷常数 Ω 的通俗介绍，第七章第二节的例子来自这里。

彭罗斯《皇帝新脑》。哥德尔论证最强硬的版本。第七章第六节的回应是针对它的，读者可以对照阅读，自己判断。

丹尼特《达尔文的危险观念》。把自然选择看作算法的最彻底论述，第五章第三节的背景。

最后一句题外话。这张清单里的书，没有一本是讲人工智能的。这不是疏忽。本书的一个基本判断是：关于智能的问题，主要不是技术问题。技术提供了新的案例，但问题的形状在哲学里已经存在了两千多年。读技术报告能知道现在能做什么，读这些书能知道"能做什么"为什么值得一问。