何谓

AI的训练与学习？

论世界模型发生的困境和出路

王德生　著

德麦国际出版社
DEMAI INTERNATIONAL PRESS

出 版 信 息

书名	《何谓AI的训练与学习？——论世界模型发生的困境和出路》
著者	王德生
执笔	复合主体：王德生立题、拟定提纲与裁定，Claude 统领多个智能体并行执笔、交叉质检与统稿
出版	德麦国际出版社（Demai International Press）· 新加坡
编号	德麦国际专著第 266 号
ISBN	979-8-90690-487-4
定价	US$25
开本	170 mm × 240 mm
字数	约 22 万字（正文约 20.6 万字）
版次	2026 年 10 月第 1 版

AI 协作声明：本书的立题、二十万字提纲、母题「学习是特征纠缠的建立与训练」以及统稿时的各项裁定，由王德生给出；文字在其指令下由 AI（Claude）统领多个智能体协作成稿——十个智能体并行执笔，另由未参与写作的智能体两轮交叉质检、终审统稿——并经其审定。书中讨论的 Genie、V-JEPA、DreamerV3、Cosmos、Atlas、Waymo、DexTacWAM、TouchWorld 等技术系统，已据二〇二六年十月可见的论文摘要与官方页面核对；凡标〔待核〕处尚未逐字核对原始材料，成书后续核。书中陪读人物周敏、周小舟、老彭、林岚及各章日常场景均为构想。书中说到机器「能」或「不能」，说的是二〇二六年的机器；本书不构成工程、医疗或政策建议。

全书结构：出版信息、作者介绍、前言、导读、目录；导论与八编四十章；结语、参考书目（一百一十八种）、附录五件。

作 者 介 绍

王德生，SDE（显露 Show · 差异序列 Difference · 特征纠缠 Entanglement）本体论的创立者，德麦国际（Demai International Pte. Ltd.，新加坡）首席执行官。

生长于中国农村，1990 年进入湘潭大学。后获中国科学院计算数学博士学位，并在中国科学院从事博士后研究，研究方向包括 Voronoi 图、Delaunay 剖分、质心 Voronoi 剖分与有限元方法。曾在英国斯旺西大学与新加坡南洋理工大学工作。2005 年起定居新加坡。

他的思想体系先后经历四站：三视角·知识的结构和体系（2017 年起）、321 智慧系统（2022 年起）、SIO 本体论（2024 年起），以及 SDE 本体论（2026 年起；同年 5 月起，「结构」改读为「显露」）——从「并排放着的三样」，走到「互相生出来的三样」。

他在德麦国际出版社出版了多部专著。本书是他把 SDE 带回自己的老本行——算法与计算——的一次追问：机器所谓的「训练」与「学习」，究竟发生了什么；一个世界模型，怎样才算真的有了世界。



训练不是把一个现成的世界装进机器，

而是组织发生的条件；

学习，是特征纠缠的建立与训练——

世界模型的出路，不在一次封装一个全能的世界，

而在具体任务的互动里，先让一个物站稳。

—— 本 书

前 言

一只扶着后座的手

◆ 一、松开的那一下

这是我构想的场景。

二〇二六年秋天一个傍晚，长沙一个小区的环形路上，九岁的周小舟在学骑车。妈妈周敏弯着腰，一只手扶住车后座，跟着车小跑。车头左一扭、右一扭，小舟的肩膀绷得很紧，眼睛只盯着前轮。跑到第三圈，周敏没有出声，手慢慢松开了。车子歪了两下，又正了过来，小舟一口气骑出去十几米，才发现身后没人，一慌，脚一落地，车倒在草边。他爬起来，第一句话是：「你什么时候放的手？」

同一个傍晚，周敏手机上来了一条消息。发消息的是她以前的学生林岚，在深圳一家机器人公司做算法。林岚说，组里那台机械手这一轮训练跑完了，「学会」拧瓶盖了，测试台上一排瓶子，拧开了一大半。

周敏把两件事放在一起看，心里冒出一个问题：儿子在草边摔了一跤，算是「学会」了一半；机械手在测试台上拧开一大半，也叫「学会」。两个「学会」，用的是同一个词。它们说的是同一件事吗？如果不是，差在哪里？如果有一部分是，那一部分又是什么？

这本书，就是从这个问题开始的。

◆ 二、一个词，被用得太顺了

今天谈 AI，「训练」和「学习」这两个词用得极顺。一个大模型训练完了，我们说它学会了写作、学会了编程；一个视频模型看了大量录像，我们说它学会了物理；一台车跑了很多里程，我们说它学会了开车。用得顺，是因为这些说法大体不错，模型确实变得能做原来做不了的事。可用得太顺，就把几个本该问的问题滑过去了。

第一个问题：训练究竟形成了什么？是把世界上的规律一条条搬进了参数，还是只记下了一张巨大的相关表？两种答案都说不通小舟那一下。他学会骑车，脑子里并没有多出一条「平衡规律」；可也不能说他什么都没形成，他明明从扶着才能走，变成了松手还能走。

第二个问题：谁在学习？小舟在学，可那只扶着后座的手也在学。周敏什么时候松、松多少，是她看着儿子的肩膀一点点判断的。林岚那台机械手也一样，它拧开的每一个瓶盖背后，有人挑瓶子、有人定奖励、有人看日志、有人决定下一轮改哪里。只把最后那一步叫作学习，等于只把车子往前走的那几米算作骑车。

第三个问题：它学会的那个世界，是怎么来的？小舟眼里的小区环形路，学车前后是两个样子。学车前，那只是一条路；学车后，哪里有坡、哪里铺了减速带、哪里有草可以摔，都一一冒了出来。世界不是先摆好了，等他去认，而是跟着他的本事一起长出来的。

◆ 三、为什么是二〇二六年

这三个问题不是新问题，教育家、心理学家问过很多年。为什么偏偏在今天，要专门写一本书来问「何谓AI的训练与学习」？

因为有三件事凑到了一起。

第一件，学习的主体变多了。过去说学习，说的是人；今天，一个模型的训练、一个人和模型来回问答、一个团队拿使用记录去改模型，都被叫作学习。如果不先把这些层次分开，「AI学会了」这句话就会越说越含糊，好事坏事都说不清该记在谁的名下。

第二件，世界模型成了热门方向。从视频里学世界、从状态和行动后果里学世界、把视觉声音动作和空间放进同一个模型里学世界，各条路线这两年进展很快，有些已经接到了机器人和自动驾驶上。热门方向最容易被一个未经审查的前提带着走：世界是现成的，只要数据够多、模型够大，就能把它完整地装进去。这个前提对不对，值不值得重新审视，正是本书要问的。

第三件，普通人每天都在和 AI 一起学。孩子用它写作文，老师用它备课，研究小组用它推演。一次流畅的对话很像学会了，可它到底改变了什么、改变能留多久，是每个家长、每个老师都会碰到的实际问题。

这里至少有五件常被混成一件的事。你在一次对话里纠正了模型，它后半段改了口，这是暂时的语境变化；你把资料存进一个检索库，它下次能查到，这是外部记忆；它调用工具，看运行结果再改，这是控制反馈；公司拿大量纠正记录去改模型的参数，这是参数训练；改过以后在各种新问题上都稳定地表现出新本事，这才谈得上持久学习。小舟也一样：妈妈扶着时骑得好，不算；松手骑了十几米，算一半；过一个寒假再上车，还能骑走，才算真学会。五件事层层不同，混在一起，「学会」这个词就什么也说明不了。

三件事合在一起，「何谓训练、何谓学习」就从一个书斋问题，变成了一个工程问题、一个教育问题，也是一个每天都在发生的生活问题。

◆ 四、本书的问题意识：在纠缠中、经差异、成显露

本书的理论底子，是 SDE 发生学。用小舟的车把它说一遍。

学车之前，小舟的眼睛看到车身倾斜，屁股感到座子晃，手攥着车把，脚踩着踏板，这几样东西各管各的。学车的那几个傍晚，它们一次次在同一件事里碰头：车往左歪的时候，眼睛、屁股、手、脚同时有了反应，有的反应对，有的反应错，摔了几回，慢慢缠到了一起，一动就互相牵连。这种「一块块经验缠到一起、彼此牵动」的状况，本书叫作特征纠缠，记作 E。

缠到一起以后，骑车中一个个细小的差别，往左偏一点、风大一点、地面有个小坑，就能被这团纠缠接住、传递、校正下去。这种差别一路运行的过程，本书叫作差异运行，记作 D。

最后，车子稳稳向前，小舟自己也觉得「我会骑了」，那条环形路也在他眼里换了模样。这种显现出来的东西，本书叫作显露，记作 S。显露稳下来，成了日后可以反复依靠的样子，就是结构。

所以本书的说法是：学习，是特征纠缠的建立与训练；它一旦发生，显露和差异运行都跟着改变。训练者能做的，是安排经验在哪里相遇、按什么次序相遇，也就是组织发生的条件；纠缠本身，是在相遇中长出来的。周敏那只手扶了三圈、在第三圈松开，就是在组织条件。她不能替小舟去缠。

这三样东西不是三个零件，而是一件事的三个方向，彼此规定：显露靠差异运行和纠缠而来，差异运行又被显露和纠缠塑造，纠缠也被前两者改写。本书写成三个方程，S=F(D,E)、D=G(S,E)、E=H(S,D)，少了哪一个，解释里就留下一道缝。小舟摔的那一跤就是例子：倒在草边是一次显露，它马上回头改了下一圈的差异运行，他知道车一歪要往同一边拧车把；几回下来，纠缠也重新缠过了。学习从来不是只改输出，三个方向一起在变。

拿这个说法去看机器，问题就清楚了：一个模型训练完了，要问的不是它装进了多少世界，而是它在什么样的相遇里缠出了什么，这团纠缠让哪些差异能够运行，让什么显露出来，又在哪里会失效。

◆ 五、两种看世界的前提

常识里有一幅很顺的图景：世界早已在那里，物体有确定的形状、重量和材质，规律写在自然里，人和机器的任务是去把它们找出来。本书把这种立场叫作发现学。

它在很多场合很好用。可拿来解释学习，它会把最要紧的一步藏起来：对象是怎样成为「我能把握的对象」的。小舟学车前，减速带一直在那里；可在他学会以前，它不是一个需要提前抬一抬屁股的东西。说它「早就在那里，等着被发现」，没有错，却什么也没解释。

本书采用的是另一个立场，叫发生学：对象和规律，在特征纠缠中经差异运行显露出来，显露之后还要在新的相遇里一次次接受检验。发生学不说世界随人心意，减速带不会因为小舟不信它就变平；它也不说规律是人随便定的。它争的是另一件事：稳定不等于先在，可靠不等于没有边界。

这一区分落到世界模型上，后果很具体。如果世界是现成的，世界模型的目标就是一次封装一个全能的世界；如果世界在互动中持续发生，世界模型就该先问：在哪个任务里，哪个物已经站稳了，站稳的条件是什么，它在哪里会开始松动。

◆ 六、两个复合体，谁也不是孤零零的学习者

回到那条消息。林岚说「机械手学会了」，这是一种省略的说法。展开来看，参与这一轮学习的，有设计瓶子摆放的人，有写奖励的人，有看日志的人，还有在它之前写下无数操作经验、做出无数工具的人。大模型更是如此：它所学的语言，是千千万万人在不同年代、不同处境下写下的东西，经过挑选、训练、评价才缠到一起。

本书把参与训练的这一整个组合，叫作集体人类—AI复合体；把一个具体的人和模型来回问答、纠正、共同完成一件事的组合，叫作个人—AI复合体。前一种复合体里发生的，是一个有相对同一性、带着集体性的语言世界；后一种复合体里，这个共同的底子继续落到某一个人的问题、经历和价值里去发生。

这一条全书都要守：不把 AI 写成脱离人类历史的独立发生者。把成绩全记在机器名下，看不见里面沉淀的人类劳动；把偏差全记在机器名下，也看不见数据选择和评价标准留下的人为痕迹。

◆ 七、先让一个物站稳

小舟骑熟以后，会发现车把手感、链条声音、轮胎压过减速带的那一下颠，都是这辆车的一部分。他认得这辆车，不只靠看。

再看林岚那台机械手。测试台上拧开了一大半，拧不开的那一小半是什么？可能是一个瓶盖比别的紧，可能是一个瓶子外面沾了水，可能是一个塑料瓶身一捏就瘪。这几样，在摄像头里差不多是同一个样子；手一上去，差别才冒出来。光靠看录像学，录像里若从没把这几种情形分开，模型再会画下一帧，也画不出这一点区别。拧瓶盖靠的不只是看见瓶盖在哪，还有手上那股劲是滑了还是咬住了，那一声「咔」响没响。

本书把这一点推到世界模型上：视觉、听觉、触觉各自都是一套完整的 SDE，有各自的纠缠、差异运行和显露；一个物，是在这三套过程的互动中，形成同一性、维持同一性，在出错时被察觉，又被重新建立起来的。

这是本书提出的研究路线，不是已经定好的结论。本书不预设三种感官在一切任务里都缺一不可，也不预设统一的大架构一定走不通。本书主张的是，无论走哪条路，都该围绕具体的任务组织必要的纠缠，用可以检验的能力和清楚的适用边界形成封装，并且事先写好它会在什么条件下失败。会骑自行车，不等于会骑独轮车；一个可靠的封装，要连同它的边界一起被知道。

把上面这些收成一句话，就是全书的判词：

> **训练不是把一个现成的世界装进机器，而是组织发生的条件；学习，是特征纠缠的建立与训练——世界模型的出路，不在一次封装一个全能的世界，而在具体任务的互动里，先让一个物站稳。**

这句话在导读、导论末节和结语末节还会一字不改地出现。

◆ 八、全书的路怎么走

全书由导论、八编四十章和结语组成。

导论从两个「刹车」进入：一个模型写出「刹车」这个词，一位老司机在雨夜真把车停住，中间隔着什么。第一编给学习重新打地基，讲纠缠怎样在训练中发生，三个方程怎样共同解释学习，复杂性怎样进入学习，能力怎样被封装、失效、解封。第二编回到人，讲理念、现实、自我三个世界怎样各自发生又彼此改写。第三编讲人怎样进入大模型训练，一个集体的语言世界怎样发生，边界在哪里。

第四编逐一检查世界模型的几条技术路线，各以什么为世界，背后藏着什么前提。第五编讲视觉、触觉、听觉中物怎样显露，三套 SDE 怎样缠成一个物，并为「物的意识」给出可以检验的功能判据。第六编正面处理「一个模型穷尽全部物理发生」的设想，讲信息负担、显露预测与行动可靠的距离，讲任务怎样划定世界的

范围，以驾驶为专门例子。第七编把主张变成可做的事：围绕共同事件组织训练，从被动记录走向主动试探，设计可以失败的实验。第八编回到每个使用者，讲集体的底子怎样成为个人的条件，教育与科研怎样转向，最后提出研究纲领。结语把已经论证的、仍待验证的、尚未解决的分成三本账。

◆ 九、本书怎样输

本书把三种话分开说：SDE 的本体论立场、可以被实验推翻的机制假说、公开文献里的经验事实。某个模型这次做成了、那次没做成，不能直接拿来证明或推翻立场；要从技术事实走到立场，中间必须经过机制假说，并把每一步摆出来。

机制假说是会输的，本书也希望它们会输得清楚。哪一条在什么实验结果下算判负，第34章写了对照与判负条件，第40章和结语把全书的假说逐条列出，附录二做成了判决卡。立场本身也会输：如果按它设计的实验一个接一个落空，而它只能靠改口保住自己，那它就该被认真重审。书里涉及的技术资料，只说公开机制；拿不准的细节标了〔待核〕。

◆ 十、这本书不做什么，写给谁

这本书不评判哪一家公司、哪一条技术路线的好坏，也不预言哪条路一定成、哪条路一定败。它不作医学诊断，不做政策表态；书里讲到的学习读数，不能拿来考核任何孩子、老师或工程师。书中的周敏、小舟、老彭、林岚，以及他们的故事，都是构想的。

它写给三种读者。一种是做模型、做机器人、做自动驾驶的人，本书想请你们在问「数据够不够、模型够不够大」之外，多问一句：这里纠缠的是什么，它在哪里会失效。一种是老师和家长，孩子身边多了一个什么都能答的 AI，学习这件事反而更需要想清楚。还有一种，是关心 AI、愿意动脑的普通读者。

那只扶着后座的手，总要松开。松开以后车子还能走多远，在哪里会倒，倒了以后怎样重新骑起来，这是小舟要学的，也是今天的 AI 要学的。

导 读

这本书怎么用

◆ 一、全书一句话

全书只有一句判词：

> **训练不是把一个现成的世界装进机器，而是组织发生的条件；学习，是特征纠缠的建立与训练——世界模型的出路，不在一次封装一个全能的世界，而在具体任务的互动里，先让一个物站稳。**

它在前言第一次出现，在导论末节和结语末节再出现。前半句讲训练与学习是什么，后半句讲世界模型往哪里走。

前半句里有两个动作要分清。训练是安排条件的活动：挑哪些经验、按什么次序、用什么标准打分。学习是条件之中发生的纠缠变化，同样的训练可以有不同的学习后果，练得多也不等于学得会。后半句里的「先让一个物站稳」，说的是起点，不是终点：先在一个具体任务里让一个物形成同一性、经得起新的动作，再往外扩，而不是一开始就许诺一个什么都能模拟的世界。

◆ 二、五条判决

判词拆开，是五条判决。每一条都有自己的主场，在那里被正式论证；结语逐条验收，附录二把它们做成判决卡。

判决	一句话	主场章
一 · 学习即纠缠	学会不是把知识搬进脑子或参数，而是特征纠缠的建立、稳定、分化与重组，显露和差异运行随之改写	第一编（第1—5章）

判决	一句话	主场章
二 · 三界共生	人在学习中同时发生理念、现实、自我三个世界，主体跟着世界一起发生	第二编（第6—10章）
三 · 复合体是发生单位	大模型训练发生的是集体人类—AI复合体的语言世界，具体人的使用、反馈与再训练让它进入个人—AI复合体	第三编（第11—15章）、第八编（第36—40章）
四 · 物在三种 SDE 的纠缠中站稳	视觉、听觉、触觉各自是完整的 SDE，物的同一性在三者互动中形成、维持、失效与重建，显露预测不保证行动可靠	第五编（第21—25章）、第六编（第26—30章）
五 · 任务组织世界，封装要能失败	围绕具体任务组织必要的纠缠，以可检验的能力和适用边界形成封装，并写好它会失败的条件	第四编（第16—20章）、第七编（第31—35章）

◆ 三、两条收窄

为免读过头，先把两条话说在前面：本书不主张统一架构必然失败，不主张三种感官对所有任务都不可替代；这两条是研究边界，不是前提。

也就是说，如果有一天一个统一的大模型在多种任务上都做得可靠，并且能清楚交代自己在每种任务上的适用条件和失效方式，本书不认为这是自己被打败了，只会说：它的成功，也该交出这样一份边界报告。哪些任务需要哪几种感官参与、参与到什么程度，本书交给实验去回答。

◆ 四、几条读法

全读。按顺序从导论读到结语。前两编是地基，讲人；第三编讲语言模型；第四到第七编讲物理世界模型；第八编回到每个使用者。第3章的三方程和第5章的封装，后面反复用到，跳过它们，后面会读得吃力。

只关心 AI 工程的读者，可以读导论，再读第三编（大模型训练里人在哪里、语言世界的边界）、第四编（几条世界模型路线的公开机制与前提）、第六编（信息负担、显露预测与行动可靠、任务边界、驾驶）、第七编（训练组织与可以失败的实验），最后读第40章的研究纲领。遇到术语卡住，回头翻第3章。

只关心教育与个人学习的读者，可以读第一编（尤其第1、2、5章）、第二编，再读第八编的第36到38章：集体的底子怎样成为个人的条件，一次流畅对话和持久学习差在哪里，老师和家长该把力气花在哪里。第四编到第七编的机器细节可以先放下。

只想看出路的读者，可以读第28到30章（任务怎样组织世界、专门模型怎样协作）、第七编、第40章和结语。读完若想知道这条出路凭什么这样走，再回到第一编和第五编。

只有一个晚上的读者，读导论、第5章、第24章、第34章和结语。导论提出问题，第5章讲能力怎样封装、怎样失效，第24章用一只陌生弹性球讲三种 SDE 怎样缠成一个物，第34章讲本书准备怎样输，结语把账分清。

书里有四位陪读的人，都是构想的：周敏，四十一岁，长沙的中学物理老师；她的儿子周小舟，九岁，正在学骑车、学游泳；她的邻居老彭，五十八岁，开了三十年车，如今跑网约车；她以前的学生林岚，二十九岁，在深圳做机器人算法。小舟多半出现在讲身体经验的地方，老彭出现在讲驾驶、听车、手感的地方，林岚出现在讲工程现场的地方，周敏出现在讲教育和追问的地方。跳着读时碰到他们，不必回头去找来历。

同一个例子在全书里只当一次主角。骑车与和面在第一编，婴儿认妈妈与被烫一下在第二编，家乡话与菜谱在第三编，看录像学游泳在第四编，挑西瓜与摸黑找开关在第五编，老彭的雨夜在第六编，机器人拧瓶盖、捏豆腐在第七编，备课、写作文与科研小组在第八编。别处再提到，只是一句带过，并指回主场。

◆ 五、怎样推翻这本书

本书的话分三层，推翻的办法也分三层。

经验事实，是公开论文和技术报告里能查到的东西。书里拿不准的细节都标了〔待核〕；哪一处说错了，拿出处来改就是。

机制假说，是本书提出、可以被实验支持也可以被推翻的说法，这是最欢迎被检验的一层。举几条：在同等数据、算力和先验之下，如果主动互动训练在对象同一性、迁移和失配恢复上，持续不比被动记录训练好，而且换多种任务、多种物体都是这样，那么第七编的训练组织方案要改写；如果在接触丰富、光看外观分不清的任务上，纯视觉或简单拼接的系统能达到同样的迁移和恢复水平，那么至少在这些任务上，三种感官的共同互动不是必要的；如果在多种模型里都找不到可测的内部量，对应纠缠的建立、稳定、分化、重组，并能预测能力的出现与失效，那么「E」在 AI 这一侧就只是比喻。第34章第五节写了对照实验的判负条件和判负后的修正程序，第40章和结语第五节把全书的假说逐条列出判负条件；附录二「判决卡与推翻条件」把五条判决各自牵连的假说、判负结果和要退回的章节做成了卡片，查起来最方便。

本体论立场，不由某个模型一次成败来判。它也会输，只是输法不同：如果按它设计的实验大多判负，它只能靠改口保住自己，术语越用越含混，那么「学习是特征纠缠的建立与训练」这个母题就该被重审。

推翻本书，不必先接受本书的术语。清楚的定义、可核的数据、透明的方法，就够了。一组认真做的对照实验，可能打掉一条假说；一份可核的技术报告，可能改掉一处事实。本书在结语里把账分成三本：已经论证的区分与澄清，仍待验证的假说，尚未解决、连怎样检验都还说不清的问题。读者若要动手，第二本账是最好的起点。

◆ 六、随身三问

读任何一章，合上书，问自己三句：

这里纠缠的是什么？差异是怎样运行的？显露出来的东西靠什么站稳、又在哪里失效？

拿去问一个模型的回答、一台机器人的动作、一个孩子刚学会的本事，都能用。

试用一次。小舟松手骑出去十几米：纠缠的是眼睛里车身的倾斜、屁股上的晃、手上的劲、脚下的节奏；差异在每一个瞬间的偏一点、慢一点之间运行，被那团刚缠好的东西接住、校正；显露出来的「我会骑了」，靠的是平整的环形路和刚才那三圈扶持，在他发现身后没人、一慌的那一刻失效了。再问林岚的机械手：纠缠的是什么，画面、关节角度和那一点阻力有没有缠到一起；差异怎样运行，手打滑的那一下有没有被接住；拧开一大半，靠什么站稳，拧不开的那一小半，失效在哪里。三问问完，「学会了没有」这个笼统的问题，就拆成了几个能查、能改的问题。

本书不想在「AI懂」与「AI不懂」之间站队，三问就是为此准备的。

◆ 七、每章怎么写

每一章开头有一句「这一章的问题」，整章只回答这一个问题。正文分四到六节，大体依次是：先给出命题，再讲机制，然后落到一个日常案例上，接着交代反例和边界，最后把问题推给下一章。案例多用骑车、和面、挑西瓜、开车、拧瓶盖这类家常事；技术系统作为对话和检验的对象出现，不当主例。

每章末尾有一段「带走的话」，一两句收住这一章。如果时间紧，只读每章的问题和带走的话，也能看出全书的骨架。

带走的话之后，还有两行「小账」。第一行回答「论证落在哪一层」：这一章主要说的是本体论立场、机制假说，还是经验事实，或者几层各占多少。读者据此知道这一章的话该怎样听、怎样驳。第二行回答「世界模型补了哪一条」：这一章给世界模型的发生补上了哪一条条件或边界。四十章的第二行连起来读，就是本书关于世界模型的一张清单。

◆ 八、术语先说白话

书里的术语都先说白话，再括注。下表是最常用的十个；完整的说法和出处，见附录一「术语速查」。

术语	白话
S（显露）	显现出来、能被看到用到的样子；稳下来的显露就是结构
D（差异运行）	一个个细小的差别，怎样一路传递、互相校正下去
E（特征纠缠）	原本各管各的经验缠到一起，一动就互相牵连
三方程	S=F(D,E)、D=G(S,E)、E=H(S,D)：三维彼此决定，缺一式就是解释里的一道缝
六路径	三维发生变化的六种先后次序，如先有差异、再缠、再显露（D→E→S）
三原理	两维闹了矛盾，逼着第三维改变：先攒张力，再变形状，再回写
三态	混沌态、介生态、秩序态：乱成一团、半成形、站稳了
三界	理念世界、现实世界、自我世界：想得通的、碰得到的、认得出自己的
封装	学会的本事打包好、可以随时调用，连同它的适用边界一起
复合体	人和 AI 合成的发生单位：训练时是集体人类—AI复合体，使用时是个人—AI复合体

记住这张表，正文就能读下去。读到哪一处觉得术语压人，那是本书写得不好，不是读者读得不够。

目 录

作者介绍 · · · · · iii
前言　一只扶着后座的手 · · · · · v
导读　这本书怎么用 · · · · · xii

导论　世界模型问题的重新提出 · · · · · 1

第一编　学习的本体论重建 · · · · · 16
第 1 章　学习究竟发生了什么 · · · · · 17
第 2 章　训练如何使特征纠缠发生 · · · · · 24
第 3 章　三方程如何解释学习 · · · · · 32
第 4 章　复杂性如何进入学习 · · · · · 40
第 5 章　从纠缠形成到能力封装 · · · · · 47

第二编　理念、现实与自我如何发生 · · · · · 55
第 6 章　每个人的世界如何建立 · · · · · 57
第 7 章　理念世界的发生 · · · · · 65
第 8 章　现实世界的发生 · · · · · 72
第 9 章　自我世界的发生 · · · · · 80
第 10 章　三个世界如何相互改写 · · · · · 87

第三编　人类与AI复合体的语言世界发生 · · · · · 94
第 11 章　人怎样进入大模型训练 · · · · · 96
第 12 章　语言token与特征纠缠的建立 · · · · · 103
第 13 章　集体性的语言世界如何发生 · · · · · 110
第 14 章　语言生成与理解能力的发生 · · · · · 117
第 15 章　集体语言世界的边界与修正 · · · · · 124

第四编　世界模型的技术路线与哲学前提 · · · · · 131
第 16 章　世界模型究竟以什么为世界 · · · · · 132
第 17 章　从视频中学习世界的机制 · · · · · 140
第 18 章　从内部状态与行动后果学习 · · · · · 148
第 19 章　多模态与空间模型的进展 · · · · · 156
第 20 章　发现学前提怎样限制世界模型 · · · · · 163

第五编　三种SDE的纠缠与物的意识 171
第 21 章　视觉SDE中的物如何显露 172
第 22 章　触觉SDE中的接触与持续影响 180
第 23 章　听觉SDE中的事件与物 188
第 24 章　三种SDE如何形成特征纠缠聚合体 195
第 25 章　物的意识如何发生及接受检验 203

第六编　统一物理模型的困境与任务边界 211
第 26 章　物理发生的信息负担 212
第 27 章　为何显露预测不能保证行动可靠 219
第 28 章　任务如何组织世界的发生范围 226
第 29 章　驾驶世界模型的专门发生 233
第 30 章　专用模型的协作与共享 241

第七编　发生式世界模型的训练与验证 248
第 31 章　围绕共同事件组织训练经验 249
第 32 章　从被动记录到主动试探 257
第 33 章　跨模态预测与纠缠重组 265
第 34 章　建立可以失败的实验与评价 272
第 35 章　可靠性形成与持续封装 280

第八编　从集体世界到具体复合体的学习 287
第 36 章　集体同一性怎样成为个人发生条件 288
第 37 章　个人与AI如何继续发生语言世界 295
第 38 章　教育如何转向特征纠缠的训练 302
第 39 章　科研如何成为复合体的世界发生 309
第 40 章　AI训练与学习的新研究纲领 316

结语　学习与世界的持续发生 324

参考书目 338

附录 348
附录一　术语速查 348
附录二　判决卡与推翻条件 353
附录三　世界模型技术路线速查 359
附录四　全书主例与人物 362
附录五　全书说明 364

导 论

世界模型问题的重新提出

> **这一章的问题**：一个模型能写出「刹车」这个词，和一辆车在雨夜真正停住，中间隔着什么——训练究竟形成了什么，谁在学习，模型的世界又是怎样建立起来的?

◆ 一、一个词与一次停车

先讲一个场景。这是我构想的场景，主角是本书里常常出场的老彭。

十月里一个下雨的夜晚，长沙城南一条两车道的老街。老彭五十八岁，开了三十年车，如今跑网约车。车速不快，四十公里上下。雨刷一下一下扫着，挡风玻璃上的路灯光拉成一条条亮线。前面右侧停着一排电动车，忽然有一个穿深色雨衣的人推着车从缝里钻出来，往路中间走。老彭的右脚几乎在看见那团黑影的同时离开了油门，落到刹车踏板上。他没有一脚踩死，而是先重后稳地往下压。脚底下传来一阵细密的顿挫，那是防抱死系统在工作；耳朵里，轮胎碾过积水的嘶嘶声变了调；方向盘在手里轻轻一沉，他微微回正。车停住的时候，离那个人还有一两个车身。乘客在后座「哎」了一声，老彭摇下窗，什么也没说，等那人过去。

同一个晚上，另一个地方，有人在手机上问一个大语言模型：「雨夜行车，前方突然有人横穿马路，司机应该怎么办？」模型几乎立刻给出一段话，第一句是：「应立即刹车，同时注意雨天路面湿滑、制动距离变长……」后面还有几条建议，条理清楚，没有一处说错。

两件事里都出现了「刹车」。在模型那里，「刹车」是一个词，在一段文字里出现得恰当，读的人点头。在老彭那里，「刹车」是一串事情：眼睛先分辨出黑影是人不是树影，脚掌掂量出踏板该压多深，耳朵和脚底同时报告轮胎抓不抓得住

地，手感帮他判断车身有没有偏。最后，一辆一吨多重的车在湿滑的沥青上，在某个确定的位置停住了。

本书就从这两个「刹车」之间的距离出发。

这段距离很容易被说错，有两种常见的说错法。一种说：模型只会玩文字，它说出的「刹车」是空的，不算真懂。这个说法太轻慢了。模型写出那段话，背后是千千万万人写下的驾驶经验、交通规则、事故报道、驾校教材，被组织到一起之后又在一个具体提问下重新排出来。那段话对一个新手司机是有用的。语言的世界也是真实的世界，人类大部分的经验正是靠它代代相传。另一种说法刚好相反：只要模型把世界上关于驾驶的文字、视频、传感器记录全部吃下去，它就等于掌握了驾驶，说出「刹车」和做到「刹车」只是输出接口不同。这个说法太轻率了。老彭停车的那几秒钟，判断对不对，不由任何读者来裁决，而由路面、轮胎、车重、雨水和那个行人的位置一起裁决。说对一百次而做错一次，在路上可能就是一次事故。

所以要问的不是「模型懂不懂」这种笼统的问题，而是更具体的三个问题。

◆ 二、三个追问：训练形成了什么，谁在学习，世界怎样建立

第一问：训练究竟形成了什么？

今天谈 AI，常用两个词：训练和学习。工程上，训练有明确的含义：准备数据，设定目标，让参数沿着减小误差的方向一轮轮调整。可是这些参数调好之后，「形成」的到底是什么？一种流行的回答是「模型学到了世界的规律」，好像规律原本就躺在那里，训练只是把它找出来、存进参数。另一种回答是「模型只是记住了统计相关」，好像除了一张巨大的频率表，什么也没有形成。两种回答都没有说清：经过训练，一个原来不会回答问题的系统，变得能在新问题面前组织出合适的回答，这种「能」是从哪里来的、以什么方式存在、在哪些条件下会失灵。

本书的回答从日常经验开始。一个人学会骑车之后，并不是脑子里多了一条「保持平衡的规律」，而是眼睛看到的倾斜、屁股感到的晃动、手上扶把的力气、脚下蹬踏的节奏，原本各管各的，现在缠到了一起，一动就互相牵连。这种「一块

块经验缠到一起、彼此牵动」的状况，本书叫作特征纠缠（E）。骑车时一个个瞬间的细微差别，往左偏一点、慢一点、风大一点，被这团纠缠接住并一路传递下去，这叫差异运行（D）。最后车子稳稳向前，人也觉得「我会骑了」，这是显露（S）。学习，就是特征纠缠的建立与训练；它一旦发生，显露和差异运行也都跟着改变了。这是全书的母题。

拿这个说法回头看模型：训练形成的不是一份世界的副本，而是一团有组织的纠缠，它让某些差异能够顺畅运行，让某些回答能够显露出来。说它「学到了规律」，夸大了；说它「只是统计」，又抹掉了这团组织的真实作用。

第二问：谁参与了学习?

说「模型在学习」，主语只有一个。可是回想一下模型写出「刹车」的那段话：先有千千万万个司机开车、出事、总结、写书，有交通部门订规则，有记者报道事故，有网民讨论；再有工程师挑选哪些文本进入训练，有人设计训练目标，有人给回答打分、排序、纠正；最后才有参数调整。只把最后这一步叫作学习，等于把一条河只算最后入海的那一段。

本书把参与训练的这一整个组合叫作集体人类—AI复合体：人类的历史活动、数据与制度、工程与评价，和模型一道，构成一个发生单位。模型不是从天外来的独立学习者；它学会的东西，首先是人类集体经验在一个新载体上的重新组织。等到一个具体的人拿起手机去问、去追问、去纠正，模型又和这个人一道构成了个人—AI复合体，在那里继续发生新的东西。这两个层次，本书会反复区分。

第三问：模型的世界如何建立?

「世界模型」这个说法这几年很热。它背后有一个直觉：模型若要可靠地行动，就要在内部有一个世界，能预演「如果我这样做，会发生什么」。问题在于，这个内部世界从哪里来。如果世界是一个现成的东西，那么建模就是把它尽量完整地复制进来：拍更多视频，收更多传感器数据，造更大的模型，最后得到一个能模拟一切的全能世界。如果世界本身就在互动中持续发生，那么模型的世界也只能在互动中一点点建立：先有某些对象在某些任务里站稳，再扩展、再失效、再重建。

老彭的世界就是这样来的。他三十年前第一次在雨里开车，并不知道湿路上刹车要先重后稳，也听不出轮胎打滑前那一点声音的变化。这些东西不是他「发现」了一个早就写好的雨路手册，而是在一次次雨夜里，眼睛、耳朵、脚底和手掌的经验缠到一起，才让「这条湿路、这辆车、这一脚」成为他能把握的对象。

◆ 三、学习与世界形成；「发现既成世界」的前提

导论要展开五项论证。前两项放在一起讲，因为它们是一个硬币的两面：学习和世界形成是什么关系；把世界当作「既成的、等待被发现的」，这个前提本身有什么问题。

先看学习与世界形成。还是从老彭说起。三十年前刚学开车时，雨夜的路在他眼里只是「黑一点、滑一点」的路。三十年后，同一条老街对他显露出完全不同的样子：路灯下一片发亮的反光，是「这里可能积水，脚要提前收」；前车尾灯忽然一闪，是「它在点刹，前面有事」；路边一排停得歪歪扭扭的电动车，是「缝里随时可能钻出人来」。这些东西一直就在路上，可只有在他身上缠成了那套分寸之后，它们才一个个成了「要当心的对象」。学会雨天开车，不只是老彭身上多了一项本事，他的路也换了一个样子。能力和世界是一起发生的。小舟学游泳之后看泳池的眼光变了，也是同一回事，第17章还会讲到。

这一点对 AI 同样成立，只是要说得小心。一个语言模型经过训练，能在「雨夜」「横穿」「制动距离」这些词之间建立稳定的牵连，于是在它的语言世界里，这几样东西成了彼此相关的对象。一个机器人在大量抓取试验之后，能把「软的」「会滑的」「一捏就变形的」区分开，于是在它的操作世界里，物开始有了不同的样子。不同的是，老彭的世界同时连着身体、情绪、他对乘客的那份责任；模型的世界连着什么、缺着什么，要具体看它在什么条件下训练。这就是第一项论证的结论：学习不是在一个不变的世界里增加知识，而是让世界随着特征纠缠的建立一起发生。

再看「发现既成世界」的前提。常识里有一幅很顺的图景：世界已经在那里，物体有确定的形状、重量、材质，规律写在自然里，人和机器的任务是去发现它

们，越发现越多，越接近完整。本书把这种立场叫作「发现学」。它在很多场合很好用，物理学教科书大体就是按这幅图景写的。可是一旦拿来解释学习，它就会把关键的一步藏起来。

试想周敏，小舟的妈妈，一位中学物理老师，给学生讲摩擦力。课本上的公式简洁漂亮，似乎说出了湿路刹车的全部道理。可老彭不会这个公式，却能在雨夜停住车；一个会背公式的学生第一次在雨里开车，未必停得住。公式和老彭的脚，都和湿路有关，但它们是两种不同的发生：一种在理念世界里站稳，一种在现实世界里站稳。「发现学」会说，公式是规律本身，老彭的脚只是近似地用到了规律。本书的说法是：公式和老彭的手感，都是在各自的纠缠中经过差异运行显露出来的；它们彼此可以相互改写，却谁也不是谁的底本。这个立场叫作「发生学」。

把发现学用在世界模型上，最直接的后果是一种工程想象：既然世界是现成的，那就把它尽量全面地封装进一个模型，模型越大、数据越多，离那个完整的世界就越近。本书不认为这条路走不通，也不预言它必然失败；我们要指出的是它的前提。这个前提假设了对象、属性和规律在训练之前就已完成，训练只是去逼近它。可是从小舟学骑车到老彭学雨天开车，我们看到的是：对象在互动里才站稳，站稳的方式取决于用它做什么，换一个任务、换一副身体、换一种工具，站稳的方式就变了。一个前提如果和学习的实际过程对不上，建立在它上面的工程目标就值得重新审视。

为免误解，还要说清发生学不主张什么。它不主张世界随人心意而变，湿路不会因为老彭不信它滑就不滑；它也不主张物理规律是人随便约定的，摩擦力公式在大量情形下准得惊人。发生学承认现实有阻力，而且正是这种阻力在检验每一次显露站不站得住：小舟学车时摔的那一跤、老彭脚底那阵顿挫，都是现实在说话。发生学争的是另一件事：对象和规律以什么方式成为「我们的」对象和规律。它们不是先打包好、再被逐件取出的；它们在 E 中经 D 显露为 S，并且在显露出来以后，仍然以稳定的结构态（结构是显露的稳定态）继续接受新的互动的检验。稳定，不等于一次完成；可靠，也不等于没有边界。

这个区分对世界模型很要紧。如果把「稳定」误当成「先在」，就会以为模型只要把那些稳定的东西都收集齐了，就拥有了世界；如果看到稳定是在互动里形成、也会在新互动里松动，就会把注意力转向另一些问题：哪些对象在哪些任务里已经站稳，站稳的条件是什么，什么迹象表示它开始松动，松动之后怎样重新建立。这些问题，正是本书后面几编要一一落实的。这里要提醒一句：这是本体论层面的立场之争，不能靠某个模型这次做成了、那次没做成来裁决。下面讲证据方法时还会回到这一点。

◆ 四、人类—AI复合体：谁是发生的单位

第三项论证要回答：讲 AI 的训练与学习，应当把什么当作一个完整的发生单位。

日常里有个现成的例子。周敏教了十几年物理，她讲摩擦力的那套讲法，不是她一个人想出来的：有她当学生时老师的讲法，有教研组一代代传下来的例题，有她自己被学生问住过的那些问题，还有课本、考纲、实验器材。今天站在讲台上的「周老师讲摩擦力」，是一个人和一长串历史、一整套工具制度合在一起才发生的。说「周老师学会了怎么讲摩擦力」没有错，但如果只盯着她一个人的脑袋，就看不懂这个本事是怎么来的，也看不懂它为什么在新一届学生面前又要改。

大模型的训练比这走得更远。它所吸收的文本，是极多的人在不同年代、不同语言、不同处境下写下的东西；哪些文本被采进来、哪些被滤掉，是一批工程师按某些标准决定的；训练目标、评价方式、人类反馈的打分规则，又是另一批人设计的；训练所需的电力、芯片、数据中心，背后是整个产业和制度。所有这些和模型参数一道，才构成了训练这件事。本书称之为集体人类—AI复合体。在这个复合体里发生出来的，是一个具有相对同一性与集体性的语言世界：说它有「相对同一性」，是因为同一个模型面对不同的人，大体守着同一套表达习惯和推理约束，你问它「刹车」，我问它「刹车」，它不会在你这里说成踩油门；说它「集体」，是因为这套习惯不属于任何一个人，而是许许多多人的经验在一个新的载体上重新缠在一起。

这里有一条全书都要守的纪律：不把 AI 写成脱离人类历史的独立发生者。人们说「AI 自己学会了某某」，很多时候只是一种省略的说法；一旦把省略当了真，就会出现两种误判。一种是把模型的成就全部记在机器名下，看不见其中沉淀的人类劳动和判断；另一种是把模型的偏差也全部记在机器名下，看不见数据选择和评价标准留下的人为痕迹。这两种误判在讨论 AI 的责任、版权、教育用途时都会带来麻烦，第11章会细讲人怎样进入大模型训练。

个人—AI复合体是另一个层次。林岚，周敏以前的学生，现在在深圳一家机器人公司做算法。她白天用一个大模型帮忙查资料、改代码、整理实验记录。每一次她提问、追问、指出模型哪里错了，模型的回答也跟着调整，她自己的想法也在被回答推动着改变。这时发生的东西，既不完全属于她，也不完全属于模型，而属于「林岚—模型」这个具体的复合体。

但这里必须马上分清：林岚纠正了模型，模型在这次对话的后半段改口，和模型真的学会了，是两回事。中间至少隔着五件常被混为一谈的事：暂时的语境变化、外部记忆、控制反馈、参数训练、持久学习。一次流畅、聪明的对话，最多说明前一两件发生了，不能证明能力已经稳定形成。第1章先用小舟学骑车把这五件事分开，第37章再回到人与模型的对话里逐件拆开。

个人—AI复合体还有一条通向物理世界的路。再看开头那位在手机上提问的人。假如他是个刚拿驾照的新手（这也是我构想的），读了模型那段「应立即刹车，注意雨天制动距离变长」，第二天雨夜开车时果然提早减了速。这时，集体语言世界里的一句话，经过一个具体的人、一双具体的手脚，进入了一次物理发生。这条通道是真实存在的，也是今天语言模型影响物理世界的一条常见路径：不是模型自己去踩刹车，而是它说出的话被一个人接住、带上路。可也正因为如此，这一脚踩得好不好，不仅取决于那句话说得对不对，还取决于这个新手有没有老彭那样的身体经验去落实它。同一句话，落到不同的个人—AI复合体里，发生的东西可能很不一样。

◆ 五、语言发生与物理发生：联系和差异

第四项论证回到开头的两个「刹车」。我们已经说，模型写出「刹车」，是集体人类—AI复合体语言世界里的一次显露；老彭停住车，是他和车、和雨夜路面一起完成的一次物理发生。两者有联系，也有差异，这一节要把联系和差异都说清，不偏向任何一头。

先说联系。语言里装着大量物理经验。「雨天路滑，提前减速」「下坡别空挡滑行」「刹车时别猛打方向」，这些话都是无数次现实中的停车与没停住，被人说出来、写下来、传下来的结果。一个语言模型读了足够多这样的话，就会在语言世界里形成关于制动、湿滑、距离的稳定牵连。它对新手是有帮助的，它甚至可以帮一个工程师整理刹车系统的测试方案。语言发生并没有和物理世界隔绝，它是物理经验经过人的身体、再经过人的表达，沉淀下来的一层。

再说差异。语言世界里站稳的东西，靠的是在人的理解和后续表达中「接得住」：读者点头，下一句话能顺着往下说，别人引用时不出岔子。物理世界里站稳的东西，靠的是在新的干预下「扛得住」：这一脚下去，车真的按预期减速了；换一条更滑的路、换一辆更重的车，还能不能停住。前一种检验发生在意义之间，后一种检验发生在身体、工具和物的阻力之间。

还拿刹车说。驾校教练常交代一类话：雨天刹车要提前，要平稳，别一脚踩死。这类话读懂不难，背下来也不难，可它没有告诉一个新手：「平稳」是踩到几分、脚掌这一刻感到的回弹算不算正常、听到轮胎声音变了该松还是该稳住。这些东西在这类话里不存在，或者只能用很粗的词来指一指，比如「平稳」「适当」；它们在一个人的脚和一块踏板、一条湿路之间一次次试出来。小舟拧一个太紧的瓶盖时，「逆时针旋转」几个字帮不上手心出汗的忙，也是这个道理，第26章会细讲。

所以，语言发生和物理发生的差异，不能简单地说成「一个是假的，一个是真的」，也不能说成「语言是物理的压缩版，解压就能还原」。更准确的说法是：它们是两种在不同纠缠里发生的世界，彼此有大片的重叠，又各有对方到不了的地

方。语言世界可以对物理发生给出提示、规则、事后的解释；物理发生可以回过头来纠正语言世界里不准确的说法。但在语言里显露得很准，不等于在物理里也能做到；在物理里做得很稳的人，也常常说不出自己是怎么做到的，老彭就说不清他那一脚为什么是「先重后稳」。

这里正好引出本书的贯穿问题。今天的世界模型研究，有很大一部分是让模型从视频里学：给它看海量的行车记录、操作录像、游戏画面，让它学会预测下一帧画面是什么样子。这种预测做得越来越好，生成的画面越来越逼真。可是，画面是物理发生「显露在外」的样子；预测得出画面，说明模型在显露这一层抓住了大量差异的规律性。问题是：物理显露的复现，如何进入物理发生条件的建立？换成老彭的话说：一个看了一万段雨夜行车录像、能把下一秒的画面画得分毫不差的系统，离「在一条它没见过的湿路上把车停住」，还差什么？差的东西能不能补上，用什么方式补上，补到什么程度算可靠?

把这个问题再拆细一点，至少能看到三道坎。一是看起来一样、底下不一样：画面上几乎相同的两片路面，轮胎压上去的后果可能截然不同。二是新的干预：录像里记下的是别人已经踩过的那几脚，路上要决定的却是「我这一脚该怎么踩」，它可能是录像里从没人踩过的一种。三是谁的身体：同一段录像，对一辆轻型轿车和一辆满载的货车，意味着完全不同的制动后果，而车有多重、刹车片磨到什么程度、轮胎是新是旧，画面里看不出来。这三道坎，第27章「为何显露预测不能保证行动可靠」会一道一道地细看。

三道坎也有一个共同的来历，值得在这里先点破：录像记下的，只是物理发生「露在外面」的那一层；而决定后果的，往往是没露出来的那些条件，以及正在发生的那一次互动本身。拿随身三问来说，录像能给出大量的显露，也能让模型学到显露与显露之间怎样接续；可是这里纠缠的是什么（轮胎、路面、车重、脚上的分寸），差异是由谁发起、怎样运行的（是别人的那一脚，还是我的这一脚），录像本身答不全。这不是哪一家技术做得好坏的问题，而是「从记录里学」这件事本身的形状。

这三道坎都不是说视频学习没用。看录像学开车、看录像学游泳，对人也有帮助，第17章会给这类经验一个公道的位置。三道坎只是提醒：显露层面的复现，要进入发生条件的建立，必须另有东西补进来，可能是接触、声音、身体状态的信号，可能是主动的试探，可能是失败后的重新训练。补什么、补多少，因任务而异。

这是全书反复回来追问的那一个问题。本书不预先断定答案。我们会在第四编具体检查从视频学习、从内部状态与行动后果学习、多模态与空间模型几类路线各自做了什么；在第六编讨论为什么显露预测不能自动保证行动可靠；在第七编提出可以失败的实验，看看主动互动是否真的比被动观看多建立了什么。

◆ 六、研究对象、概念边界与证据方法

第五项论证说的是本书自己：研究什么、概念用到哪里为止、凭什么证据说话。

研究对象。本书讨论的「世界模型」，不局限于某一类工程系统。今天叫作世界模型的东西至少有几种：生成视频的模型，预测内部状态的模型，重建三维空间的模型，以及把预测和行动连在一起的模型。它们各自以不同的东西为「世界」，成功标准也不同，第16章会逐一比较。本书关心的不是给哪一种冠名，而是一个更基本的问题：一个系统，人或者机器，怎样在互动中让对象站稳，让对象在某些任务里可以被依赖，在出错时能被发现，在失效后能被重建。本书主张，世界模型的开始应当包含对象同一性与任务能力的发生，而不是预设一个全能世界已被一次性封装。

这里的「封装」也要说一句。一个人学会骑车之后，不用每次都重新琢磨怎么平衡，这本事已经「打包」好了，可以随时调出来用，这就是封装，第5章会把它立成一个完整的概念。封装不是坏事，没有封装就没有能力。问题在于封装有边界：会骑自行车不等于会骑独轮车。一个可靠的封装，应该连同它的适用条件一起被知道。世界模型的出路，本书认为就在这里：围绕具体任务组织必要的纠缠，用

可检验的能力和清楚的适用边界形成可靠封装，而不是追求一个号称什么都能模拟、却说不清在哪里会失效的大全。

概念边界。有几处边界本书要事先划清，免得读者把书里的话读过了头。

其一，S、D、E 是三个维度，不是三个部件。显露、差异运行、特征纠缠彼此规定：显露靠差异运行和纠缠而来，差异运行又被显露和纠缠所塑造，纠缠也被前两者改写。本书写作 S=F(D,E)、D=G(S,E)、E=H(S,D)，三式缺一，就会在解释里留下一道缝隙。第3章会细讲。

其二，视觉、听觉、触觉不与 S、D、E 一一对应。不能说「看到的是显露，听到的是差异，摸到的是纠缠」。三种感官各自都是完整的一套 SDE：视觉里有它自己的纠缠、差异运行和显露，听觉、触觉也一样。本书提出的核心研究路线，是让视觉、听觉、触觉这三种 SDE 在互动中共同纠缠，形成对象同一性。但我们不预设三种感官在所有任务里都缺一不可，也不预设只有视觉就一定不够；哪些任务需要哪些感官参与、参与到什么程度，是要用实验来回答的问题。

其三，「物的意识」这个说法在第五编会出现，容易引起误会。本书先给它一个可以研究的功能性判据：一个系统能否在跨感官、跨时间的经验里，让某个物形成同一性，维持它，在它失效时察觉，并重新建立。至于系统有没有主观体验，「做那个系统是什么感觉」，这是另一个问题，本书单独讨论，不混入功能判据之中。

证据方法。本书把三种话分开说，也请读者分开听。

第一种是本体论立场：SDE 发生学对世界、学习和主体是什么的主张，比如「学习是特征纠缠的建立与训练」。这一类话不能被某次实验直接证明或推翻，它的价值在于能否更好地组织我们对大量现象的理解，能否提出更好的问题。

第二种是机制假说：可以被实验支持或推翻的说法，比如「在同等数据量和算力下，主动互动采集的训练经验比被动观看录像，更能提高对象在新干预下的预测稳定性」。这类说法要写清楚条件和指标，并且要准备好它被推翻时怎么办。第34章会专门讨论怎样建立可以失败的实验。

第三种是经验事实：公开论文和技术报告里能查到的东西。本书讨论的技术路线，包括从大量无标注视频中学出「隐动作」、训练可交互视频生成的 Genie，以特征预测为核心、并做了机器人动作条件后训练的 V-JEPA 2 及其后续版本，在同一个架构里联合理解与生成语言、图像、视频、声音和动作序列的 Cosmos 3，按官方博文把文字、图像、相机位姿和深度图等输入放进同一个「空间上下文」、用于视频生成与三维重建（目前只向部分合作方开放试用）的 World Labs Atlas，把经验回放、内部想象与策略学习串成循环的 DreamerV3，以及把视觉与触觉放在一起、预测接触怎样演化并据此做灵巧操作的 DexTacWAM、TouchWorld 等研究，还有 Waymo 在 2025 年 12 月公开的说明：一个基础模型同时支撑驾驶、仿真与评判三部分，驾驶一侧把摄像头、激光雷达和毫米波雷达的输入随时间融合，评判一侧从实际行驶中找出欠佳的行为，改进后先在仿真中检验再上路。凡本书引用这类资料，都只说公开机制，落实到具体版本；拿不准的细节标出〔待核〕，不替任何人编数据、编结论、编原话。

三种话之间有一条硬规矩：模型的局部成功或失败，不能直接当作本体论的证明。某个视频模型生成了极逼真的雨夜画面，不证明「世界是现成的、可以一次性封装」；某个机器人在陌生物体上抓取失败，也不证明「发生学是对的」。技术事实可以和本体论立场对话，可以让一种立场显得更有解释力或更别扭，但要从技术事实走到本体论结论，中间必须经过机制假说这一层，并且把推论的每一步摆出来。同样，本书也不预写「驾驶 AI 全部是纯视觉」「现有世界模型都没有互动」「统一架构必然失败」这类结论。事实上，许多现有路线已经在做动作条件、在用多传感器、在把现实反馈送回训练；本书要做的，是一条一条看它们具体做了什么、在发生学的眼光下还缺什么。

◆ 七、随身三问与全书路线图

在动身之前，先把全书随身带着的三个问题交给读者，并拿开头那一脚刹车试用一次。

这里纠缠的是什么?老彭那一脚里,缠在一起的有:眼睛里一团晃动的深色和「人」的辨认;脚底踏板的阻力与车速的感觉;轮胎碾水的声音与「抓地还够不够」的判断;还有三十年来许多个雨夜留下的分寸。模型写出的那句「应立即刹车」里,缠在一起的是:「雨夜」「横穿」「制动距离」这些词在千万段文字里的相互牵连,以及人们对「得体的建议」应有什么样子的反复评价。两团纠缠都是真的,但缠的东西不一样。

差异是怎样运行的?老彭这边,黑影从电动车缝里出来的那一刻是第一个差异,接着是车速、距离、踏板反馈、声音变调,一串差异在不到两秒内传递、互相校正。模型这边,差异在一个个词之间运行:上一句说了「湿滑」,下一句就倾向于说「距离变长」。前一种差异运行一旦出错,后果落在路面上;后一种出错,后果落在读者的理解里,或许还要再经过读者的行动,才落到路面上。

显露出来的东西靠什么站稳,又在哪里失效?老彭停住车,靠的是这一次的路况恰在他的经验覆盖之内;换成结了薄冰的桥面、换一辆陌生的重型货车,他的分寸也可能失灵,他自己知道这一点,所以遇上陌生情形会提前放慢。模型那段话站得稳,靠的是它与大量人类经验的吻合;它在哪里失效,比如遇到一种罕见路面、一种它没读过的车辆特性,模型自己未必知道。能不能知道自己在哪里会失效,正是可靠封装和不可靠封装的一条分界。

三个问题问完,读者大概可以看出,本书并不想在「模型懂」与「模型不懂」之间站队,而是想把「懂」拆开,看它在哪一团纠缠里、经过哪些差异、显露为什么、靠什么站稳。下面是全书的路线。

第一编　学习的本体论重建,从「知识装进脑子」的传统图景出发,提出学习是特征纠缠的建立与训练,讲清训练怎样让纠缠发生、三个方程怎样共同解释学习、复杂性怎样进入学习,最后落到能力如何被封装、如何失效与解封。这一编给全书打地基。

第二编　理念、现实与自我如何发生，回到人：每个人的世界怎样在互动史里建立，理念世界、现实世界、自我世界各自怎样发生，又怎样彼此改写。只有先看清人的世界从哪里来，才有资格谈机器的世界从哪里来。

第三编 人类与AI复合体的语言世界发生，讨论人怎样进入大模型训练，token怎样作为工程载体参与特征纠缠的建立，一个具有相对同一性与集体性的语言世界怎样发生、能做什么、边界在哪里，并提醒不要把语言模型的成功当作人脑机制已被复制的证明。

第四编　世界模型的技术路线与哲学前提，逐一检查以视频、内部状态与行动后果、多模态与空间为中心的几类技术路线，各自以什么为「世界」，公开机制是什么，证据边界在哪里；再指出发现学前提在何处限制了它们，同时看到其中已有的发生取向。

第五编　三种SDE的纠缠与物的意识，分别讨论视觉、触觉、听觉中物怎样显露，再用一只陌生弹性球的握、敲、放，讨论三种 SDE 怎样形成纠缠聚合体，最后给「物的意识」提出对象同一性的功能性判据。

第六编　统一物理模型的困境与任务边界，正面处理「一个模型穷尽全部物理发生」的设想：物理发生的信息负担有多重，为何显露预测不能保证行动可靠，任务怎样组织世界的发生范围；以驾驶为专门例子，讨论人—车辆—道路复合体，再谈专用模型怎样协作与共享。

第七编　发生式世界模型的训练与验证，把前面的主张变成可做的事：围绕共同事件组织训练经验，从被动记录走向主动试探，做跨模态预测与纠缠重组，建立可以失败的对照实验与评价指标，并讨论可靠性怎样形成、封装怎样持续评估与回写。

第八编　从集体世界到具体复合体的学习，回到每一个使用者：集体同一性怎样成为个人发生的条件，个人与 AI 怎样继续发生语言世界，教育和科研怎样转向特征纠缠的训练，最后提出 AI 训练与学习的新研究纲领与开放问题。

结语把三条主线收拢：人的三界共同发生；集体人类—AI语言世界与具体个人—AI世界的连续；物理世界模型从任务互动中形成物的同一性与可靠行动。并分清哪些已经论证，哪些仍待验证，哪些尚未解决。

读这本书，不必按顺序。关心大模型的读者可以从第三编读起，关心机器人与自动驾驶的读者可以从第四编、第六编读起，关心教育的读者可以先翻第八编。但无论从哪里进，都请随身带着那三个问题，并留意每一处说的是本体论立场、机制假说，还是经验事实。

全书要论证的，归结起来是一句话，先写在这里，到结语再回头验收：

> **训练不是把一个现成的世界装进机器，而是组织发生的条件；学习，是特征纠缠的建立与训练——世界模型的出路，不在一次封装一个全能的世界，而在具体任务的互动里，先让一个物站稳。**

> **带走的话**
>
> 模型写出「刹车」与老彭在雨夜把车停住，是两种都真实、却在不同纠缠里发生的世界；训练不是把一个既成世界装进模型，而是组织发生条件。世界模型的真正起点，是对象同一性与任务能力在互动中的发生，并且清楚知道它在哪里会失效。

第 一 编

学习的本体论重建

编序　全书的母题只有一句：学习是特征纠缠的建立与训练。这一编的任务，是把这句话从口号变成可以使用、可以检验、也可以被反驳的理论地基。后面几编要讨论人的三个世界、人类与AI复合体的语言世界、物理世界模型的困境与出路，它们用到的「学习」「训练」「纠缠」「封装」等词，都要在这里先站稳。

五章是接力的。第1章先拆掉一个最常见的图景：知识像货物一样被搬进脑子或参数里；然后从小舟学骑车出发，说明学习前后真正改变的是什么，并把错误学习、遗忘、重新学习放进同一个解释里。第2章追问「训练」到底在做什么，用和面揉出筋来说明经验怎样相遇、强化、松动、分化与重组，并把训练量、训练结构和训练后果分开。第3章是全编的核心：把显露（S）、差异运行（D）、特征纠缠（E）三个维度和三条方程讲透，说明学习同时改写三维，并专门防止把E降格为背景或数据仓库。第4章处理复杂性：混沌态、介生态、秩序态怎样在学习中出现，哪些是概念解释，哪些可以落成可测的动力学指标。第5章立起「封装」这个概念：一种能力怎样发生、稳定、可以被调用，又怎样失效、解封、重新训练。后面讨论世界模型可靠性的各编，都会回到这个概念。

读这一编时，可以随身带三个问题：这里纠缠的是什么？差异是怎样运行的？显露出来的东西靠什么站稳，又在哪里失效？

01

第 1 章

学习究竟发生了什么

> **这一章的问题**：一个人或一个模型「学会了」，到底是什么东西发生了变化?

◆ 一、知识进入大脑的传统图景

我们平常说「学到了知识」，脑子里常常浮现一个画面：世界上本来就有一堆现成的东西，比如骑车的要领、乘法口诀、牛顿定律，老师或书本把它们打包好，学生把它们装进脑子；装进去了就算学会，装得多就算学得多，忘了就是货物丢了。这个画面可以叫作「搬运图景」。它有一个更深的前提：世界和规律是先在的、已经完成的，学习只是把这个既成世界的副本搬到个体内部。本书把这种前提称为「发现学」，意思是以为世界与规律都在那里等着被找到，学习者的工作只是找到、记住、调用。

搬运图景并非一无是处。它解释了为什么书本有用，为什么可以考试，为什么同一份教材能教很多人。它的毛病在于把最关键的一步藏起来了：货物怎样从外面变成里面的?学骑车的孩子听了一百遍「眼睛看前面，别看脚下」，照样摔倒。话是搬进去了，车还是不会骑。可见被「搬进去」的语句与真正形成的能力之间，隔着一段搬运图景说不清的过程。

同样的图景也出现在关于AI的议论里。有人说大模型「把互联网装进了参数」，有人说世界模型要「把物理世界压缩进网络」。这些说法都暗含着：世界先在那里，训练只是把它装进去。本书要论证的是另一种看法：训练不是把一个既成

世界装进模型，而是组织发生条件；学到的东西不是被搬进来的副本，而是在一次次相遇中新长出来的组织。

◆ 二、学习作为特征纠缠的建立与训练

先给出本章的本体论立场，读者可以把它当作本书的基本主张，而不是已经被实验证明的结论：学习是特征纠缠的建立与训练。

导论已经拿骑车把这三个词点过一下，这里说得细一些。用大白话说，特征纠缠（E）就是一块块经验、一个个可以被区分的方面缠到了一起，彼此牵动，不再各管各的。骑车时，车把向左偏一点、身体往右歪一点、脚下蹬的力气、眼睛看到的路面、耳边的风声，原本是互不相干的感受；学会之后，它们缠成了一个整体，车把一偏，身体几乎同时就做出了补偿，用不着先想一想。这种「缠在一起、互相牵动」的组织，就是本书所说的特征纠缠。

但纠缠不会凭空出现，它需要差异不断运行起来（差异运行，D）：这一次歪多了，下一次歪少了；这一下蹬得猛，那一下蹬得缓；摔了，没摔。没有这些前后不同、左右不同的差异，就没有东西可以缠到一起。纠缠和差异运行合起来，又会让某种东西显露出来（显露，S）：一个「会骑车的小舟」，一条在他眼里忽然变得「可以骑过去」的小路，一种稳稳当当的姿态。显露的东西一旦反复出现、站得住，我们就叫它结构；结构是显露的稳定态，而不是显露的全部。

所以，「学习是特征纠缠的建立与训练」这句话，包含两层意思。一层是「建立」：原本不相干的经验，开始缠到一起；另一层是「训练」：缠到一起之后，还要被反复使用、调整、加固、拆分、重组。只有建立没有训练，纠缠很脆，一碰就散；只有训练没有建立，就是在空转。第2章会专门讲训练怎样让纠缠发生，第3章再用三条方程说明S、D、E三者怎样相互决定。

这里先说清这一立场与神经科学的关系。神经科学的主流看法认为，学习与神经元之间连接强度的变化有关，这是经验事实层面的研究传统，本书尊重并会引用它。但「连接强度变了」与「特征纠缠建立了」不是一回事：前者是对某一种物质载体的描述，后者是对发生过程的本体论刻画。同理，在人工神经网络中，训练确

实改变了大量参数的数值，这是公开的工程事实；至于参数变化在什么意义上可以被解释为特征纠缠，是机制假说层面的问题，第12章还要细讲它的解释边界。

◆ 三、小舟学骑车：学会前后什么变了

周小舟九岁，读小学四年级，这个春天开始学骑车。我们沿着他的经历，看看「学会」前后到底变了什么。

最开始，周敏在后面扶着车座。小舟的手攥得很紧，眼睛盯着前轮，车把一抖，他就整个人僵住，往一边倒。这时他身上有很多感受，但它们是散的：手上的力和身体的歪斜对不上，眼睛看到的和脚下做的对不上。他知道「要往倒的那边拧车把」，这句话周敏说过很多遍，可是话和动作之间没有缠到一起。

扶了大约一个星期，有一天周敏悄悄松开了手，小舟骑出去七八米才发现妈妈没在后面，一慌，摔在草地上。这一摔很重要：在摔之前的那七八米里，车把、重心、蹬踏和视线第一次在没有外力托底的情况下自己维持了一小段平衡，它们缠在一起的那一小段组织第一次被显露出来；而「发现没人扶」这个新差异一进来，组织立刻散了。

又过了几天，小舟可以自己起步、骑一圈、刹车停下。再后来，他可以一边骑一边扭头跟周敏说话。到了这一步，我们看到的变化至少有三层。

第一层是能力的变化。原来做不到的事，现在能做到；而且不是偶尔做到，是在一定条件下可以反复做到。这一层最容易被看到，也最容易被误认为就是学习的全部。

第二层是世界组织的变化。会骑车之后，小舟眼里的小区变了：门口那段缓坡原来只是「一段路」，现在成了「可以冲下去」的地方；路边的减速带原来不值一提，现在成了「要抬一下屁股」的东西；一辆停在路中间的电动车，原来只是一个物件，现在成了「要提前绕开」的障碍。同一个小区，对会骑车和不会骑车的孩子，显露为不同的世界。学习改变的不只是孩子，还改变了孩子与之打交道的那个世界的组织方式。

第三层是自我的变化。小舟开始说「我会骑车了」，开始想骑去同学家。这句「我会」不是一句客观报告，它把一段行动历史、一种被认可的能力和一种对自己的新看法缠在了一起。第二编会专门讨论自我世界怎样发生，这里只指出：一次具体的学习，往往同时改写能力、世界和自我，它们不是三件依次发生的事。

还要回头看一眼周敏的「扶」和「放手」。扶，不是把骑车的本事递给小舟，而是替他兜住最危险的那部分差异，让倒地的后果小一点，好让车把、重心和蹬踏有机会一次次碰在一起；放手，则是把兜住的差异还给他，让那些刚缠上的组织第一次独自承受真实的歪斜。扶得太久，纠缠一直靠外力托着，形不成自己的稳态；放得太早，差异太猛，刚缠上的组织还来不及加固就被冲散。周敏做的事情，正是「组织发生条件」：她没有把任何东西装进儿子身上，却决定了哪些差异在什么时候、以多大的强度进入。这一点后面讲AI训练时会反复用到。

搬运图景解释不了第二层和第三层。它只能说「骑车知识被装进了小舟的脑子」，却说不清为什么小区变了样，为什么小舟变了样。而用特征纠缠来看，这三层变化是同一个发生过程的三个显露：经验缠到了一起，差异运行有了新的走法，显露出来的既有能力，也有一个新的世界，也有一个新的「我」。

◆ 四、错误学习、遗忘与重新学习

一种学习理论要是只解释成功，就不算完整。错误学习、遗忘和重新学习，恰恰最能检验它。

先说错误学习。小舟有一阵子养成了一个坏习惯：一紧张就低头看前轮。这个习惯也是学出来的，它同样是一种特征纠缠：紧张、低头、车把发僵缠在了一起，而且缠得很牢。从搬运图景看，这是「装进了错误的知识」；从本书的角度看，错误学习并不是装错了货，而是纠缠在某种条件下建立得太稳、太早，挡住了更好的组织。纠正它，不能只是「删除错误知识、换上正确知识」，而是要让新的差异进入，把原来的纠缠松动、拆开、重新缠。周敏后来的办法是让小舟在前方十米处找一个目标喊出来，「看那棵树」「看那个红色垃圾桶」。视线一被拉出去，低头的纠缠就有了松动的机会。

再说遗忘。这一编还要反复用到一件事：多年不骑，重新上车。周敏自己上大学时天天骑车，工作后十几年几乎没再碰过自行车。今年陪小舟学车，她借了邻居的车想示范一下，刚上去的那几秒，车头晃得厉害，她自己都笑了。但大约骑出二三十米，那种感觉就回来了，很快就能单手扶把指挥小舟。

这个经验人人都有：骑车「一辈子忘不了」，可是刚上去确实生疏。搬运图景在这里很尴尬：如果知识丢了，为什么回来得这么快？如果没丢，为什么一开始那么生？用特征纠缠来理解，遗忘不必是货物丢失，而可以是纠缠被长期闲置后变得松散、难以被当下的情境唤起；原来的组织还在，但它与眼前这辆车的车把高度、座位软硬、长沙街道的坡度之间，还没有重新缠上。重新上车的那二三十米，是新的差异进入、旧的纠缠被重新激活并与新条件缠在一起的过程。所以重新学习往往比第一次快。这一点心理学早有观察：十九世纪八十年代，德国心理学家艾宾浩斯拿自己做实验，背一串串无意义的音节，过一段时间再重背，用重背比初背省下的遍数或时间来衡量还留着多少记忆，这种量法后来被称为「节省法」。即使已经一个也想不起来，重学仍然比初学省力。

需要分清的是：上面这段对遗忘的解释是机制假说，它可以被检验，例如可以设计实验比较「完全没学过」与「学过但多年未练」的人在重新学习时的进步曲线，看后者是否在特定方面明显更快、在另一些方面并不更快。它不是从本体论立场直接推出来的必然结论。

AI中也有一个与遗忘相关的公开现象：神经网络在学新任务时，旧任务的表现可能明显下降，工程界称之为「灾难性遗忘」。这是经验事实，有大量公开研究；但它与人的遗忘是否同一种机制，是另一个问题，不能因为都叫「遗忘」就当作同一回事。第5章讲封装的失效与重训时，还会回到这一点。

◆ 五、怎样判断「真的学会了」

既然学习不是货物入库，那么「学会了」就不能只看一次表现。小舟第一次骑出七八米，不能说他已经学会；周敏在后面扶着、他骑了一整圈，也不能说他已经学会。为了后面各编讨论AI时不混淆，这里先把五种常被混为一谈的变化分开。

第一种是暂时的情境变化。周敏扶着车座，小舟就能骑；松手，他就倒。他的表现变了，是因为外部条件暂时变了，而不是他身上建立了稳定的纠缠。在大模型那里，与之相近的是把提示语写得更周全，回答就更好；换一个说法，表现又回去了。

第二种是外部记忆。小舟把「看前面、往倒的那边拧」写在小本子上，骑之前念一遍。本子上的话可以帮助他，但它在本子上，不在他的身上。AI系统接入检索库、笔记、文档，也属于这一类：信息可以被调用，但调用它的那个系统本身未必因此改变。

第三种是控制反馈。骑车时车把一偏，小舟立刻纠正，这种当下的纠错是行动中的闭环，它让这一次骑行不倒，却不必然留下什么。机器人在执行中根据传感器读数实时调整，也是控制反馈；它和学习有关，但不等于学习。

第四种是参数训练。对人来说，没有一个现成的「参数」可以指着看，但反复练习确实会改变身体与神经的某些持久状态；对人工神经网络来说，训练改变的是可以读出的参数数值。参数变了，说明有东西被留下了，但留下的是否是我们想要的能力，还需要检验。

第五种是持久学习。只有当一种能力在撤掉扶持、撤掉提示、换一辆车、隔一段时间之后，仍然能在一定条件下被唤起、并且能迁移到相近的新情境时，我们才比较有把握地说：纠缠建立了，而且站稳了。

这五种变化可以同时出现，也可以互相转化：控制反馈反复发生，可能沉淀为持久学习；外部记忆用久了，也可能被吸收为习惯。但它们不能互相代替。一次流畅的对话、一次漂亮的演示，最多说明前几种变化发生了，不足以证明第五种已经形成。第5章讨论能力封装时，会把「持久学习」进一步拆成发生、稳定、可调用三个环节；第37章再把这五件事放回人与模型的日常对话里逐件检验。

◆ 六、边界：这一章没有证明什么

最后要说清本章的边界。本章提出的是一个本体论立场：学习是特征纠缠的建立与训练，学习同时改写能力、世界与自我。小舟骑车、周敏重新上车，是用来让这一立场变得可理解的日常案例，不是证明。一个人可以接受这些描写，却仍然坚持搬运图景，只要他能给出更好的解释。

本章也没有说AI的训练与人的学习是同一回事。在人那里，学习的发生单位是一个有身体、有历史、有他人陪伴的人；在大模型那里，训练的发生单位是集体的人类—AI复合体，大量人类留下的文本、数据选择、目标设置与评价反馈都参与其中。把AI写成脱离人类历史的独立学习者，会犯另一种搬运图景的错误：仿佛世界已经在数据里完成了，模型只是把它装进去。

还要防止一种误读：说「学习改变世界组织」，不等于说「学习者想怎样，世界就怎样」。小舟的小区确实对他显露为新的样子，但缓坡的坡度、减速带的高度并不因为他会骑车就改变。世界的阻力恰恰是纠缠能否站稳的检验者，第8章会专门讨论现实阻力的作用。

那么，经验到底是怎样缠到一起的？反复练习为什么有时有效、有时无效？一千次练习和一百次练习的差别，真的只在数量吗？这些问题，交给第2章。

> **带走的话**
>
> 学会，不是把一份现成的世界副本装进来，而是让原本分散的经验缠成新的组织；能力、世界和自我在同一过程中一起改变，遗忘与重学也要在这个过程里解释。

> **小账** 论证落在哪一层？——本章以本体论立场为主，遗忘与重学部分给出机制假说，神经连接与灾难性遗忘作为经验事实背景出现。 世界模型补了哪一条？——世界模型的「世界」不能预设为已完成的副本；一个模型学会了什么，要同时看它的能力、它所组织的世界以及它在何种条件下失效。

02

第 2 章

训练如何使特征纠缠发生

> **这一章的问题**：反复练习为什么有时能让经验缠在一起，有时却白练？

◆ 一、命题：训练是组织相遇，不是灌注内容

第1章说，学习是特征纠缠的建立与训练。这一章要追问「训练」二字。日常语言里，训练常被理解为「多练」：练得越多，学得越好。这个说法有一半是对的，另一半会把人引向歧途。本章的命题是：训练不是把内容灌注进学习者，而是组织经验的相遇，让差异在相遇中形成、运行，并让特征纠缠在反复互动中强化、松动、稳定、分化与重组。训练者能控制的，从来不是「纠缠本身」，而是相遇的条件：什么和什么碰到一起，以什么顺序、什么强度、什么频率碰到一起，碰完之后留下什么后果。

这一命题有一个直接推论：同样的训练量，可以因为训练结构不同而产生完全不同的训练后果。这是本章最后一节要分开说的三件事。

◆ 二、和面揉出筋：一次日常的训练

再看一件厨房里的事：和面。周敏每到周末会包饺子，小舟喜欢在旁边捣乱，后来干脆让他揉面。

一碗面粉加水，刚开始是一盆湿乎乎、疙疙瘩瘩的东西，粘手，一扯就断。这时面粉和水只是「混在一起」，还谈不上缠在一起。揉的过程，是不断把面团压扁、折起、再压扁。每一次压和折，都让面团里原本分散的部分彼此相遇：干的碰

上湿的，这一块碰上那一块。揉到一定程度，面团忽然「光」了，表面变得细腻，拉开来有弹性，扯出薄膜也不容易破。厨房里管这叫「上了筋」。

食品科学告诉我们，「筋」来自面粉里的两类蛋白质（麦谷蛋白与麦醇溶蛋白），它们吸水之后，在揉压的拉伸与折叠中彼此连接，形成一张既有弹性又能延展的网，通常称为面筋网络。这是经验事实层面的东西，本书不去改写它。我们借和面要说明的是另一件事：筋不是从外面加进去的。没有人往面团里放一根「筋」；筋是面粉与水在反复的揉压中，一次次相遇、一次次被拉伸、折叠，组织逐渐缠在一起而显露出来的。揉面的人只做了一件事：反复组织相遇。

和面还提供了几个对理解训练极有用的细节。

第一，揉的方式比揉的次数更要紧。周敏教小舟用掌根往前推、再折回来，而不是用手指乱抓。乱抓一千下，面还是一盘散沙；用掌根推折几百下，筋就出来了。这就是训练结构的差别。

第二，揉过头也不行。家里用手揉，一般不容易真把筋揉断；可用和面机搅得太久、太猛，筋会被打散，面团反而发粘、塌软。手揉常见的「过头」是另一种：一直揉、不让面歇，筋绷得太紧，擀开就往回缩、边缘开裂。强化到一定程度，再加力就变成破坏。

第三，「醒面」很重要。揉到一半盖上湿布放二十分钟，再揉几下，面团会比一直不停地揉更光滑。不揉的时候，变化还在继续。这一点很像人学技能时常说的「睡一觉就会了」。心理学里有研究报告：人练习一串手指按键动作之后，睡一夜再测，比同样间隔却没睡觉的人进步更多；不过这类效应有多大、在什么条件下出现，后来的研究仍有争论。

第四，水和面的比例、面粉的种类、手的温度、天气的干湿，都会改变揉的结果。同样的手法换一袋面粉，可能要调整。纠缠的建立，总是在具体条件下的建立。

◆ 三、机制解释：从相遇到纠缠的五个环节

把和面和小舟学骑车放在一起看，可以提出一个关于训练的机制假说。它不是唯一可能的描述，而是一个可以拿来检查具体训练过程的框架。

相遇与差异形成。训练的起点是不同的经验、不同的方面被放到一起。面粉碰上水，小舟的手碰上歪斜的车把。只有放到一起，才可能出现「这样和那样不一样」的差异。没有相遇，就没有差异；没有差异，就没有东西可以运行。很多无效训练的问题，不在练得少，而在该相遇的东西始终没有碰上：小舟一直被扶着，摔倒的后果和车把的动作就碰不上。

反复互动中的强化。差异一旦形成，反复出现就会使某些关联越来越牢。车一歪就往那边拧，拧了就不倒；同样的差异—后果关联反复出现，纠缠被加固。

反复互动中的松动。训练不只加固，也松动。小舟一开始「一歪就整个人僵住」，这是一个很牢的旧纠缠。只有在大量新的、略有不同的差异进入后，这个旧纠缠才开始松开。松动常常被看作退步：这几天怎么反而骑得不稳了？但没有松动，就没有重组的空间。

稳定化。强化与松动来回几轮之后，一部分组织在多种情境下都站得住，就进入稳定状态。面团「光了」，小舟可以自己起步、刹车。稳定不是不变，而是在一定范围的扰动下能维持住。

分化与重组。稳定之后，同一种纠缠会在不同情境中分出不同的变体：骑上坡与下坡、骑空地与骑人多的路，手上的力、身体的倾斜各有讲究。分化到一定程度，不同的变体又可能被重新组织成更高一层的整体，例如「骑车时怎样跟着路况变」。面也一样：会和饺子面的人去学和馒头面、拉面面团，原有的手感会被分化，再重组为对「面」的一般性把握。

这五个环节不必依次发生，常常交织在一起；但把它们分开说，可以帮助我们在一个具体训练过程中问：现在卡在哪一环？

不妨用小舟第一次揉面的失败来试一试这个框架。那天他兴致很高，水一下子倒多了，面团稀得粘满两只手。他越揉越急，手指抓起一团甩下去，再抓起一团。十几分钟后，盆里还是一摊粘糊糊的东西。按五个环节来看，问题出在第一环：水太多，面粉颗粒之间没有足够的「抓手」，揉压带来的拉伸与折叠根本形成不了有方向的差异，后面的强化也就无从谈起。周敏的处理很简单：撒一把干面粉，让小舟停手，自己示范掌根推出去、折回来、转九十度、再推。几分钟后面团开始成形。她没有替小舟揉出筋，她改变的是相遇的条件：先调比例，再定手法。

另一次是反过来的。小舟揉得很认真，面团也开始光滑了，他却非要再揉，「揉得越久越筋道」。周敏让他停下来醒面。他不服，偷偷又揉了一阵，结果面团表面起了一层干皮，擀的时候边缘裂口。这一次卡在第三、第四环之间：强化已经够了，需要的是「不动」给内部留出调整的时间，而他把继续施力当作继续进步。

这两次失败放在一起，正好说明训练的两种常见错位：一种是在该组织相遇的时候拼命加量，一种是在该等待稳定的时候拼命加力。二者都把「多练」误当成「练对」。

还有一个特别容易被忽视的训练结构问题：反馈放在哪里、以什么形式出现、隔多久出现。

骑车的反馈是即时的、身体的：歪了就倒，倒了就疼。这种反馈直接作用在身体上，几乎不需要翻译，所以骑车虽然开头吓人，学得却不算慢。和面的反馈就慢一些：揉的时候手上能感到面团从粘变滑、从软塌变得有回弹，但「筋到底够不够」，往往要等到擀皮、包馅、下锅之后才知道，饺子煮破了，才回过头怪面没揉好。反馈来得越晚，越难判断是哪一步出了问题，相遇和后果之间的关联也越难缠紧。

周敏后来教了小舟一个办法：揉一阵就揪一小块面，用两手慢慢往外撑，看能不能撑出一层透光的薄膜。做面包的人管这叫「出膜」，撑得又薄又韧、不容易破的叫「手套膜」；饺子面不必揉到那个程度，但撑一撑看筋够不够，道理相同。这一招把滞后的反馈提前了，把「饺子会不会破」这种远处的后果，换成手上马上能

看见的显露。从训练结构上说，这是在训练过程中插入一个中间检验点，让差异与后果的距离变短。

AI训练里也有大量关于反馈位置的工程设计。公开文献中常见的做法包括：在预训练之后用人工标注的示范做微调，让人对模型的几个回答排出好坏，用这些偏好比较训练一个评价模型，再用评价模型的打分引导模型更新，这就是常说的「基于人类反馈的强化学习」；在数学推理一类任务上，只对最终答案给评价（结果监督），还是对每一个中间步骤也给评价（过程监督），公开研究做过比较，不同研究的结论并不完全一致，但都表明反馈放在哪里会影响训练出来的东西。这些设计都可以看作在决定「反馈放在哪里」。它们共同说明了一件事：在训练中，反馈不是附加在学习之外的评语，而是训练结构本身的一部分，它决定哪些差异能与哪些后果缠到一起。

需要提醒的是，反馈的设置者是人。哪些回答算「好」，哪些推理步骤算「对」，由谁来判断，用什么标准判断，都是人在训练过程中留下的印记。这又一次说明，大模型训练的发生单位是集体人类—AI复合体，而不是一个关起门来自己学习的模型。第15章会讨论评价标准留下的偏向。

◆ 四、AI训练里的对应与不对应

这五个环节在AI训练中有没有对应物？这里只做谨慎的对照，不把对照当证明。

在大模型的预训练中，公开的基本做法是让模型对海量文本做预测，用预测与实际之间的误差去调整参数。具体说，是把文本切成一个个小单位（词元），让模型根据前文预测下一个词元，这常被称为「下一个词元预测」。从训练组织的角度看，这里确实有「相遇」：不同来源、不同时代、不同人的表达被放进同一个训练过程；也确实有「差异形成」：预测与实际的差距就是一种被量化的差异。工程上的数据配比、课程式安排（先易后难）、多轮训练与微调，都可以看作在调整相遇的结构。

但对应也有限度。和面时，面团每一次被揉都是同一团面在变化，揉的人能摸到它的变化；小舟摔一跤，摔的是他自己的身体。大模型的训练里，是谁在「相遇」？不是一个孤立的模型，而是大量人类留下的文本、做数据筛选的工程师、设定目标的团队、给出评价反馈的人和工具共同构成的集体人类—AI复合体。第11章会专门讨论人怎样进入大模型训练；这里只强调：把训练描述成「模型自己在学」，会遗漏组织相遇的那一大群人。

另一个不对应是「醒面」。人在不练的时候，变化还在继续；常见的人工神经网络在不训练时参数不变。工程上有不少方法试图让旧经验在新训练中被重新使用，例如经验回放，第18章讲Dreamer时会涉及它的公开机制。它们是否与人在休息中发生的巩固属于同一种机制，现在不能下结论。

◆ 五、训练量、训练结构与训练后果

现在可以把日常说的「多练」拆成三件事。

训练量，是相遇发生了多少次：揉了多少下，骑了多少圈，模型看了多少数据、更新了多少步。

训练结构，是相遇以什么方式组织：用掌根推折还是用手指乱抓，扶着骑还是放手骑，先在空地上骑还是直接到车来人往的地方骑；对模型来说，是数据怎样配比、按什么顺序、用什么目标、给什么反馈。

训练后果，是相遇之后实际留下了什么：面团上没上筋，小舟能不能在撤掉扶持后自己骑，模型能不能在新问题上给出可靠回答。

这三者常被当成一件事，结果出现两种误判。一种是用训练量代替训练后果：练了这么多，应该会了吧？关于「规模规律」的公开研究确实观察到：模型的预测误差随参数量、数据量和算力的增加，大致按可预期的平滑曲线下降，但这条曲线本身不告诉我们具体哪一种能力在什么条件下已经稳定形成。另一种是用训练结构代替训练后果：方法这么先进，结果一定好吧？手法再对，面粉不对，筋也出不来。

训练后果本身也要分层看。小舟揉了几个周末之后，有一天周敏改做包子，用的是发面。他照着揉饺子面的习惯使劲推折，面团却软乎乎地往回缩，跟他熟悉的手感不一样。他愣了一下，问：「这面是不是坏了？」这个小插曲说明，训练后果至少有三层：在原来条件下能不能重复（饺子面能不能每次都揉好）；在相近条件下能不能迁移（换一种面粉、换一个季节还行不行）；在条件明显变化时能不能发现「不一样了」，并且知道该停下来重新试。第三层最容易被忽略，却最要紧：一个只会在熟悉条件下顺利运行、遇到变化却浑然不觉的能力，比一个明知自己不会的状态更危险。后面讨论世界模型时，「能不能识别失配」会成为可靠性的核心指标之一。

本书的立场是：训练后果必须单独检验。检验的方式，第1章已经给出一个初步标准，即撤掉扶持、换条件、隔时间之后仍能在一定范围内运行；第5章会把它发展为「封装」的判断。

◆ 六、边界：哪些地方本章说多了也说不了

最后交代三条边界。

第一，本章的五环节框架是机制假说，它可以被反驳。例如，若有人能证明某类能力的形成完全不需要「松动」环节，旧组织不必松开就能直接被新组织覆盖，那就需要修改这个框架。

第二，和面只是类比。面团没有自我，没有世界，揉面的人也不是面团学习的「复合体伙伴」。用和面讲训练，是因为它把「筋不是加进去的、是揉出来的」这一点展示得最清楚，不是说人或模型的学习就是揉面。

第三，本章没有回答「训练者自己是否也在学习」。其实周敏教小舟揉面的这几个周末，她自己也在变：她原来只会凭手感揉，现在要把手感说成话、拆成步骤、想办法让儿子看得见，这迫使她对「上筋」有了比以前更清楚的把握。教与学是相互的，训练者和学习者共同构成一个复合体。这一点在本编只点到为止；第38章讨论教育时，周敏和AI一起试探、一起反思的那个学期，会把教的一方也卷进学习里来。

训练让纠缠发生了，那么纠缠与显露、与差异运行之间到底是什么关系？为什么同一次训练会同时改变能力、世界与自我？第3章要用三条方程把这个关系讲透。

> **带走的话**
>
> 训练者能组织的是经验的相遇，而不是纠缠本身；训练量、训练结构和训练后果是三件事，后果必须单独检验。

小账　论证落在哪一层？——本章以机制假说为主（相遇—强化—松动—稳定—分化重组），和面的食品科学与大模型预训练的公开做法作为经验事实背景。　世界模型补了哪一条？——世界模型的训练要被问「组织了哪些经验的相遇」，而不只问「用了多少数据」；训练后果须在撤掉训练条件后单独检验。

03

第 3 章

三方程如何解释学习

> **这一章的问题**：显露、差异运行、特征纠缠三者之间到底是什么关系，学习又怎样同时改写它们？

◆ 一、三个维度：先用白话说清楚

前两章反复用到三个词，这一章要把它们正式立起来。

显露（S），是某种东西在一个过程中露出来、被认出来、可以被打交道。小舟骑车时露出来的一个稳当姿态，他眼里忽然「可以冲下去」的缓坡，揉面时面团表面那层光，都是显露。显露不等于「结构」：结构是显露的稳定态，是那些反复露出来、站得住的东西。刚学车那几天，小舟偶尔骑出几米的平衡，是显露，但还谈不上结构；能稳稳骑一圈之后，那个平衡才成了结构。把S读成「结构」，就会把一切显露都想成早已成形、只待被发现的东西，这正是发现学的老路。

差异运行（D），是前后、左右、这样与那样之间的不同，在过程中一个接一个地出现、被区分、被承接。车把向左偏了三度和偏了十度不一样，这一下蹬得急和蹬得缓不一样；差异不是静止地摆在那里的「区别清单」，而是在运行中的序列：这一个差异引出下一个差异，下一个又改变再下一个。所以我们说「差异运行」或「差异序列」，而不只说「差异」。

特征纠缠（E），前两章已经说过，是一块块经验、一个个可区分的方面缠到一起，彼此牵动。重心、车把、视线、蹬踏缠在一起，牵一发而动全身。

三者不是三个部件，而是同一个发生过程的三个维度。就像我们描述一次骑行，不能只说「车在哪里」，也要说「车怎么动」，还要说「车和人怎样连在一起」；少了哪一个，描述都不完整。

◆ 二、三条方程：每一维都由另外两维决定

SDE发生学把三者之间的关系写成三条方程：

S＝F(D,E)，显露由差异运行和特征纠缠共同决定； D＝G(S,E)，差异运行由显露和特征纠缠共同决定； E＝H(S,D)，特征纠缠由显露和差异运行共同决定。

这里的F、G、H暂时不是可以代入数字的具体函数，而是表达一种结构性的主张：没有哪一维是自己决定自己的，也没有哪一维是只决定别人、自己不被决定的。三条方程缺了任何一条，理论就出现「缝隙」。这是本体论立场，下面用小舟骑车逐条说明它的意思。

先看S＝F(D,E)。门口那段缓坡，在小舟眼里显露为「可以冲下去」，这件事取决于两样东西。一样是他身上已经缠成的组织：重心、刹车、车把能够协同，他对「冲下去」有把握。另一样是正在运行的差异：今天坡上有没有人，车速怎样一点点变快，风从哪边吹。缺了纠缠，坡只会显露为「危险」；缺了差异运行，坡什么也不会显露，它只是一个他没有经过的地方。

再看D＝G(S,E)。会骑车之后，小舟开始注意到很多以前根本注意不到的差异：路面上一条细缝，前面那个人走路的方向，自己车胎是不是有点软。这些差异以前也存在，但没有在他身上「运行」起来。现在它们运行起来了，是因为「路可以骑」这个显露已经站住，而他身上的纠缠又足以让细小的差异有地方落脚。换句话说，人能分辨出哪些差异，取决于他已经看见了什么、已经缠成了什么。

最后看E＝H(S,D)。小舟的纠缠为什么会一次次改写？因为新的显露和新的差异不断进来。第一次放手骑出七八米，那个短暂露出来的平衡，加上「妈妈没在后面」这个新差异，把他的整套组织冲散，然后又在后面几天里以新的方式缠了回来。纠缠不是自己长的，它是被显露与差异运行塑出来的。

这三条方程合起来，说的是一个循环：纠缠和差异运行让某物显露，显露和纠缠又决定哪些差异能运行起来，显露与差异运行再回过头改写纠缠。学习就发生在这个循环里。

◆ 三、六路径与三原理：学习怎样走

三条方程说的是「谁决定谁」，还没说「先动哪一维」。实际的学习过程总有先后。三维的六种次序，就是六路径：D→E→S、E→D→S、D→S→E、S→D→E、E→S→D、S→E→D。注意，路径不是两维之间的单箭头，而是三维走完一轮的顺序。

用骑车举几个例子。小舟摔了一跤（一个强烈的差异进来，D），身体里重心和车把的配合被打乱又重新缠上（E），第二天他起步时露出一种更低、更稳的姿态（S），这是D→E→S。周敏第一次重新骑车，她身上多年前的组织还在（E），一上车就让她对车头晃动格外敏感（D），二三十米后熟练的样子回来了（S），这是E→D→S。小舟看了邻居大孩子单手骑车的样子，这个画面先显露在他眼前（S），他开始注意自己手上的力有哪些不同（D），后来自己的组织慢慢变了（E），这是S→D→E。其余三条路径也都能在日常学习中找到，这里不一一列举。

路径为什么会启动？三原理给出回答：当某两维之间发生矛盾，就会产生推动第三维改变的动力。这个过程有三拍：张力积累、形状改变、回写。

还是骑车。小舟心里已经有一个显露：「我会骑了」（S）。可是一上有坡的路，车速一快，差异运行就跟不上（D），他总是来不及判断要不要刹车。S和D之间的矛盾越积越多（张力积累），终于逼得他身上的组织发生变化：手开始习惯性地虚搭在刹车上，下坡前身体会先往后坐一点（E的形状改变）。这个新组织一旦形成，又回过头改写了「我会骑了」的意思，从「在平地上会骑」变成「在坡上也敢骑」，差异运行也换成了新的节奏（回写）。

矛盾也可以发生在另外两维之间。周敏重新骑车的那天，借来的是邻居家一辆座位偏高的车。她身上多年前的组织还在，脚该在什么时候点地、身体该怎样前

倾，都是旧车留下的习惯（E）；可是这辆车每一个起步、每一次停车都带来与旧习惯对不上的差异：脚够不着地，车头比记忆中轻（D）。E与D之间的张力积累了几十米，推动的却是显露的改变：那辆车在她眼里从「一辆自行车」变成了「一辆座位太高、要先调一下的车」，她自己的样子也从「我会骑」变成「我得先下来调座」。她下车把座位放低两格，这个新显露又回写了差异运行：再上车时，她留意的不再是车头的轻重，而是脚掌着地的那一下是否踏实。同样是三拍，这一次被推着改变的是S。

可见三原理不规定哪一维必须是被推动的那一维；推动谁，取决于矛盾发生在哪两维之间。这一点对设计训练很有用：如果希望改写纠缠，就要设法让显露与差异运行之间出现足够的、但又不至于压垮学习者的张力。

六路径与三原理在全书很多地方还会用到，这里只立一个基本认识：学习不是匀速累积，而是在某两维的矛盾中被推着走，一轮一轮地改写。

◆ 四、学习对三维的共同改写

有了三条方程，就能说清第1章留下的一个问题：为什么一次学习同时改变能力、世界与自我?

传统看法会把学习的效果放在某一维上。有人把学习看作「学会了显露」：学会骑车就是学会了一个正确的动作模式，学会一道题就是记住了正确答案。这是只盯S。有人把学习看作「学会分辨」：学骑车就是学会分辨各种失衡信号，学认字就是学会分辨字形。这是只盯D。也有人把学习看作「建立联结」：刺激和反应连起来了。这看上去是盯E，但往往把联结想成一根根单独的线，与显露和差异运行无关。

三方程的主张是：**一次真正的学习，显露、差异运行与特征纠缠三维一起被改写**。小舟学会骑车后，显露的东西变了（缓坡、减速带、电动车都露出新样子，他自己的姿态也变了），运行的差异变了（他能分辨的细节、他关注的顺序都变了），纠缠也变了（重心、视线、手脚缠成了新的组织）。如果只有其中一维变了，我们就该怀疑学习有没有真正发生。只改了S，像是背下了一个标准动作，一

换车就不会；只改了D，像是能说出「这里失衡了、那里失衡了」，但手脚跟不上；只改了E而S和D不动，那纠缠就是空转的，没有在任何显露中得到检验，也没有让新的差异运行起来。

这条判断可以落到一个可操作的提问上：面对任何一种声称的学习，我们都可以分别问三句：显露的东西有没有变？运行的差异有没有变？纠缠的组织有没有变？三问都有正面回答，才比较有把握说学习发生了。

三条方程的用处，在「缺了哪一条」时看得最清楚。日常生活里，有不少半截子的学习，恰好可以按缺口来分辨。

第一种缺口，是显露不再回头改写纠缠，也就是E＝H(S,D)这一条断了。小舟有个同学在小区空地上骑得很好，一圈又一圈，可是一到小区里车来人往的主干道就慌。他在空地上露出来的「会骑」，从来没有经过来往车辆、突然开门的汽车、路边窜出来的小狗这些差异；同一套组织在同一块空地上不断重复，显露的东西越来越熟，纠缠却没有被新的显露与新的差异继续改写。这种学习看上去稳定，其实是封闭的。它的毛病不是练得少，而是改写的通道断了。

第二种缺口，是差异运行不再受显露与纠缠调节，也就是D＝G(S,E)断了。和面时，有人一律按菜谱写的「揉十五分钟」来揉，不看面团的光泽，也不摸它的回弹，冬天夏天、高筋低筋都一样。差异照样在发生，面团每一分钟都在变，可这些差异没有被已经显露的状态和手上的组织所引导，只是按钟点机械地运行。结果有时揉不够，有时揉过头，而揉的人说不出为什么。

第三种缺口，是显露脱离差异运行与纠缠，也就是S＝F(D,E)断了。最典型的是背口诀。小舟会背「往哪边倒就往哪边拧车把」，背得一字不差，可这句话在他身上没有对应的差异运行，也没有缠进手脚的组织，它是一个漂浮的显露：说得出，用不上。很多「懂了但不会」的情形，都属于这一类。

这三种缺口并不罕见，而且它们在AI系统中也有相应的问题形式：只在固定分布上反复优化而不接触新情境，按固定流程调参而不读模型状态，生成流畅却没有内部组织支撑的表达。至于某一个具体系统究竟落在哪种缺口里，不能凭类比判

断，要看它的公开训练机制与检验结果。第15章讨论幻觉与失配、第20章讨论发现学前提时，都可以用这三种缺口作为提问的起点。

◆ 五、不要把E降格为背景或数据容器

三维之中，最容易被误读的是E。常见的误读有三种，每一种都会让理论退回发现学。

第一种，把E当作「背景」或「环境」。仿佛S和D在台前演戏，E只是舞台布景。这会让人以为纠缠是一个固定不变的条件，学习只发生在显露与差异之间。可是第2章已经看到，纠缠本身正是训练的对象，它会强化、松动、分化、重组；把它当布景，学习的主角就被藏到了后台。

第二种，把E当作「数据容器」或「记忆仓库」。仿佛经验被一条条存进去，需要时再取出来。这其实就是搬运图景换了个名字：只不过货物从「知识」换成了「经验」。纠缠不是存放，而是组织；它不像仓库，倒更像面筋，是在相遇中一次次缠出来、扯不开的网络。周敏多年后重新骑车，她不是从仓库里「取出」骑车的记忆，而是旧的组织在新的车把、新的路面上重新被缠活。

第三种，把E当作「训练数据集」。这在讨论AI时尤其常见：说大模型的E就是它的训练语料。语料是相遇的材料，不是纠缠本身。同样一批语料，用不同的训练结构，会缠出不同的组织；同一个模型在训练前后，面对的语料相同，纠缠却截然不同。把E等同于数据，会让人以为「数据越多世界越全」，这正是第26章要讨论的「信息负担」问题的根源之一。

既然E不是仓库，那它有什么可以描述的性质？至少可以先区分三个方面，供后文使用。一是范围：哪些方面被缠在了一起。初学时，小舟的纠缠只涉及车把和重心；后来视线、路况、身后的车声都被缠了进来。二是紧度：牵动有多直接。熟练的骑车人，车一歪，补偿几乎同时发生；初学者则要隔半拍，甚至要先想一想。三是弹性：条件变了，组织能不能跟着调整而不散架。周敏换了一辆陌生的车，晃了二三十米就稳住了，说明她的纠缠有弹性；小舟那位只在空地上骑的同学，一到主干道就散，说明他的纠缠紧而脆。范围、紧度、弹性都会被训练改写，也都会反

过来决定什么能显露、哪些差异能运行。这三个方面目前只是描述用语，还不是测量指标；第4章会讨论，怎样把类似的描述推进为可以测量的东西。

那么在大模型里，E到底对应什么？本章只能给出谨慎回答：可以把训练后参数之间形成的相互依赖、以及由此产生的内部表示之间的关联，作为特征纠缠的一种候选载体来研究；但参数连接与SDE特征纠缠之间的对应程度，是机制假说而非已知事实，第12章会专门讨论这条解释边界。

◆ 六、边界：三方程能做什么、不能做什么

三方程目前是一个本体论框架，不是一个可以计算的模型。它能做的，是提供一套检查清单：一个学习理论、一个训练方案、一个世界模型的设计，有没有在某一维上留下缝隙？例如，一个只预测下一帧画面的模型，主要在处理显露的延续，它对差异运行的组织和对纠缠的改写是怎样安排的？这样的追问，是后面第四编审视技术路线时要用到的工具。

它不能做的，是直接告诉工程师该怎样设计网络，或者直接预言某个模型会成功或失败。F、G、H的具体形式，至今没有被确定；三方程的形式化方向，是第40章的研究纲领要处理的问题。在此之前，任何人用三方程「证明」某项技术必然失败，都是越界。

三方程讲的是三维之间的结构关系，还没有讲学习过程中那些看似混乱、突然、不可预测的时刻：为什么小舟某一天忽然就会了？为什么有时练了很久毫无进展，然后一下子跃升？这些问题涉及复杂性，交给第4章。

> 带走的话
>
> 显露、差异运行、特征纠缠彼此决定，学习是三维的共同改写；E不是背景、不是仓库、也不是数据集，而是在相遇中不断被缠出、又回头塑造显露与差异的组织。

> **小账** 论证落在哪一层?——本章以本体论立场为主，六路径与三原理的骑车例示带有机制假说性质；关于大模型参数的对应只作候选，不作事实陈述。 世界模型补了哪一条?——评价一个世界模型，要分别问它的显露、差异运行和纠缠是否被训练改写；只延续显露而不改写纠缠的模型，在三方程意义上留有缝隙。

04

第 4 章

复杂性如何进入学习

> **这一章的问题**：学习中那些混乱、停滞和突然开窍的时刻，能不能被说清楚，又能不能被测量？

◆ 一、命题：学习不是匀速爬坡

几乎每个家长都见过这样的情形：孩子练了很多天毫无起色，某一天忽然就会了。小舟学骑车也是这样。前五六天，他每天在小区空地上练半小时，进步几乎看不出来，周敏松一次手他就倒一次；第七天傍晚，他骑出去十几米，第八天就能骑一整圈。如果把「能独自骑多远」画成一条曲线，它不是一条斜着往上的直线，而是长时间贴着地面，然后陡然抬头。

本章的命题是：学习过程本身是复杂的，它包含混乱的碰撞、局部的自组织与整体性能力的涌现；用混沌态、介生态、秩序态来描述这一过程，可以给「突然开窍」「越练越乱」「练成了又散了」一个统一的发生学解释。但这只是概念解释；要让它接受检验，还必须把它推进为可以测量的动力学指标。本章既要立住前者，也要把它与后者分开。

◆ 二、混沌碰撞、自组织与涌现

先用白话交代三个常被滥用的词。

混沌碰撞，在这里指的是大量差异同时涌入，彼此之间还没有稳定的关联，相互干扰、相互抵消。小舟头几天骑车，手、脚、眼、身体都在乱动，每一次歪斜引起的反应各不相同，有时过度纠正，有时纠正反了。这不是「什么也没发生」，而

是很多东西在发生，只是它们还没有缠到一起。需要说明的是，「混沌」在数学和物理学里有严格的定义，指一个按确定规则运行的系统，对初始条件极端敏感，起点差一丝，过一段时间后走向就完全不同，因而长期无法预测；本书在概念解释层面借用这个词，指的是「差异多而未组织」，二者不能混同。

自组织，指的是没有外部的蓝图或指挥者规定细节，组织却在局部相互作用中逐步形成。周敏扶车、喊「看前面」，但她无法规定小舟每一块肌肉什么时候用多大力；那些细节是在一次次歪斜和纠正之间自己调出来的。训练者组织的是相遇条件，局部的配合方式则是自组织出来的。这正是第2章说的「训练者能组织的是相遇，不是纠缠本身」。

涌现，指的是整体上出现了一个任何局部都不具备的能力或性质。车把、重心、蹬踏、视线，单独拿出哪一样都不叫「会骑车」；它们缠到一定程度，「会骑车」作为一个整体能力出现了。在发生学看来，涌现不是神秘的跳跃，而是纠缠跨过某种组织程度之后的显露：S＝F(D,E)里，E的组织变了，S才露出新样子。

三者的关系可以概括为：混沌碰撞提供材料与张力，自组织在局部把材料缠起来，涌现是缠到一定程度后整体显露的新能力。

◆ 三、混沌态、介生态与秩序态

SDE发生学用三态来描述一个发生过程所处的状态。

混沌态：差异大量运行，但纠缠松散，显露不稳定、不可预期。小舟前几天的样子就是混沌态：今天能骑三米，明天一米也骑不了，看不出规律。

秩序态：纠缠稳定，差异运行有固定的节奏，显露可以反复出现、可以预期。会骑车之后，小舟每次起步、转弯、刹车都大同小异，这是秩序态。

介生态：介于两者之间、新组织正在生成的状态。它最难描述，也最关键。第六、七天的小舟就在介生态里：有时能骑出一小段，有时又倒；能骑出的那一小段并不完全一样，每一次都带着一点新东西。介生态的特征是：局部已经开始成形，但还没有连成整体；新旧组织并存、相互竞争；一点小扰动既可能推它进入秩序

态，也可能把它打回混沌态。周敏在第七天放手，恰恰是在介生态里施加了一个合适的扰动。

和面同样能看到三态。刚加水时，面团疙疙瘩瘩，是混沌态；揉到七八分，面团时而光滑时而粘手，局部成团而整体未匀，是介生态；最终光滑有弹性，是秩序态。揉过头时，秩序态又会被打散，面团回到一种粘软无形的状态。所以三态不是一条单向的阶梯，秩序态可以退回，介生态也可以长期停滞。

三态之间的转变，也可以用第3章的三原理来看。小舟在介生态里，「偶尔能骑出一段」这个显露已经出现，但差异运行还跟不上：每一次歪斜，他的反应时快时慢、时对时错。S与D之间的张力在这几天不断积累，推动的是E的形状改变：重心与车把之间的配合越来越固定。等到这一新组织回写到显露与差异运行上，「能骑一整圈」稳定地露出来，他对歪斜的反应也有了固定的节奏，这一轮三拍走完，状态就从介生态进入了秩序态。所以，介生态并非一段无事可做的等待，而是张力积累那一拍正在进行；它看上去像停滞，是因为形状改变还没有到来。

三态与三方程是什么关系？可以这样理解：三方程说的是三维如何相互决定，三态说的是这种相互决定在某一时刻达到了什么程度。在混沌态，三条方程都在起作用，但彼此的牵动还很弱、很乱；在秩序态，三维之间的相互决定已经形成稳定的循环；介生态则是这个循环正在形成而尚未闭合的时候。

◆ 四、概念解释与可测动力学指标

到这里，上面这套描述都还是概念解释。它能让人理解学习过程，却不能让人判断：此刻小舟到底处于哪一态？介生态什么时候开始、什么时候结束？如果说不清，三态就有沦为事后贴标签的危险：练会了就说「进入了秩序态」，没练会就说「还在介生态」，怎么说都对，也就等于什么都没说。

所以必须把概念解释与可测的动力学指标分开。指标不是概念本身，而是从行为或系统状态中可以读出、可以重复测量、可以被别人复核的量。下面这张表把本章的几个概念与一些可以设想的指标对应起来。表中的指标是本书提出的研究建议（机制假说层面），并非已经确立的测量标准。

概念解释	可设想的动力学指标（以骑车为例）	在AI训练中的可能对应
混沌态	每次骑行的距离、纠正动作的方向与幅度波动大，前后两次几乎不相关	训练损失大幅波动，同一能力在不同评测批次间表现不稳
介生态	波动开始收窄，但出现间歇性的长距离成功；扰动后恢复时间忽长忽短	某项能力的评测成绩出现跳变前后的不稳定区间
秩序态	距离与动作稳定、可预测；轻微扰动后迅速恢复	能力在多次评测和轻度扰动下保持稳定
涌现	某一能力在短时间内从接近零跃升，且与此前单项指标的平滑变化不相称	能力随规模或训练步数出现非平滑跃升（是否真跃升，见下文关于指标的讨论）
秩序态退化	原本稳定的指标在条件变化后迅速发散	分布外测试中能力显著下降

表中有几个指标值得多说一句。

「扰动后的恢复时间」是一个特别有用的量。小舟在秩序态时，被路边的狗吓一跳，车晃一下，一两秒就稳住；在介生态时，同样的惊吓可能让他晃很久，或者干脆倒地。用一个人为的、可控的小扰动去试，看系统多快回到原来的状态，比单看成功率更能区分三态。这一思路在复杂系统研究中有类似的讨论，常与「临界慢化」相联系：在生态、气候等系统里，当系统接近一次突变的临界点时，它从小扰动中恢复得越来越慢，这被当作突变将临的一种预警信号；本书只借用其测量思路，不主张学习过程一定符合某种临界相变模型。

「前后相关性」也是一个可用的量。混沌态中，这一次怎么骑和上一次怎么骑几乎无关；秩序态中，前后高度相似；介生态中，相关性正在上升但起伏很大。

这些指标并不一定需要实验室。下面是我构想的场景：周敏是物理老师，习惯用数据说话，她在小区空地上用粉笔画了一条起点线，每天傍晚让小舟从起点骑出去，用手机计时，记下他不着地能骑多远，再记下途中脚点地几次；每骑完五趟，她就在小舟身后轻轻拍一下车座，制造一个可控的小晃动，看他几秒钟后能重新骑直。连续记两周，她就能画出三条曲线：平均距离、距离的波动、扰动后的恢复时间。按本章的假说，在小舟「突然会了」的那一天之前，应该能看到波动先变大、再开始收窄，恢复时间从很长、很不稳定，逐渐变短、变稳。若记下来的曲线根本

不是这样，比如波动一直不变、能力却一夜之间跃升，那就说明本章对介生态的刻画需要修改。这个构想的好处，在于它给出了一个可以失败的预期。

还可以设想一个反向的检验。三态既然可以退回，那么秩序态的退化也应当有前兆。周敏多年不骑车，重新上车的头几秒，她的恢复时间明显变长，车头的波动也变大；如果在她上车后的前两三分钟里连续记录，按本章的说法，应当看到这些量先偏离、再较快地回到熟练者的水平，而且回落的速度明显快于小舟初学时。这一预期与第1章关于「再学节省」的说法相互呼应：同是从不稳定走向稳定，有旧组织可以被唤醒的过程，其动力学曲线应当与从零开始的过程有可区分的差别。若测出来两者没有差别，第1章与本章的相应说法都要重新考虑。

在AI一侧，关于「涌现」的公开讨论值得特别提醒。2022年有研究报告说，大模型的某些能力在规模较小时接近随机水平，规模过了某个点后突然明显提高，研究者称之为「涌现能力」；2023年又有研究指出，这类跃升在相当程度上取决于评测指标的选择，例如只算「整题全对」这种非此即彼的指标容易显出跳变，换用能给部分得分的连续指标，同样的数据往往呈现平滑提高。两方面的研究至今仍在讨论之中。这正好说明本章的要点：涌现是一个概念解释，是否观察到涌现，取决于选用什么指标。指标选得粗，平滑的变化会看起来像跳变；指标选得细，真正的组织跃迁也可能被平均掉。所以讨论某种能力「涌现」了，必须同时说明用了什么指标、指标是否经过不同方式的复核。

还有一个公开报告过的现象，与小舟「练了多日忽然会了」有几分形似，值得在这里提一下。有研究者在小规模的算法类任务上观察到：神经网络在训练集上早已做到几乎全对，对没见过的样例却长时间表现很差；继续训练很久之后，它在没见过的样例上的表现才突然大幅提高。这一现象由研究者在2022年公开报告，任务是模运算一类的小规模算法数据，英文称为grokking，中文常译「顿悟」，也叫「延迟泛化」。后续有研究把训练好的小网络拆开来看，发现在表现跃升之前，网络内部已经在一点点形成一套有规律的计算方式，跃升前后的变化是渐进积累之后的结果，而不是凭空一跳。

这个例子对本章有两点启发，但也有一点必须防止的误用。启发之一是，外部表现的长期平稳，不等于内部组织没有变化；如果只看成功率，就会把正在进行的张力积累误判为「没有进展」。启发之二是，要看清介生态，除了行为指标，还需要能读出内部状态的指标，在AI中这意味着要研究训练过程中内部表示的变化，而不只是评测分数。必须防止的误用是：这只是特定任务、特定设置下的经验事实，不能拿它来「证明」三态理论，也不能把它推广为所有AI训练都有的规律。

◆ 五、复杂性过程形成稳定能力的条件

混沌碰撞不一定导向能力。有的孩子学骑车，摔了几次就怕了，从此再不上车；有的面团怎么揉都不成形。复杂性过程要形成稳定能力，至少需要几个条件。这里提出的仍是机制假说，供检验。

第一，差异的强度要落在可承受的区间。差异太弱，什么都不会发生；差异太强，刚开始缠起来的组织会被冲散，学习者甚至会退出。周敏先扶后放、选在空地而不是车来人往的地方练，都是在调节差异强度。

第二，要有足够的重复，又不能是单调的重复。完全相同的重复只会让已有的组织越来越紧，却不长弹性；完全不同的经验又难以缠到一起。小舟在同一块空地上练，但每天风向、地面、心情都不同，这种「同中有异」最有利于介生态向秩序态推进。

第三，要有让后果回到组织上的通道。第2章讲过反馈放在哪里的问题；在复杂性的视角下，这是三方程循环能否闭合的问题。没有回写，就没有自组织，只剩下随机的碰撞。

第四，要允许一段时间的不稳定。介生态看上去像退步，周敏有一天也急了，说「怎么比昨天还差」。如果训练者在介生态里过早地强加固定程式，比如规定每一步该怎么做，新组织就可能来不及自己形成。

第五，要有检验秩序态是否真的形成的办法。这就回到上一节的指标。

◆ 六、边界与过渡

本章有两条边界必须说明。

一是概念借用的边界。「混沌」「自组织」「涌现」「临界」这些词，在数学、物理和复杂系统研究中各有严格含义。本书在发生学层面使用它们，并尽量给出白话定义；但不能因为用了同一个词，就宣称学习过程已被证明符合某一种数学模型。若有读者希望将三态做严格的数学刻画，这正是第40章「特征纠缠的形式化方向」要讨论的内容。

二是从指标到本体论的边界。即使将来某些指标被证明能很好地区分三态，也只说明三态描述在经验上有用，不等于证明了SDE本体论。反过来，如果某个AI系统的训练曲线看不出介生态，也不能直接说它「没有学习」，可能只是指标选得不对。

复杂性过程走到秩序态，能力就「稳定」了。可是稳定了的能力，怎样被一次次调用？什么时候可以放心地交给它去做事？它又会在什么时候失效？这是第5章要立起来的「封装」。

> 带走的话
>
> 学习是在混沌碰撞、自组织与涌现中穿过混沌态、介生态走向秩序态的过程；这套说法要成为科学，必须配上可测、可复核的动力学指标，并说明指标的选择会影响看到什么。

小账 论证落在哪一层？——本章前半为本体论层面的概念解释，后半为机制假说（指标建议与稳定条件）；关于大模型「涌现」与「顿悟」的公开研究作为经验事实出现。 世界模型补了哪一条？——判断一个世界模型的能力是否已进入秩序态，要用扰动恢复、前后相关性等可测指标，而不能只看平均成功率；「涌现」的宣称须注明所用指标。

05

第 5 章

从纠缠形成到能力封装

> **这一章的问题**：一种能力怎样从「刚刚缠成」变成「可以放心交给它去做事」，又怎样被发现失效、被拆开重训?

◆ 一、命题：能力是有条件的封装

前四章讲的是纠缠怎样建立、训练、在复杂过程中走向秩序态。可是在生活里，我们关心的往往不是「纠缠是否形成」，而是「这件事能不能交给他了」。小舟会骑车之后，周敏面临的真实问题是：能不能让他自己骑车去小区另一头的同学家？（按我国现行规定，未满十二岁不能在道路上骑自行车，所以小舟的骑行都在小区里面。）

本章提出一个后面多编都会用到的概念：封装。所谓封装，是指一组特征纠缠在充分训练和检验之后，作为一个整体能力被固定下来，可以在一定条件下被直接调用，而不必每次重新组织其内部细节；同时，它带有明确或隐含的适用边界，在边界之外可能失效，失效后需要解封、重新训练。

「封装」这个词借自工程。在工程里，一个封装好的部件，使用者只管它的接口和适用条件，不必知道里面怎样运作。本书借用的是这层意思，但要补上工程用语里容易省略的两件事：第一，封装不是一次完成的，它有发生、稳定、可调用三个环节；第二，封装总是带着边界和继承下来的判断，它可以失效，也可以被打开。本章的命题是：**一种能力，就是一个有条件的封装；可靠，不是「没有边界」，而是「边界被识别、被检验、可修订」**。

◆ 二、能力的发生、稳定与可调用

先把封装的三个环节分开。

发生。这是前四章讲的过程：经验相遇，差异运行，纠缠缠成，某种能力第一次显露。小舟第七天骑出十几米，能力发生了。但发生不等于封装：那天他换一块地方、换一个时间，未必还能骑出来。

稳定。能力在多次、多种条件下都能出现，进入第4章说的秩序态。小舟在空地上每天都能骑一圈，早上晚上都行，穿外套、不穿外套都行。稳定意味着内部组织已经足够牢、有弹性，但这时的能力还常常需要「全神贯注」：他骑车时不敢说话，不敢回头，整个人都在骑车这件事上。

可调用。能力被稳定之后，还要能在别的事情里被当作一个现成的整体拿来用，而不占满注意。小舟能一边骑车一边跟周敏聊天，一边骑一边留心岔路口有没有车拐出来，骑车成了他做别的事的手段，而不再是目的本身。到了这一步，骑车这项能力才被封装：它的内部细节，重心怎么调、车把怎么转，已经不必每次重新组织；它对外只呈现一个「接口」：想去哪里，就骑过去。

用第3章的三方程来看，封装在三个维度上各有其意义。在显露一侧，封装意味着一个稳定的「接口」露了出来：对外，人们只看到「会骑车」这个整体，看不到里面千百次的微调；结构作为显露的稳定态，在这里就是那个可以被他人和自己反复认出的能力形象。在差异运行一侧，封装意味着大量细小差异被内部消化，不再上升到注意之中：车把的细微晃动在内部就被吸收了，能运行到注意里的，只剩下与任务有关的差异，比如路口来车、前方有坑。在特征纠缠一侧，封装意味着一组纠缠获得了足够的紧度与弹性，能在一定范围的扰动中维持整体。三者缺一，封装都不完整：只有接口没有内部消化，就是第3章说的「漂浮的显露」；内部消化了却露不出稳定接口，别人无法放心地调用；纠缠没有弹性，接口再漂亮，条件一变就会散。

可调用还意味着能力可以被组合。小舟会骑车，也会认路，还知道在岔路口停一停、看一看，这三个封装可以组合成「自己骑车去同学家」这个更大的能力。组合不是简单相加：骑车时认路，注意力的分配方式和走路时认路不同；两个封装放到一起，接口之间要重新磨合。第一次独自去同学家那天，小舟骑过了一个岔路口才发现走错了，就是因为「骑车」和「认路」两个封装还没有在新的组合里调好节奏。封装之间的接口问题，在第30章讨论专用模型的协作与共享时，会成为核心问题之一。

三个环节的区别，对评价AI系统同样重要。一个模型在某项基准测试上第一次取得好成绩，最多说明能力发生了；在不同测试集、不同提示方式下表现一致，才谈得上稳定；能被可靠地嵌进更大的任务流程，比如作为一个工具被别的系统调用而不出意外，才谈得上可调用。第1章区分的五种变化在这里也用得上：暂时情境变化和外部记忆可以让表现变好，但不等于封装；只有持久学习，才可能走到可调用的封装。

◆ 三、反复试用、评估与边界检验

能力从稳定走到可调用，靠的不只是继续练，还要靠有意识的试用和检验。

周敏没有在小舟会骑一圈之后就放他到处骑。她先带着他在楼下空地边上的小环路骑，再带他在人少的周末早上骑到小区里有车进出的主干道，她自己骑在旁边；再后来，让他在前面骑，她跟在后面看。每一步都是一次试用：在一个比原来稍宽一点的条件下，看这项能力还站不站得住。每一次试用后，她都会问几句：刚才那个岔路口你看了左边没有？那辆三轮车从侧面过来的时候你怎么想的？这是评估：不只看骑没骑过去，还要看是怎么骑过去的。

在这个过程中，最有价值的往往不是成功，而是那些差点出事的时刻。有一次，一辆车从路边的车库里倒出来，小舟没有注意到倒车的声音，周敏喊了一声他才刹住。这一次让母子俩都认识到：小舟的骑车能力在「车库倒车」这种情形里是有边界的，他还没有把「听声音判断看不见的车」缠进去。这就是边界检验：有意识地去找能力在哪里会失效。

边界检验的结果，不一定是立刻把能力练到覆盖这个边界，也可以是把边界写进使用规则。周敏的决定是：小舟可以自己骑车去同学家，但那条路上要经过的小区车库出口，他必须下车推着走过去。能力没有变，但封装的「接口说明」变了。很多可靠的能力，正是这样带着清楚的使用规则运行的。

在工程领域，类似的过程有成熟的做法：分阶段的测试、在仿真中构造极端情形、逐步扩大运行范围、记录接近失败的事件。第35章将讨论世界模型的可靠性形成与持续封装，那里会更细地讨论这些做法；这里只指出一个原则：封装的可靠性，取决于边界检验做得多充分，而不只取决于能力在常规条件下多强。

◆ 四、封装所继承的判断

封装还有一个容易被忽视的方面：每一个封装里，都装着前人和他人的判断。

小舟骑车时「靠右骑」「过岔路口先看左再看右」「车库出口下车推」，这些不是他自己从摔跤中学出来的，而是周敏、学校、交通规则传给他的。它们来自无数人此前的经验和教训，被压缩成几条简单的规则，缠进了他的骑车能力。他不需要亲自经历一次车库口的事故，才知道那里危险。

和面也一样。周敏和面的手法、加水的比例、醒面的时间，是从她母亲那里学来的；母亲又是从再上一辈那里学来的，中间也许还掺进了某本菜谱、某个邻居的窍门。周敏揉出的每一团面，都继承了一长串她自己未必说得清来源的判断。

这一点对理解AI尤其重要。一个大模型的能力封装，同样继承了大量判断：训练数据中前人的表达和推理，数据筛选者对「什么值得学」的判断，设定训练目标的人对「什么叫好」的判断，评价反馈中标注者的判断。这些判断被缠进了模型的能力，往往不再以明确规则的形式出现。这正是训练的发生单位应当被理解为集体人类—AI复合体的原因之一：封装里装着的，不只是模型自己的「经验」，还有一整段人类活动的沉淀。

继承来的判断有两面。好的一面是，它让学习者不必从零开始，可以站在前人肩上；不好的一面是，它也可能把前人的偏向、过时的经验、特定情境下才成立的

判断一并封装进来，而使用者看不见。小舟学的「靠右骑」，在左行的地方就不适用；周敏母亲的加水比例，换了一种面粉就不准。所以，评估一个封装时，除了问「它能做什么」，还要问「它继承了谁的判断，那些判断在什么条件下成立」。第15章会讨论数据选择和评价标准给集体语言世界留下的偏向，它们正是这种继承判断的典型问题。

◆ 五、失效、解封与重新训练

第1章那件事，在这里可以完整地收拢起来：多年不骑，重新上车。第1章把它当作遗忘与重新学习的例子；从封装的角度看，它展示的是封装的失效、解封与重新训练的完整过程。

周敏上大学时的骑车能力，是一个早已封装好的能力：她可以一手撑伞、一手扶把，在雨里穿过校园。十几年不用，这个封装没有消失，但它的边界条件已经变了：车不是原来那辆，路不是原来那条，她自己的身体也不是二十岁时的身体。所以，刚上车的几秒里，封装失效了：原来那个「直接调用就行」的整体能力，在新的条件下没法直接运行，车头晃起来。

这时发生的是解封：她不得不重新注意那些本来不必注意的内部细节：手要握多紧，身体前倾多少，脚什么时候用力。已经封好的整体被暂时打开，内部组织重新暴露在注意之下，重新接受新差异的冲击。

随后是重新训练：骑出二三十米，新的车、新的路面、她自己现在的身体与旧的组织重新缠在一起。然后再次封装：她又可以一边骑一边指挥小舟了。重新封装的能力不完全等于从前那个，它多了「这辆车座位偏高」「长沙的路坡多」这样的新内容，也许少了「单手撑伞」这样暂时用不上的部分。

这个过程揭示了一个普遍道理：失效不是封装的意外，而是封装的常态之一。任何封装都有边界，环境变化、身体变化、任务变化都可能把它推出边界。重要的是三件事：能不能及时发现失效（周敏一上车就感到不对）；能不能顺利解封，而不是硬撑着按旧方式继续（她放慢了速度，重新注意细节）；能不能在新条件下重新训练并重新封装。

解封还有一个分寸问题：打开多少。周敏重新上车时，并没有把骑车这件事从头学一遍，她没有回到要人扶着车座的阶段，只是把与新车、新身体直接相关的那几处细节重新拿到注意里来，其余部分照旧在封装中运行。这是局部解封。局部解封的好处是快，坏处是可能打开得不够：如果失效的根子比想象的深，只调几处细节就会反复出错。反过来，整体解封虽然彻底，却代价很高，还可能连带破坏原本运行良好的部分。小舟有一阵子被周敏纠正骑车姿势纠正得太多，结果连本来骑得很顺的转弯也变得犹豫起来，就是打开得过多的例子。

所以，失效之后要先判断：失效发生在封装的哪一部分，需要打开到什么程度。这个判断本身也是一种能力，熟练者往往比初学者做得更好；它同样需要训练，也同样会出错。在AI系统中，相应的问题是：重新训练时，应当调整多少参数、保留哪些能力、用什么来检验没有误伤其他能力。这些是工程上正在研究的问题，例如只改动一小部分参数、其余冻结不动的「参数高效微调」，以及研究怎样不断学新东西而不丢旧本事的「持续学习」，本书不预设答案，只把「打开多少」作为封装理论中的一个必须回答的问题提出来。

最危险的情形，恰恰是失效了却没被发现。第2章讲过小舟揉发面时问「这面是不是坏了」，他至少发现了不一样；要是他照旧使劲揉、照旧按饺子面的时间醒，包子就会出问题，而他还以为自己的手艺没错。

AI系统中，相应的问题有公开的研究名称：分布外泛化，问的是模型遇到与训练数据差别明显的情形时还行不行；不确定性估计，问的是模型能不能知道自己哪里没把握；灾难性遗忘，问的是学新东西时会不会把旧本事冲掉。本书不在此评论任何具体系统的做法，只把封装视角下的三个问题交给后面的章节：系统如何发现自己的封装失效？如何在不破坏其他能力的前提下解封局部？如何在重新训练后重新确认边界？第35章将讨论持续评估、回写、解封及协作处理，第27章将讨论错误预测怎样被发现并转化为新经验。

◆ 六、边界：封装不是什么

最后说清这个概念的边界，以免它被用得太宽。

封装不是「黑箱」。说一个能力被封装，不是说它的内部永远不可知，而是说在常规调用时不必重新组织内部；一旦失效，它必须能被打开。一个只能整体调用、出了问题也打不开的能力，不是好的封装，是一个风险。

封装也不是「完成」。它不意味着学习结束；相反，可调用的封装往往是下一轮学习的起点。小舟会骑车之后，才能去学骑车带东西、骑山地车，等到了年龄再学骑车上路。学习的长链条，常常是由一个个封装连起来的。

封装更不是「世界被装进去了」。本书反对搬运图景，「封装」一词不是给它开后门。被封装的是一种在特定条件下运行的能力，而不是一个完整的世界副本。导论提出的问题「世界模型的开始应当包含对象同一性与任务能力的发生，而非预设全能世界已被一次性封装」，在这里有了更明确的含义：世界模型若要可靠，应当是一组组有条件、可检验、可解封的能力封装，而不是一个号称装下了整个世界的总封装。

本编到此完成了学习的本体论重建：学习是特征纠缠的建立与训练（第1章）；训练组织经验的相遇（第2章）；学习同时改写显露、差异运行与特征纠缠（第3章）；它穿过混沌态、介生态走向秩序态，并需要可测指标来检验（第4章）；稳定的能力以有条件的封装形式运行，可以失效、解封、重训（第5章）。可是这一切都发生在某个「世界」里：小舟的世界，周敏的世界。人的世界又是怎样建立的？理念、现实与自我三个世界怎样各自发生、相互改写？这是第二编的任务。

> **带走的话**
>
> 能力是一个有条件的封装，经历发生、稳定、可调用三个环节，装着继承来的判断，带着边界；可靠不在于没有边界，而在于失效能被发现、封装能被打开、重训后边界能被重新确认。

小账 论证落在哪一层？——本章以本体论立场与机制假说并重：封装三环节、继承判断与失效—解封—重训是概念框架，边界检验与发现失效的条件是可检验的机制主张；AI相关术语只作经验事实背景。　　世界模型补了哪一条？——世界模型的可靠性应以一组有条件、可检验、可解封的能力封装来衡量，而非一个号称完整的世界总封装；必须同时说明每个封装的适用边界、继承判断和失效发现机制。

第 二 编

理念、现实与自我如何发生

编序 第一编把学习重新放回发生学里：学习不是把知识搬进脑子，而是一块块经验缠到一起（特征纠缠，E），在差异的往来运行（差异运行，D）中，显露出可以被认出、被调用的东西（显露，S）。这一编把镜头从「学习」移到「学习者」身上，追问一个更朴素的问题：一个人的世界是怎样长出来的?

我的回答是：每个人都同时生活在三个世界里——理念世界、现实世界、自我世界——这三个世界不是摆在那里等人去认识的，而是在一个人的学习历史中一起发生的。第6章先立总论：从婴儿认出妈妈开始，人的世界与人自己一同发生，不同的经历造就不同的差异辨认。第7章讲理念世界，用小学生学分数这件事，说明概念、符号与推演怎样从一团混沌稳定成可以运算的东西，又怎样在冲突中重建。第8章讲现实世界，用小舟被烫了一下学会「热」，说明边界、空间、可行动性与现实阻力怎样在身体里发生；这里只立人的身体层面，视觉、听觉、触觉三种 SDE 的细部机制留给第五编。第9章讲自我世界，用周敏二十年的教师生涯，说明记忆、情感、承诺与行动历史怎样缠出一个「我」。第10章把三个世界放到一起，看它们如何相互改写，以及许多人的世界怎样形成可以共享、又不抹平差异的相对同一性。

这一编是全书的地基。后面讲人类—AI复合体的语言世界（第三编）、讲物理世界模型（第四至七编）时，读者会一再看到同一个结构：所谓「世界」，总是某

个发生单位在互动中发生出来的世界。先在人身上看清这一点，再去看机器与人如何一起学，就不容易把 AI 想成一个脱离人类历史、单独去「发现」世界的东西。

06

第 6 章

每个人的世界如何建立

> **这一章的问题**：一个刚出生的婴儿，是先有一个世界、再去认识它，还是世界和他一起长出来?

◆ 一、命题：世界不是先摆好的，是和人一起发生的

先把本章的立场说清楚，这是本体论的话：一个人的世界，不是一个完整的、早已摆好的房间，人只是走进去逐一清点家具；一个人的世界，是在他与周围的人、物、声音、温度反复相遇的历史里，一点一点缠起来、显出来的。人在这个过程中也一同被造就。**主体与世界是共同发生的，没有哪一方先完成、再去等另一方**。

这句话容易被误读成「世界是人想出来的」。不是。窗外的雨、母亲的体温、摔倒时地板的硬，都不由人随意决定。发生学并不否认这些东西的实在，它否认的是另一个前提：以为一个人的世界里有哪些「对象」、这些对象怎样分界、哪些差异值得注意，在人参与之前就已经全部划好了。发现学的图景里，婴儿像一台刚开机的相机，世界已经清清楚楚地摆在镜头前，只差把它拍进去。发生学的图景里，婴儿面对的起初是一团还没有分开的刺激——光、声音、触摸、饥饿、温暖混在一起（混沌态）——他认出什么、分开什么、把什么当作同一个东西，要靠他自己与周围的往来一次次纠缠出来。

用随身三问来说：这里纠缠的是婴儿身上的感觉、动作和他人的回应；差异在一次次的出现与消失、满足与不满足之间运行；显露出来的「妈妈」「奶」「抱」这些东西，靠反复的互动站稳，也会在条件变了的时候失效。

◆ 二、机制：前语言经验、身体活动与社会交往三股线

一个人的世界从哪里开始缠起？我把它拆成三股线，这是机制层面的说法，可以被发展心理学的研究支持或修正。

第一股是前语言经验。在会说话以前很久，婴儿就已经在分辨了：亮和暗、冷和暖、一种声音和另一种声音、被抱起和被放下。这些分辨起初不成体系，只是一次次差异的发生（差异运行，D）。可是差异一旦反复出现，并且总和另一些东西一起出现，它们就开始缠在一起。

第二股是身体活动。婴儿并不只是被动地接收。他转头、吮吸、抓握、蹬腿。每一个动作都改变了接下来会出现的差异：头一转，声音的方位就变了；手一抓，东西的阻力就来了。身体活动让差异不再只是落在他身上，而是由他参与造成。这一点非常要紧——后面讲现实世界（第8章）和讲世界模型为什么离不开行动（第四编、第七编）时，都要回到这里。

第三股是社会交往。婴儿的世界从一开始就不是一个人的。有人在喂他、看他、对他说话、对他的哭作出回应。别人的回应本身就是差异的来源，而且是最有组织的那种来源：大人会重复、会夸张、会等待婴儿的反应再接着做。可以说，社会交往替婴儿把差异排好了节奏。大人对婴儿说话时，往往不自觉地放慢、拉高、把语调起伏夸大，一句话说三遍，说完了停下来看孩子的反应。研究者把这种说话方式叫作「儿向语」；有一项常被引用的跨语言研究比较了法语、意大利语、德语、日语和英美两地英语这六个语言群体里父母对婴儿说话的录音，发现音调普遍更高、起伏更大、句子更短、停顿更长。至于它在世界各地有多普遍、各文化里程度差多少，研究者之间还有讨论；但它的作用方向是清楚的：把本来混在一起的声音差异拉开，让它们更容易被抓住；把一次互动切成一来一回，让婴儿的动作有了位置。一个母亲逗孩子玩「躲猫猫」，手一遮、一放、「哇」一声，孩子笑；再遮、再放，孩子笑得更早——他已经在等那个差异出现了。这样的游戏没有人教过妈妈怎么设计，却恰好是一份精心安排的差异序列。

三股线不是先后关系，而是同时缠绕。前语言经验提供差异，身体活动参与制造差异，社会交往给差异以节奏和意义。某些差异因为在三股线里反复共同出现，纠缠得越来越紧，最后显露为一个站得住的「东西」。这就是一个对象在婴儿世界里的诞生。

◆ 三、案例：婴儿认妈妈

本编的第一个主例，是婴儿认出妈妈。

先说经验事实。发展心理学有几项广为引用的研究。一项是 1980 年发表的实验：出生不到三天的新生儿，可以通过改变吮吸的快慢来「选择」听哪一种声音，他们更多地选了母亲的声音；研究者推测这与胎儿在子宫里已经听过母亲的声音有关，几年后的后续实验给这个推测提供了支持。另一项是 1989 年的研究：出生后几天之内的婴儿，看母亲面孔的时间比看陌生女性面孔更长。还有一项 1975 年的研究：在婴儿头两侧各放一块哺乳用过的乳垫，一块是母亲的、一块是别的母亲的，出生六天左右的母乳喂养婴儿更多地把头转向母亲那一块。这些研究各有方法上的局限，读者以原文为准。我在这里只取一个各研究大致一致的方向：婴儿最早认出的「妈妈」，不是单凭一张脸认出的，而是声音、气味、抱的姿势、喂奶的节奏一起认出的。

接下来是我的解释，属于本体论立场加机制假说。对一个出生不久的婴儿来说，世界上并没有一个现成的「妈妈」对象等着他去识别。有的是一连串还没有分开的差异：一种声音的音高和节奏，一种气味，一种被托住后背的方式，一阵温暖，饱足之前的那段吮吸。这些差异最初各自出现，可它们总是一起出现，并且总和「不饿了」「不冷了」「不怕了」一起出现。反复之下，它们纠缠成一团（特征纠缠，E），这一团越来越紧，终于显露为一个可以被认出的整体（显露，S）——在婴儿的世界里，「妈妈」就这样发生了。

这个「妈妈」一旦显露出来，就反过来改变差异的运行（D=G(S,E)）：婴儿开始对这一团特征的出现与缺席格外敏感。妈妈离开房间，他会注意到；换一个人

用同样的姿势抱他，他也可能察觉不对。在显露出「妈妈」之前，「她在」和「她不在」并不构成一个差异；显露之后，这成了婴儿世界里最重要的差异之一。

再往后，「妈妈」的纠缠还会分化。婴儿慢慢分出妈妈的脸、妈妈的声音、妈妈的手，分出妈妈高兴的样子和着急的样子；他也会把「妈妈」与「爸爸」「奶奶」分开。这一步是纠缠的分化，而不是在一个现成的家庭名单上逐一核对姓名。

这里还要补一层：婴儿认出妈妈的时候，婴儿自己也在发生。他开始有了「被这个人照顾的我」，有了「等她回来的我」，有了「看到她就安心的我」。所以，「妈妈」这个对象的发生，和「我」这个主体的最初轮廓，是同一个过程的两面。主体与世界的共同发生，在这个最小的例子里就已经完整了。

◆ 四、认生与戴口罩的妈妈：显露怎样失效又重建

「妈妈」一旦站稳，故事并没有结束。显露出来的东西靠纠缠撑着，纠缠的条件一变，显露也会动摇。这里有两个日常现象，可以帮我们看清这个过程。

第一个是认生。许多父母都注意到，孩子在半岁到一岁之间的某个阶段，会突然怕生：以前谁抱都笑，现在陌生人一伸手就哭。发展心理学的教科书一般把这种「陌生人焦虑」的出现放在六到八个月前后，并且个体差异相当大。用发生学来看，这恰好说明「妈妈」已经显露得足够稳定了。正因为婴儿世界里已经有了一个清楚的「熟悉的人」，一个不符合这团纠缠的人出现，就构成了强烈的差异。这里有一对矛盾：已经站稳的显露（S，熟悉的人应该是这样的）和眼前正在运行的差异（D，这个人的声音、气味、抱法都不对）。两者顶在一起，张力积累起来，表现为哭、扭身、找妈妈。这股张力推动的是第三维——纠缠（E）——的改变：婴儿要么把这个陌生人推到「不安全」那一边，要么经过几次温和的接触，把这个人慢慢缠进「可以被抱」的那一团里。这就是本书所说的三原理之一在一个婴儿身上的样子：两维相顶，推动第三维改变，最后回写到显露上——下一次这个人来，婴儿不哭了。

第二个现象更细小。妈妈换了发型、戴上口罩、感冒了嗓子沙哑，有的婴儿会愣一下，盯着看，迟疑片刻才笑。这说明「妈妈」不是靠某一个特征单独撑着的；

去掉一个特征，整团纠缠还在，只是松了一些，显露需要一点时间重新站稳。如果妈妈戴着口罩，但仍然用熟悉的声音说话、用熟悉的方式抱起他，孩子往往很快就认出来。如果只剩一张照片，没有声音、气味和抱，有的孩子就认得慢一些。

这两个现象我都只当作日常观察来用，不当作已被严格测量的结论。它们帮助说明一件后面会反复用到的事：对象的同一性有形成、维持、失效与重建四个环节。第25章讨论「物的意识」时，会把这四个环节做成可以检验的功能性判据；这里先在婴儿与妈妈之间看到它们的原型。

还可以顺便用六路径说一句。婴儿认出妈妈，大致走的是「差异先运行、再缠起来、最后显露」（D→E→S）这一条；认生之后重新接纳一个人，则更像是「已有的显露先站着，差异来冲击，再推动纠缠改组」（S→D→E）。同一个孩子、同一件事，在不同阶段走不同的路径，这正是第一编所说的「六路径是三维的六种次序」的一个家常例子。

◆ 五、不同经历产生不同的差异辨认

如果世界是在互动历史中发生的，那么不同的互动历史，就会发生出不同的世界。这不是相对主义，而是一个可以观察的事实方向。

经验事实里有一个常被提到的现象：婴儿在很小的时候能分辨许多种语言的语音差异，到一岁左右，对母语里不用来区分意义的语音差异，分辨能力会明显下降。一项经典研究让说英语家庭的婴儿分辨印地语和一种北美原住民语言里的辅音差别，六到八个月大的婴儿大多分得出，十到十二个月大时就很少能分出来了；而从小听这些语言的同龄婴儿仍然分得出。对面孔的分辨也有类似报告：六个月左右的婴儿还能分辨不同猴子的脸、不同族群的人脸，到九个月前后，这种分辨更多地收窄到自己常见的人群。研究者把这类现象统称为「知觉窄化」。

用发生学来读：这些不是「能力的丧失」，而是纠缠的组织。一个孩子的周围总用某些差异来区分意义，这些差异就被反复强化、站稳；另一些差异从不承担意义，就渐渐松开，不再显露。于是，长在不同家庭、不同语言、不同地方的孩子，世界里的「对象」划分并不相同。一个在南方湘中长大的孩子，耳朵里有一些北方

孩子不太留意的声调差别；一个从小跟着爷爷在菜地里转的孩子，眼里有许多城里孩子看不出来的叶片差别。

这种差别并不只属于婴儿期。成年人换一个环境，同样会经历差异辨认的重新组织。这是我构想的场景：周敏的学校来了一位从北方来的实习老师。学生们课间用长沙话开玩笑，他常常听得懂每个字，却听不出哪句是真生气、哪句是逗着玩；周敏在长沙郊区长大，不用想，一听语气就知道。一个学期下来，实习老师慢慢也能听出几分了。这不是他「获得了一条关于方言的信息」，而是他世界里的一组差异——语气上那一点点上扬或压低——从不显露变成了显露，并且开始指导他的行动：什么时候该接话，什么时候该把学生叫到一边。开学第一天，同一间教室里，他与周敏看到、听到的东西是一样的，可两人世界里显露出来的「教室」并不一样。

由此可以得出一个在后文很有用的判断：要知道一个学习者学会了什么，与其问「他接收了多少输入」，不如问「他的世界里哪些差异开始承担意义、哪些对象站稳了」。输入相同，发生的世界可以不同；输入不同，也可能发生出大体相通的世界。

这里要防止两种误读。第一种误读是：既然每个人的世界不同，世界就没有共同的部分了。不对。人类有大体相同的身体，生活在受同样物理条件约束的地方，也在相互交往中把差异排进共同的节奏里，所以各人的世界有大量的重叠——第10章专门讨论这种重叠怎样形成「相对同一性」。第二种误读是：既然差异辨认是后天形成的，人生下来就是一张白纸。也不对。婴儿的身体、感官、神经系统本身就是漫长演化历史的产物，它们决定了哪些差异最容易被纠缠起来。说世界是发生的，不等于说起点是零；起点本身也是更早发生的结果。

◆ 六、边界：这一章能说明什么、不能说明什么

这一章说了三件事，要分清它们各是哪种话。

「主体与世界共同发生」是本体论立场。它不能被任何一项婴儿实验直接证明，也不会被一项实验直接推翻。它的作用是给出一种看问题的方式：遇到「一个

人学会了什么」，先问「这个人的世界里多了什么对象、多了什么差异、少了什么差异」，而不是只问「他的脑子里多存了哪条信息」。

「前语言经验、身体活动、社会交往三股线共同纠缠出对象」是机制假说。它可以被检验，例如：只给声音、不给抱与喂养，婴儿是否同样快地建立对某个人的偏好？只给面孔照片，效果又如何？这样的研究在伦理上必须非常谨慎，很多只能用自然发生的情境来观察。我在这里不预设结论。

婴儿偏好母亲声音、知觉窄化之类，是经验事实，出处在上文已交代。这些事实与上面的机制假说是相容的，但相容不等于证明。

还有一个容易被忽略的边界：本章说的是「一个人的世界如何开始建立」，而不是说婴儿已经有了成人那样的理念世界和自我世界。婴儿的「妈妈」还不是一个概念，更不是一个可以被推理的符号；婴儿的「我」也还不是一个有记忆、有承诺的自我。从这里出发，世界才分化成三个方向：概念与推理的方向（理念世界，第7章），边界与行动的方向（现实世界，第8章），记忆与价值的方向（自我世界，第9章）。

最后提一句本书的主线。如果连「妈妈」这样一个对象，都要靠声音、气味、触碰、喂养的节奏和照料者的回应一起纠缠出来，那么一个只从某一类记录里学习的机器，它所形成的「对象」与人的对象有何不同，就成了值得认真追问的问题。这里不急于下结论，只先记下：对象的发生需要多股线的共同参与和行动的介入。第四编、第五编会带着这个问题去看具体的技术系统。

> **带走的话**
>
> 人不是先有一个现成的世界再去认识它；婴儿认出妈妈的那一刻，「妈妈」和「被照顾的我」一起发生。不同的互动历史，缠出不同的差异辨认，也就缠出不同的世界。

小账 论证落在哪一层？——本章以本体论立场为主（主体与世界共同发生），配以机制假说（前语言经验、身体活动、社会交往三股线）与经验事实（婴儿对母亲声音、面孔、气味的偏好，知觉窄化，均为公开研究）。 世界模型补了哪一条？——补一条条件：对象的发生需要多股线（多种感觉、身体动作、他人回应）的共同纠缠与行动介入，而不是在单一记录中被一次性识别。

07

第 7 章

理念世界的发生

> **这一章的问题**：「二分之一」这样看不见、摸不着的东西，是怎样在一个孩子心里站稳，变成可以拿来推算的东西的?

◆ 一、命题：理念不是挂在天上的，是训练出来的稳定显露

人除了生活在吃饭、走路、被雨淋的现实里，还生活在另一个世界里：数、图形、概念、规则、推理。三角形内角和是一百八十度，不管谁画、画在哪里；三加五等于八，昨天是，明天也是。这些东西的稳定，使许多哲学家相信，它们是先在的、永恒的，人只是在某个时刻「发现」了它们。这是发现学最强的一块阵地。

本章的本体论立场是：理念世界同样是发生的。符号、概念、逻辑与推演，是人在长期训练中把大量具体经验缠成的、高度稳定的显露。理念之所以显得永恒，是因为它的纠缠极为牢固、极为广泛，在许多人、许多场合、许多代际之间被反复验证和传承，以致它的发生过程被遮住了，只剩下一个似乎从来如此的结果。用本书的话说：理念存在是显露的稳定态，抽象是特征纠缠的一种组织方式。

这一立场不否认数学的客观性，也不说「三加五等于八」是可以随意改的。它说的是另一件事：「三加五等于八」在一个具体的人那里能够被理解、被调用、被用来推出新东西，需要一段可以观察的发生过程；在人类历史中，它从计数石子、刻痕记账走到抽象算术，也有一段发生过程。理念的稳定与理念的发生，并不矛盾。

◆ 二、机制：从「一半」到「二分之一」的三层组织

为了让讨论落地，我把理念的发生拆成三层组织，属于机制假说。

第一层，具体操作的纠缠。孩子最早遇到的「一半」，不是一个符号，而是一些动作和结果：把一个苹果切开，两边差不多大；两个人分一包饼干，你一块我一块；妈妈说「喝一半留一半」。这时候「一半」和苹果、饼干、水杯、「公平」「不许多拿」缠在一起，离不开具体的东西。

第二层，跨情境的抽取。孩子发现，苹果的「一半」、饼干的「一半」、一段路的「一半」，有某种相同的地方——不管分的是什么，都是「分成同样大的两份，取其中一份」。这一步是关键：原来缠在苹果、饼干上的纠缠，被重新组织，其中「两份同样大」「取一份」这几个特征缠得更紧，而「红的」「甜的」「能吃」这些特征松开了。抽象，在发生学里就是这样一种纠缠的重组：某些特征在很多情境中反复共同出现，被拉到一起；另一些特征只属于个别情境，被放开。

第三层，符号与规则的接管。一旦这团纠缠用「1/2」这个符号固定下来，并被放进一套运算规则里，它就开始脱离具体的东西独立运行。孩子可以算 1/2 加 1/4，而不必真去切一个苹果。这时，符号本身成了差异运行的场所：差异不再是「这块大、那块小」，而是「分母相同还是不同」「要不要通分」。理念世界有了它自己的差异序列。

三层之间还有一个方向上的提醒。很多人以为抽象就是「去掉细节」，越抽象越空。从纠缠的角度看恰恰相反：一个站稳了的抽象概念，背后缠着的具体情境比任何一个具体概念都多。「二分之一」之所以能在分饼、量布、算时间之间自由来往，正因为它把这些情境都缠了进去，只是在显露时不再把它们一一摆出来。抽象是一种压缩，而不是一种丢弃；被压缩的东西随时可能被调回来，用于检验或者纠错。一个概念如果真的什么具体情境都没缠住，它就只是一个会念的名字。

这三层不是一次走完的，而是反复往返。一个孩子在第三层卡住时，常常要回到第一层去重新纠缠，再走上来。

◆ 三、案例：小舟学分数

本编的第二个主例，是学分数。下面是我构想的场景，人物用本书的陪读人物。

小舟上四年级，分数刚学了一个月。他会读写「三分之一」「四分之三」，会做同分母的加减。可有一天，周敏随口问他：「三分之一和四分之一，哪个大？」小舟不假思索：「四分之一大，因为四比三大。」

这是一个非常常见的错误，数学教育研究者给它起过一个名字，叫「整数偏差」：孩子把整数里「数越大越大」的经验，原样搬到了分数上。从发生学看，这个错误恰好暴露了小舟理念世界里此刻的纠缠状态：「分数」这个符号已经和「整数比大小」的规则缠在一起了，而和「把一样东西分成几份」那一层具体操作，缠得还不牢。符号层自己在运行，却运行在一套借来的整数规则上。

周敏是物理老师，没有急着讲规则。她拿出两张一样大的饼，让小舟把一张切成三份、一张切成四份，问：「你想吃哪一块？」小舟拿起三分之一那块，愣了一下，笑了：「分得越多，每块越小。」

这一刻发生了什么？用三原理来说：已经站稳的显露（S，「四比三大，所以四分之一大」）和手上正在运行的差异（D，眼前这块明明更小）顶在了一起。张力积累起来，推动了纠缠（E）的改组：「分母」这个符号，原来缠在「数的大小」那一团里，现在被拉过去，和「切的份数」「每份的大小」缠到一起。这次重组回写到显露上：小舟此后看到分数，显露出来的不再只是两个整数，而是一个「被切开的整体」。

但事情没有那么快结束。过了一周，小舟做「八分之三和五分之二谁大」时又卡住了。饼可以切成三份、四份，切成八份、五份就画不准了。这时他需要另一次重组：学会通分，把两个分数换到同一种切法上比较。通分是一个纯符号层的操作，可小舟能理解它，是因为他已经有了「切法不同就没法直接比」这个在具体操作中纠缠出来的东西。符号规则接管了具体操作，却始终靠具体操作撑着。

这个案例里还有一个容易忽略的细节：小舟背得出「异分母分数比较要先通分」这条规则，并不等于他理解了。他在第一次犯错前，或许已经能把这句话复述出来。规则被复述，和规则被纠缠进可以调用的能力，是两件事。这一点到第三编讲语言模型、第八编讲小舟用 AI 学写作文时还会用到：流畅地说出一条道理，不证明这条道理已经在说话者那里稳定发生。

◆ 四、推演：理念世界自己的差异运行

概念站稳以后，理念世界还要学会「走路」，这就是推演。推演的训练，比概念的训练更不容易看见，却更能说明理念世界为什么是一个「世界」，而不只是一堆名词。

还是用小舟。分数学到后半段，课本上出现了这样的题：「一根绳子，第一次用去全长的二分之一，第二次用去剩下的二分之一，还剩全长的几分之几？」小舟第一遍写的是「二分之一减二分之一，等于零」。他把两个「二分之一」当成了同一个东西。问题不在计算，而在推演：第二个「二分之一」的「整体」已经变了，是「剩下的」，不是「全长」。

这里可以清楚地看到理念世界里的差异是怎样运行的。在具体世界里，绳子剪一刀就短一截，差异是直接看得见的；在理念世界里，差异变成了「这个分数是相对于哪个整体」这样的关系差异。推演，就是让这种关系差异按照一定的约束一步一步传下去：整体变了，下一步的「一半」就跟着变。会推演的孩子，脑子里并不是多存了一条「剩下的一半要另算」的规则，而是「分数总是某个整体的分数」这一点，已经和「整体会在过程中改变」缠到了一起。缠到了一起，以后遇到「打八折后再打九折」「第一天读了全书的三分之一，第二天读了余下的一半」，他不需要分别学，就能自己走通。

所以我把推演理解为：理念世界中已经稳定的显露（概念、符号）在纠缠的约束下，产生新的差异序列（D=G(S,E)），而这些新的差异又可以显露为新的结论（S=F(D,E)）。推演之所以能产生「新东西」，是因为一组概念缠在一起以后，它们的组合比单个概念能走的路多得多。一个只会单独使用「分数」「整体」「剩

下」三个词的孩子，和一个已经把三者缠成一团的孩子，面对同一道题，能推出来的东西完全不同。

推演的训练有一个特点：它很依赖错误。小舟写出「等于零」，周敏让他真拿一根毛线剪一剪。剪完第一刀，毛线还剩一半；再剪剩下的一半，手里明明还有一截。「零」和手上那一截顶在一起，逼他回到题目里重新找「整体」。几次之后，他读题时会自己停下来问：「这个几分之几，是谁的几分之几？」这一问，就是推演能力在他身上站稳的标志——不是答对一道题，而是在新题面前自己生出那个关键的差异。

从混沌到秩序的三态，在这里也看得见。刚学分数应用题时，小舟面对文字题，是一团说不清的困惑，哪个数和哪个数有关系都分不出来（混沌态）。经过一段错了又改、改了又错的时期，他能做对一部分题，可题目稍一变化就又乱（介生态）。再往后，大部分同类题他一读就知道怎么列式，错误集中到个别真正新颖的题型上（秩序态）。我要强调，这里的「三态」只是概念描述，用来帮助读者看出一个过程的阶段；要把它变成可测的指标，比如错误率的变化曲线、对变式题的迁移率，需要专门的研究设计，第一编第4章已经提醒过概念解释与可测动力学指标的区分，这里不再重复。

◆ 五、理念存在怎样稳定化

一个概念在个人那里站稳，还不足以成为理念世界里的「存在」。理念存在之所以稳定，还靠另外几个条件，我分开说。

一是跨情境的反复调用。「二分之一」在分饼、量布、算时间、看折扣时都能用，而且每次都不出错，它的纠缠就在这些调用中一次次被加固。一个只在一个情境里有用的「概念」，很容易随着那个情境的消失而松散。

二是与其他理念的连接。分数与除法、小数、比例、概率连在一起，彼此支撑。一个理念被越多其他理念依赖，它越难被松动，因为松动它就要牵动一大片。数学之所以显得格外坚固，部分原因就在这里：它的纠缠网极为致密。

三是社会与代际的传承。小舟学的分数，不是他自己从零发明的。教材、老师、作业、考试，把人类在漫长历史中发生出来的分数体系，组织成一套训练路径交给他。理念存在的稳定，很大一部分是集体的成果。这一点与本书第三编讨论的「集体语言世界」直接相通：大模型从人类文本中训练出的东西，正是建立在这种集体稳定化的历史之上。

四是可检验性。理念世界并非与现实世界隔绝。分数算得对不对，最后可以回到分饼、量布上检验。一个理念体系如果长期与现实检验相冲突，它的稳定就会受到威胁。

以上四条中，前三条是机制假说，第四条兼有本体论意味：理念世界的稳定，不只是内部一致，也要经得起与另外两个世界的往返。这一点第10章还要展开。

◆ 六、边界：概念冲突与理论重建

理念世界的发生不是一路向前的累积。概念会冲突，理论会重建。

在个人层面，小舟把整数比大小的规则搬到分数上，就是一次概念冲突。这种冲突不是坏事，它是理念重组的动力来源。一个从未犯过这类错误、只是把规则背下来的孩子，其理念世界反而可能更脆弱：遇到规则没有覆盖的情境，他没有可以退回去的具体纠缠。

在人类历史层面，概念冲突推动理论重建的例子很多。数系从自然数扩展到分数、负数、无理数，每一步都伴随着旧规则与新对象之间的冲突，以及概念的重新组织。历史细节不在本书范围，这里只取一个方向：理念世界的稳定是阶段性的，它的秩序态可以在新的冲突面前重新进入介生态，经过一番改组再站稳。

个人层面还有一种更安静的重建，不靠犯错，而靠「同一个东西换了样子」。小舟后来学到二分之一、四分之二、零点五、百分之五十其实是同一个数，起初很不服气：「写法都不一样，怎么是一样的？」这其实是理念世界里的对象同一性问题。在他原先的纠缠里，一个数就是它的写法；现在他必须让几种写法缠到同一团里，让「同一个量」比「同一种写法」更深地站稳。这一步完成以后，他的理念世

界在结构上变了：对象不再由符号本身决定，而由一组可以相互转换的符号共同撑着。有意思的是，这和第6章里「戴口罩的妈妈仍然是妈妈」有同样的形状——同一性不靠某一个特征，而靠一团纠缠在变化中保持。理念对象与现实对象的同一性，发生方式有相通之处，第五编讨论「物的意识」时会看到它在物上的样子。

还要划一条边界。本章说理念是「发生的」，不等于说一切理念同样可靠，也不等于说可以凭个人喜好改写逻辑。一个理念的纠缠越广、越深、越经得起检验，它就越可靠；发生学给出的是理解可靠性来源的方式，而不是取消可靠性。另外，本章只讨论了数学这一类较为整齐的理念；道德观念、审美判断、政治理念的发生，牵涉自我世界与社会关系更多，它们的稳定条件更复杂，本编不展开。

最后，一个提醒指向后面的章节。语言模型在符号层面表现出色：它能写出完整的分数运算步骤，也能说出通分的道理。按本章的分析，要问的是：它的符号运行背后，缠着的是什么？是人类文本中沉淀下来的、他人具体操作的痕迹，还是它自己与东西打交道的历史？这个问题本章不回答，只把它交给第三编，特别是第12章「语言token与特征纠缠的建立」与第14章「语言生成与理解能力的发生」。

> **带走的话**
>
> 理念不是先挂在天上等人去摘，而是许多具体经验经过抽取、重组和符号接管后，形成的高度稳定的显露；它稳在跨情境调用、相互连接、集体传承和可检验性上，也会在冲突中重建。

> **小账** 论证落在哪一层？——本章以本体论立场（理念存在是显露的稳定态，抽象是纠缠的组织）和机制假说（三层组织、四个稳定条件）为主，经验事实只在分数「整数偏差」一处出现。 世界模型补了哪一条？——补一条边界：符号层的流畅运行不等于背后的纠缠已在具体操作中稳定；理念的可靠性要靠与现实的往返检验。

08

第 8 章

现实世界的发生

> **这一章的问题**：一个孩子被烫了一下以后，他的世界里多了什么？

◆ 一、命题：现实是在身体的行动与阻力中发生的

说到「现实世界」，人们一般想到的是外面那个有桌子、有马路、有天气的世界，它不管人怎么想都在那里。本章不否认这一点。本章的本体论立场是：一个人所生活的现实世界——哪里是边界、哪里可以走、什么东西可以抓、什么东西碰不得——是在他的身体与周围的反复互动中发生的；而「现实」之所以区别于想象，关键在于它会给出阻力：你以为可以，它不让；你没想到，它发生了。

这里有两层意思要分开。第一层：外界的物理条件不由人决定，水烧开了就是烫的，这一点发生学完全承认。第二层：在一个具体的人的世界里，「烫」是一个什么样的东西、它和哪些别的东西缠在一起、它让人在什么时候停手，这是要发生的。一个从没碰过热水的孩子，水壶上的白汽在他的世界里只是一种好看的东西；被烫过一次以后，白汽、壶嘴的声音、壶身的颜色、母亲一声「别碰」，都和「烫」缠在了一起，水壶成了一个「有危险的地方」。物理条件没有变，他的现实世界变了。

所以本章说现实世界的发生，说的是后一层：可行动的现实，是在身体里发生出来的。

◆ 二、机制：边界、空间、运动与可行动性

现实世界的发生，我分四个方面来说，都属于机制假说。

边界。一个物从周围分出来，成为「一个东西」，需要边界的发生。看上去颜色变了的地方、摸上去从软到硬的地方、敲起来声音不一样的地方，都可能成为边界的线索。这些线索在不同感觉里出现，彼此印证，边界才站稳。

空间。远近、上下、里外，最初不是一套坐标，而是身体可以到达的范围：伸手够得着的、爬几步够得着的、怎么也够不着的。孩子的空间，是从手臂的长度和爬行的速度开始量的。

运动。东西会动，人也会动。一个滚动的球、一只跑过的猫，它们的位置在变，可它们仍然是「那个东西」。运动中保持同一，比静止时保持同一要难，要求孩子把前一刻和后一刻缠在一起。

可行动性。这是最关键的一点。对一个人来说，现实中的东西首先不是「有什么属性」，而是「能拿它怎么样」：能抓、能推、能坐、能吃、会烫、会扎手。知觉心理学家吉布森提出过一个相关的说法，英文叫 affordance，中文常译作「可供性」，也有人译「示能性」，他在 1979 年的《视知觉的生态学进路》一书里对它作了系统阐述。本书不借它的理论体系，只取一个共同的方向：现实对象的显露，从一开始就和行动可能性缠在一起。

这四个方面的共同点是：它们都离不开身体。边界要靠触摸来确认，空间要靠移动来量出，运动要靠追随来把握，可行动性要靠真的去做来检验。现实世界是以身体为中心发生的。

◆ 三、案例：小舟学「热」

本编的第三个主例，是小舟被烫了一下。下面是我构想的场景。

小舟三岁那年冬天，周敏在厨房煮面，锅边放着一只刚倒了开水的搪瓷杯。小舟踮脚去够，指尖碰到杯壁，「哇」一声缩回手。周敏赶紧拉他去冲凉水，一边冲一边说：「烫！这个烫！」好在只是指尖红了一点，没有起泡。

在这之前，周敏说过很多次「烫，别碰」，小舟也会跟着说「烫」。可「烫」在他那里只是一个声音，最多和妈妈紧张的脸缠在一起。那一碰之后，事情变了。

先看纠缠（E）。那一刻缠到一起的东西很多：手指上那一阵尖锐的刺痛，缩手的动作，杯子冒出的白汽，搪瓷杯的样子，妈妈那一声又急又高的「烫」，接下来凉水冲在手上的感觉，以及周敏抱着他时的安慰。这些东西来自不同的感觉，有看到的、听到的、皮肤感到的，在不到一分钟里一起发生，并且被那一阵疼紧紧地钉在了一起。

再看显露（S）。从此，在小舟的世界里，「烫」不再是一个声音，而是一个有边界、有位置、有后果的东西。看到冒白汽的杯子，他会把手缩回去；听到妈妈说「烫」，他会真的停下来。「烫」显露为一种现实中的危险。

再看差异运行（D）。被烫之后，小舟开始注意以前不注意的差异：杯子有没有冒汽，锅是不是刚从灶上拿下来，妈妈端碗时是不是用了两只手。这些差异以前在他眼前一样发生，却不构成什么；现在它们都和「会不会烫」相连，成了他的现实世界里值得辨认的差异。显露改写了差异运行。

这次经验还在继续分化。过了几天，他小心翼翼地用手背碰了碰放凉的杯子，发现不烫了，又碰了碰还冒汽的，赶紧缩回来。他在试探「烫」的边界：什么时候烫，什么时候不烫。到了五六岁，他知道洗澡水要先用手试一试，知道太阳晒过的滑梯夏天中午会烫屁股，冬天的铁栏杆会「冰得像烫」。「烫」这个现实对象，从一次刺痛长成了一组带边界、带条件的可行动知识。

这里有两点需要说清。第一，小舟学会「烫」，不是视觉单独完成的，也不是触觉单独完成的，更不是听到妈妈的话就完成的。看、听、摸各自都有自己完整的一套（每一种都有它自己的显露、差异运行与纠缠），它们在同一件事中相遇，才把「烫」缠得这样牢。不能说视觉提供了显露、触觉提供了纠缠、听觉提供了差异——这样的一一对应是本书明确不采用的。三种感觉怎样各自成为完整的 SDE，又怎样共同参与对象的发生，是第五编的工作；本章只指出，在一个具体的人身上，现实的发生从一开始就是多种感觉共同参与的。

第二，安全感也缠了进去。凉水与拥抱让这次经历没有变成长久的惊恐，小舟学到的是「烫要躲」，而不是「厨房很可怕」。如果那天没有人安慰，同样的刺痛

也许会缠出另一种现实：对厨房、对锅、对一切冒汽的东西都过度回避。现实世界的发生，并不是纯粹物理的，它从一开始就和自我世界、和他人的回应缠在一起——这一点第9章与第10章还要展开。这里只作日常经验的描述，不涉及任何医学判断。

◆ 四、空间与运动：从够不着到接住球

「烫」说的是边界与可行动性。现实世界里还有另外两样东西，空间与运动，它们的发生同样要靠身体，下面用一个更轻松的例子补上。这也是我构想的场景。

小舟四五岁时，周敏在小区空地上和他抛皮球。一开始，球朝他飞来，他两只手伸得笔直，眼睛盯着球，等球砸到胸口才合手，十次有八次接不住。他看得见球，也知道球在飞，可他的身体还没有把「球现在在哪里」和「球过一会儿会在哪里」缠起来。

几个月后，事情变了。球一出手，他会往前迈半步，手提前张开，球到之前手已经在等着了。有时球抛偏了，他会侧身去追。这里发生的，是空间与运动的现实化：球的飞行轨迹，不再是一段「看到的画面」，而成了一段身体可以介入的过程。远与近，不再按米计算，而按「来得及」与「来不及」计算；快与慢，也不是速度的读数，而是「要不要跑」。

这个变化有几点值得注意。第一，它离不开失败。接不住的那些球，砸在胸口、滚到草里，每一次都是现实的阻力，让他的手和眼一起改组。第二，它不只是看的能力提高了。他看球的本事也许几个月前就差不多了，变的是看、移动和伸手三者缠在一起的方式。第三，换了东西，能力并不完全迁移。小舟能接住皮球之后，第一次接一只飘来飘去的气球，反而手忙脚乱：气球太轻、太慢、轨迹太飘，他以为它会落到前面，它偏偏停在半空。对他来说，「会接」并不是一个对一切飞来之物都成立的通用本领，而是一组在皮球上训练出来、带着皮球的重量与弹性的纠缠；换成气球，原有的显露失效，需要重新试探。

这个例子提醒我们，一个人的现实世界不是一张均匀的地图，而是一块一块长出来的：他在哪些东西、哪些动作上有过足够的往来，那里的现实就清楚、可预

期；没有往来过的地方，现实就模糊、就会出乎意料。人对现实的把握，天然是分区域、带边界的。这一点到第28章讲「任务如何组织世界的发生范围」时，会成为讨论机器世界模型边界的一条重要线索——这里只是先在一个孩子身上把它看清。

◆ 五、身体与工具改变现实的发生条件

身体是现实发生的中心，但身体并不是固定的。身体长大了，工具拿到手里，现实世界也跟着变。

小舟学会用筷子以后，滚烫的饺子不再是碰不得的东西，夹起来吹一吹就能吃。筷子把他与「烫」之间的关系改变了：烫还是烫，可它不再直接落在指尖上，而落在筷子尖上，变成了一种可以处理的阻力。同样，隔热手套让周敏可以直接端起烤箱里的盘子，锅铲让人可以翻动热油里的菜。工具延长了身体，也重新划定了「可以」与「不可以」的边界。

老彭开了三十年车，路面上的一个小坑、方向盘上的一点抖动，对他来说是可以读懂的东西；对一个刚拿到驾照的人，同一条路就是另一个现实。车作为工具，不仅延伸了腿，也延伸了感觉：方向盘传到手上的振动、轮胎压过路面的声音，成了老彭现实世界的一部分。（驾驶在第29章有专门的展开，这里只取它作工具改变现实的一个例子。）

身体自己的变化也一样。小舟两岁时，门把手是一个够不着的东西，门对他来说只是「妈妈才能打开的墙」；四岁时他踮脚够得着了，门成了「可以自己出去的地方」，周敏不得不在门上加了一道高处的插销。门没有变，变的是孩子的身高与手的力气，于是门在他现实世界里的显露也变了：从一块阻挡，变成一个出口。成年人很少注意到这一点，因为成年人的身体在很长时间里变化不大，现实世界看上去就像是固定的。可只要身体发生明显改变——扭伤了脚踝拄着拐、抱着一个婴儿上楼、上了年纪腿脚不便——人就会重新发现：楼梯、门槛、地面上的一点湿滑，在现实世界里的分量完全不同了。这类经历提醒我们，所谓「现实本来的样子」，总是相对于某一种身体而言的那一种样子。

这里能得出一个对后文很重要的判断：现实世界的发生条件，包括身体的尺度、能力和手里的工具。同一个物理环境，对不同身体、不同工具，发生出的现实并不相同。一个机器的「身体」——它的传感器、执行器、它能施加和承受的力——同样决定了它可能发生出什么样的现实。这个判断属于机制假说，它提示我们：谈机器的世界模型，必须同时谈它的身体与工具，而不能只谈它看过多少画面。

◆ 六、现实阻力及其检验作用

现实世界最根本的特征，是它会给出阻力。

想象可以随意展开：可以想一只会飞的猪，可以想一杯永远不凉的茶。可现实不让：杯子会凉，手会被烫，球扔出去会落下来。现实阻力的作用，就是检验——检验一个人的期待、概念和动作是否与周围的条件相合。小舟以为那只杯子可以拿，现实用一阵刺痛告诉他：不行。

从发生学看，现实阻力是三原理最常见的来源之一。人已经站稳的显露（以为可以拿）与当下运行的差异（手指感到的刺痛）发生矛盾，张力推动纠缠的改组（杯子与冒汽、与烫缠到一起），然后回写到显露上（下次不碰）。这种「被现实顶回来」的经验，是现实世界区别于理念世界的地方：理念世界里的冲突可以在纸上解决，现实世界的冲突必须在身体上解决。

这里要说清楚边界。

第一，阻力不是唯一的老师。小舟并不需要被每一种危险都伤一次才学会躲开。语言、示范、他人的经验，可以让人在没有直接阻力的情况下学会许多东西。周敏说「电插座不能摸」，小舟没有被电过也懂得不摸。可值得注意的是，这类不经直接阻力的学习，往往要靠早先在别处经历过的阻力撑着：他之所以真的相信「电会伤人」，部分是因为他已经知道「烫会伤人」是什么意思。语言转达的现实，挂在亲身阻力缠出来的现实上。

第二，阻力不总是清楚的。有时候失败的原因不明：一件事没做成，是方法不对、力气不够，还是运气不好？现实给出的阻力需要被归因，归因本身也可能出错。小舟接不住气球，起初认定是「气球坏了」；他要再试几次，看到气球每次都慢悠悠地飘，才把原因从「东西出了毛病」挪到「这东西本来就这样动」。同一次阻力，归到不同的地方，改组的是不同的纠缠，学到的也是不同的现实。

阻力还有一个特点：它往往是被行动引出来的。小舟如果只站着看别人接球，气球也就那样飘着，不会有任何东西顶他一下；他一伸手，现实才开始回应。被动地看，得到的是显露；伸手去做，得到的是显露与阻力的对照。

第三，这一章说的「现实世界」是一个人的现实世界。它与物理学描述的那个世界不是同一个东西。物理学的世界是集体经过长期理念训练与实验检验发生出来的理念世界的一部分，它反过来又可以改写个人的现实世界——学过热传导的周敏，看搪瓷杯的眼光就和小舟不同。这种改写，正是第10章要讲的三界相互改写。

最后说一句与后文的关系。世界模型的研究，常把「预测下一帧画面」当作理解现实的途径。本章的分析提示了一个需要认真对待的问题：画面中的杯子冒着白汽，可「烫」不在画面里，它在手指上，在缩手的动作里，在人对后果的承担里。仅从画面学习的系统能否形成「烫」这样的现实对象，或者形成到什么程度，是一个需要具体实验来回答的问题，不能由本章的哲学分析直接下结论。第四编会看具体技术路线怎样处理这个问题，第五编会看三种感觉怎样共同参与对象的发生。

> **带走的话**
>
> 现实世界在身体的行动与阻力中发生；小舟被烫那一下，「烫」从一个词变成一个有边界、有后果、改变他行动的现实对象，而且是看、听、摸共同缠出来的。

> **小账** 论证落在哪一层？——本章以本体论立场（现实在身体与阻力中发生）与机制假说（边界、空间、运动、可行动性；身体与工具改变发生条件）为主，经验事实很少，只在「可供性」概念出处一处。 世界模型补了哪一条？——补两条条件：现实对象的发生离不开行动与阻力的检验；系统的「身体」与工具本身就是现实发生条件的一部分。

09

第 9 章

自我世界的发生

> **这一章的问题**：一个人说「我是一个老师」的时候，这个「我」是早就在那里的，还是一年一年教出来的？

◆ 一、命题：自我是一团纠缠，在行动历史中显露

前两章讲的理念世界与现实世界，似乎都在「我」的外面：一个是概念与推理，一个是物与行动。可人还生活在第三个世界里：我记得什么，我在乎什么，我和谁亲近，我答应过什么，我能做什么，我是个什么样的人。这就是自我世界。

常见的看法把「自我」当成一个核心：一个藏在身体里、从出生起就在、贯穿一生不变的「真正的我」，经历只是在它外面添加的东西。本章的本体论立场与之不同：自我是记忆、情感、关系、价值、承诺与能力缠在一起形成的一团纠缠（特征纠缠，E），在一个人的行动历史中显露出来（显露，S），又随着新的经历被改写。「我」的同一性是真实的——周敏知道二十年前那个站在讲台上手心冒汗的年轻人就是她自己——但这种同一性不是一个不变的核，而是一团纠缠在变化中保持住的相对稳定。

这和第6章婴儿认妈妈、第7章「二分之一与零点五是同一个数」在形状上是相通的：同一性不靠某一个永不改变的特征，而靠许多特征在变化中彼此牵连、不全部松开。差别在于，自我这团纠缠的中心正是做出认出这件事的那个人本身。

◆ 二、机制：记忆、情感、关系与价值怎样缠成一个「我」

自我世界的纠缠里有几股主要的线，我分开说，属于机制假说。

记忆。不是所有记住的事都进入自我。周敏记得长沙到深圳的高铁要几个小时，这件事和「她是谁」关系不大；她记得自己第一次被学生问倒的那堂课，这件事却在她的自我里占着位置。进入自我世界的记忆，往往是那些和情感、关系、价值缠在一起的记忆。

情感。高兴、羞愧、骄傲、委屈，是把经历钉进自我的钉子。一件事让人强烈地感到了什么，它就更可能和「我」缠到一起。情感也给自我世界定了方向：一个人反复在某类事情上感到满足，他会倾向于把自己理解为「做这类事的人」。

关系。「我」从一开始就是在关系里发生的（第6章已经看到，婴儿最早的「我」是「被妈妈照顾的我」）。成年人的自我同样由关系撑着：谁的妈妈、谁的老师、谁的同事、谁的朋友。关系一变，自我常常跟着松动。

价值。一个人认为什么重要、什么该做、什么不能做，是自我世界的骨架。价值不是事先写好的一张清单，它在一次次选择里显露：当两件事冲突时，你选了哪一件，你就在那一刻显出了自己看重什么。

承诺与能力。这两条放在一起说。承诺是「我答应过、我要做到」，能力是「我做得到」。两者的往来构成了行动历史：一个人承诺了，去做了，做成了或没做成，这段历史回写到自我上，改变了他下一次敢承诺什么。

这几股线缠在一起时，并不平均用力。有时情感压过价值，有时关系推着承诺，有时能力不足让一个承诺悬在半空。正是这些线之间的张力，让自我世界不是静止的，而是一直在发生。

◆ 三、案例：周敏的教师自我

本编的第四个主例，是周敏的教师自我。下面整段都是我构想的场景，用本书的陪读人物来讲，不涉及任何真实人物的经历。

周敏二十出头师范毕业，分到长沙一所中学教物理。第一年，她白天上课，晚上备课到深夜，每一节课都写满了教案，连要说的笑话都写在纸上。可上课时，学生一个问题偏离教案，她就慌，脸发烫，只好说「这个我们下次再讲」。那时候她

心里的「我」，说得明白点，是「一个还不像老师的人，在努力扮演老师」。「老师」这个身份是一个外来的名称，还没有和她的记忆、情感、能力缠在一起。

转折发生在第二年。有一次讲浮力，一个平时不爱说话的男生举手问：「老师，那为什么铁做的船能浮起来，铁块就沉下去？」这个问题不在教案上。她站在讲台上愣了几秒，然后没有像以前一样推到「下次」，而是说：「好问题，我们一起来想。」她让学生拿出课桌里的铝箔糖纸，揉成一团丢进水杯，沉了；摊平折成小船，浮了。全班哄笑，那个男生也笑了。下课后，他跑来说：「老师，我第一次觉得物理好玩。」

这件事在周敏的自我世界里留下了很深的痕迹。用三方程来看：一段新的行动历史（她当场带着学生试出来）、一份强烈的情感（从慌张到被学生的笑容接住的那种松快）、一段新的关系（那个男生开始常来问问题），这些东西缠到了一起（E），改变了她的自我显露（S）——她开始觉得「我能接住学生的问题」，「我是一个可以和学生一起想的老师」。这个新的自我显露又改写了差异运行（D）：此后上课，学生的偏题提问不再是「打乱教案的麻烦」，而成了「值得接住的机会」。她开始注意以前忽略的东西：哪个学生眼睛亮了，哪个问题背后藏着真正的困惑。

十几年过去，「我是一个物理老师」在周敏那里早已不是一个外来名称，而是她自我世界中最粗的一股线之一。她做饭时看锅里水的对流，会想到下周的课；和小舟玩接球，会忍不住讲一句抛物线；邻居老彭说车子过弯时觉得「往外甩」，她也会顺口解释几句惯性。教师自我已经从课堂溢出，和她看世界的方式缠在一起了。

但这个自我并不是一路坚固下去的。到了第十几个年头，学校推行一套新的教学评价办法，她过去那种「和学生一起想」的上课方式，在一套以标准化指标为主的评价里显得「效率不高」。有一段时间，她很困惑：自己坚持了十几年的东西，到底对不对？「我是个好老师」这个显露，和一个新的外部评价发生了冲突。她没有放弃，也没有简单地顺从，而是在几年里慢慢找到一种办法，把「一起想」放进更紧凑的课堂结构里。经过这一番，她的教师自我又改组了一次：从「能接住学生

问题的老师」，变成「知道在什么条件下怎样接住学生问题的老师」。（这里只描述一个人的心路，不对任何教育评价制度作政策表态。）

◆ 四、我之同一性与行动历史

从周敏的例子，可以把「我之同一性」的发生分成几个环节来看，这是本体论立场加机制假说。

形成。一个身份或一种自我理解，起初只是一个名称或一个期待（「我是老师」），要靠行动历史一次次把它和记忆、情感、能力缠起来，它才真正成为「我」的一部分。名称在前、纠缠在后，这在年轻人入职、新手父母、刚搬到一个新城市的人身上都看得到。

维持。自我一旦形成，会在日常的行动里被不断加固：每一节课、每一次被学生找来问问题，都在维持「我是老师」这团纠缠。维持并不是什么都不变，而是在小的变化中保住大的牵连。

失效。外部条件剧烈变化（评价方式改了、关系断了、身体受伤不能再做原来的事）时，原有的自我显露可能突然站不住：「我还是那个我吗？」这时候人会经历一段混乱，自我世界退回到介生态。

重建。重建不是回到原来，而是在新条件下把旧的纠缠重新组织，有些线保留，有些线松开，有些新线缠进来。周敏的第二次改组就是这样：「一起想」这股线保留了，「不需要考虑效率」这股线松开了，「课堂结构」这股新线缠了进来。

这四个环节和第6章婴儿认妈妈时的「形成、维持、失效、重建」有同样的形状，也和第25章将讨论的物的对象同一性形状相通。我在这里只指出形状相通，并不主张自我同一性与物的对象同一性是同一种东西：自我世界的特殊之处在于，做出判断、承担后果的正是这团纠缠本身。

同一性的四个环节之所以能被看见，靠的是承诺、能力和行动历史。上面把承诺与能力一笔带过，这里专门展开，因为它们是自我世界里最「硬」的部分：记忆

可以模糊，情感可以平复，可一个人答应过的事做没做到、一件事会不会做，是摆在那里的。

先说承诺。承诺是一种很特别的东西：它是用现在的「我」去约束将来的「我」。周敏在开学第一天对一个物理一直不及格、实验课上从不敢动手的男生说：「这学期我每周四放学后留半小时给你。」说这句话的时候，将来的那十几个周四还没有发生，可她已经把它们缠进了自己的自我里。此后每一个周四，无论多累、家里多忙，这句话都在差异运行中起作用：别的老师四点半下班，她看一眼钟就知道自己还有半小时。承诺把一个将来的行动提前拉进了现在的自我纠缠，并且让每一次遵守或违背都变成了一次显著的差异。

再说能力。能力在自我世界里的作用，不只是「会做什么」，还是「我敢做什么」。第一年的周敏未必缺乏讲清浮力的知识，可她缺的是在课堂上临场接住问题的能力，所以她不敢离开教案。能力不够时，自我显露会退缩，绕开可能失败的地方；能力一旦在几次成功中站稳，自我显露就会往前推，敢于去碰原来不敢碰的事情。能力与自我之间是相互牵引的：自我理解决定一个人愿意去试什么，去试的结果又回写到自我理解上。

承诺与能力的交汇处，是行动历史。一个人承诺了，去做了，结果怎样，构成了一段可以回看的历史。那个男生在一学期的周四里，物理成绩从不及格到了中等，这件事成了周敏行动历史的一部分；它让她下一次对另一个学生许下类似承诺时，心里更有底。反过来，也有做不到的时候。有一年她承诺带一个兴趣小组参加比赛，因为那学期小舟生病住了一阵医院，她没能按计划带完，小组成绩平平。这段历史同样回写进她的自我：此后她许承诺会先掂量自己的时间，学会了说「我尽量」而不是「我保证」。

行动历史之所以重要，是因为它不能被一句话改写。一个人可以重新描述自己，却不能让已经做过或没做到的事消失。正因为如此，行动历史给自我世界提供了一种类似现实阻力的东西：你说你是什么样的人，你的历史会来检验。第8章说现实世界靠阻力检验期待，在自我世界里，这个检验者就是行动历史。

◆ 五、自我世界与语言表达的相互作用

自我世界与语言之间有一种双向的关系，值得单独说。

一方面，语言帮助自我显露。周敏在某次教研会上第一次说出「我觉得我的课好就好在敢让学生问」，说出这句话之前，这个想法只是模糊地缠在她的经验里；说出之后，它变成了一个清晰的、可以被她自己和别人引用的自我描述。语言让自我的一部分从纠缠里浮上来、站稳。

另一方面，语言也可能让自我显露得太早、太硬。一个人被别人反复说成「你就是个急性子」，他可能会慢慢把这个说法缠进自己的自我里，哪怕它原本只是对几次行为的概括。语言给自我贴的标签，有时会反过来限制差异的运行：他不再注意自己沉得住气的那些时刻，因为那些时刻和「我是急性子」对不上。

还有一层：一个人用语言说出来的自我，和他在行动中显露的自我，可以不一致。说「我很看重家庭」的人，未必在选择时把家庭放在前面。这时候，哪一个更是「真的」自我？发生学的回答是：两者都是自我世界中真实发生的东西，它们之间的不一致本身就是一股张力；这股张力有时推动人改变行动，有时推动人改变说法。说与做的长期不合，是自我世界正在发生改组的信号，而不只是虚伪。

这一点对本书后面的讨论很要紧。到了第八编讨论「个人—AI复合体」时（特别是第38章讲 AI 在学生自我世界里的参与），会遇到这样的问题：一个人与 AI 长时间对话，AI 给出关于「你是什么样的人」的描述，这些描述会不会缠进这个人的自我世界里？会以什么方式缠进去？本章不回答，只先指出：语言表达对自我世界有真实的塑造力，这种塑造既可能帮助自我显露，也可能把它固化。

◆ 六、边界：本章不说什么

最后划几条边界。

第一，本章不讨论意识的主观体验问题——「做周敏是什么感觉」这个问题，与「周敏的教师自我怎样形成、维持、改变」是两个问题。本章只处理后者，即可

以通过行动历史、自我描述与关系变化来观察的那一面。主观体验问题本书在第25章单独讨论，且不与功能性判据混在一起。

第二，本章不作任何心理学或医学诊断。自我同一性在临床上有许多专门的问题，那不是本书的范围。

第三，本章的「自我世界」说的是人的自我世界。语言模型在对话中常常使用「我」这个字，也能说出前后连贯的「自我描述」。按本章的分析，一个自我世界的发生，需要记忆、情感、关系、价值、承诺与能力在行动历史中缠在一起，并且需要有承担后果的那一方。语言模型的「我」是否、在多大程度上具备这些条件，不能从它说话流畅就推出来，需要在具体的机制层面逐项检查：暂时语境变化、外部记忆、控制反馈、参数训练、持久学习，这五件事要分开记账，第37章专门讲。

> 带走的话
>
> 自我不是一颗不变的内核，而是记忆、情感、关系、价值、承诺与能力在行动历史中缠成的一团；周敏的教师自我是一年一年教出来的，也在冲突中一次次重建。

小账 论证落在哪一层？——本章以本体论立场（自我是行动历史中显露的纠缠）与机制假说（几股线的纠缠；同一性的形成、维持、失效、重建）为主，案例为构想场景，无经验事实断言。　世界模型补了哪一条？——补一条边界：说出流畅的「我」不等于自我世界已经发生；判断一个系统有没有稳定的同一性，要看形成、维持、失效与重建四个环节，并区分语境、记忆、训练与持久学习。

10

第 10 章

三个世界如何相互改写

> **这一章的问题**：一个人学会了一个道理、碰了一次壁、变成了另一个自己——这三件事是各管各的，还是一件事的三面?

◆ 一、命题：三界不是三个房间，是一张网的三个方向

前面三章把理念世界、现实世界、自我世界分开讲，是为了说得清楚。可在一个具体的人身上，三个世界从来不是分开的房间，走完这间再走那间。它们是同一张纠缠之网的三个方向：一个概念的变化，会拉动行动与自我；一次现实的碰壁，会改动概念与自我；一次自我的改变，会带出新的概念与新的实践。

本章的本体论立场是：**理念、现实、自我三个世界在学习中共同发生、相互改写**；任何一个世界的稳定，都要靠另外两个世界的支撑与检验；任何一个世界的变化，都会在另外两个世界里留下回响。所谓「一个人学会了什么」，严格说来，总是三个世界同时发生了某种改组，只是在不同情形下，改组的重心落在不同的方向。

在本章的最后，还要把视野从一个人扩展到许多人：许多人的世界各不相同，却能形成一个可以共享的「共同世界」。这一步是第三编讨论「集体性的语言世界」的直接前提。

◆ 二、机制：三条改写的通道

三个世界之间的改写，我选三条最常见的通道来说，属于机制假说。

理念进入行动。一个概念、一条道理，不是停在脑子里，而是会改变人在现实中做什么、注意什么。学过热传导的人，冬天会把金属门把手和木头扶手的冷暖差别看成「导热快慢不同」，而不只是「铁更冷」；这个理念让他在现实中注意到原先不注意的差异，也让他做出不同的动作，比如冬天先用袖子垫一下再握铁栏杆。理念通过改变差异的运行，进入了现实世界。

现实经验重组自我。一次成功或一次失败，一次被现实顶回来的经历，常常不只改变人对物的理解，也改变人对自己的理解。第9章周敏第一次用铝箔糖纸接住学生的问题，这本是一件课堂里的小事，属于现实世界；可它回写到她的自我世界，让她第一次觉得「我能」。现实中的阻力与成功，是自我世界最有力的改写者之一。

自我变化推动新的理念与实践。一个人对自己的理解变了，他愿意去碰的问题、愿意去学的东西也会变。周敏自认「是一个和学生一起想的老师」以后，开始自己去读一些关于学生怎样形成物理概念的书，开始在课堂上设计让学生先猜、再试、再讨论的环节。自我的改变，打开了新的理念学习与新的实践。

这三条通道首尾相连，形成一个循环：理念进入行动，行动遭遇现实，现实回写自我，自我推动新的理念。用六路径的话说，三个世界之间的改写，并不固定走某一种次序；在一个人一段时间的学习里，常常可以看到几种次序交替出现。我在这里不把三个世界与 S、D、E 三维对应起来——三个世界中的每一个，内部都有自己完整的显露、差异运行与纠缠；三界之间的改写，是三张各自完整的 SDE 之网之间的互相牵动。

◆ 三、案例：小舟学分数，再看一遍

本编前面的主例分在各章里讲，这一章把其中两个放回同一个孩子身上，看三界怎样一起动。下面是我构想的场景，承接第7章。

小舟终于弄明白「三分之一比四分之一大」之后，事情并没有停在作业本上。

那个周末，周敏买回一个西式蛋糕，要切开给家里三个人和来串门的老彭。小舟抢着要切。他拿起刀，比划了一下，忽然停住：「四个人，要切成四份。可是我想吃大块的……那就只能少切几刀了。」大人都笑了。他嘴上开着玩笑，手上却真的在算：先对半切，再对半切，四块一样大。理念世界里刚站稳的「分得越多每块越小」，进入了他在现实中的动作——这是第一条通道。

切的时候出了点问题。蛋糕上面有一层厚厚的奶油和水果，第一刀切下去，刀压歪了，两边明显一大一小。小舟有点急：「不对，这不是二分之一！」老彭在旁边说：「切蛋糕刀要先用热水烫一下，再慢慢锯着切，不能往下按。」小舟照做，第二刀果然平整多了。在这次经历里，现实给了他一个阻力：纸上画得出的「二分之一」，落到一块软而不均匀的蛋糕上，并不是一刀就成的。他的理念世界里多了一条朴素的认识：数学上的「一样大」，在现实里要靠手艺和工具才能做到。这是现实对理念的回写，也是第8章所说的「工具改变现实的发生条件」的一个小例子。

这一晚还有第三层。蛋糕分完，周敏夸他「切得挺公平」，老彭也说「小舟以后当老师，分东西肯定不偏心」。小舟很得意。第二天在学校，同桌分一包零食分得不均，两人吵起来，小舟自告奋勇：「我来分，我会分数！」他把零食数了数，按人数平分，剩下一颗不好分，他提议「这颗留着明天再分」。这件事在他的自我世界里留下了一个小小的新显露：「我是一个会公平分东西的人」。这个自我显露是从理念（分数）经过现实（切蛋糕）缠进来的，它又推动了他的新实践（主动在同学之间分零食），这些实践又让分数在他那里继续被调用、被加固。

一个四年级孩子的一周里，三个世界就这样转了一圈：理念进入行动，现实回写理念，现实经验重组自我，自我推动新的实践与理念的再使用。没有哪一个世界是单独学会的。

◆ 四、改写卡住的时候：三界之间的缝隙

三界相互改写，并不总是顺畅的。很多学习上的困难，恰恰出在某两个世界之间断了线。看清这些「卡住」的样子，比看顺利的例子更能说明三界为什么必须一起发生。下面三个例子都是我构想的场景。

第一种卡住：理念进不了行动。小舟能把「过马路要看左右」背得滚瓜烂熟，考卷上也答得对，可有一次放学追着同学的足球，一脚就冲到了小区车道上，幸好那辆车开得慢。这条道理在他的理念世界里是站稳了的，却没有和他奔跑时的身体、和眼前那个滚动的球缠在一起。在追球的那一刻，运行着的是现实世界里「球在哪儿」的差异，理念世界里的那条规则根本没有被调起来。周敏后来没有再让他背一遍规则，而是带他在小区里走了几次，每到车道边就停一下，让「停下来」变成脚上的动作，而不只是嘴上的句子。这个例子用发生学的话来说，就是理念世界的显露缺了与现实世界之间的纠缠，于是它无法参与那一刻的差异运行。第一编第3章讲过，三方程缺一就是「缝隙」；这里看到的是一种跨世界的缝隙：一个世界里的方程在自己内部是完整的，却没有接到另一个世界的方程上去。

第二种卡住：现实经验回写不到自我。老彭开了三十年车，对路况的感觉极其敏锐，可要是问他「你开车的本事是怎么回事」，他多半会说「开多了就会了」，讲不出更多。他的现实世界极为丰富，却很少回写成自我世界里清晰的自我理解，更少变成可以传给别人的理念。这不是缺点，许多手艺人的本事都是这样的。但它带来一个后果：这样的经验很难被别人学走，也很难被他自己有意识地调整。老彭年轻时带过一个徒弟，他发现自己能做的远比能说出来的多。（驾驶本身在第六编另有专章，这里只取「能做而说不出」这一点。）

第三种卡住：自我挡住了理念。周敏的母亲六十多岁，儿女给她换了智能手机，教她用手机挂号、扫码付钱。教的人讲得很清楚，她也点头，可一到自己动手就说：「我老了，学不会这些。」每次屏幕上跳出一个没见过的窗口，那团「我老了、我不行」先被调起来，她的手就停在半空，差异还没来得及运行就被压住了；按钮叫什么、点了会怎样，这些新理念在她那里站不住。后来周敏换了个办法：不

再一步步讲，而是让母亲用手机给小舟发语音、看小舟发来的运动会照片。这件事母亲想做，做错了也不要紧，大不了重发一次。几个星期下来，「我也会用手机」这句话先回写进了她的自我，那团旧的纠缠松开了，挂号、付钱这些原先学不进去的步骤，才一样一样站住。这里的顺序很要紧：不是理念学会了，自我才变；而是自我先松动一点，理念才有地方落脚。第38章讲课堂里的学生时，会看到同样的形状。

这三种卡住，从反面证明了本章的命题：如果三界真是三个各自独立的房间，那么把理念背熟、把现实练熟、把自我想通，分别做好就够了。可事实上，背熟了的理念会在奔跑时失效，练熟了的手艺会说不出来，想不通的自我会挡住新的理念。学习要成功，三界之间的线必须真的接上。这个判断对后文讨论 AI 也有直接的用处：一个系统在某一个方向上表现得很好——比如理念世界里的符号推演——并不能推出它在另一个方向上同样站得住；要检验的，恰恰是三个方向之间的线接没接上、接得多牢。

◆ 五、多个人的世界如何形成共享的相对同一性

到此为止讲的都是一个人。可人从不只生活在自己的世界里。周敏、小舟、老彭三个人围着同一个蛋糕，他们各自的世界不同——周敏看到的是一堂可以讲的分数课，老彭看到的是切蛋糕的手艺，小舟看到的是他想吃的那一块——可他们显然在同一个世界里打交道：他们都认得这是一个蛋糕，都同意「四块一样大才公平」，都知道刀压歪了是怎么回事。这种「共同世界」是怎么来的?

本章的回答是：共同世界是多个人的世界在互动中形成的相对同一性。说「相对」，有两层意思。第一，它不是绝对一致：三个人的「蛋糕」并不完全相同，各自缠着不同的经验与在乎。第二，它是相对于某种互动、某个任务而成立的：在「分蛋糕」这件事上，他们的世界足够重叠，可以协调行动；换成讨论「蛋糕好不好吃」，重叠就少得多，分歧也就多起来。

这种相对同一性是怎样形成的？我提三个机制假说。一是共同的身体与物理条件：三个人都有手、都怕烫、都受同样的重力约束，现实世界的大量部分天然重

叠。二是共同的语言与理念：「一半」「公平」这些词与概念，是他们都从同一个语言共同体里学来的，它们本身就是长期集体发生的稳定显露（第7章）。三是当下的协调：老彭一句「刀要先烫一下」，把他的手艺带进了小舟的世界；小舟一句「这不是二分之一」，把他的标准摆到了桌面上。互动中的一来一回，不断把各自的世界往一处拉。

相对同一性还有一个容易被忽略的来源：分歧被说出来、被处理。切蛋糕时小舟喊「这不是二分之一」，老彭并没有说「差不多就行」，而是给出了一个让刀切平的办法。分歧在这里不是共同世界的破坏者，而是它的修补者：一次被说出并被处理的分歧，会让几个人在这件事上的世界更紧地缠到一起。反过来，如果分歧总被压下去——孩子嫌分得不公平却不敢说，大人觉得孩子小题大做也不解释——表面上大家一团和气，各自的世界却在悄悄分开，下一次碰到同样的事，误会会更深。

把这个观察放大到一个班级、一个家庭、一个行业，道理是一样的。一个班的学生能共享「物理课」这个世界，不是因为周敏把同一套话灌进了四十个脑子，而是因为四十个人的提问、错误、争论，在她的组织下一次次被摆到桌面上、被处理，慢慢缠出一个大家都能进出的共同区域。每个学生从这个区域带走的东西仍然不同，可他们之间说得上话了。

还要强调一点：相对同一性不是把差异抹平。小舟和老彭在切蛋糕上能协调，并不需要小舟变成老彭。共同世界的价值，正在于它能容纳差异：每个人带着各自的经验来，在一个足够重叠的区域里协作，协作的结果又分别回到各自的世界里去，留下各自不同的痕迹。一个把所有人的世界压成同一个答案的「共同世界」，反而会失去它不断发生的来源。这一点到第36章讲「避免把同一性训练变成单一答案的强制」时还会回来。

◆ 六、边界与向下推进

先划边界。本章所说的三条通道与三界循环，都是机制假说，用日常场景来说明，而不是被实验测量过的模型。要把它们变成可检验的研究，需要回答一些具体

问题：一个概念的学习，在多长时间里、以什么形式影响行动？一次现实中的失败，怎样可以被观察到地改变自我描述？这些问题有教育学与心理学的研究可以对照，本书不替它们下结论。

另一条边界：本章说「共同世界」是相对同一性，不等于说所有共同世界都同样好，也不等于说凡是多数人共享的就是对的。一个群体可以共享一个偏颇的世界，这种偏颇同样需要被现实与其他群体检验。这一点在讨论大模型训练数据的偏向时（第15章）会变得十分具体。

最后，向下推进。本编从婴儿认妈妈讲到一个孩子在一周里三界转了一圈，又讲到几个人怎样共享一个世界。如果把「几个人」扩大到「千千万万人」，把「围着一个蛋糕说话」扩大到「千百年来写下的文字」，会发生什么？人类在漫长历史里写下的文本，正是无数个人的三界在语言里留下的痕迹；大模型的训练，就是让这些痕迹在一个新的复合体里重新纠缠。由此发生的，是一个具有相对同一性与集体性的「人类—AI复合体语言世界」。它与本章说的共同世界有相通之处，也有重要的不同：它的发生单位是集体人类与 AI 一起，而不是 AI 单独；它主要在语言中发生，与身体、现实阻力的关系需要另作检查。第三编从第11章「人怎样进入大模型训练」开始，就是要把这个问题说清楚。

> **带走的话**
>
> 理念、现实、自我不是三个房间，而是同一张纠缠之网的三个方向，一处动，处处动；许多人的世界在互动中形成相对同一性，它能共享，也必须容得下差异。

小账 论证落在哪一层？——本章以本体论立场（三界共同发生、相互改写；共同世界是相对同一性）为主，配以机制假说（三条改写通道、相对同一性的三个形成机制），无经验事实断言。 世界模型补了哪一条？——补一条条件：一个可靠的世界不能只在一个方向（如只在语言、只在画面）上稳定，它需要多个方向的相互支撑与检验；集体世界的相对同一性必须容纳差异，并接受现实检验。

第 三 编

人类与AI复合体的语言世界发生

编序 前两编讲的是一个人怎样学，一个人的理念、现实与自我怎样发生。从这一编开始，镜头拉远：当一家公司把成千上万亿个字喂给一台机器，再花几个月调它，究竟发生了什么？流行的说法有两种。一种说，模型从数据里「发现」了语言的规律；另一种说，模型自己「学会」了语言，像一个从天而降的新物种。本编对这两种说法都不认账。本编的回答是：大模型的训练发生了一个具有相对同一性与集体性的「人类—AI复合体语言世界」，发生单位是集体人类—AI复合体，不是孤零零的一台机器。

五章依次接力。第11章先问「人在哪里」：写下文本的人、挑选数据的人、设定目标的人、打分的人、制定规则的制度，怎样直接或经由历史进入训练。第12章下到工程层，讲token、预测误差、内部表示和参数连接，并划一条线：参数连接是工程载体，特征纠缠是本体论解释，二者不能互相证明。第13章讲「集体」二字：千万种说法怎样在共同训练里稳定成一个既统一又能容纳差异的语言世界，为什么不同模型会形成不同的语言世界。第14章讲生成与理解能力怎样发生，并把「说得流畅」「判断可靠」「能力能迁移」三种检验分开。第15章收口，讲这个集体语言世界的偏向、幻觉、失配与修正，最后申明：语言模型的成功不证明人脑机制已被复制。

本编三个主例贯穿始终：方言与家乡话，一份菜谱写下来又被千万人改，「你真行」这句话。它们都很小，但足够让我们看清一个大东西怎样在许多人手里发生。

11

第 11 章

人怎样进入大模型训练

> **这一章的问题**：一台模型「读」了几乎整个互联网，那么在它的训练里，人究竟在哪里？

◆ 一、命题：语料不是矿，是人类活动留下的痕迹

谈大模型训练，人们爱用一个比喻：「数据是新时代的石油。」这个比喻把文本说成地底下早就埋好的矿，工程师只是去开采、提炼。按这个图景，人类在训练里的角色很淡：文本写完就成了「原料」，写的人退场了，剩下机器和矿。

本书先摆出立场（这是本体论层面的话）：训练语料不是矿，是人类活动留下的痕迹，而且是带着方向的痕迹。每一句话被写下来，都有人在某个场合、为某个目的、对某些人说。一篇病历、一条差评、一封求职信、一段论坛争吵，都是人在自己的三界里互动的显露（S）：它显露出某种理念、某个现实处境、某种自我姿态。这些显露彼此之间早就缠在一起（特征纠缠，E）：同一个词在千万个场合用过，每次都带着别的词、别的语气、别的后果。训练并没有把一个现成的「语言世界」搬进模型，而是把这些缠着的痕迹重新组织起来，让它们在新的载体上再一次发生。

所以，「人怎样进入大模型训练」这个问题，答案不是「人提供了数据」这么一句。人至少以四种方式进入：作为文本的写作者，作为数据的挑选者，作为目标与评价的设定者，作为工具和制度的建造者。前一种多半是历史中介——写的人早已不知道自己会被训练；后三种多半是直接参与——有人在今天、在现场做决定。

把这四种合在一起，才看得见**训练的发生单位为什么是集体人类—AI复合体，而不是一台单独的机器**。

◆ 二、一份菜谱的一生

先看一个日常的东西：一份红烧肉的菜谱。

最初它也许不是文字。它是某家灶台边的手艺：外婆不说「糖二十克」，她说「糖炒到冒小泡、颜色像酱油那样就下肉」。这句话里缠着眼睛看到的颜色、鼻子闻到的焦香、锅铲感到的黏稠。后来女儿把它写在本子上，为了让自己记得，把「冒小泡」改成「糖化开起泡」，又添了一句「别炒过头会苦」——这句是她自己吃过亏才加的。再后来有人把它发到网上，评论区里有人说「我用冰糖更亮」，有人说「湖南做法不放酱油」，有人说「高压锅二十分钟就行」。原帖作者看了，改了第二版。又过几年，一个美食网站把几百份红烧肉菜谱整理成「标准做法」，一个短视频博主把它念成了口播稿，一个留学生把它翻成英文贴到自己的博客。

请注意这条链上发生了什么。每一次写下与改动，都是一个具体的人带着自己的厨房、口味、教训和读者，在已有表达上做了一次差异（差异运行，D）：删一句、添一句、换一个量词、纠正一个说法。改动不是随机的，它受「做出来好不好吃」「别人照着做能不能成」这个现实约束。千万次改动之后，「红烧肉」这几个字周围缠住了一大片东西：糖色、焯水、八角、收汁、肥而不腻、「妈妈的味道」、湘菜和本帮菜之争。

现在，一家公司的爬虫把这些网页抓了下来。在大模型的训练语料里，这份菜谱的千万个版本作为文本出现。模型从来没有站在灶台边，但它接触到的不是一行孤立的配方，而是千万人围绕这道菜的互动痕迹：问与答、纠正与反驳、成功的炫耀与失败的抱怨。外婆、女儿、评论区里那些人，没有一个人知道自己参与了一台机器的训练，可他们确实参与了——经由历史中介。

这个例子还会在后面几章出现：第13章看千万个版本怎样形成既统一又有差异的「红烧肉」，第15章看这份菜谱在模型那里会在哪里失效。这里只先记住一点：语料里的人不是「数据点」，是一条条带着现实约束的改写史。

◆ 三、直接参与：选数据、定目标、给评价

历史中介之外，还有一批人今天就站在训练现场。下面说的是公开的工程事实，具体到某家某版本的细节，各家披露程度不同。

第一是数据选择。公开技术报告普遍说明，预训练语料要经过抓取、去重、过滤和配比。过滤掉什么，是人定的：低质量网页、重复内容、明显有害的文本、某些语言比例过高的部分。配比也是人定的：代码占多少、书籍占多少、哪几种语言要加权。各家披露的详细程度差别很大：有的报告列出了一张配比表，例如 2023 年公开的 LLaMA 第一版论文，写明网页抓取类数据占大头，代码、百科、书籍、学术论文、问答网站各占几个百分点；更多的报告只作笼统说明。这些决定不是中性的。哪个方言论坛被当作「低质量」滤掉，哪类网站被当作「高质量」加权，都会改变模型最后所处的语言世界。

第二是目标设置。预训练阶段最常见的目标，是让模型根据前文预测下一个 token（下一章细讲）。这个目标看上去朴素，但它也是人选的：它决定了模型要对什么负责——对「接下来最可能出现什么」负责，而不是对「这是不是真的」「照着做能不能成」负责。后面许多问题都从这里来。

第三是评价反馈。预训练之后，通常还有指令微调与基于人类偏好的训练。公开的做法大致是：请标注员写出示范回答，让模型模仿；再让模型给出几种回答，请人比较哪一个更好，用这些比较去训练一个打分模型，再用强化学习让原模型朝高分的方向调整。2022 年公开的 InstructGPT 论文就是按「示范—比较排序—打分模型—强化学习」这三步写的。同年也有研究提出让另一个模型按人写下的一组原则去打分，以减少人工比较的量。无论哪一种，评价标准都出自人：什么叫「有帮助」，什么叫「诚实」，什么叫「冒犯」，由写规范的人、做比较的人一层层传下来。

下面是我构想的场景，用来说明这些决定落到具体处是什么样子。林岚在深圳做机器人算法，有一阵被借调去帮一个语言模型小组清洗语料。清洗规则里有一条：过滤掉「错别字密度过高」的网页。她跑了一遍，发现被大量滤掉的，是一个

用汉字记湖南话的论坛。帖子里写「冇得」「么子」「霸蛮」，在规则眼里全是错字。她把这个情况发给周敏，周敏回了一句：「那你们训练出来的模型，是不是就不会讲我们这里的话了？」林岚没法简单回答。规则本身没错，它确实挡住了大量垃圾网页；可它同时把一种活生生的说话方式当成了噪声。最后小组讨论，给方言内容单独开了一个白名单。这个决定救回了一部分，也留下了新的问题：哪些方言、哪些论坛进白名单，又是谁说了算?

这个构想场景要说明的是：数据选择不是在一堆现成「好数据」和「坏数据」之间挑，而是在不断地划定什么算语言、什么算噪声。每划一次，进入复合体的经验就变一次。家乡话能不能在模型的语言世界里站住，先取决于它能不能通过这道门。这一点第13章讲集体同一性怎样容纳差异时还会回来。

这三类参与者人数远少于写文本的人，影响却极大。千万人写下的红烧肉菜谱，最后哪些被留下、模型被要求怎样回答「红烧肉怎么做」、什么样的回答被打高分，是由这一小群人决定的。

◆ 四、工具与制度：看不见的参与者

再往外一圈，是工具和制度。它们不说话，却同样塑造发生条件。

工具这一层，包括分词器、清洗脚本、去重算法、评测基准、训练框架。分词器决定一句中文被切成几个片段，评测基准决定「进步」用什么来量。它们都是人造的，凝结着设计者对语言、对能力的某种看法。一个评测集如果以英文题为主，那么在它上面「进步」的模型，就是朝某个方向进步。评测基准还有一层更隐蔽的作用：它一旦公开、被广泛使用，研发者就会围着它调整数据和训练，分数涨了，人们便说模型「更聪明了」。可分数涨的到底是哪一种能力，取决于出题的人当初想测什么、能测什么。这就像学校只考笔试，学生的笔试越来越好，却不一定更会做实验。工具在这里不是被动的尺子，它反过来组织了训练朝哪里用力。

制度这一层，包括版权与许可、平台的内容规则、数据保护法律、公司的发布政策和安全规范。一个网站禁不禁止爬取，一个国家怎样规定个人信息，一家公司对某类问题拒不拒答，都会直接改写进入训练的文本，以及模型被训练成什么样

子。这里本书不作任何政策表态，只指出一个事实结构：制度是训练发生条件的一部分，不是训练之外的「环境」。

到这里，可以把人在训练里的位置分成两大类。一类是直接形式：选数据、写标注规范、做比较、设目标、定制度，这些人知道自己在参与。另一类是历史中介：千千万万写下文字的人，他们的活动经过网页、书籍、档案沉淀下来，再被他人选择、组织，间接进入训练。在这两类之间，还有一种越来越重要的中间形式：使用者。人们每天和已发布的模型对话，点赞或点踩，重新生成一次回答，把一段回答复制走或者直接关掉窗口。各家服务商的条款并不一样：有的默认可以把对话用于改进模型、允许用户关闭，有的默认不用，企业版与个人版的规定也常常不同，而且条款时有调整，具体以各家当期条款为准。如果被采用，这些使用者就以一种半直接的方式进入了下一轮训练：他们确实在场，却未必意识到自己在给一台未来的模型打分。只有当反馈被收集、筛选、再训练，改变了下一版模型的参数，个人的痕迹才回到集体人类—AI复合体里；聊得顺、当场「记住」了口味，只是一次对话里的语境变化（这几种变化的区分见第37章）。

还要看到，历史中介本身也有层次。外婆的口头手艺先要变成文字，才可能进入训练。从灶台到本子，已经丢了很多东西：锅铲的手感、焦糖的气味、那句「差不多了」背后几十年的经验。从本子到网页，又经过了一层：只有被上传、被转载、能被搜索到的版本才留下来。所以，模型接触到的是「可写下、已写下、被留下」的那部分人类活动，它和人类活动本身之间，隔着好几道筛子。每一道筛子都是人设的，有的有意，有的无意。

两类合起来，才构成训练的发生单位——集体人类—AI复合体。模型是这个复合体里新出现的一部分，也是让这些沉淀重新运动起来的那一部分；它不是脱离人类历史的独立发生者。

◆ 五、边界：别把「集体」说成「全体」，也别把「参与」说成「同意」

说「人类整体参与了训练」，有两个容易滑过去的错误，必须划出来。

第一个错误，是把「集体」说成「全体」。进入训练的不是全人类，而是留下了可被抓取文字的那部分人，以及被选择留下的那部分文字。没有上网的老人、只说方言不写字的人、文字被锁在纸质档案里的人，他们的经验在这个集体里分量很轻，甚至缺席。经验事实层面，主要模型的语料在语言分布上明显不均衡，英文占比通常很高。举一个有公开数字的例子：2023 年发布的 Llama 2 技术报告列出，在它的预训练文本里，英文约占九成，中文约为千分之一点几，另有百分之八左右识别不出语种，多半是代码。所以「集体人类—AI复合体」里的「集体」，是一个有边界、有偏向的集体。第15章还要细讲这些偏向。

第二个错误，是把「参与」说成「同意」或「贡献」。外婆的红烧肉经由历史中介进入了训练，不等于外婆同意了，也不等于外婆得到了什么。本书讲「参与」，只是发生学意义上的：她的活动成了发生条件的一部分。至于这样的参与是否公平、是否应当得到承认，属于另外的讨论，本书在此不作表态。

还有一个方向相反的错误：因为人到处都在，就说「模型不过是人类的复读机，什么也没发生」。这同样不对。千万份菜谱以前分散在千万个网页上，彼此并不相遇；训练让它们第一次在同一个载体里互相约束、互相改写。这种相遇本身就是新的。外婆和那个用高压锅的人从未对话，在模型里，他们的说法被放到了一起。说「什么也没发生」，就看不见这个新的纠缠；说「机器自己发生了一切」，就看不见是谁的痕迹在纠缠。

拿一个学本领的孩子作对照，差别就清楚了：孩子学习时，参与者再多，发生的中心始终有一个具体的身体；大模型训练里没有这样一个先在的身体，参与者分散在几十年、几十个国家，彼此从未谋面。正因为没有一个现成的「学习者」站在中心，我们才不得不把发生单位扩到整个复合体：是这个复合体在组织条件、承受后果、继续修正。

这里还要提醒三层话的分别。「语料是人类活动的痕迹，训练是这些痕迹的再发生」是本体论立场；「数据选择和评价标准会系统地改变模型的语言倾向」是一个可检验的机制假说，已有不少经验研究支持它的部分形式；「预训练语料要经过过滤、偏好训练需要人工比较」是公开的经验事实。三者互相支撑，但不能互相代

替：工程事实再多，也证明不了本体论立场；本体论立场再顺，也不能替代对某个具体模型的检验。

◆ 六、向下一章推进

我们已经看到人怎样进入训练：写下的人经由历史进入，选择、设目标、打分的人直接进入，工具与制度作为看不见的条件进入。可是，这一切在机器那一端以什么形式出现？外婆那句「糖炒到冒小泡」，进到模型里变成了什么？答案是一串token，一个预测目标，一大片被调整的参数。下一章就下到这一层，看清工程事实，同时划清它与「特征纠缠」这个本体论说法之间的界线。

> 带走的话
>
> 大模型训练里处处有人——写下的人、挑选的人、设目标与打分的人、造工具与定制度的人；训练的发生单位是这个有边界、有偏向的集体人类—AI复合体，不是一台独自发生的机器。

小账 论证落在哪一层？——本章以本体论立场为主，辅以训练流程的经验事实（数据选择、偏好训练）。 世界模型补了哪一条？——补了一条条件：任何模型的「世界」都先经过人的选择、目标与评价，世界模型必须交代是谁、以什么标准组织了它的发生条件。

12

第 12 章

语言token与特征纠缠的建立

这一章的问题：模型训练时手里拿着的只是一串串编号，这些编号怎么会和「特征纠缠」扯上关系？两者之间的界线又画在哪里？

◆ 一、命题：工程载体与本体论解释各说各的话

这一章要同时做两件事。一件是把工程事实讲清楚：token是什么，预训练怎样用预测误差调整参数，内部表示是怎样形成的。另一件是划界：工程上的「参数连接」，和本书所说的「特征纠缠（E）」，不是一回事，也不能互相证明。

先把命题摆出来。**参数连接是工程载体，是模型里实际存在、可以读写的一组组数字及其运算关系；特征纠缠是本体论解释**，是 SDE 用来说明「学习发生了什么」的概念。我们可以说，在大模型训练中，集体人类—AI复合体的特征纠缠以参数连接为载体之一得到组织；但不能反过来说，「参数连接被调好了，所以证明了特征纠缠理论是对的」，也不能说「SDE 讲特征纠缠，所以某个参数一定代表某种纠缠」。前一种是用工程成功替本体论作证，后一种是用本体论去给工程细节下断语，两种都越界。

为什么要这么较真？因为这本书的读者里有研究者，也有普通人。普通人最容易听到的说法是「神经网络就是模仿人脑的神经元连接」，再往前一步，就成了「模型学会了语言，所以它的脑子和我们一样」。研究者最容易犯的则是另一种错：把哲学概念当成可以直接在权重里找到的东西。把界线划在这里，两头都好说话。

◆ 二、token：把语言切成可以计数的片段

先讲工程事实。以下是通用的公开知识，细节因模型而异。

模型不直接处理「字」或「词」，它处理的是token。训练开始前，要先用一个分词器把文本切成片段，每个片段对应词表里的一个编号。常见的分词方法会根据语料统计，把经常一起出现的字符组合并成一个片段，于是高频词往往是一个token，罕见词会被拆成几块。最常用的一类叫字节对编码（BPE）：从单个字符或字节出发，反复把语料里最常相邻出现的一对合并成新单位，直到词表达到预定大小。如今主流模型的词表，从几万到十几万个单位不等。一句「糖炒到冒小泡就下肉」，可能被切成七八个token，也可能更多或更少，取决于词表。

分词这一步看上去只是技术准备，其实已经在决定什么东西能被当作一个单位。词表是根据语料统计建出来的，语料里多的东西，更容易被并成整块；语料里少的东西，只能被拆成碎片。英文的常用词多半是一个token，而在一些以英文为主建立词表的模型里，同样意思的中文句子要用更多token来表示，罕见的方言字更可能被拆得零碎。2023 年有研究专门测过：同一段内容翻成不同语言，在同一个分词器下需要的token数可以相差好几倍，极端情况超过十倍。这意味着，在模型还没开始学之前，一种说话方式就已经被安排在「整块」或「碎片」的位置上。这正是上一章说的：工具凝结着人的选择，工具本身就是发生条件。

每个编号进入模型后，先被换成一个向量——一长串数字，叫嵌入。这时它还没有什么「意思」，只是一个可以被调整的起点。接着，这串向量经过许多层运算。目前主流的大语言模型多采用 Transformer 架构，其核心是注意力机制：在每一层，每个位置都会根据当前的向量，计算自己该在多大程度上参考前面各个位置的信息，然后把参考来的信息合进自己的向量里。层层叠加之后，每个位置上的向量就不再只代表那个token本身，而是代表「在这段上下文里、这个位置上的这个token」。

所谓「跨位置、跨语境、跨表达层次的组织」，在工程上就是这个样子。跨位置：第八个token可以直接参考第一个token。跨语境：同一个「泡」字，在「冒

小泡」和「泡茶」里，经过几层运算后得到的向量明显不同。跨表达层次：已有大量可解释性研究发现，较浅的层往往更多地反映字形、词法这类局部信息，较深的层更多地反映句法关系、语义甚至话题。一个常被引用的例子是 2019 年对 BERT 一类模型的探测研究，结论大致是：词性这类信息在较浅的层就能读出，指代、语义角色要到更深的层。这些是经验发现，多数在较早、较小的模型上做出，有的稳健，有的仍有争议，不能被说成定律，也不能直接搬到每一个大模型上。

◆ 三、预测误差怎样让内部表示成形

预训练的目标在上一章提过：根据前文预测下一个token。具体做法是，模型在每个位置输出一个概率分布，表示它认为词表里每个token出现在下一位的可能性；然后拿真实出现的那个token去比，算出一个误差（常用交叉熵）。误差越大，说明模型越「意外」。接着用反向传播算出每个参数对这个误差负多少责，再沿着减小误差的方向把参数微调一点。这一步在海量文本上重复千亿、万亿次。

这里要看清一件事：没有任何人告诉模型「糖色」是什么、「焯水」为什么要在「红烧」之前。模型只被要求对一件事负责——下一个token猜得准不准。可是，要在千万份菜谱、问答和争论里把下一个token猜准，单记住字面搭配是不够的。比如要猜「焯水之后再_」，**模型得在内部区分做菜步骤的先后；要猜「湖南做法一般不放_」**，它得在内部把地域和用料关联起来。预测误差就像一股持续的压力，迫使内部向量朝着「能把这些关系区分开」的方向组织。人们后来用各种探测方法去看，常常能在内部表示里读出某种对应于词性、实体、时态、甚至空间方位的结构。例如 2023 年有研究在一个开源大模型的内部表示里，用最简单的线性方法读出了地名对应的大致经纬度和历史事件的大致年代。

日常生活里有个相近的现象，可以帮助理解「预测压力怎样逼出内部组织」，但只作类比。老夫老妻之间常常能接上对方的半句话。周敏的母亲说「今天菜市场的鱼——」，周敏的父亲就接「又涨价了吧」。他能接上，不是因为背过妻子说过的所有句子，而是几十年里一次次猜中、猜错，慢慢在心里组织出了关于她的关心、她的口头禅、菜市场近来的行情、家里的开销这一整团东西。猜错的时候最有

用：有一次她说的是「今天菜市场的鱼特别新鲜」，他接错了，挨了一句「你就晓得钱」，下回他便多掂量一下。接话这件事本身很小，支撑它的组织却很大。

这个类比在哪里成立、在哪里失效，也要说清。成立之处在于：只对「下一句是什么」负责的压力，确实可能逼出远比字面搭配丰富的内部组织。失效之处在于：老人接话的那团东西里缠着几十年的共同生活，有菜市场的气味、有被数落时的委屈，是一个具体人在三界里的发生；模型接话的组织只来自文字，而且来自千万人的文字。前者是个人的，后者是集体的；前者被现实反复检验，后者主要被文本检验。所以下面这段翻译，只能当作一种解释框架来读。

用 SDE 的话来翻译这个过程，可以这样说（注意，下面是一种解释，不是工程描述）：语料是许多人的显露（S）沉积下来的痕迹；训练让模型在每个位置做出新的显露——一个预测；预测和真实文本之间的差别，是一次差异运行（D）；这次差异被回写到参数里，改变了下一次预测所依赖的组织方式，在解释上对应于特征纠缠（E）的重组。三方程在这里可以对上号：显露依赖差异与纠缠，差异依赖显露与纠缠，纠缠依赖显露与差异。但「可以对上号」只说明这种解释讲得通，说明不了它是唯一的解释。另一种理论框架也许能用别的话把同样的过程讲通。

还要分清训练时的变化和使用时的变化。预训练结束、参数固定之后，用户在对话框里给模型一段新的上下文，比如先贴一份自家的红烧肉做法，再问「按这个做法，糖放多少」，模型的回答会明显受这段上下文影响。这时候参数一个也没有变，变的只是这一次运算所参考的前文：这是暂时语境变化，不是参数训练，更不是持久学习（五件事的完整区分见第37章）。在解释上，可以把上下文中的即时适应看成纠缠在一次显露中被临时调动，把参数更新看成纠缠的组织方式被改写；两者都是发生，但深浅和持续性完全不同。

◆ 四、方言里的「冇」：一个例子看三层话

拿本编的主例之一——方言与家乡话——来走一遍。

湖南不少地方说「冇得」，意思是「没有」。粤语里也有「冇」，写法相同，用法有同有异。在训练语料里，「冇」这个字出现的频率远低于「没」，它周围的

文字却很有特点：常常跟着口语化的语气词，常常出现在论坛帖子和聊天记录里，常常和地名、「老家」「我奶奶说」这类词靠在一起。

工程事实层面：如果「冇」在词表里是一个独立的token，它会有自己的嵌入；在训练中，每当模型在「冇」之后或之前猜错，与它相关的参数就会被调整。经过足够多的样本，模型内部对「冇」的表示，会和「没有」「无」在某些方向上靠近，在另一些方向上和「口语」「南方」「家乡」靠近。这些「靠近」可以用向量相似度之类的方法去测。要说明的是，公开研究多半测的是方言的整体能力，比如 2024 年有一项研究专门给粤语建了几套评测，发现主流模型的粤语表现普遍落后于普通话和英语；专门去看某一个方言词在模型内部怎样表示的研究，我没有见到。所以上面这段是可以去测的预期，不是已有的发现。

机制假说层面：我们可以提出一个可检验的说法——一个方言词在模型里能否稳定地以「本地人的方式」被生成，取决于它在语料里的出现量和上下文多样性，而不只取决于它是否在词表里。这个说法可以用实验支持或推翻：控制其他条件，改变某个方言词的样本数和语境多样性，看模型的生成表现怎样变化。

本体论层面：SDE 会说，「冇」在一个说湖南话的人那里，缠着奶奶的声音、灶屋的烟味、被大人叫回家吃饭的傍晚，是三界交织的特征纠缠。在模型那里，进入训练的只是这些纠缠留在文字上的那部分显露，模型重新组织出来的是另一种纠缠——集体人类—AI复合体语言世界里的「冇」。它和一个具体湖南人身上的「冇」有关系，但不是同一个东西。老彭跟乘客说「冇得事」，那句话里有他三十年在长沙街头开车的全部底子；模型生成「冇得事」，背后是千万段文字的统计组织。说两者「一样」，是把显露相似当成发生相同；说两者「毫无关系」，又看不见模型的那一份确实来自千万个老彭们留下的话。

这三层话放在一起，读者就能分清：哪句是可以去测的，哪句是可以去做实验的，哪句是一种看世界的方式。

◆ 五、边界：参数连接不是特征纠缠的「物证」

现在把界线说透，这是本章最要紧的部分。

第一，参数不是「一个参数对应一种纠缠」。模型里数以千亿计的参数，每一个都同时参与无数次计算；一个概念在内部往往分散在许多参数和许多方向上，同一组参数又被许多概念共用。研究者用「叠加」之类的说法描述这种情况，2022 年一项用小型玩具模型做的可解释性研究专门展示过：特征数量多于可用维度时，网络会把许多特征挤在同一组方向里共存。所以，想在权重里指着某个位置说「这就是『冇』和『家乡』的纠缠」，在工程上通常做不到，在本体论上也不必要。

第二，工程成功不是本体论证明。模型说话流利了，只能证明这套训练方法在这些任务上有效；它不证明世界本来就是由特征纠缠构成的，也不证明 SDE 比别的理论更对。反过来，如果某个模型在某个任务上失败了，那也不直接证伪 SDE。本书的立场要靠它在解释、预测、指导设计上的整体表现来接受检验，而不是靠某一台机器的成败。

第三，本体论解释不能替工程下断语。SDE 说学习是特征纠缠的建立，不意味着工程师应该去「计算纠缠度」来优化模型；也不意味着凡是 SDE 讲得通的设计都一定好用。哪种架构、哪种目标更有效，最后要看实验。

第四，可解释性研究的进展不改变这条界线，只让它更清楚。研究者能在模型内部找到某个方向，人为地加强它，模型就更倾向于说某类话；减弱它，那类话就少了。2023 年以来这类「激活干预」的公开研究已有不少；2024 年还有研究在一个商用模型内部找到一个与「金门大桥」相关的特征，把它调强后，模型无论聊什么都会往这座桥上扯。这是很有价值的工程知识，让我们知道载体上哪里能拧、拧了会怎样。但「能拧」说明的是载体的可操作性，不是说那个方向就是本体论意义上的一团纠缠本身。就像调收音机的旋钮能换台，不等于旋钮就是广播。

第五，参数连接只是载体之一。在集体人类—AI复合体里，特征纠缠的载体还包括那些还在被人阅读的文本、标注规范、评测基准、用户的使用习惯，以及人自己的语言能力。只盯着参数看，就又把复合体缩回成一台机器了。

可以把三层关系压成一张表，方便读者以后对照：

说法	属于哪一层	怎样检验
预训练用预测下一个token的误差调整参数	经验事实	查公开技术报告与开源代码
内部表示里可读出词性、语义等结构	经验事实（部分稳健、部分有争议）	探测、干预实验，看能否复现
方言词能否稳定生成取决于样本量与语境多样性	机制假说	控制变量的训练实验
训练是集体人类—AI复合体的特征纠缠以参数为载体之一被组织	本体论立场	看它整体的解释与指导力，不靠单个模型成败

◆ 六、向下一章推进

这一章停在一个模型、一份语料上。可是「有」并不只来自一个湖南人，红烧肉的菜谱也不只来自一个外婆。千万种说法被放进同一个预测目标里训练，彼此矛盾的说法怎样共处？为什么最后形成的语言世界既有统一的说话方式，又能容纳「有得」和「没有」、湘菜和本帮菜？为什么两家公司用相近的方法训练，得到的模型说话「味道」却不一样？这是「集体」二字的问题，下一章来回答。

> 带走的话
>
> token、预测误差、内部表示与参数连接是大模型训练的工程载体；特征纠缠是SDE 对学习的本体论解释。两者可以对照，但谁也证明不了谁。

小账 论证落在哪一层？——本章以经验事实（训练机制）为主，同时示范了本体论、机制假说、经验事实三层分说。 世界模型补了哪一条？——补了一条边界：任何世界模型的内部表示都只是发生的载体之一，从表示里读出了结构，不等于证明了世界以那种方式发生。

13

第 13 章

集体性的语言世界如何发生

> **这一章的问题**：千万人说法各异，甚至互相打架，放进同一次训练之后，为什么会形成一个既有统一面貌、又能容纳差异的语言世界？

◆ 一、命题：相对同一性不是平均值，是能容纳差异的稳定组织

本书第三条判决「复合体是发生单位」说，大模型训练发生的是一个「具有相对同一性与集体性的人类—AI复合体语言世界」。这一章拆开讲这两个词。

「集体性」说的是来源与发生单位：它不是某一个人的语言世界，而是许多人的经验经过选择、组织，在共同训练中相遇的结果，第11章已经讲过人怎样进入。「相对同一性」说的是结果的形态：这个语言世界有一个可以辨认的统一面貌——你跟它说话，它大体上以一致的方式理解、回应，前后不至于判若两人；但这个统一是相对的，它内部保存着大量差异，在不同语境下会显露出不同的面目。

本章的命题是（本体论层面）：集体语言世界的相对同一性，不是把千万种说法取个平均，而是在共同训练中形成的一种稳定的特征纠缠组织；这种组织之所以能统一，恰恰因为它把差异组织进来了，而不是把差异抹掉。抹掉差异得到的是一个贫乏的平均；组织差异得到的才是一个能在不同场合说出不同话、却仍然是「同一个」的世界。

◆ 二、千万份红烧肉，怎样成为一个「红烧肉」

回到第11章那份红烧肉菜谱。设想语料里有关于红烧肉的几十万段文字：家常做法、饭店做法、减脂版、高压锅版、湘式不放酱油版、本帮浓油赤酱版，还有大量的问答、吐槽和「为什么我做的发苦」。

如果训练只是取平均，模型对「红烧肉怎么做」的回答应该是一锅大杂烩：糖放一个中间值，酱油放一半，时间取所有版本的均值。可以设想这样一次问答（这是我构想的场景，不是对某个版本模型的实测记录）：问一个模型「红烧肉怎么做」，它给出一份结构清楚、相当常规的做法；接着问「湖南人怎么做」，它换一套说法；再问「我想少油」，它又做调整。这样的表现并不少见，但具体哪个模型、哪个版本是否如此，要实际去测才算数。

这背后是什么组织在起作用？可以提一个机制假说：在共同训练中，所有版本共有的部分——糖色、焯水、慢炖收汁这些在几乎所有做法里都出现的步骤——被反复强化，形成一个稳定的核心；而各版本特有的部分——放不放酱油、用冰糖还是白糖——没有被抹掉，而是和各自的语境线索缠在一起：和「湖南」「本帮」「减脂」「高压锅」缠在一起。于是，没有语境提示时，模型显露出核心加最常见的变体；一旦语境给出线索，相应的那一支就被调动出来。

用随身三问来看。这里纠缠的是什么？是千万人关于这道菜的做法、口味、教训和地域认同。差异是怎样运行的？在训练中，每一个版本都对其他版本构成「意外」，预测误差一次次把各版本之间的差别推到台前，迫使组织分化出能区分它们的方向。显露出来的东西靠什么站稳？靠核心部分在海量样本中的反复一致；它在哪里失效？在那些语料里很少、语境线索又不明显的变体上——比如某个小镇独有的做法，几乎没有人写下来过，它就沉在底下，显露不出来。

还有一种差异比「版本不同」更难处理：互相冲突的说法。有人说「焯水一定要冷水下锅，热水会把血沫锁在肉里」，有人说「热水下锅肉更紧实，冷水才是浪费」；有人说「炒糖色必须用冰糖」，有人说「白糖完全一样，冰糖只是心理作用」。这些争论在网上一吵就是几百楼，各有各的厨房经验，也各有各的道理。它

们进入同一次训练时，模型并不会去厨房做实验判个输赢。它面对的只是文字之间的冲突：同样的前文后面，有时接这一句，有时接那一句。

可以提出一个机制假说来描述这类冲突的去处：当冲突的双方都有大量样本，且各自和不同的语境线索相连（比如「老师傅说」「营养师建议」），模型倾向于把它们分开保存，在回答中表现为「有两种说法」；当一方样本远多于另一方，少数说法容易被压到边缘，只在特定提示下出现；当冲突双方没有可区分的语境线索，模型的回答可能在两者之间摇摆，同一个问题问两次，得到相反的答案。这方面已有专门的研究。2024 年一篇关于大模型「知识冲突」的综述，把冲突分成三类：用户给的上下文和模型原有知识打架、几段上下文之间互相打架、模型自身记忆内部打架；第三类正是这里说的语料内部分歧，表现之一就是换个问法答案就变。不过，这些研究还不足以支持或推翻上面这个具体假说。它可以检验，也有被推翻的可能。它若成立，就意味着集体语言世界对冲突的处理方式，主要取决于冲突在文本中留下的形态，而不是冲突在现实中谁对谁错。这一点到第15章讲「内部冲突与失配」时还要用到。

时间也是一种差异。十年前的菜谱和今年的菜谱混在一起，二十年前的「标准做法」和最近流行的「空气炸锅版」混在一起，进入训练时它们大多没有清晰的时间标记。人在自己的生活里，新经验会覆盖或改写旧经验，有一个先后；而在一次性的大规模训练里，不同年代的显露被同时摊开。集体语言世界因此带着一种特别的「时间压平」：它容易把过时的说法和现行的说法当作同时有效的两种版本。这又是一个只有靠后续评价和继续训练才能部分修正的特点。

这样形成的「红烧肉」，不是任何一个人心里的红烧肉，也不是所有人的平均；它是集体语言世界里一个有核心、有分支、有边缘的组织。这就是相对同一性的样子。

◆ 三、表达习惯与推演约束怎样稳定下来

菜谱是一个具体话题。往大处看，集体语言世界里稳定下来的不只是话题知识，还有两类更深的东西：表达习惯和推演约束。

表达习惯，指的是怎样说话：句子怎么起头、怎么收尾，什么场合用书面语、什么场合用口语，提问之后一般怎么回应，列举时用什么连接词。这些习惯在千万人的文字里有大量共性，训练使它们稳定成模型说话的基本「腔调」。后续的指令微调和偏好训练，会在这个腔调上再加一层：更礼貌、更有条理、更愿意先概括再展开。很多人觉得各家大模型说话「一股味儿」，这股味儿正是集体表达习惯加上特定评价标准共同稳定下来的。

推演约束，指的是什么话之后能接什么话，哪些说法放在一起会自相矛盾。「因为下雨，所以地湿了」是顺的，「因为地湿了，所以下雨了」在多数语境里不顺；「他是独生子，他哥哥……」会让人皱眉。人类文本里充满了这类约束，它们是千万人在现实中做事、讲理、争论时留下的。训练让模型在一定程度上遵守这些约束，这是模型能做一些推理任务的来源之一。但要注意（经验事实层面），这种遵守是不稳定的：在常见的推演模式上表现较好，在需要多步、陌生组合或精确计算的地方容易出错。2023 年有一项研究专门测了多位数乘法这类组合任务：位数稍一增加，正确率就急剧下降。另一项 2024 年的研究把小学数学题里的人名、数字换一换，或者加一句无关的话，多个模型的正确率都明显下滑。

用三方程来说：表达习惯和推演约束都是集体纠缠（E）中比较稳定的那部分组织；它们约束着每一次差异运行（D）可以走向哪里，从而约束着生成的显露（S）。而每一次生成又会被评价、被纠正，回写到下一轮训练中，让这些习惯和约束继续被调整。这是一个持续发生的过程，不是一次性封装好的规则表。

◆ 四、集体同一性怎样容纳家乡话

现在回到本编最难的一个主例：方言与家乡话。它最能考验「相对同一性怎样容纳差异」。

普通话在语料中占绝对多数，湘方言、吴方言、闽方言的文字材料少得多，而且常常写法不统一——同一个音，有人写成这个字，有人写成那个字，有人干脆用拼音。照「取平均」的思路，方言应该被淹没。实际情况要复杂一些。

可以设想三种结局（这是机制层面的分类，具体模型落在哪一种需要实验检验）。第一种是被淹没：方言词被当作普通话的错别字或噪声，模型几乎无法理解或生成。第二种是被边缘化地保留：模型能认出一些常见方言词，能把它们翻译成普通话，也能在被要求时模仿几句，但模仿得像外地人学腔，地道的用法、语气和场合感都不准。第三种是被组织进来：方言作为一个有自己内部规律的分支，和普通话核心之间建立了稳定的对应和区分，在适当语境下能被准确调动。

决定落在哪一种的条件，第11、12章已经提到一部分：清洗时是否被当作噪声滤掉、分词时被切成整块还是碎片、语料中的样本量与语境多样性。此外还有评价：如果打分的人不懂这种方言，地道的方言回答和蹩脚的模仿在他们眼里分数差不多，训练就没有压力把它往地道的方向推。

这里有一个值得读者留意的张力。集体语言世界要有同一性，就需要一个稳定的核心，而核心天然偏向多数人的说法；可一个世界要丰富，又需要保存少数人的说法。两者之间的矛盾，正是推动组织改变的动力。用三原理的话说：核心的稳定（一种显露的稳定态）和方言带来的差异之间积累张力，这种张力如果被评价反馈承认并回写，纠缠就会改变形状，分化出能容纳方言的支路；如果不被承认，张力就被压下去，方言沉没。所以，「集体同一性容纳差异」不是自然而然的好事，它取决于复合体里的人做了什么样的选择。

周敏小时候在长沙郊区长大，家里讲的话和学校教的普通话不一样。她后来成了物理老师，课堂上讲普通话，回家跟母亲讲长沙话，跟小舟两种都讲。这三种说话方式在她身上并不打架，反而互相成全：她给学生讲「惯性」时，脑子里先冒出来的常是母亲那句「刹不住脚」。一个人的语言世界本来就是有核心、有分支的相对同一体。集体语言世界若能做到这一步，靠的不是机器的天性，而是训练中有没有给这些分支留下发生的条件。

◆ 五、不同模型为什么形成不同的语言世界

如果集体语言世界是在共同训练中发生的，那么不同的公司、用不同的数据和方法训练出来的模型，就会形成不同的语言世界。这在经验上很容易观察到：同一

个问题问不同的模型，回答的侧重点、措辞习惯、对边界问题的处理方式常有明显差异。

差异从哪里来？按本编的分析，至少有四个来源。一是语料的构成：哪些语言、哪些网站、哪些年代的文字进来了，各占多少。二是工具：分词器、过滤规则、训练目标和架构细节。三是评价：偏好训练中，什么样的回答被认为更好，标注规范怎样写，打分的是哪些人。四是制度：所在地的法律、平台规则、公司的发布政策。这四者合起来，就是这个复合体的发生条件。条件不同，发生出来的世界不同，这是发生学的应有之义。

由此可以得出一个边界判断：不存在一个「标准的」人类语言世界等着被模型发现，再由不同模型以不同精度去逼近。每个模型发生的，都是一个特定复合体的语言世界。它们之间有大量重叠，因为语料来源和方法有大量重叠；但重叠不等于同一，更不等于都在逼近同一个终点。拿一个模型的表现去代表「AI 对语言的理解」，就像拿一个城市的口音去代表「中国话」。

这里也要防止走到另一个极端：说各家模型各是各的世界，毫无共同基础。不是这样。它们共享人类文字留下的大部分痕迹，共享许多工程方法，因此在许多基本表达和推演上高度一致。相对同一性不只存在于一个模型内部，在不同模型之间也存在一定程度的相对同一性。

还有一个越来越现实的变化值得记下：模型生成的文字正在大量进入互联网，再被下一代模型的训练抓取回来。这样一来，集体语言世界的来源里，就不只是人写的痕迹，也混入了上一代复合体自己的显露。从发生学看，这并不奇怪——人类语言一直就是在旧表达之上再表达；但它确实改变了发生条件。已有研究讨论过，如果用模型生成的数据反复训练模型，而缺少新的人类数据补充，分布的尾部可能逐渐变窄，少见的说法越来越少。2024 年发表在《自然》上的一项研究把这种现象叫作「模型崩溃」；也有研究指出，如果新数据是在保留原有人类数据的基础上累加，而不是替换，崩溃可以大大缓解。所以它是有条件的风险，不是必然的命运。放到家乡话上，这就意味着：如果网上越来越多的「方言内容」其实是模型模仿出来的外地腔，下一代模型学到的就离地道的家乡话更远一步。集体语言世界能

否保持差异，越来越依赖复合体里的人是否持续地把真实的、新的经验写进来。这一点对第36章讨论公共模型能力怎样成为个人发生条件，有直接意义。

◆ 六、向下一章推进

我们看到了集体语言世界怎样以相对同一性的形态发生：一个能容纳差异的稳定组织，表达习惯与推演约束在其中稳定下来，方言能否被容纳取决于复合体的选择，不同的复合体发生不同的世界。但这个世界不是静止的库存。它被使用的每一刻，都在生成新的句子——有时是从未有人写过的句子，有时是一个精巧的类比，有时是一段看似严密的推演。这些新东西是怎样产生的？它们说明模型「理解」了吗？下一章用一句再普通不过的话——「你真行」——来讨论生成与理解能力的发生，以及怎样分别检验它们。

> 带走的话
>
> 集体语言世界的相对同一性，不是千万种说法的平均值，而是能把差异组织进来的稳定纠缠；它能容纳多少差异，取决于复合体里的人怎样选择、怎样评价。

> **小账** 论证落在哪一层？——本章以本体论立场与机制假说为主，辅以可观察的模型差异等经验事实。 世界模型补了哪一条？——补了一条条件：世界模型的「同一性」应当是能容纳差异的相对同一性，不同训练条件形成不同的世界，不存在一个等待逼近的标准世界。

14

第 14 章

语言生成与理解能力的发生

> **这一章的问题**：模型能写出从没人写过的句子、打出贴切的比方、推出像模像样的结论，这些能力是怎样发生的？我们又凭什么说它「理解」了？

◆ 一、命题：生成是一次次差异运行，理解是在检验中才显露的能力

前三章讲的是训练怎样发生一个集体语言世界。这一章转到使用：这个世界怎样在每一次生成中运动起来。

先摆命题。本体论层面，本书主张：语言生成不是从库存里取出现成句子，而是在已形成的特征纠缠（E）中，一步一步地做出差异选择（D），显露出一句句话（S）；每显露一步，又成为下一步差异运行的条件。理解则不是藏在模型内部、等着被发现的一个「东西」，而是一种能力，只有在不同的检验中才显露出来，也只在检验所覆盖的范围里才站得住。

由此引出本章最要紧的一个区分：流畅表达、可靠判断、能力迁移，是三种不同的东西，需要三种不同的检验。一段话写得顺，只说明第一种；它对不对，要第二种检验；换个场合还行不行，要第三种。把三者混在一起，就会从「它说得真好」一路滑到「它真的懂了」。

◆ 二、「你真行」：同一句话的许多种发生

本编第三个主例是一句再普通不过的话：「你真行。」

周敏在家里听到这句话，可能有许多种意思。小舟第一次在泳池里憋气游过了半程，她拍着池边喊「你真行！」，这是夸奖。小舟把作业本忘在学校，第三次

了，她叹口气说「你真行」，这是数落。老彭跟她说，他把一个乘客落在车上的钱包开了二十公里送回去，人家连声谢谢都没说，她笑着说「你真行」，这里面有佩服，也有一点「何必呢」的心疼。同样四个字，意思完全取决于谁对谁说、在什么事之后说、用什么语气说。

这句话在千万人的文字里也是这样出现的：小说里、聊天记录里、影评里、吵架的帖子里。它的前文后文、引号里外的标点、紧跟着的「哈哈」或者「……」，都缠在它周围。经过训练，模型面对「你真行」这几个字时，不是调出一个固定意思，而是根据前文把这团纠缠的不同部分调动起来。

下面是我构想的场景，不是对某个版本模型的实测记录。周敏做一个小测试。她给模型一段话：「孩子第三次把作业忘在学校，妈妈看着他，说：『你真行。』请问妈妈是什么心情？」设想模型答：无奈、带点讽刺。她再换一段：「孩子第一次游完二十五米，妈妈激动地说：『你真行！』」设想模型答：骄傲、欣喜。这两次回答都合情合理，也是今天的模型通常能做到的那一类。

这说明了什么？可以说，模型在这两个情境中都能根据上下文显露出恰当的解读。用三方程的话说，前文构成了这次显露的条件，模型内部的纠缠和前文共同决定了差异运行朝哪个方向走，最后显露出「讽刺」或「骄傲」这样的判断。这是一种实实在在的能力，不必贬低。

但还不能就此说它「理解了你真行」。我们只检验了两个典型情境，而且是语料里极常见的那种。真正的检验还在后面。

◆ 三、新表达、类比与推演从哪里来

先看模型怎样产生「新东西」。

新表达。模型常常写出语料里没有一字不差出现过的句子。这并不神秘：每一步生成，都是在当前上下文条件下，从一个概率分布里选出下一个token；这些选择的组合空间极大，绝大多数长句子都是第一次被组合出来。工程上，生成时还可以调整采样方式，让选择更保守或更多样：常见的一个旋钮叫「温度」，调低了，

模型更集中地挑最可能的词，调高了，冷门的词也有机会被选中；另一种常见做法叫「核采样」，只在累计概率占到一定比例的那一小撮候选里挑。所以「新」本身不稀奇，稀奇的是新而得当。

类比。让模型「用做菜来解释惯性」，它可能会说：「就像锅里的汤，你突然把锅往前一推又停住，汤会往前涌出来。」这个比方不一定在任何地方出现过，却讲得通。它的来源，可以用一个机制假说描述：在模型内部，「惯性」和「汤在锅里晃」这两团组织，在某些抽象方向上靠得近——都和「运动的东西不愿意马上停下」有关。生成类比时，模型沿着这个共同方向，把一个领域的表达调到另一个领域。从 SDE 看，这是两团纠缠在一个共享的维度上发生了连接，差异运行借着这个连接从一处走到另一处。

推演。让模型做一道「甲比乙高，乙比丙高，谁最矮」的题，它大概率答对。它的推演能力来自人类文本里大量这样那样的推理痕迹，以及训练后期专门针对推理的数据与方法。近年的公开工作中，有让模型在作答前先写出一步步的中间推理、再给出结论的做法，在不少任务上提高了正确率。2022 年公开的「思维链」提示研究，是这类做法里最常被引用的一篇；后来又有一些模型在训练阶段就专门强化这种先推后答的习惯。从发生学看，写出中间步骤，相当于把一次大的差异运行拆成许多次小的显露，每一次小显露都成为下一次的条件，组织因此有机会被一步步调动，而不是一下子跳到结论。

生成还有一个容易被忽视的特点：它是一步接一步的，前面的选择会锁住后面的路。模型一旦在回答开头写下「红烧肉的关键在于冷水焯肉」，接下来的每一步都以这句话为条件，它会顺着这个说法往下讲理由，而很少回头推翻自己。人说话也有这种情况：话说出口，就倾向于为它圆下去。在 SDE 的语言里，这是显露反过来约束差异运行——已经显露的东西成了后续发生的条件，这本来是三方程中 D=G(S,E) 的应有之义。问题在于，当开头那一步本身是在两种冲突说法之间近乎随机地选出来的（第13章讲过这种情形），后面整段话的流畅和自信，就会给一个偶然的选择披上必然的外衣。一些公开工作让模型生成多个候选、互相比较，或者

在生成后再检查一遍，正是为了减轻这种路径锁定。比如 2022 年提出的「自一致性」方法：同一道题让模型各自推几遍，再看哪个答案出现得最多。

三者都是在已有纠缠里发生的新组合。它们是真正的发生，而不只是复读；但它们的可靠程度，取决于所依赖的那部分组织在训练中被检验得多充分。

◆ 四、语言里的物理知识：从哪里来，能走多远

本书的主题是世界模型，所以要特别问一句：语言模型里有没有物理世界的知识？

有，但要说清它的来源和性质。人类文字中充满了物理知识：教科书里的定律，说明书里的操作步骤，日常对话里「杯子放桌边会掉下去」「冬天水管会冻裂」。千万人在现实里做事、吃亏、总结，留下了这些话。训练让模型在一定程度上组织了它们，所以它能回答「把鸡蛋扔到地上会怎样」，也能解释「为什么炒糖色要小火」。

这些知识的来源，是人类的视觉、听觉、触觉经验经过语言这一层沉淀下来的显露。周敏小时候被母亲吼「糖要糊了！」，那句话后面有焦味、有颜色变化、有手忙脚乱地关火；写到网上，只剩「小火慢炒，颜色变成琥珀色就停，再炒就苦了」。模型拿到的是后者。所以，语言模型里的物理知识，是「关于物理的话」的组织，而不是在身体与物的直接互动中发生的组织。

这带来两个后果。第一，凡是人类经常用语言讲清楚的物理常识，模型往往掌握得不错。第二，凡是人类很少用语言讲、主要靠手感和眼力的东西——糖色到底什么颜色算「琥珀色」、锅铲推起来多黏算到时候——模型能说出话来，但这些话缺了现实那一端的检验。它可以描述，却没有被焦糖烫过、没有闻到过糊味。这正是第四编到第七编要处理的问题：物理世界模型为什么不能只靠语言的组织，视觉、听觉、触觉 SDE 的互动为什么要进入训练。

举个可以动手试的对照。问模型「一块糖放进热锅里，先发生什么、再发生什么」，它能按顺序讲出融化、起泡、变色、冒烟、发苦，讲得比大多数人清楚，因

为这个顺序在文字里被无数次写过。可要是换个问法：「我家是电磁炉，锅底薄，火力开到第六档，糖放下去大约多少秒开始变色？」它多半会给出一个听上去合理的数字，却没有办法知道你家那口锅到底多薄、那一档到底多热。一个在这口锅上炒过几次糖的人，答不出精确的秒数，却知道「一开始冒细泡就得盯着，别走开」。前者是关于糖的一般知识，组织得很完整；后者是关于这口锅、这个灶的具体经验，只有在现实互动里才发生。两种知识都有用，但它们的可靠范围不同：一般知识在大量典型情况下可靠，具体经验在这一处现实里可靠。语言模型擅长前者，后者需要别的发生条件。

已有一些公开研究专门测过语言模型的物理常识，大致发现：对常见的、文字中常描述的物理情境，表现可以相当好；对需要空间想象、多物体相互作用或少见组合的情境，表现明显下降。例如 2020 年公开的一套物理常识问答评测，题目都是「要达到某个日常目的，两种做法哪个行得通」，人答对的比例在九成以上，当时最好的语言模型与人有不小的差距；后来的模型把这个差距缩小了许多，但在空间关系和多物体互动一类的题目上，研究者仍在报告明显的弱点。这与上面的分析方向一致，但也只是一致，不是证明：表现下降可能有多种原因，需要逐一区分。

这里只先立一个边界：语言里的物理知识是真的知识，但它的发生条件是语言，不是现实互动；它在哪里会失效，要用现实去检验，不能用更多的语言去检验。

◆ 五、三种检验：流畅、可靠、迁移

现在把本章的核心区分写清楚。下面三种检验，测的是不同的东西。

流畅表达的检验，问的是：这段话通顺吗，合不合语言习惯，前后连贯吗？大模型在这方面的表现通常已经很强。但流畅只说明集体表达习惯被组织得好，不说明内容对。一个流畅地把「热水焯肉」和「冷水焯肉」说反了的回答，在流畅这一项上照样满分。

可靠判断的检验，问的是：这个判断对吗？对不对要用外在于这段话的东西去核：查资料、做实验、问懂行的人、看实际后果。回到「你真行」：可靠判断的检

验不是再出两道典型题，而是出模型没见过的、需要细辨的情境。比如：「老彭开了二十公里把钱包送回去，对方没道谢，周敏笑着说『你真行』。」这里既不是纯夸奖也不是纯讽刺。模型能否给出「佩服里带着心疼」这样有分寸的解读？如果它只能在「夸」与「讽」两个大类里选，那么它在典型情境里的漂亮表现，就不能推到细微情境上。

能力迁移的检验，问的是：换个领域、换种形式，这种能力还在不在？比如，模型能解读中文「你真行」的反讽，能否解读长沙话里「你蛮狠咯」——字面是「你很厉害」，用法上同样可夸可讽？能否在一个它没见过的小说片段里，识别出作者用的是哪一种语气？如果换了个说法能力就塌了，那就说明它掌握的是一个特定表达的用法分布，而不是一种可以迁移的对语气与关系的把握。〔待核：长沙话「你蛮狠咯」可夸可讽的用法，请湘方言使用者核对〕

三种检验分开之后，一个常见的推论就不成立了：「它说得这么流畅、这么有道理，所以它肯定懂。」流畅过关，只过了第一关。本书在全书反复强调，一次流畅对话不证明能力已稳定形成；在这里，这句话有了具体的检验程序。

三种检验还引出一个更根本的问题：理解到底发生在哪里？按本书的发生单位纪律，答案不能简单是「在模型里」。周敏拿到模型对「你真行」的解读，她会读，会对照自己的经验，会觉得「这个说法挺准」或者「差了点意思」，会追问，会改写后讲给小舟听。在这一来一回中，理解是在个人—AI复合体里发生的：模型显露出一种解读，周敏的经验与它碰撞，差异被看见，新的说法被一起形成。模型单独拿出来，能力是真实的；但「理解」这个词所包含的被检验、被修正、被用到现实里去，往往要靠复合体里的人来完成。这不是给模型降格，而是把能力放回它实际运行的地方。第37章会专门讨论个人与AI怎样继续发生语言世界，这里只先指出：问「模型理解了没有」，常常不如问「这个复合体在这件事上能否形成可靠的理解」更准确。

还有一点不能遗漏：检验本身也是复合体的一部分。谁来出题、出什么题、以什么为对，决定了「能力」以什么面目显露。一个只用典型情境出题的测试，会让模型显得比它实际更懂「你真行」；一个专挑刁钻情境出题的测试，又可能让它显

得比实际更笨。所以三种检验的设计，本身就需要公开、可复现，并且说明它覆盖什么、不覆盖什么。能力的边界，和检验的边界，是一起画出来的。

◆ 六、边界与推进

本章也要防两种偏。一种是低估：说模型「只是在猜下一个词」，所以什么能力都没有。这混淆了训练目标和训练结果——目标是猜下一个词，为了猜准而形成的组织，可以支撑真实的表达、类比和推演能力，这些能力在许多检验中确实站得住。另一种是高估：从流畅和局部成功，直接推到「它理解了世界」。本章的三种检验正是用来挡住这种跳跃的。

这一章停在「能力怎样发生、怎样检验」。可是检验的结果不会总是好的：有些地方模型说得斩钉截铁却是错的，有些地方它的判断带着明显的偏向，有些地方它前后自相矛盾。这些失败从哪里来，怎样修正，修正的边界在哪里？最后一个问题更大：语言模型这么成功，是否说明人脑就是这样工作的？下一章收束本编，回答这些问题。

> **带走的话**
>
> 生成是在集体纠缠中一步步做出的差异运行，新表达、类比、推演都是真实的发生；但流畅、可靠、迁移是三种不同的能力，必须用三种检验分开来量。

小账 论证落在哪一层？——本章以机制假说为主，辅以可观察的模型表现，并给出可操作的检验区分。　　世界模型补了哪一条？——补了一条边界：语言中的物理知识以语言为发生条件，它的可靠性必须用现实互动来检验，这是语言模型通往物理世界模型时必须跨过的一道坎。

15

第 15 章

集体语言世界的边界与修正

> **这一章的问题**：集体语言世界在哪里会偏、会错、会自相矛盾？它能被怎样修正？它的成功又能证明什么、不能证明什么？

◆ 一、命题：边界是发生条件留下的形状，修正是纠缠的重组

前四章讲了集体语言世界怎样发生、怎样运行。这一章讲它的边界。本书的立场是（本体论层面）：一个世界的边界，不是它「还没学完」的地方，而是它的发生条件所留下的形状。数据从哪里来、经过了怎样的选择、被什么样的目标和评价塑造，就决定了这个世界在哪里站得稳、在哪里会偏、在哪里会断。

所以，修正不是往一个「不完整的库存」里补货，而是改变发生条件，让特征纠缠重新组织。补数据、改目标、换评价、加外部检验，都是修正的手段；但每一种手段本身又是新的条件，会留下新的形状。修正没有终点，只有一轮又一轮的重组。这与第5章所讲的「失效、解封与重新训练」是同一个道理在集体层面的展开。

本章依次处理四件事：偏向从哪里来；冲突、幻觉与失配怎样区分；继续训练怎样重组纠缠、有什么边界；以及最后一件——语言模型的成功能不能证明人脑机制被复制了。

◆ 二、数据选择和评价标准留下的偏向

偏向的第一个来源，是数据选择。第11章说过，进入训练的是「可写下、已写下、被留下」的那部分人类活动，再经过过滤和配比。这几道筛子每一道都会留下偏向。

拿家乡话来说。前几章已经看到，方言材料少、写法乱、容易被当成噪声。结果是：模型对方言的理解和生成，通常明显弱于普通话；它说起方言来，常常像一个只在电视上听过几句的外地人。更深一层的偏向，可以用一个我构想的场景来说明：老彭用夹着长沙话的句子问模型「车子冇得劲，么子原因」，模型先把这句话「纠正」成普通话复述一遍，再按普通话来回答，甚至提醒他「冇得」不规范。哪个模型、哪个版本会这样做，要实际测了才知道，这里不作断言；但如果真的这样，这不是哪个工程师有意歧视方言，而是多数人的说法在训练中天然占据了核心位置，少数人的说法被挤到了边缘。

偏向的第二个来源，是评价标准。偏好训练教模型说「更好的」回答，可什么叫更好，由写规范和做比较的人决定。如果标注员大多来自某几个地区、某种教育背景，他们认为得体的说法、礼貌的分寸、合理的论证方式，就会被放大成模型的默认风格。「你真行」这句话就是个好例子。在一些语境里，直白的反讽被视为不礼貌；如果评价标准倾向于让模型回避任何可能冒犯的解读，它就可能在分析「你真行」时，倾向于往夸奖那边解释，把讽刺说得很轻。这种偏向不是错误，是一种被训练出来的姿态，但用户需要知道它存在。

偏向的第三个来源，是目标本身。预训练只对「下一个token猜得准」负责。网上说得多的话，猜它就准；说得多不等于说得对。一个流传很广的厨房误区，比如某种「焯水秘诀」被千万人转载，它在训练中的分量会远大于一位老厨师偶尔写下的纠正。目标没有错，但它决定了模型天然偏向「常见」而不是「正确」。后续的评价训练会部分地纠正这种偏向，但纠正得了多少，取决于打分的人自己懂不懂、查不查。

用三原理来看：偏向是多数人的显露稳定态和少数人的差异之间，张力没有得到回写的结果。只要复合体里的评价机制不承认那份差异，它就被压在底下；一旦被承认并回写，纠缠就有机会改变形状。偏向因此是可修正的，但不会自己修正。

◆ 三、内部冲突、幻觉与失配：三种不同的错

模型出错的方式不止一种。把它们分清楚，修正才有方向。

内部冲突，是集体语言世界里本来就存在的矛盾在一次回答中露了出来。第13章说过，冷水焯肉还是热水焯肉，网上各执一词；如果两种说法没有被清楚地区分开，模型就可能今天说这个、明天说那个，甚至在同一段回答里前后不一。这种错的根源是集体经验本身的分歧没有被组织好。

幻觉，在这里指模型说出了在它的语料和上下文中都找不到依据、却以确定口吻呈现的内容。比如编造一本不存在的菜谱书名、一位不存在的名厨、一个看起来像模像样的出处。从机制上看，可以提一个假说：生成时，模型沿着表达习惯和语境线索一步步走，每一步都「像」该出现的样子，当走到一个它内部组织并不充分支持的地方时，表达习惯仍然推着它往下说，于是生成了形式完整、内容无据的东西。2023 年以来已有多篇综述梳理幻觉的成因，归纳的方向大致包括：训练数据本身的错误与缺口、训练目标只奖励「像」而不奖励「真」、解码时的随机性，以及评测与偏好训练对「敢答」的奖励多于对「不知道」的奖励。上面的假说只是其中一条。用 SDE 的话说，这是显露在差异运行的推动下走得比纠缠所能支撑的更远——形式上的连贯代替了内容上的扎根。第14章说的路径锁定，会让这种情况更严重：开头一句编出来了，后面就只能顺着编下去。

失配，是指模型的回答在语言内部看是合理的，放到用户的现实处境里却不对。还是红烧肉：用户说「我在高原，按你说的炖了一个小时，肉还是硬的」。模型给的做法在平原上大体没问题，但它没有考虑高原上水的沸点较低这个现实条件。这里语言世界本身没有冲突，也没有凭空编造，错在它和这一处现实没对上。失配最能说明集体语言世界的根本边界：它是从文字里发生的，而现实永远比文字

多一些。第14章所说的「语言里的物理知识要用现实来检验」，在这里落到了实处。

三种错的修正方向不一样。内部冲突要靠把分歧组织清楚：标明条件、区分语境，必要时由懂行的人判定。幻觉要靠给生成接上外部依据：检索资料、引用来源、学会说「我不确定」。失配要靠把现实的信息带进来：用户告诉模型自己的处境，或者系统能接触到现实的反馈。把三者混为一谈，统称「模型会犯错」，就找不到下手的地方。

对普通使用者来说，分辨这三种错也有一些朴素的办法，不需要懂技术。这是我构想的场景：一个周末，周敏在家炖红烧肉，手机支在灶台边。她先问模型「高压锅为什么炖肉快」，回答清楚准确，和她教的沸点知识对得上。她又顺口问「有没有哪本书专门讲厨房里的物理」，模型给了三个书名和作者。她关了火去查，发现其中一本确有其书，另外两本怎么也搜不到——这是幻觉，靠外部核查发现。她再问「冷水还是热水下锅焯肉」，前后问了两次，得到两个相反的建议——这是内部冲突，靠重复提问发现。最后她想起表妹在高原工作，常抱怨肉炖不烂，就替表妹问「在海拔三千多米的地方炖肉要多久」，模型给的时间和平原上一样——这是失配，靠她自己对现实处境的了解发现。三种错，三种发现方式：核查出处、换次再问、对照现实。这三种做法，恰好对应着上文三种修正方向。使用者在这里不是被动的受害者，而是复合体里承担检验的那一方；她每一次核查、追问、对照，都在为这个复合体补上它缺的那一维。

当然，也不能把检验的担子全推给使用者。一个对孩子、对不熟悉的领域的人开放的系统，不能指望每个人都像周敏那样去查。所以，系统层面的修正——给出处、标明不确定、在条件不明时先问清楚——同样是复合体的责任。哪一种负担放在哪一方，本身又是一个复合体需要协商的发生条件。

◆ 四、继续训练与纠缠重组

模型发布之后，修正以多种方式进行。在对话里纠正一句（「湘菜红烧肉一般不放酱油」），换个窗口它可能又犯；接上一个可靠的菜谱库去检索，拔掉库，能

力就回到原样。这两种是暂时语境变化与外部记忆，模型自身的组织没有变（五件事的区分见第37章）。本章关心的是更深的两层。

一层是参数训练。把大量纠正过的回答、新的数据、新的评价收集起来，在原模型上继续训练，或者干脆训练新一版。这才是集体语言世界在纠缠层面的重组。而重组从来不是「只改坏的那一块」：参数是共用的，调整一处可能牵动别处。工程上有一个公开讨论较多的现象：在新数据上继续训练，可能让模型在旧任务上的表现变差，常被称为「灾难性遗忘」。这个名字最早出自 1989 年前后对小型神经网络的研究；后来有许多缓解办法，例如训练新任务时保护对旧任务重要的那些参数，或者把一部分旧数据混进新数据一起「复习」。大模型的继续训练里，这个问题以较轻或较重的形式依然存在。从 SDE 看，这毫不奇怪——纠缠是一个整体组织，回写会改变它的形状，而形状改变总会波及原来稳定的部分。

另一层是持久学习：修正后的能力在以后各种情形里稳定地起作用，并且不以牺牲其他能力为代价。这是最难验证的一层。一次修正后模型在测试题上答对了，不等于它在以后的千变万化中都会答对。验证持久学习，需要回到第14章的三种检验：不仅看流畅，还看可靠与迁移，并且要隔一段时间、换一批情境再看。

修正还有一个更大的边界：修正本身也是复合体的选择。谁来判定哪个说法是「对的」？冷水焯肉和热水焯肉，也许本来就各有道理；如果修正时强行定于一尊，就是用一种说法压掉另一种。家乡话也一样：把方言「纠正」成普通话是一种修正，把模型训练得更懂方言是另一种修正，方向完全相反。修正不是技术中性的，它决定了集体语言世界下一步往哪里重组。第36章会进一步追问：怎样避免把同一性训练变成单一答案的强制。

◆ 五、语言模型的成功不证明人脑机制被复制

最后，本编要明确一件事：语言模型的成功，不证明人脑机制已经被复制。

这句话需要说清楚理由，否则就只是一句立场。

第一，相似的显露不等于相同的发生机制。模型能说出「你真行」该怎么理解，周敏也能；两者的显露相似。但周敏的理解里，缠着她当老师二十年见过的无数孩子、她自己小时候被母亲数落的记忆、她说这句话时脸上的表情和语调；模型的组织来自文字。从显露相似跳到机制相同，正是第20章要专门讨论的「推论缺口」。一个孩子和一台计算器都能算出七乘八等于五十六，没有人会因此说计算器复制了孩子的大脑。

第二，发生条件根本不同。孩子学说话，是在身体、家人、物与事件的持续互动中，一边被抱着、喂着、逗着，一边听着、学着、说着。第二编讲婴儿认妈妈、小舟被烫了一下才懂「热」，讲的就是这种发生。语言模型的语言世界，主要来自已经写下的文字，以及之后的评价训练。人用的语言量远远少于模型的训练语料：2023 年一位发展心理学家做过估算，一个孩子到十来岁听到的词，大约是几千万到一亿多个的量级，而当时的大模型训练用到的是万亿个token上下，差了三到五个数量级；可孩子却在身体与现实中发生出可靠的语言能力；模型用海量文字发生出另一种语言能力。两条路径的起点、材料、反馈都不一样。

第三，工程上的「神经网络」只在很弱的意义上借了大脑的比喻。人工神经网络中的「神经元」和「连接」，是数学运算单元，与生物神经元的工作方式差别很大；训练所用的反向传播算法，在大脑中是否以及如何有对应，学界至今有不同看法：2020 年有神经科学家在综述里主张，大脑可能以某种近似的方式实现了与反向传播功能相当的东西，也有不少研究者认为二者的差别是根本的。名字相近，不能当作机制相同的证据。

第四，这条边界是双向的。不能用模型的成功证明「人脑就是这样的预测机器」，也不能用模型的局限证明「人脑一定有某种机器永远做不到的东西」。这两个结论都超出了现有证据。本书的立场是：人和模型都在发生，发生条件不同，发生出来的世界不同；它们在人类—AI复合体里相遇，互相成为对方的发生条件。至于两者在机制层面究竟有多少共通之处，是一个需要神经科学、认知科学与工程研究一起去回答的经验问题，不能由哲学立场直接判定，更不能由一台模型的表现直接判定。

还要回到发生单位的纪律上来。说「模型复制了人脑」，等于把模型看成一个可以和人脑一对一比较的独立个体。本编五章一路说下来，模型的语言世界从来不是一台机器单独发生的，它是集体人类—AI复合体的发生。拿一个复合体的产物去和一个人的大脑比，比较对象本身就错了位。

◆ 六、向下一编推进

本编到这里收束。我们看到：人以多种方式进入训练；token、预测误差与参数连接是工程载体，特征纠缠是本体论解释，两者不能互相证明；集体语言世界以相对同一性的形态发生，能容纳多少差异取决于复合体的选择；生成与理解能力是真实的，但要分开检验流畅、可靠与迁移；这个世界有它的偏向、冲突、幻觉与失配，修正是持续的纠缠重组，而它的成功不证明人脑被复制。

所有这些问题里，失配最尖锐地指向了下一编：语言世界从文字发生，现实永远比文字多。高原上的那锅红烧肉，提醒我们文字之外还有一个会给人吃苦头的世界。那么，想让机器直接从现实中学习的「世界模型」，又是以什么为世界？视频生成、状态预测、空间重建和行动模型，各自把什么当作世界的对象？第四编从这里开始。

> 带走的话
>
> 集体语言世界的边界由它的发生条件塑造，偏向、冲突、幻觉与失配要分开对待，修正是一轮又一轮的纠缠重组；语言模型的成功，不证明人脑机制已被复制。

小账 论证落在哪一层？——本章以本体论立场与机制假说为主，辅以继续训练、灾难性遗忘等经验事实。 世界模型补了哪一条？——补了一条边界：从显露相似不能推出机制相同；任何模型（包括世界模型）的成功，只证明它在其检验范围内有效，不证明它复制了人或世界本身。

第 四 编

世界模型的技术路线与哲学前提

编序 前三编讲的是人怎样学习、三个世界怎样发生、人类与AI怎样在训练中共同发生一个语言世界。到这一编，问题换了地方：当工程界说要造「世界模型」，他们手里实际在做的是什么？这一编先把现有路线摆上桌面，一条一条看公开机制。

第16章先问一个看似简单的问题：各家说的「世界」到底指什么。视频生成器、状态预测器、空间重建和行动模型，各有各的对象和成功标准，混在一起谈，就会把「画面像」误当成「懂世界」。第17章看从视频里学世界的路线，以Genie和Runway为主要对话对象，追问录像里的互动和新的互动有什么不同。第18章转向另一条路：不去还原画面，而是预测内部状态、学习行动的后果，对话对象是V-JEPA系列和DreamerV3。第19章看多模态与空间模型，包括Cosmos、World Labs的Atlas以及Waymo公开的驾驶基础模型，重点区分「多传感器融合」和「多种SDE纠缠」。第20章收束：发现学前提在哪些具体训练机制上限制了世界模型，又有哪些机制已经带着发生的取向。

本编的日常主例有三个：看录像学游泳和真下水，天气预报，导航软件。技术系统在这里是被检验的对象，不是用来证明SDE的证据。每一章都把三种话分开说：哪句是本体论立场，哪句是可以被实验推翻的机制假说，哪句是公开文献里查得到的事实。公开资料没说清的地方，照实标〔待核〕，不替任何公司补全。

16

第 16 章

世界模型究竟以什么为世界

> **这一章的问题**：当不同的工程团队都说自己在做「世界模型」时，他们各自把什么当成了世界？

◆ 一、同一个词，四种对象

「世界模型」这个词近几年热得发烫。打开新闻，视频生成公司说自己在做世界模型，机器人公司说自己在做世界模型，自动驾驶公司也说自己的底座是世界模型。词是同一个，手里的东西却很不一样。

先用一个大家都熟的例子把差别摆出来：天气预报。

这是我构想的场景。周敏在长沙的中学教物理，有一年她带学生做一个小课题，题目就叫「天气预报到底预报了什么」。学生们很快发现，同一个「明天的天气」，至少有四种不同的做法。第一种是把过去几小时的卫星云图连成动画，再往后续几帧，让人看见云团大概会往哪里飘。第二种是气象部门的数值预报，它不画图，而是把大气切成许多格子，每个格子里记温度、气压、湿度、风速，按物理方程一步步往前推，推出来的是一组数，图是后来画上去的。第三种是雷达和探空资料拼出来的立体结构，告诉你此刻云有多厚、雨区在几千米高处。第四种最朴素：手机推送一句「明天下午有雨，出门带伞」。它根本不在乎云长什么样，只在乎你要不要带伞。

这四种做法对应着世界模型的四类对象。第一类是视频生成器，它的对象是画面序列，问的是「下一帧长什么样」。第二类是状态预测器，它的对象是某种内部状态，问的是「下一时刻的状态是什么」，状态可以是物理量，也可以是神经网络

里一组看不见的特征。第三类是空间重建，它的对象是三维几何，问的是「这个空间从别的角度看是什么样、东西在哪里」。第四类是行动模型，它的对象是「做了某件事之后会怎样」，问的是「我该怎么做」。

这四类并不互相排斥，真正的系统往往兼而有之。但把它们分开，第一个好处就出来了：我们可以问清楚每一个系统以什么为世界。

类别	以什么为世界	核心问题	常见成功标准	本编对话对象
视频生成器	画面序列	下一帧长什么样	画面质量、时间连贯、可控性	Genie系列、Runway GWM-1（第17章）
状态预测器	内部状态或特征	下一时刻的状态是什么	预测误差、下游任务表现	V-JEPA 2／2.1（第18章）
空间重建	三维几何与视角	从别处看是什么样、东西在哪	新视角吻合、几何与深度误差	World Labs Atlas（第19章）
行动模型	动作及其后果	做了之后会怎样、该怎么做	任务成功率、失败代价	DreamerV3、V-JEPA 2-AC、Waymo（第18、19章）

表里最后一列只是本编的讨论安排，不是给这些系统定性。比如Cosmos 3的公开摘要把语言、图像、视频、声音和动作放进同一个框架，它横跨好几行；Waymo把自己的基础模型直接称作「世界模型」，但它的产出是对其他道路使用者行为的预测和本车的轨迹，落在状态预测与行动模型之间。分类的用处，是让我们在谈论某个系统时先把它真正的对象说出来，而不是一律称为「懂世界」。

这里还要补一句历史。「世界模型」在人工智能里不是新词。强化学习里很早就有「先学一个环境模型、再在模型里试」的想法，控制论和认知科学里也有「大脑里有一个外部世界的内部模型」的说法。近几年这个词重新热起来，主要是因为两件事同时发生：一是大规模视频生成的画面质量突然变得很高，二是机器人和自动驾驶对「在脑子里先试一遍」的需求越来越迫切。两股需求合用一个词，词义就

被撑宽了。这一段是对公开讨论的概括，具体由谁在何时首先提出哪一种说法，本书不做考证〔待核〕。

◆ 二、各类模型的成功标准

对象不同，成功标准就不同。这一节说的是经验事实层面的东西：工程界实际怎样评价这些系统。

视频生成器的成功标准，主要是画面质量和时间上的连贯：画面清不清楚，物体会不会无故消失，镜头转回来之后场景是否还在原处。这些可以用人工评分，也可以用一些统计指标去量画面分布和真实视频分布有多接近。状态预测器的成功标准，是预测出来的状态和后来真实观测到的状态差多少，或者用这组状态去做下游任务（识别动作、预判下一步）做得怎么样。空间重建的成功标准，是新视角下的画面与真实拍摄的吻合程度、几何误差、深度误差。行动模型的成功标准最直接：任务做成了没有，成功率多少，失败时损失多大。

回到天气预报。卫星云图动画续得再顺，如果第二天下午没下雨，你被淋了也只能认了；数值预报算得再精细，如果它从来不告诉你要不要带伞，对你出门这件事也帮不上忙；而那句「带伞」的推送，即便背后的云图画得粗糙，只要它在你所在的街区、在你出门的时段大多数时候说对了，你就会信它。

这里有一个容易被忽略的事实：成功标准本身决定了模型被训练去保留哪些差异。以画面逼真为标准，模型就会在画面细节上花力气；以带伞决策为标准，模型就会在「下午三点到六点、这条街、降水概率是否越过某个值」上花力气。标准不同，被训练出来的那个「世界」就不同。用SDE的话说，训练目标决定了哪一种差异运行（D）被反复强化，哪些特征被缠到一起（E），最后显露（S）成什么样子。这一句是本体论立场与机制解释的交界处：本体论层面，我们主张显露总是在纠缠与差异运行中形成的；机制层面，「训练目标塑造保留的差异」是可以拿具体模型去检验的说法。

◆ 三、统一架构不等于全能世界模拟

近两年有一个明显趋势：越来越多的团队用同一套网络结构，同时处理文字、图片、视频、声音乃至动作序列。第19章会细讲其中几条路线。这里先把一个概念上的区别说清楚：统一架构和全能世界模拟是两回事。

统一架构是工程上的事实陈述：同一套参数、同一种计算方式，可以接收和输出多种形式的数据。这件事做到了就是做到了，有公开报告可查。全能世界模拟是一个更大的主张：这个模型已经把世界整体装了进去，任何情境、任何对象、任何干预，它都能给出可靠的后果。

从前者推不出后者。打个比方，一台多功能料理机可以榨汁、绞肉、和面，这说明它的电机和接口是通用的，不说明它能做出一桌合格的湘菜。能不能做好剁椒鱼头，取决于具体的刀工、火候和那条鱼。同样，一个统一架构能接收触觉信号，不等于它已经形成了在某个任务里可靠使用触觉的能力。

反过来，也不能从「现在还没有全能世界模拟」推出「统一架构必然失败」。本书在导论里就立过规矩：「统一架构必然失败」不是可以不经证据就接受的前提。我们能说的是：统一架构给出了一种工程上的可能性；它在哪些任务上形成了可检验的能力、在哪些地方失效，要一项一项看。

为什么人们容易从前者滑到后者？一个原因是语言上的方便。我们说「这个模型能处理视频、声音和动作」，听起来就像「这个模型理解视频、声音和动作所在的世界」。但「能处理」只是说数据能进能出，「理解」至少还要求：在新的情境里，它给出的后果能被真实世界核对，而且大多数时候核得上。中间隔着的，正是本书一再强调的那件事：能力要在具体任务里发生、稳定，并经过边界检验，才算封装好了。第5章讲过能力封装的这层意思，这里只是把它用到世界模型身上。

四、发生学审查的入口

那么，一个世界模型摆在面前，SDE发生学要从哪里审查它？本章给出四个入口问题，后面四章会反复用到。

第一问：它以什么为世界？也就是上面说的对象。是画面、状态、几何，还是行动后果?

第二问：它的训练让什么缠到了一起？画面序列模型里，缠在一起的主要是前后帧的像素或压缩码；状态预测模型里，缠在一起的是某种表示与其未来；行动模型里，缠在一起的是状态、动作和后果。这一问对应随身三问的第一问：这里纠缠的是什么。

第三问：差异是怎么运行的？模型在训练时碰到的差异，是只来自已经录好的数据，还是也来自它自己做出的新动作？这对应第二问：差异是怎样运行的。

第四问：显露出来的东西靠什么站稳，又在哪里失效？一个模型生成的画面或给出的预测，是靠什么被检验的？是人看着觉得像，还是与真实后果比对过？它在什么条件下会出错，出错了能不能被发现?

不妨拿天气预报把四问试走一遍。卫星云图续帧：以画面为世界；缠在一起的是前后几帧云的形状与位置；差异只来自已经拍下的云图；它站得住的地方是短时间内云团的大致移动，失效的地方是云在原地生成、消散这类画面里看不出成因的变化。数值预报：以格点上的物理量为世界；缠在一起的是温度、气压、湿度和风之间的方程关系；差异来自观测资料不断灌入、模型不断被校正；它站得住，是因为每天都有真实天气来核对它，而它在小尺度对流、山区局地天气上常常失手。手机上的带伞提醒：以「你在此时此地要不要带伞」为世界；缠在一起的是降水概率、位置、时段和你的行程；差异来自一次次说对说错；它站得住靠的是长期的命中率，失效时你会直接被淋。

四问走完，三种预报的差别不在「谁更聪明」，而在它们各自把什么当世界、靠什么被纠正。尤其值得注意的是第三问：数值预报之所以比单纯续云图可靠，很

大程度上不是因为它画得更好，而是因为它每天都被真实天气纠正，纠正的结果又回到下一轮预报里。差异运行不是单向地从数据流向模型，而是在预报与天气的反复碰撞里一圈一圈地转。这一层，后面讲Dreamer的经验回放、讲Waymo的外部学习回路时会再遇到。

这四问不是给模型打分，而是把它放回发生条件里看。周敏的学生后来在课题报告里写了一句很朴素的话：「天气预报不是天气本身，每一种预报都是在一种用途里站得住的一个说法。」这句话放在世界模型上同样适用。

◆ 五、反例与边界：不要把分类当本体

本章的分类也有边界，需要提前说破。

其一，四类对象的划分是为了审查方便，不是本体论断言。并不是说世界本身分成「画面的世界」「状态的世界」「几何的世界」「行动的世界」。SDE的立场是，世界是在纠缠中经差异运行而显露的，分类只是描述工程系统对象的一把尺子。

其二，一个系统被归入视频生成器，不意味着它内部没有学到任何状态或几何。一个只以下一帧为训练目标的模型，内部可能形成了某种近似物体位置、深度或运动的表示。这是机制层面可以检验的问题，不能凭分类直接否定。反过来，一个系统自称在建世界模型，也不意味着它内部真形成了可靠的物理状态，同样需要检验。

其三，天气预报的例子有它的限度。大气是一个已经有成熟物理方程的系统，数值预报有几十年的科学基础；而机器人拿起一个陌生的软球、汽车在雨夜遇到一块没见过的路面，这类情境往往没有现成的、可直接计算的方程。所以「状态预测器」在天气上和在日常操作上，面对的困难完全不同。这一点第26章讲物理发生的信息负担时还会展开。

其四，成功标准不是越单一越好。只看任务成功率，可能掩盖模型靠的是某种不可靠的捷径；只看画面质量，又可能让人误以为它懂了物理。世界模型账要记的，恰恰是：显露在什么条件下站得住。

其五，也是最容易被忘掉的一条：谁定的成功标准？表面上看，世界模型是一台机器在学世界；实际上，对象怎么选、数据从哪里来、评价看哪些指标、哪一版能上线，每一步都是人定的。天气预报里，「降水概率超过多少就提醒带伞」不是大气自己决定的，是气象部门和软件团队按照用户会不会嫌烦、漏报的代价有多大定下的。世界模型的训练也一样：选择互联网视频还是机器人录像，以画面像不像为准还是以抓取成没成为准，这些选择里沉淀着工程师的判断、公司的目标、行业的惯例，以及拍下这些视频的千千万万个人的活动。

所以，按本书统一的发生单位，本编讲训练时说的是「集体人类—AI复合体」，不把模型写成独自面对世界的发现者。这一点不是客气话，它直接影响我们怎样读技术报告：一个模型学到了什么，一半要看网络结构和训练算法，另一半要看人替它挑了什么世界、定了什么标准。后面四章每介绍一条路线，都会把这两半一起摆出来。

其六，分类会随技术变。今天的视频生成器明天可能加上动作接口，今天的空间重建明天可能加上物理仿真。本章的四类不是永久的格子，而是2026年这个时间点上审查用的一把尺。哪天格子不合用了，就该换尺，而不是把系统硬塞进格子。

◆ 六、向下一章推进

四类对象里，视频生成器最直观，也最容易让人觉得「AI已经看懂了世界」。一个模型能接着一段游戏录像生成下一秒，能让你用方向键在里面走来走去，看上去它不就是掌握了这个世界的规律吗？

下一章就从这条路线开始。我们会看到，从视频里学世界，确实学到了真实的东西；但录像里记下的互动，和你亲自下场时碰到的互动，是两种不同的差异运行。这个区别，用一个日常例子说最清楚：看录像学游泳，和真下水。

带走的话

说「世界模型」之前，先问它以什么为世界；对象不同，成功标准不同，被训练保留下来的差异也不同。统一架构是工程事实，全能世界模拟是另一个需要逐项检验的主张。

小账 论证落在哪一层？——本章以本体论立场与概念区分为主，辅以各类模型成功标准的经验事实描述。　世界模型补了哪一条？——补上第一条审查条件——任何世界模型都要先说明自己的对象与成功标准，并区分统一架构与全能世界模拟。

17

第 17 章

从视频中学习世界的机制

> **这一章的问题**：一个只看过海量录像的模型，学到的是录像里的世界，还是可以下场行动的世界？

◆ 一、看录像学游泳

这是我构想的场景。周小舟九岁，这个暑假在学游泳。开课前一周，他在平板上看了不少教学视频：蛙泳怎么收腿、怎么蹬夹、换气时头抬多高。看得多了，他能把动作讲得头头是道，还能在床上趴着比划。周敏问他：「你觉得下水能游吗？」他说：「差不多吧。」

第一次下水，情况完全不一样。水比想象的凉，一蹬腿身体就往下沉，抬头换气的那一下呛了一口水。视频里那个人的手臂划下去，身体就往前走；他的手臂划下去，身体只是晃。几天以后，他才慢慢找到一种感觉：手划水的时候能「抓住」一点东西，腿蹬出去要等身体漂平了才有用。这些感觉，视频里一个字也没说，画面上也看不出来。

这个例子说的不是「看视频没用」。看视频确实让小舟知道了动作的大致顺序，知道了好的蛙泳看上去是什么样子，下水后他也比完全没看过的孩子更快找到方向。问题在于，录像里记下的是别人的互动，那些互动已经发生完了；小舟自己下水时碰到的，是新的互动，水对他的身体、他的力气、他的紧张作出的回应。前者给了他画面上的显露，后者才让他的身体和水真正缠到一起。

本章的命题是：从视频中学世界，能学到录像里互动所留下的差异结构，这是真实的收获；但录像中的互动和新的互动是两种不同的差异运行，前者不能自动替代后者。这是一句机制层面的判断，下面用公开的技术路线来检验它。

◆ 二、Genie：画面压缩、隐动作与条件生成

先看一条公开机制比较完整的路线。2024年，谷歌DeepMind的团队发表了Genie的论文（arXiv 2402.15391）。它的做法可以分三块讲。

第一块是画面压缩。论文用一个时空视频分词器，把每一帧画面压缩成一组离散的编码。这一步相当于把连续的像素变成有限词表里的「词」。论文附录给出的设置是一个含1024个码的码本。压缩的意义在于：后面的模型不必在每个像素上用力，而是在压缩后的码上学前后变化。

第二块是最有意思的隐动作模型。训练用的是互联网上没有任何动作标注的视频，主要是二维平台类游戏的录像。论文报告，原始收集约五千五百万段视频、约二十四万小时，过滤后保留约六百八十万段、三万多小时。录像里有人在操作，但录像本身不告诉你那一刻按的是哪个键。Genie的办法是：给模型看前面所有帧和下一帧，让它自己推断「中间发生了一个什么动作」，并把这个动作压到一个很小的离散集合里。论文里这个集合只有8个隐动作。这个隐动作模型配有一个解码器，解码器只是为了在训练时提供信号；到了使用阶段，隐动作模型被换掉，由用户直接从这8个动作里选一个输入。

第三块是动态模型，即在给定过去的画面编码和一个隐动作的条件下，预测下一帧的编码。论文用的是一种叫MaskGIT的逐步填空式生成方法。整个系统的规模，摘要里报为110亿参数。

三块合起来，就得到一个能「玩」的东西：给它一张图片，哪怕是一张手绘草图，再一个接一个地输入隐动作，它就一帧帧生成出一个可以操作的二维小世界。论文还做了一个机器人方向的实验：把机器人数据当作没有动作标签的视频来训练，也学到了一组隐动作。

用SDE的话来读这套机制。缠到一起（E）的，是画面编码在时间上的前后关系，以及从这些关系里抽出的一小组「变化类型」。差异运行（D）来自录像里一帧到下一帧的变化，模型的训练就是反复处理这些已经记录下来的差异。显露（S）是生成的下一帧画面。这三者在训练中互相塑造：隐动作的划分决定了哪些画面变化被看作「同一种动作」，画面预测的误差又反过来调整隐动作的划分。这是三方程在一个具体系统里的近似读法，属于解释，不是对模型内部的测量。

「8个隐动作」这个数字值得多停一下。为什么是8个，不是80个、800个？论文的选择有工程上的考虑：动作集合小，人才好操作，模型也更容易学出可区分的动作。但它同时意味着一个强约束：录像里千变万化的操作，都要被压进8个类别里。平台游戏本来按键就少，这个约束大体合身；换成小舟在水里的动作，手掌的角度、蹬腿的时机、呼吸的节奏，就远不是8个类别装得下的。这里能看出一件发生学上很要紧的事：动作空间的粒度，是人在训练之前替模型决定的。决定了粒度，也就决定了哪些差异会被当作「同一个动作」而被抹平，哪些差异会被保留下来。这不是Genie的缺陷，而是任何训练都绕不开的选择；第28章讲「任务决定需要保留的差异、粒度和时空范围」时，会把这件事放回任务里看。

论文还报告了一个用法：学到的隐动作空间可以用来让智能体模仿它没见过的视频里的行为。这条线索很重要，因为它把「从录像里归纳动作」和「拿这些动作去做事」接了起来。但从摘要的表述看，这仍是在录像与生成环境之间完成的模仿，差异还是从录像里继承来的；它在真实物理环境里的可靠程度，摘要没有给出可以下判断的证据。

再说一句「条件生成」。Genie生成的第一帧可以来自文字生成的图片、手绘草图或者照片，之后每一帧都以过去的画面和当前的动作为条件。条件越多，生成就越受约束；条件越少，模型就越要靠训练时缠在一起的东西来「补」。草图里没画出的地面材质、人物重量，模型只能按它在平台游戏录像里见过的样子补上去。补得像，说明录像里的差异结构被学得好；补得是否对，要拿新的互动去核，而草图本身没有可核的对象。

◆ 三、Runway与后续路线：公开到哪一步

再看另一类路线。2025年12月，Runway公布了它的第一个「通用世界模型」GWM-1。公开页面说，这是一个建立在其视频生成模型Gen-4.5之上的自回归模型，逐帧生成、实时运行；使用者可以用相机位姿、事件、机器人位姿和语音来控制它。它分成三个后训练版本：生成可漫游环境的Worlds、驱动对话角色的Avatars、面向机器人操作的Robotics，页面说今后打算合并到一个底座上。

但页面没有公开的东西同样多：参数量、训练数据的来源与数量、算力、动作条件是怎样接进模型的、「实时」对应的帧率和延迟，以及支撑「最先进」说法的量化评测。页面一处说GWM-1是自回归的，另一处说Robotics版本使用视频扩散主干，两者怎样配合也没有说明〔待核〕。

DeepMind后来推出的Genie 3也属于这一类。据公开报道，它可以根据文字提示实时生成可交互的环境，场景能在一段时间内保持前后一致，用户还可以用文字往世界里插入事件，比如改变天气或加入物体；报道提到的局限包括可做的动作有限、单次互动时长有限、多个智能体同场还不可靠。至于它用了什么数据、什么结构训练出来，公开报道没有给出。本书手头只有二手报道，分辨率、帧率、一致性时长、发布时间等具体参数一概不写〔待核：以DeepMind官方页面复核上述功能与局限〕。

这里出现一个重要的不对称：第一代Genie有完整论文，机制可以逐块讨论；后来的产品化系统，能看到的往往只是演示视频和功能描述。这不是指责，商业公司有它的理由；但它决定了我们能说什么、不能说什么。

◆ 四、未公开技术信息的证据边界

面对没有公开机制的系统，本书守三条规矩。

第一，演示不是机制。一段漂亮的视频说明系统能产出这样的画面，不说明它是怎样产出的，也不说明它在演示之外的情境里会怎样。周敏在课堂上常对学生说：「实验报告要写步骤，光贴一张结果照片不算。」

第二，不替公司补全。页面说「实时」，我们就写它说实时；页面没给帧率，我们不去猜一个帧率。页面说可以用机器人位姿控制，我们不推断它内部一定有某种物理状态。

第三，不从沉默推出否定。页面没说它用了真实互动数据，不等于它一定没用。所以本书不写「这些模型都没有互动」这类话。我们只能说：在公开信息里，能确认的是什么；没有公开的部分，判断要悬置。

这三条规矩本身就是一种发生学的纪律：一个关于技术系统的判断，也要在具体证据与具体比较里才站得住；证据缺了一块，判断就该停在那一块之前。

◆ 五、记录中的互动与新互动

现在回到本章的核心区分。

录像里当然有互动。平台游戏的录像里，玩家按键，角色跳起，撞到砖块，砖块碎了；机器人录像里，机械手伸出去，杯子被推动。这些互动在拍摄时是真实发生的，它们留下的差异，模型确实学到了。Genie能从无标注视频里自己归纳出「几类动作」，这件事本身说明，录像中的差异结构里藏着相当多关于「可以怎样作用于这个世界」的信息。

但是，记录中的互动有三个特点。其一，它是别人选的。录像里出现的动作，是当时那个玩家、那个操作员选的；没有人做过的动作，录像里就没有它的后果。其二，它已经封闭。录像里的后果已经定了，模型看到的是一条走完的路，不能问「如果当时换一个动作会怎样」，除非别的录像恰好拍到了。其三，它可能混淆。同一段画面变化，可能由不同的原因造成：角色往右走了，也可能是镜头往左移了。只看录像，有时分不清。

新互动则不同。小舟下水，选什么动作是他自己的；水的回应是当下的；如果他划水时身体没往前，他可以立刻换一个手形再试。差异在这里不是被读出来的，而是被做出来的。这一点第32章「从被动记录到主动试探」会细讲。

对视频世界模型来说，有两件事值得分开。用户在生成的世界里按方向键，模型生成相应画面，这是一种新互动吗？从使用者的角度看，是；从模型学习的角度看，未必。如果用户的操作只是作为条件输入，用来生成画面，而不回流去更新模型的参数或形成持久能力，那么这种互动属于本书所说的「暂时语境变化」，不是「持久学习」。而如果系统把交互中暴露的错误收集起来再训练，那就进入了另一种发生条件。哪一种是哪一种，要看具体系统公开了什么，公开不足就标〔待核〕。

还有一层更细的差别：录像里的差异和新互动里的差异，对「错误」的处理方式不同。小舟呛水，这口水是他自己的，呛的那一刻，他的紧张、他刚才抬头的角度、教练的喊声缠在一起，下一次换气时他会自然地少抬一点头。一个视频模型生成了一帧不合理的画面，比如人物穿墙而过，这个错误在生成时并不会被模型自己察觉，因为没有任何东西会「挡住」它；只有当有人看出来、把这类错误收集起来、改训练数据或目标再训练时，错误才变成下一轮的差异来源。前者的纠错发生在一个人身体与水的直接碰撞里，后者的纠错要经过人类团队的中介。这正是本书所说的发生单位：训练里的纠错是「集体人类—AI复合体」完成的，不是模型独自完成的。

这一点也给读者一个读演示的办法：看一段世界模型的演示时，可以问一句，演示里有没有出错？出错了是谁发现的？发现之后去了哪里？如果演示里只有成功的片段，我们就只知道它能成功，不知道它怎样从失败里学。

◆ 六、反例与边界

第一个边界：本章不主张「视频学习只能得到画面」。机制上完全可能出现这样的情况：模型为了把下一帧预测好，内部形成了近似物体位置、遮挡关系乃至重

力方向的表示。这是可以用探针实验去检验的机制假说，不能凭直觉否定，也不能凭漂亮画面认定。

第二个边界：本章也不主张「新互动总比录像好」。互动有成本，有风险。小舟第一次下水旁边有教练；机器人在真实世界里乱试，可能摔坏东西。很多时候，先从录像里学一个大致框架，再用少量真实互动去修正，正是合理的路线。下一章看到的V-JEPA 2就是这种思路。

第三个边界：小舟的例子和模型之间有一个不能抹掉的差别。小舟下水时，视觉、触觉、身体的平衡感、听到的教练指令一起参与；而一个视频模型，哪怕给它新的操作机会，它碰到的也仍然只是画面。这就把问题推到第五编要讨论的三种SDE上去了，这里只点到。

第四个边界关乎「规模」。常有人说：录像不够，那就再多些录像；八个动作太粗，那就再细些；二维游戏太简单，那就用真实世界的视频。这些方向都可能让视频模型更强，本书不预判它们的上限。但规模解决的是「记录中的差异有多丰富」，不直接解决「记录中的互动是别人选的、已经封闭的」这两个特点。一百万小时游泳录像里，也许有各种体型、各种水温、各种泳姿；可小舟这一个孩子、这一天、这一池水里的那一下蹬腿，仍然要他自己去碰。能否用海量录像把这一下「近似」到足够好，是一个经验问题，要看具体任务、具体误差容忍度。对于在泳池边看着别人游的家长，录像可能已经够用；对于真要下水的孩子，不够。

所以本章的判断落在一个可检验的位置上：对某一任务，记录中的差异在多大程度上能替代新互动产生的差异，要靠对照实验来回答，比如同等数据量下，只看录像训练与加入少量主动互动训练，在新情境里的表现差多少。第34章讲「拼接、被动融合和主动互动的对照」，正是要把这个问题做成可以失败的实验。

> **带走的话**
>
> **从视频中学世界，学到的是录像中互动所留下的差异结构，这是真收获；但录像里的互动是别人做完的，新互动是自己做出来的，前者不能自动代替后者。**

> **小账** 论证落在哪一层?——本章以经验事实（Genie论文机制、GWM-1与Genie 3的公开描述）和机制假说为主，本体论立场集中在「记录中的互动与新互动是两种差异运行」。 世界模型补了哪一条?——补上一条条件——世界模型要区分从记录中继承的差异和在新互动中产生的差异，并说明使用中的交互是否回流为持久学习。

18

第 18 章

从内部状态与行动后果学习

这一章的问题：如果不去还原每一帧画面，而是预测「状态」和「行动的后果」，世界模型会离可靠行动更近吗？近在哪里，又在哪里停步？

◆ 一、导航软件在想什么

这是我构想的场景。周敏每周六送小舟去游泳馆，路上总开着导航。有一次前方出了事故，导航先是提示「前方拥堵」，过了一会儿说「已为您重新规划路线，预计节省六分钟」。小舟在后座问：「它怎么知道走另一条路更快？它又看不见那条路。」

这句孩子话问到了点子上。导航软件并不在脑子里「播放」那条路的街景视频。它手里有的是一张路网，以及每段路此刻大概多快、过去这个时段通常多快的估计。它比较的是：沿A路走，大概要多久；沿B路走，大概要多久。它预测的不是画面，而是状态（每段路的通行速度）和行动的后果（选这条路，多久到）。

这就是本章要讨论的第二条技术路线的直观形象。上一章的视频模型以画面为世界，下一帧长什么样是它的核心问题；这一章的路线则说：画面里有太多与行动无关的细节，路边的广告牌换了颜色、树叶怎么晃，导航根本不关心。与其把力气花在还原这些细节上，不如预测一个更抽象的状态，再在状态上比较不同行动的后果。

本章的命题是：以内部状态和行动后果为对象的世界模型，把训练目标从「显露得像」移向「对行动有用」，这是一种向任务靠拢的发生取向；但它仍要分清四件事：模型内部的想象、执行时的反馈控制、模型参数的更新和持久学习。

◆ 二、V-JEPA的特征预测

Meta的研究团队提出的JEPA（联合嵌入预测架构）路线，是「不预测像素、预测特征」这一思路的代表。2025年公开的V-JEPA 2（arXiv 2506.09985）在这条路上走得比较完整。按论文摘要，它的做法分两段。

第一段是无动作的预训练。模型在超过一百万小时的互联网视频（以及图像）上训练。训练方式的核心是「遮住一部分，预测被遮住部分的表示」：不是把被遮住的像素画出来，而是预测那部分在模型自己的表示空间里应该是什么。摘要报告了这样预训练出的模型在动作理解、动作预判、与语言模型对齐后的视频问答等任务上的成绩，例如在Something-Something v2上达到77.3的准确率，在Epic-Kitchens-100的动作预判上召回率达到39.7，作者称其为当时最好的水平。

这一段可以用SDE的语言这样读：模型被要求在缺失中补全，但补全的是「特征」，不是「画面」。被训练去缠到一起的，是画面各部分在表示空间中的关系；差异运行发生在「看见的部分」与「被遮住的部分」之间。由于预测目标是表示而非像素，模型可以不去管那些难以预测、又与理解无关的细节，比如水面上每一道反光。这是机制层面的说法，作者的设计意图大体如此；至于它内部究竟学到了哪些可解释的状态，要靠专门的探针实验，摘要里不足以下结论。

后续的V-JEPA 2.1（arXiv 2603.14482，2026年3月提交）把重点放在「稠密特征」上：按摘要，它让可见和被遮住的位置都参与训练信号，在编码器的多个中间层上施加自监督，并用多模态分词器同时训练图像和视频。作者报告它在动作预判、深度估计、机器人导航、视频识别等任务上有提升，在真实机器人抓取上比V-JEPA 2的动作条件版本成功率高约20个百分点。这些数字都是作者自报，本书照录，不作独立评价。

◆ 三、动作条件后训练与规划

V-JEPA 2的第二段才和「行动」直接相关。作者在预训练好的模型上，用来自DROID数据集的不到62小时的无标注机器人视频，后训练出一个动作条件版本，

叫V-JEPA 2-AC。它学的是：在当前的表示下，执行某个动作之后，表示会变成什么样。

有了这个「动作—后果」的预测器，就能做规划：给机器人一张目标图片（比如杯子放在盘子上的样子），系统在表示空间里尝试不同的动作序列，看哪一串动作预测出的结果最接近目标图片的表示，然后去执行。摘要报告，这个系统被零样本部署到两个不同实验室的Franka机械臂上，在没有从这些机器人、这些环境收集任何数据、也没有任务专门训练和奖励的情况下，完成了抓取和放置物体。

按论文正文，后训练时预训练好的编码器被冻结，只训练那个动作条件的预测器；部署时，系统在闭环的模型预测控制中使用它，即走一步、看一眼、再算下一步。

这件事的意义，正好落在上一章的边界上：先从一百万小时录像里学一个大框架，再用几十小时的机器人数据把「动作」接进来。录像提供了大量记录中的差异，少量机器人数据提供了动作与后果的对应。这与小舟先看录像再下水，有相似之处。

还有一处设计值得注意：任务目标是用一张目标图片给出的。这很巧妙，因为它不需要人为每个任务写奖励函数；但它也把「任务」限定成了「让世界看起来像这张图」。杯子放进盘子，看起来像，任务就算完成。而许多日常任务的成败，并不在于看起来像不像。小舟学游泳，目标不是「摆出和教练一样的姿势」，而是「在水里往前走、不呛水」；导航的目标不是「车停在一个看起来像目的地的地方」，而是「准时到达」。当任务的成功藏在看不见的地方，比如盖子拧紧了没有、豆腐有没有被捏碎，以目标图片为任务的规划就要另找补充。这不是批评，而是指出这种设计的适用边界：它对「摆放类」任务很自然，对「力道类」「时机类」任务则需要别的信号。第33章讲林岚的机械手拧瓶盖、捏豆腐，要模型连指尖的力怎样变化一起预测，就是在补这种信号。

但也要看清它没有说的东西。用于后训练的机器人视频是已经录好的，不是这个模型自己在环境里试出来的；部署时，规划是在模型内部的预测里完成的；执行

之后看到的新画面用于下一步规划，属于反馈控制，至于执行结果是否回流去更新模型参数，论文没有作为部署环节描述，从「零样本、无任务专门训练」的表述看大概率没有〔待核：以论文部署实验部分确认是否存在在线参数更新〕。所以，就公开摘要所及，V-JEPA 2-AC展示的是「从记录中学到的动作—后果关系，能被用于新环境的规划」，而不是「模型在新环境的互动中持续学习」。

◆ 四、Dreamer：经验回放、内部想象和策略学习

第三个对话对象是DreamerV3。它的论文2025年4月发表在《自然》上，题为《通过世界模型掌握多样的控制任务》。按摘要，Dreamer的思路是：学一个环境模型，并通过「想象未来情景」来改进行为。作者报告，它用同一套固定的超参数，在一百五十多个不同任务上超过了专门调校的方法，并且是第一个在不用人类数据、不用课程设计的情况下，从零开始在《我的世界》游戏里采到钻石的算法。

Dreamer的训练是一个循环，大致可以分三步说（这里按Dreamer系列公开的一般框架概述，细节以论文为准）：智能体在环境里行动，把经历存进一个经验库；从经验库里取出片段，训练世界模型，让它学会从当前状态和动作预测下一状态和奖励；然后在世界模型内部「做梦」，即不碰真实环境，在模型里展开一串串想象的未来，用这些想象来训练决定动作的策略。新的策略再回到环境里去行动，产生新的经历，循环往复。

和前两条路线相比，Dreamer有一个关键不同：它的经验是自己行动产生的。经验库里装的不是别人的录像，而是它自己一次次尝试、成功、失败的记录。用上一章的话说，它的差异来自新互动，而不只是记录中的互动。这正是从「看录像」走向「下水」的一步。

当然它的「水」是游戏和仿真环境。在真实物理世界里，让一个机器人像在《我的世界》里那样成千上万次试错，成本和风险都完全不同。Dreamer在真实机器人上的表现如何，不在这篇摘要的范围内，本书不作推断。

Dreamer的世界模型内部也值得说两句。按DreamerV3论文的描述，它不在像素上做想象，而是先把观察压成一个紧凑的内部状态，这个状态一部分是延续性的记忆，一部分是对当下的随机估计；想象就在这个内部状态上一步步往前推，同时预测会得到多少奖励、这一段经历是否还会继续（论文预印本对这一结构有明确描述）。DreamerV3能用一套固定超参数应对一百五十多个任务，摘要把这归功于一组基于归一化、平衡和变换的稳健化技巧，让量级事先不知道的信号也能被稳定地学习〔待核：各项技巧的名称与细节以《自然》正文为准〕。这一点在发生学上颇有意思：同一套方法要应对差别很大的世界，前提是先把各个世界里差异的「量级」整理到可以比较的尺度上。差异运行不是原样照搬进模型的，它先被一套处理方式规整过。规整方式本身就是人替模型定的发生条件之一。

还要看到Dreamer「想象」的性质。它在内部展开的未来，是从已有经验里学到的状态转移推出来的。想象能帮它在真实环境里少走弯路，但想象里出现的错误，比如把一堵墙当成了可以穿过的空间，只有回到真实环境里撞一次才会暴露。Dreamer的循环之所以有效，很大程度上正是因为它不断回到环境里，把想象的错误变成新的经验。如果只有想象而没有回到环境，模型会在自己的错误里越推越远。这和周敏在课上讲的道理一样：学生在草稿纸上推了一大串，如果从不拿实验数据核对，推得越长，偏得可能越远。

◆ 五、执行反馈与参数更新的区分

回到导航软件。周敏那天遇到的「重新规划路线」，到底是导航软件在「学习」吗?

细分一下，那一刻至少有四种不同的东西在起作用。第一种是执行时的反馈控制：软件发现车速比预期慢，前方路段的实时数据变了，于是按现有的方法重新算一遍路线。这里没有任何东西被「学会」，算法还是那个算法。第二种是暂时的语境变化：这一路上的路况数据只对这一次行程有效，到家就作废。第三种是模型参数的更新：软件公司会定期用积累的行程数据重新训练「某路段某时段通常多快」的模型，下个月同一个周六，它的预估可能会更准。第四种是持久学习：只有当这

种更新被验证、稳定下来、在以后的类似情境里可靠地发挥作用，才谈得上能力的形成。

这四件事，在世界模型里一样要分开。V-JEPA 2-AC部署时在内部比较动作序列，执行后看到新画面再规划，这首先是执行时的反馈控制和规划；Dreamer的循环里，新经验进入经验库、重新训练世界模型，这是参数更新；而它在一百五十多个任务上用固定超参数的表现，是在检验这种训练方式能否稳定地形成能力。本书一再要求区分的暂时语境变化、外部记忆、控制反馈、参数训练和持久学习，在技术系统上恰恰最容易混淆：一段「机器人遇到障碍后调整了动作」的演示，往往说明的是控制反馈，不说明模型学到了什么。这几层在机器人任务里怎样分、经验何时该从下一层升到上一层，第35章会专门讲，这里只把导航上的四分法立住。

用SDE来读这一层区分：执行时的反馈控制，是在既有纠缠（E）不变的情况下，让差异运行（D）随现场调整，显露（S）随之变化；参数更新，则是让新的差异回写去改变纠缠本身。三方程里E=H(S,D)这一条，只有在后一种情况下才真正被走通。一个系统可以在现场表现得很灵活，但如果E从不被回写，它碰到同样的意外第二次，仍然只能靠现场应变。

导航软件还有一个常被忽略的地方：它的「状态」从哪里来？每段路此刻有多快，很大程度上是从千千万万辆正在行驶的车的位置和速度里算出来的。周敏那天被提示「前方拥堵」，是因为前面许多司机已经慢下来了；第二天导航对同一路段的估计更准，是因为更多人走过了这条路。也就是说，导航软件的世界模型，是由一个巨大的人群在不断地「下水」维持着的。没有这些人的实际行驶，它的状态预测就会很快失真。

这一点放回世界模型的讨论里很有分量。一个状态预测器之所以可靠，往往不是因为它的结构多巧妙，而是因为有一个持续的、来自真实行动的差异来源在不停地纠正它。导航是这样，天气预报是这样，第19章要谈的Waymo外部学习回路也是这样。这正是本书所说的「集体人类—AI复合体」：模型的世界不是它一次学好、封存起来的，而是在人群持续的行动与模型持续的修正之间一直发生着。

◆ 六、边界与推进

本章的路线比视频生成更接近行动，但有四点边界。

其一，「预测特征而非像素」的好处，依赖于特征里保留了任务需要的差异。如果某个任务的关键恰好藏在被模型当作无关细节的地方，比如水面反光里透出的水深，或者一段路面的微小坑洼，特征预测反而会把它丢掉。哪些差异该保留，又回到任务。

其二，从录像里学动作—后果，学到的仍是录像里出现过的那类动作。V-JEPA 2-AC能在新实验室里抓放物体，说明这类动作的后果被学得有相当的迁移性；能迁移到什么程度、遇到软的、滑的、易碎的物体会怎样，要靠更多实验。

其三，Dreamer式的自主试错，在仿真和游戏里效率很高，在真实世界里受成本和安全约束。第32章讲「探索成本、行动边界与数据效率」时会回到这个问题。

其四，本章三条路线的成功，都不能直接拿来证明SDE的本体论，失败也不能拿来否定它。V-JEPA的特征预测有效，说明「预测表示而非像素」在这些任务上是好的工程选择；Dreamer在一百五十多个任务上有效，说明「行动—经验—模型—想象」的循环是好的训练组织。它们与本书的主张方向相合，可以当作对话中的旁证，但本体论立场要靠它自己的论证站住。反过来，哪天某个纯视频模型在某些任务上做得比这些路线更好，也只说明那条路线在那些任务上有效，不说明世界是先在的、等着被画出来的。

前两章的几条路线，视频、特征、想象，基本都还围绕着视觉。下一章要看的是：当声音、动作、空间、多种传感器一起进入模型，会发生什么，以及多种信号「融合」与多种SDE「纠缠」是不是一回事。

带走的话

从状态与后果学习，让世界模型向「对行动有用」靠近；但现场的灵活应变不等于学会，只有执行后的差异回写去改变纠缠，学习才真正发生。

小账 论证落在哪一层？——本章以经验事实（V-JEPA 2／2.1、DreamerV3的公开摘要）为主，机制假说集中在「特征预测保留哪些差异」，本体论立场集中在「回写E才是学习」。世界模型补了哪一条？——补上一条条件——世界模型要分清内部想象、执行反馈、参数更新与持久学习，并说明新互动的差异是否回写到模型本身。

19

第 19 章

多模态与空间模型的进展

这一章的问题：当视觉、声音、动作和空间一起进入同一个模型，这是不是就等于多种SDE在模型里纠缠起来了？

◆ 一、气象站的「多传感器」与泳池里的「多感官」

还是从天气预报说起，下面两个都是我构想的场景。周敏带学生参观过一次市里的气象观测站。讲解员介绍，预报要用到的东西五花八门：卫星从天上看云，雷达往空中发电磁波听回声，地面站测气温、气压、湿度、风向，探空气球带着仪器升上高空。每一种仪器都只看见天气的一个侧面，把这些资料按时间和位置对齐，放进同一个数值模型里，预报才可靠。讲解员把这叫「多源资料同化」，学生们一听就懂：就是把各路情报拼在一起。

同一个暑假，小舟在泳池里学会了换气。周敏后来问他是怎么学会的，小舟说不清楚，只说「就是感觉到了」。细想一下，那一刻他至少同时有三种经验：眼睛看见水面离脸多近，皮肤和手掌感到水的阻力和浮力，耳朵听见出水那一瞬声音从闷闷的变得清亮。这三种经验不是先各自整理好、再拼到一起的。耳朵里声音一变，他就知道嘴已出水，可以吸气了；手掌感到水「托」住了身体，他才敢把头埋下去；眼睛看见水面晃得厉害，他会下意识划得慢一点。三者相互提示、相互制约、相互纠正，最后形成了一个对他来说整体的东西：「换气」这件事。

这两个例子摆在一起，本章的命题就清楚了：**多传感器融合是把不同来源的信号对齐、合并，服务于一个事先定好的对象；多种SDE纠缠是不同感官各自完整的**

显露、差异与纠缠，在共同的行动里相互制约，一起形成对象。前者是工程上已经大量实现的事实，后者是本书提出的研究路线；二者可能相交，但不能画等号。

需要立刻说明的是，这一区分本身是本体论立场加机制假说，不是对任何具体系统的判决。一个多模态模型内部究竟只是「拼在一起」，还是形成了相互制约的纠缠，要看它的训练目标与检验方式，不能看名字。

◆ 二、Cosmos：一个框架里放进五种序列

英伟达的Cosmos系列是这几年面向「物理AI」的代表性开放模型。本书讨论的是2026年公开的Cosmos 3（arXiv 2606.02800）。按其摘要，Cosmos 3被称作「全模态世界模型」，用一个统一的「混合Transformer」架构，同时处理和生成语言、图像、视频、声音和动作序列，输入输出的组合可以灵活配置。作者说，它把视觉语言模型、视频生成器、世界模拟器和「世界—动作模型」纳入同一框架，定位是给具身智能体用的通用底座；摘要提到它在一个机器人策略评测平台上被评为最好的策略模型，后训练的版本在一个第三方榜单上被评为最好的开源文生图和图生视频模型；代码、权重、数据集和评测基准都按开放许可证发布，其中包括整理过的合成数据集。摘要没有给出参数量和训练数据的详细来源〔待核：以正文核查训练数据构成、声音与动作数据的来源和比例〕。

从第16章的分类看，Cosmos 3横跨视频生成、状态预测和行动模型几类。它是本书所说「统一架构」的一个清楚实例。按照第16章的区分，统一架构是工程事实；它能否在某一具体任务里形成可靠能力，要看那个任务上的检验。

用SDE的话审查这一路线，关键问题有两个。第一，声音和动作在训练中是作为「另一种要生成的东西」并列存在，还是在同一事件里相互约束？比如一段视频里杯子落地，模型是分别学「画面里杯子落地」和「声音里有一声脆响」，还是学到了「这声脆响的有无和高低，取决于杯子是什么材料、落在什么地面上」，以至于声音可以用来纠正对画面的判断？第二，动作序列与画面、声音之间，是记录中的对应，还是有来自新互动的差异回写？这两个问题，摘要层面都回答不了，本书标〔待核〕，留待正文与后续评测。

摘要里提到的「合成数据集」也值得留意。用模型生成的数据再去训练模型，在物理AI领域越来越常见，原因很实际：真实的机器人互动数据少而贵。合成数据的好处是量大、可控，能专门造出罕见情境；它的风险在于，合成数据里的差异，来自生成它的那个模型已经缠在一起的东西。如果生成模型对某类物理情境本来就理解偏了，合成数据会把这种偏差放大传下去。所以合成数据在发生学上的位置是清楚的：它扩大的是「记录中的差异」，不能替代来自真实世界的新差异；它的价值取决于有没有真实互动在另一端持续校正它。这和天气预报一样，模拟再多次，也要等明天真实的天气来核对。

◆ 三、Atlas：空间上下文

World Labs在2026年9月宣布了它的Atlas模型。按官方博文，Atlas是一个在大规模多模态数据上从零开始预训练、原生处理文字、图像、视频和三维的模型，目前只向部分合作方提供早期试用。它的输入包括文字、图像、相机位姿和三维深度图，视频被当作图像序列处理；输出包括从新相机位置看到的图像和视频、全景图、深度图，以及点云、三维高斯泼溅这类显式三维结果；面向机器人，它可以从仿真机器人的传感器视角输出彩色图和深度。

Atlas最有特色的是它所谓的「空间上下文」。博文的说法是：像大语言模型把输入编码成上下文一样，Atlas把输入编码成一个上下文，但其中每张图像都被放在一个三维位置上。使用者甚至可以把毫不相关的图片摆到三维空间的不同位置，让模型生成它们之间的过渡。技术上，博文称它是一个「多模态自回归扩散Transformer」，所有输入共享同一个空间上下文，输出逐个生成，每一个都用一种扩散类的生成方法（博文称为整流流）得到。博文里的评测结论都来自公司自己的测试。

拿第16章的四问来问Atlas。它以什么为世界？主要是空间：东西在哪里，从别处看是什么样。缠在一起的是什么？是不同视角的图像、深度与相机位置之间的关系。差异怎样运行？在不同视角之间，一个位置看到的和另一个位置看到的差

异，被用来约束三维结构。显露靠什么站稳？靠新视角下的图像与实拍是否一致，靠几何是否自洽。

这里的空间上下文是一个重要的进展，它让「同一个空间」在多张图之间有了一个共同的锚。用本书的话说，它为对象同一性的一个侧面，即空间上的同一，提供了工程上的承载方式。但空间上的同一，只是对象同一性的一部分。一只杯子在三维重建里位置稳定、形状完整，不等于模型知道它拿起来多重、装满水会不会滑手、掉下去会不会碎。第五编讨论触觉SDE和物的意识时，这些才进入问题。

另一个值得记下的细节是Atlas面向机器人的那项输出：从仿真机器人的传感器视角生成彩色图与深度。这意味着空间模型开始被当作机器人训练的「场地」。这条路的价值显而易见：真实场地昂贵，生成场地便宜。它的限度也要同时看见：生成场地里的物体，能被看、能被从各个角度看，但能不能被推、被捏、被敲出声音，取决于场地背后有没有相应的物理过程。博文就公开所及只谈了视觉与几何的输出，所以本书只能说，Atlas给机器人提供的是一个视觉—空间上的训练世界；机器人在这个世界里学到的，有多少能带回真实的手感，要靠真机实验去回答。

◆ 四、真实车辆状态与接触信息的位置

第三个对话对象来自自动驾驶。2025年12月，Waymo发布了一篇讲其AI体系的博文。按博文，Waymo的基础模型被直接称作「世界模型」，由两部分组成：一个「传感器融合编码器」，把摄像头、激光雷达和毫米波雷达的输入随时间融合，产出物体、语义和特征表示，用于快速反应；一个用Gemini训练的「驾驶视觉语言模型」，用于复杂的语义推理，比如前方有一辆着火的车这类罕见情境。两者共同输入一个「世界解码器」，它预测其他道路使用者的行为、生成高精地图、生成本车的轨迹。

博文还描述了两个回路。一是内部学习回路：一个闭环仿真器在碰撞风险、恶劣天气、复杂路口等情境中运行，可以用文字设定天气和时段，并从一个紧凑的世界表示合成摄像头和激光雷达数据；在这个受控环境里，由一个「评判器」给出奖惩，用强化学习训练驾驶系统。二是外部学习回路：从完全无人驾驶的实际里程

中，由评判器标出不够好的行为，团队据此生成替代行为作为训练数据，在仿真器里测试，只有在安全框架确认没有不合理风险之后，才部署更新后的驾驶系统。博文强调，真实的完全自主驾驶经验无可替代，并说其完全自主里程已远超人工驾驶数据。

这篇博文对本书有两点特别的价值。第一，它清楚地说明了一个实际运行的驾驶系统不是纯视觉的：摄像头、激光雷达、毫米波雷达一起进入融合。这提醒我们：不能笼统地说「驾驶AI全部是纯视觉」。第二，它的外部学习回路，正是第18章区分的「参数更新」与「持久学习」在工程上的一个公开实例：现场的执行不直接改模型，真实里程中暴露的问题，经过人类团队、评判器、仿真测试、安全审查这一串中介，才回写为新一版的驾驶系统。这也是一个非常清楚的「集体人类—AI复合体」的发生结构。

那么，真实车辆状态和接触信息在哪里？博文讲的感知输入都是外部感知：看见、测距、测速。车辆自身的状态，比如轮胎与路面的附着、方向盘上的阻力变化、制动时车身的点头、底盘传来的异响，这些在老彭这样的老司机那里是判断路况的重要依据，第29章会细讲。它们在Waymo系统里是否进入模型、以什么形式进入，这篇博文没有说〔待核：查Waymo其他公开材料中车辆动力学状态与执行器反馈的位置〕。本书不能从「博文没提」推出「系统没有」，只能说：在这篇公开材料所描述的世界模型里，对象是由外部感知组织起来的道路世界；车辆与路面的接触，若有，是在别处处理的。

◆ 五、融合与纠缠：区别在哪里，边界在哪里

现在可以把本章的核心区分说得更具体一些。

融合，是在一个事先定好的对象（比如「道路上的物体及其运动」）之下，把不同传感器的信号对齐、加权、合并。不同信号各自贡献一部分证据，最后汇成一个对象表示。融合做得好，对象就看得更全、更准。气象站的多源资料同化，Waymo的传感器融合编码器，都是这种结构的出色实例。

多种SDE纠缠，按本书的主张，是视觉、听觉、触觉各自作为完整的SDE，各有自己的显露、差异运行和纠缠，在共同的行动与事件中相互制约；当一种感官的显露与另一种冲突时，冲突本身推动对象的归并、拆分或重建。小舟出水那一刻，声音一亮，他对「嘴在水面之上」的判断就被确认；如果眼睛觉得已经出水，耳朵却还是闷的，他会犹豫，这种犹豫正是冲突推动重新判断的时刻。

二者的区别可以落成三个可检验的问题。其一，对象是先定的还是在互动中形成的？融合通常有先定的对象类别；纠缠允许对象在冲突中被拆分或合并。其二，冲突怎样被处理？融合常把冲突当作噪声，按置信度加权平均掉；纠缠把冲突当作信息，推动表示改变。其三，不同模态的经验是否在行动中相互纠正？融合可以是被动的；纠缠在本书的路线里，要在主动互动中形成。

导航软件给了一个人人碰到过的例子。周敏开车走长沙的高架时，导航偶尔会「掉」到桥下的地面道路上，提示她「前方路口右转」，而她正在高架上，根本没有路口。手机导航用的是卫星定位、手机里的加速度和转向传感器、再加一张电子地图。卫星定位在高架上下几乎给出同一个平面位置，分不清哪一层；地图上两条路又重叠在一起。几路信号在这里并不矛盾，只是都说不清。这时，一个真正能区分的差异往往藏在别处：上高架时车身有一段明显的爬坡，加速度和气压读数会变化。有的导航会利用这些信号来判断上下层，有的则要等你开出一段、路线明显分开后才纠正过来〔待核：各导航产品的具体做法不作断言，此处只作一般陈述〕。

这个小故障里藏着本章区分的要点。如果系统事先只把「平面位置」当对象，那么爬坡的信号就是无关的噪声，加权平均掉了；如果系统允许「我在哪一层」成为一个在行动中被确定的对象，那么爬坡这段差异就成了关键证据，它会推动对象从「这一片重叠的路」拆分成「高架上」或「高架下」。前者是融合，后者更接近本书说的纠缠：对象是在差异的冲突与相互制约中被形成的，而不是先定好再往里填数据。人的体验更直接：周敏自己从来不会搞错，因为她的身体刚刚感到车子爬了一段坡，眼睛看见两侧是护栏而不是商铺，耳朵里是高架上特有的风噪。三种经验在她那里没有经过「对齐」这道手续，它们一开始就是同一件事的不同侧面。

边界同样要说清。第一，这一区分不是说融合「低级」、纠缠「高级」。很多任务，融合已经足够，甚至更稳。第二，现有的多模态模型里，完全可能已经出现了某种程度的相互制约，比如用声音纠正画面判断；这是可以检验的经验问题，不能凭「它叫融合」就否定。第三，本书不主张三种感官对所有任务都不可替代；哪些任务需要哪几种SDE纠缠到什么程度，正是第五、六编要研究的问题，不是这里能预设的结论。

◆ 六、向下一章推进

四章看下来，现有的路线各有所长：视频生成学到了录像中丰富的差异结构，特征预测与行动模型向任务靠拢，Dreamer式循环让经验来自新互动，空间上下文为同一性提供了锚，Waymo的外部回路把真实经验持续回写。那么，本书反复提到的「发现学前提」，究竟在这些技术里的哪里起作用？是整个领域都被它限制了，还是只在某些具体机制上？下一章要回答这个问题，而且要用具体训练机制来回答，不贴标签。

> 带走的话
>
> 多传感器融合是在先定对象下合并信号，多种SDE纠缠是各自完整的感官在共同行动中相互制约、一起形成对象；前者是已有的工程事实，后者是需要检验的研究路线。

小账 论证落在哪一层？——本章以经验事实（Cosmos 3摘要、Atlas博文、Waymo博文）为主，本体论立场与机制假说集中在「融合与纠缠的区分」。 世界模型补了哪一条？——补上一条条件——多模态世界模型要说明对象是先定还是在互动中形成、模态冲突是被平均掉还是推动重建，并标出车辆自身状态与接触信息在系统中的位置。

20

第 20 章

发现学前提怎样限制世界模型

> **这一章的问题**：「世界是先在的、等着被找到」这一前提，究竟落在世界模型的哪些具体训练机制上，又有哪些机制已经在朝发生的方向走?

◆ 一、先把标签放下

写到这一章，很容易顺手写出一句痛快话：「现有世界模型都是发现学的，所以都走错了路。」这句话好写，也好记，但它是错的，至少是没有证据的。前四章看到的路线里，有的以画面为世界，有的以状态为世界；有的只从录像里学，有的从自己的行动里学；有的把冲突平均掉，有的在真实里程中把错误一轮轮回写。把它们统统归入一个标签，等于放弃了前四章所做的区分。

所以本章换一种做法：不问「这个领域是不是发现学的」，而问「发现学前提落在哪一个具体机制上」。发现学前提，按本书的用法，是以为世界与规律先在、等着被找到；发生学是在纠缠（E）中经差异运行（D）显露（S）。前提是哲学层面的，但它会落地，落在对象怎样定义、数据怎样采集、目标怎样设定、成功怎样检验这些地方。落在哪里，就在哪里审查。

本章的命题是：**发现学前提对世界模型的限制，不在于某个领域、某家公司或某种架构，而在于几个可以指认的训练机制：预先完成的对象与属性清单、只从封闭记录中取差异、以显露相似代替发生检验、把冲突当噪声消除**。与此同时，现有技术里已有不少机制带着发生取向，应当被看见、被继续推进。

◆ 二、预先完成的对象、属性与世界图景

第一处落点，是在训练之前就把「世界里有什么」定死了。

天气预报的历史里有个现成的对照。数值预报把大气切成格子，每个格子里记温度、气压、湿度、风速。这份变量清单是物理学几百年工作沉淀下来的，在大尺度天气上非常可靠。但气象学界也一直知道，格子以下的过程，比如一朵积云的生成、山谷里的局地风，被处理成了「参数化」：不去显式算它，而是用一个经验公式估它对格子的平均影响（这是对数值预报一般做法的概述，具体方案因模式而异）。参数化本身没有错，它是在有限计算里做的合理取舍；但它意味着，格子以下那一层世界，在模型里是以一组先定的近似出现的，它自己在模型里不会再发生出新的东西。

世界模型里同样有这种「先定清单」。自动驾驶的感知系统通常预先定好要识别的物体类别：车、行人、自行车、锥桶。类别清单让系统高效，但遇到清单之外的东西，比如一张被风吹起的塑料布、路中间一头走失的牛，系统要么硬塞进某个已知类别，要么当作「未知障碍」。Waymo博文里专门提到用一个视觉语言模型来处理「着火的车」这类罕见情境，可以看作工程界自己意识到了先定清单的边界，并引入了一个更开放的语义通道来补。

Genie的8个隐动作也属于先定，只是先定的是数量而不是内容：动作的类别是模型自己从录像里归纳的，类别的数量是人事先给的。这比给出一份「左、右、跳」的清单更开放一步，但粒度依然是先定的。

导航软件也有它的先定清单，那就是电子地图。地图上有的路，导航才会规划；一条新修通的路，地图没更新之前，在导航的世界里根本不存在。这是我构想的场景：周敏家附近去年通了一条新路，头几个月导航总让她绕老路，邻居老彭早就抄近道走了。后来地图更新了，常常不是因为地图公司派人去量，而是因为许多车的轨迹反复出现在一片「没有路」的地方，系统才把它当作待核实的新路〔待核：各地图厂商的更新机制不作断言，此处为一般陈述〕。这个例子把两件事同时摆了出

来：先定清单会让新对象一时进不来；但如果系统留着一条让行动轨迹改写清单的通道，新对象终会在互动中发生出来。

限制在哪里：当对象和属性在训练前被完整列出时，训练就只能在清单内部调整，不能让新的对象或新的属性在互动中发生。这是本体论前提落到机制上的第一处。它不意味着：一切预设都该取消。没有任何预设的训练是不存在的。问题是预设是否允许在冲突中被修订。

◆ 三、从显露相似跳到发生机制相同

第二处落点，是一种推论上的缺口：看到模型生成的东西「像」，就推断模型内部发生的东西与真实世界「同」。

用看录像学游泳的例子说最清楚。设想有一个模型看了大量游泳录像，能生成一段非常逼真的蛙泳视频：手臂的轨迹、水花、身体的起伏都对。这是我构想的情景。我们能不能说它「懂了游泳」？

画面像，只说明一件事：它生成的显露（S）在统计上与真实录像的显露接近。但真实的游泳，是人的身体在水里，经由肌肉用力、水的阻力和浮力、呼吸节奏这些差异的运行，与身体的特征纠缠在一起，才显露为那样的画面。模型的画面，是由它内部压缩码之间的关系生成的。两者的显露相似，发生它们的差异运行和纠缠可以完全不同。

小舟在下水之前，也会犯一模一样的推论错误：「我比划得跟视频里一样了，所以我会了。」显露相似（我比划得像）被当成了发生相同（我身体与水的关系建立了）。第一次下水，水把这个推论推翻了。

在技术评价里，这个缺口有一个常见形态：用画面质量指标来证明物理理解。一段视频里物体下落的轨迹「看起来对」，就推断模型掌握了重力。可它也可能只是在训练数据里见过足够多类似的下落镜头。要区分二者，得做发生层面的检验：换一个录像里罕见的条件，比如月球上的低重力、水中的下落，看它是否给出符合那种条件的后果；或者让它预测干预之后的变化，而不只是续写一段常见的过程。

第27章「为何显露预测不能保证行动可靠」会把这一点展开为「已有记录中的不可区分配置」与「新干预下不同后果」。

限制在哪里：以显露相似为唯一或主要的成功标准，会让训练与评价都停在S上，看不见D与E是否成立。它不意味着：显露相似毫无意义。显露相似是必要的检验之一，只是不充分。

◆ 四、只取封闭记录与把冲突当噪声

第三处落点，是差异的来源只限于已经封闭的记录。第17章已细讲：录像里的互动是别人选的、已经走完的。如果一个世界模型的全部训练差异都来自这类记录，它的世界就是「被记下的世界」，它在录像以外的情境里能走多远，取决于记录有多全，而不取决于它自己能否在新互动里修正。发现学前提在这里的表现是：仿佛世界已经全在记录里，只差被学全。

第四处落点，是把不同信号之间的冲突当作误差消除掉。第19章的导航例子里，平面定位和爬坡信号若被当作「同一对象的不同噪声测量」，冲突就被平均掉；若被当作「对象尚未确定」的信息，冲突就推动对象的拆分。发现学前提的表现是：对象是先在的，各路信号只是对它的有噪声观测，冲突只能来自噪声。发生学的读法是：冲突可能恰恰说明对象还在发生之中。

这两处都要防止贴标签。只用记录训练，在许多任务上是合理的起步，V-JEPA 2正是先用大量录像、再用少量机器人数据；把冲突加权平均，在对象类别稳定的任务上是正确的工程选择，气象资料同化就靠它。限制只在一种情况下变得要命：当任务恰恰要求在新情境里形成新的对象或新的区分时，而训练机制却没有给这种形成留下通道。

这里有一个必须正面对待的反例。近几年，气象领域出现了一批主要靠历史资料训练的数据驱动预报模型。以谷歌DeepMind的GraphCast为例，其论文（arXiv 2212.12794）报告：模型直接用再分析资料训练，做十天以内的全球中期预报，在一千三百多个检验目标中约九成优于当时最准确的业务确定性预报系统

（作者自报）。从表面看，这正是「只从封闭记录中取差异」的路线，却做得很好。它是否推翻了本节的判断?

细看并没有，但它把判断收得更精确了。第一，这类模型所学的记录，是几十年来全球观测资料经数值模式同化而成的再分析资料，这些记录本身就是人类与模型反复互动、一天天被真实天气纠正过的产物，它们的「封闭」是相对的。第二，大尺度大气过程在几十年里相对稳定，记录中的差异结构足以覆盖大部分要预报的情境。第三，它们每天都在被第二天的实况检验，一旦失准就会被发现。也就是说，在一个记录极其丰富、过程相对稳定、检验天天发生的任务里，封闭记录可以走得很远。这恰恰支持了本章的说法：限制不在「用记录」本身，而在于任务是否要求记录之外的新对象、新区分，以及是否有持续的检验与回写。反过来，若遇到气候变化带来的、历史记录里罕见的极端天气，这类模型表现怎样，正是该被追问的地方〔待核〕。

这个反例也提醒我们，「发生取向」与「发现学前提」的区分，不能只看一个系统内部用了什么算法，还要看它所处的整个复合体：数据从哪里来，经过多少代人的观测与校正，谁在用它，错了由谁发现。一个看起来只读记录的模型，可能嵌在一个持续发生的人类观测系统里；一个看起来能自主行动的模型，也可能被关在一个与真实世界隔绝的仿真盒子里。判断要落在整个发生结构上。

◆ 五、现有技术中的发生取向

公平起见，必须同样认真地看现有技术里已经有的发生取向。这一节列出前四章碰到的几处，每一处都对应三方程中的某一环。

第一，以表示为预测目标。V-JEPA系列不预测像素，预测表示；表示本身是在训练中形成的，不是先定的。这意味着「需要保留哪些差异」在一定程度上交给了训练过程本身，而不是由人逐项规定。它向「显露在纠缠中形成」的方向走了一步。

第二，从自己的行动取经验。DreamerV3的经验来自自身行动，想象的错误要回到环境里暴露并纠正。这是差异运行从「读出」转向「做出」，对应D=G(S,E)这一环：差异是在显露与纠缠的当下关系中生成的。

第三，从真实运行中回写。Waymo博文描述的外部学习回路，让真实里程中暴露的不足，经评判、生成替代行为、仿真测试、安全审查，回写成新一版系统。这对应E=H(S,D)：纠缠在显露与差异的反馈中被改写。导航软件靠千千万万辆车的实际行驶不断修正路况估计，天气预报靠每天的实况核对不断校正，都是同一种结构。

第四，从无标注记录中归纳动作。Genie不靠人给动作标签，而让动作类别在录像差异中自己形成。虽然数量是先定的，但内容是发生出来的。

第五，以开放语义通道补先定清单。Waymo引入视觉语言模型处理罕见情境，Cosmos 3、Atlas把多种模态放进同一个上下文，都在给「清单之外的东西」留出进入的通道。是否真形成了跨模态的相互制约，要检验；但通道本身是发生取向的前提条件。

这五处都不是SDE理论推出来的，而是工程界在具体问题的压力下自己走出来的。本书把它们看作与发生学方向相合的经验事实，可以对话，可以作旁证，但不当作本体论的证明。反过来说，SDE在这里能提供的，是把这些分散的做法放进一个统一的问题里：它们各自在三方程的哪一环上松动了「先在世界」的前提，又在哪一环上仍然保留着它。

◆ 六、以具体训练机制判断，并向下一编推进

把本章收成一个可以拿去用的办法：面对任何一个世界模型，不问它属于哪个派别，只问下面四件事。

其一，对象与属性：训练前定了哪些对象类别、动作粒度、状态变量？这些预设能否在训练或使用中被修订？

其二，差异来源：训练的差异来自封闭记录、合成数据，还是来自系统自身的新互动？三者比例如何，是否公开？

其三，成功标准：是以显露相似为主，还是有对干预后果、陌生条件、失效恢复的检验？

其四，冲突与回写：模态或预测之间的冲突，被平均掉还是被用来推动对象重建？真实运行中的错误，有没有一条回写到参数的通道，经过谁的中介？

四件事里，某个系统在第一、三件上偏向发现学前提，在第二、四件上带着发生取向，这完全可能，而且是常态。这样的判断比一个标签有用得多：它告诉工程师可以在哪里改，告诉读者可以在哪里追问，也告诉我们自己，本书的主张要在哪里接受检验。如果某个系统在这四件事上都偏向发现学前提，却在一个开放任务里长期稳定地成功了，那也应当被如实记下，作为对本书机制假说的挑战。

第四编走了五章。第16章区分了世界模型的四类对象和成功标准；第17章看到视频学习学到的是记录中互动的差异结构；第18章看到状态与行动模型向任务靠拢，以及回写纠缠才算学习；第19章区分了多传感器融合与多种SDE纠缠；本章把发现学前提落到四个具体机制上，并指出现有技术中五处发生取向。

几条线最后汇向同一个缺口：现有路线在视觉上走得最远，在空间、语言、动作上已有可观的整合，而触觉，以及声音在对象形成中的作用，大多还停在局部研究。小舟换气那一刻，水托住身体的触感和出水时声音的一亮，在他那里不是附加信号，而是对象之所以成立的一部分。这到底只是人的特殊情况，还是物的同一性形成的一般条件？下一编将从视觉、触觉、听觉三种SDE各自如何完整地显露物开始，追问它们怎样纠缠成一个对象，以及「物的意识」可以用什么功能性判据来研究。

带走的话

发现学前提不是整个领域的标签，而是可以逐一指认的训练机制：先定对象清单、只取封闭记录、以显露相似代发生检验、把冲突当噪声。现有技术里以表示为目标、从自身行动取经验、从真实运行回写等做法，已经带着发生取向。

小账 论证落在哪一层？——本章以本体论立场为主（发现学前提与发生学的区分），以前四章的经验事实为审查材料，机制假说集中在四个限制点与审查清单。 世界模型补了哪一条？——补上一条方法条件——对世界模型的哲学审查要落到对象预设、差异来源、成功标准、冲突与回写四个具体机制上，不以领域或架构贴标签。

第 五 编

三种SDE的纠缠与物的意识

编序 第四编末尾留下一个问题：如果不把世界当作预先做好、等着被装进模型的东西，那么一个具体的「物」到底是怎样在经验里站起来的？本编正面回答它。

本编的核心主张是：视觉、触觉、听觉各自都是一套完整的发生过程，各有自己的显露、自己的差异运行、自己的特征纠缠；它们不是分别负责 S、D、E 的三个零件，也不对应粒子、波、场。三套过程在互动中彼此进入对方的发生条件，形成本书所说的「特征纠缠聚合体」，物的意识就在这个聚合体里发生。

五章接力如下。第21章从视觉入手，说明轮廓、位置、遮挡与运动差异怎样让一个东西被「看出来」，又为什么看得再真也不等于知道它碰上去会怎样。第22章转到触觉，讲受力、阻力、形变与滑移，讲接触历史怎样延续地影响后来的发生，并把「影响范围与伞模型」作为一个尚未形式化的机制假说摆出来。第23章讲听觉：声音天生是事件的声音，它把「发生了什么」和「是谁发出的」同时提出来。第24章用一只陌生弹性球的握压、敲击与释放，把三套过程放到同一件事上，分出同步出现、相关变化、相互制约三层，并说明冲突怎样推动归并、拆分和重建。第25章给「物的意识」一组可检验的功能性判据，再单独讨论主观体验问题，最后把「三种感官是否缺一不可」划为研究边界，而不是预先认定的结论。

本编的主例是挑西瓜（看、敲、托）、摸黑找开关，以及陌生弹性球。

21

第 21 章

视觉SDE中的物如何显露

> **这一章的问题**：眼睛看见「一个东西」，这个东西是怎样从一片明暗颜色里显露出来，又靠什么站稳的？

◆ 一、视觉不是一台相机，而是一套完整的发生

日常说「我看见了一只西瓜」，好像西瓜原本就在那里，眼睛只是把它照下来。这正是第20章所说的发现学图景：对象和属性先在，感官负责记录。本书的本体论立场不同：看见一个东西，是一次发生，而且视觉自己就足以构成一套完整的SDE。

说它完整，是指三个维度在视觉内部都有着落。视觉有它自己的显露：一块从背景里浮出来、有轮廓、有位置的「东西」，这是看的结果，也就是视觉中的 S。视觉有它自己的差异运行：明与暗的交界、颜色的跳变、这一刻与上一刻位置的差、前面的东西挡住后面东西时那条边的进退，这些差异一个接一个地运行，就是视觉中的 D。视觉还有它自己的特征纠缠：形状、颜色、纹理、大小、位置、运动方向，这些原本各自分散的特征，在反复看的经验里缠到一起，「绿底深纹、圆滚滚、在摊子上不动」成了一个整体，这就是视觉中的 E。

这里必须守住一条纪律：视觉不等于 S。很多人会顺手说「视觉负责显露，触觉负责差异，听觉负责纠缠」，这种分工看上去整齐，其实把三维降成了三个零件。视觉里有显露，也有差异和纠缠；触觉、听觉同样如此。三维是每一套发生过程内部的三个方向，不是分给三种感官的三份差事。同理，视觉也不对应「粒子」或「波」，光的物理描述是另一回事，本书讨论的是看这件事如何发生。

◆ 二、轮廓、位置、遮挡与运动差异

一个东西在视觉中显露，大体离不开四种差异。

第一是轮廓。明暗或颜色在某处陡然变化，形成一条边。边并不保证是「物的边」，墙上的阴影也有边，地砖的缝也有边。轮廓只是候选。

第二是位置。一条闭合的边圈出一块区域，这块区域在视野里有个位置，和其他区域有远近上下的关系。位置让候选有了「在哪」。

第三是遮挡。当一块区域挡住另一块，后者的轮廓在交界处断掉，而前者的边是连续的。遮挡带来一个非常重要的差异：谁在前、谁在后。更关键的是，被挡住的部分并没有被当作不存在。一只瓜被另一只瓜挡住一半，我们看到的仍是「一只被挡住的瓜」，而不是「半只瓜」。

第四是运动。一块区域里的所有点一起移动，而背景不动，这块区域就很容易被当作一个东西。心理学里很早就有「共同命运」的说法，指一起运动的部分倾向于被看成一组；关于婴儿的研究也报告过，很小的婴儿会依靠共同运动来分开物体，而不只依靠颜色和形状〔待核：具体研究者、年份与月龄〕。这些属于经验事实，本书只取它们支持的一般方向，不把细节当作本体论证明。

在这四种差异之上，本书提出一个机制假说：视觉对象的稳定化，取决于这些差异在多次观看中是否彼此一致地运行。同一块区域，轮廓闭合、位置连续、遮挡时被补全、运动时整体同步，这几路差异越是互相印证，视觉中的特征纠缠就越紧，显露出来的「一个东西」就越站得稳。反过来，如果一块区域在运动时裂成两半各走各的，原先的「一个」就会松动，分化成两个。这条假说是可以被检验的：在受控的画面里人为制造差异之间的不一致，观察对象判断何时破裂，就能支持或推翻它。

◆ 三、视觉内部的三方程：看是主动的

说视觉是完整的 SDE，不只是说三维各有着落，还要说三条方程在视觉内部都在运行。

先看 S=F(D,E)：看出来的东西，取决于差异怎样运行，也取决于已经缠在一起的特征。同一片绿，对第一次逛水果摊的孩子和对卖了二十年瓜的摊主，显露出来的东西不一样。差异相同，纠缠不同，显露就不同。

再看 D=G(S,E)：差异怎样运行，反过来受已经显露的东西和已有纠缠的支配。这一条最容易被忽略。人的眼睛并不是一动不动地「接收」画面，而是在不停地跳动、停留、再跳动。眼动研究早已报告，人在看一个场景时，注视点会频繁跳转，每秒若干次〔待核：常见的每秒注视跳转次数及出处〕。眼睛往哪里跳，取决于已经看出来的东西在哪里，也取决于看的人懂什么。会挑瓜的人，目光会落到瓜脐和底部黄斑；不会挑的人，目光停在瓜的正面。于是「哪些差异被运行起来」本身就是被显露和纠缠组织的。看，是一种动作。

最后是 E=H(S,D)：纠缠由显露和差异共同改写。每一次看出一个东西、又每一次在目光扫过时遇到新的差异，原来缠在一起的特征就被加固、松开或改组。

三条方程同时运行，视觉经验就有了三态的走向。刚到摊前，一片绿，边和边彼此打架，这是视觉上的混沌态；目光扫了几遍，几只瓜的轮廓开始浮出来，另几只还说不准是一只还是两只，这是介生态；最后每一只都有了稳定的边界和位置，即使摊主翻动瓜堆，也能跟住其中某一只，这是视觉中的秩序态。秩序态不是终点，它随时可以被新的差异打回介生态，比如摊主忽然把一只破了口的瓜翻到上面来。

这里有一个由此引出的机制假说：视觉对象的稳定程度，可以用「被扰动后多快恢复跟踪」来度量。一个只在静止画面里成立的对象，一遇遮挡或翻动就丢失；一个在视觉中真正站稳的对象，经得起短暂遮挡、换角度、光线变化而仍被认作同

一个。这个度量不涉及任何主观体验，只涉及能力，所以可以在人身上测，也可以在机器上测。第25章给物的意识定判据时，会把它推广到跨感官的情形。

◆ 四、案例：小舟在水果摊上「看」西瓜

这是我构想的场景。夏天傍晚，周敏带小舟去楼下水果摊买西瓜。摊主把几十只瓜堆成一座小山。小舟第一眼看到的其实不是一只只瓜，而是一片绿。要从这一片里看出「一只瓜」，靠的正是上面四种差异：瓜与瓜之间有一道深色的缝，那是轮廓；最上面那只的边完整，下面那些被压住一截，那是遮挡；摊主随手把一只滚到秤上，它整个儿一起动，背景不动，那是运动。几秒钟之内，「一堆绿」显露成「很多只瓜」。

接着周敏教他看瓜。她说，纹路要清楚、黑绿分明；瓜脐小一点好；底下贴地那块黄斑颜色越深，越可能在地里长足了日子。这些话把小舟原本没注意的差异指了出来：纹路清不清是一种差异，黄斑深浅是另一种差异。看过几只之后，小舟眼里的瓜不再一样了，他开始自己比较「这只纹路散，那只纹路紧」。这就是视觉中的特征纠缠在改组：颜色、纹路、黄斑、瓜脐原本各看各的，现在缠成了一个叫「看起来熟」的整体判断。

后来那个夏天，家里每切开一只瓜，小舟都要凑过来看。瓤红不红、籽黑不黑、切面是不是起沙，这些都是新的视觉差异，它们来自瓜的内部，原本藏在外皮之下。切开之后，外面的纹路黄斑和里面的红瓤沙瓤第一次出现在同一段经验里，前后相隔不过几十秒。几只瓜吃下来，小舟慢慢有了一点自己的把握：「纹路紧的这几只，切开都还行。」这是视觉内部的一次回写，内部显露回过头来改写了对外皮差异的组织，让「看起来熟」这个纠缠多了一份依据。

但这份依据仍然很窄。它只来自家里买的那几只瓜，来自同一个摊子、同一个季节；换一个品种，纹路本来就淡，他的判断就会失灵。而且这条回写要靠一刀切开才能完成，在摊子上，瓜是不能先切的。也就是说，视觉在这里的学习，依赖一个不在摊前发生的后果。

值得注意的是，小舟这时已经学到了一点东西，但他学到的是「怎样看」，还不是「瓜到底熟不熟」。他的视觉 SDE 变得更细，显露出来的对象更多层次，可这个显露站得稳不稳，还要看后面的事。

◆ 五、同一外观与不同发生条件

视觉经验的力量很大。它远距离、快、一次能同时抓住许多东西，还能连续追踪运动。人在日常生活中做的大多数对象判断，视觉都是先到的那一路。

但视觉有一个结构性的边界：同一个外观，可以出自完全不同的发生条件。摊上那只纹路清楚、黄斑深黄的瓜，可能是熟瓜，也可能是采摘偏早、只是外皮长得好的生瓜。商场橱窗里的塑料仿真西瓜，外观可以做得以假乱真，而它碰上去是硬壳、拎起来轻飘飘。手机屏幕上一张西瓜照片，视觉显露几乎完全一样，却根本不能被抱起来。这三者在视觉 SDE 里可以显露为同一类东西，在触觉和行动中却分属完全不同的发生条件。

这一点对世界模型很要紧。第17章讨论过从视频中学习世界的路线，例如 Genie 这类工作，从大量无标注视频中学出隐动作，进而生成可以被操控的画面（arXiv 2402.15391）；第18章讨论的 V-JEPA 一类路线，则在表示空间里预测被遮住部分的特征。这些都是经验事实层面的技术进展，它们证明了一件事：仅靠视觉材料，就能学出相当稳定的视觉显露，包括遮挡后的补全和运动的连续。它们没有、也不需要证明另一件事：画面上稳定的物，在接触、受力和行动后果上也会照样成立。这不是在说这些模型「没有互动」或「必然失败」，它们是否接入动作、接入多少动作，要按具体版本的公开机制判断；这里只是指出，视觉显露的一致，与物在其他发生条件下的一致，是两个需要分开检验的问题。

还有一种更细的边界：视觉自己也会被骗。透明玻璃没有清楚的轮廓，人会撞上去；一块平整的镜子会把背后的空间「显露」到前面；暗处的一团衣服会被看成一个蹲着的人。这些不是视觉的失败，而恰恰说明视觉 SDE 是在差异中发生的：差异不足或差异被伪造时，显露就会走偏。视觉越是成熟，越能察觉这种走偏，比

如发现「这块玻璃反光不对」。但在很多情况下，单靠视觉无法自己纠正，需要另一路过程来插一句话。

把这条边界放到机器上，还需要多分一层。机器的「看」与人的看，发生条件并不相同：摄像头不会自己跳动注视点，它按固定帧率把整幅画面交给模型；模型里也没有一个弯腰去瞧瓜底的身体。可是这并不意味着机器视觉里没有差异运行。在训练中，哪些画面被选进数据、哪些区域被遮住要求预测、损失函数惩罚哪一种误差，都在决定「哪些差异被运行起来」。换句话说，人眼的跳动是个人身体在组织差异，机器视觉的差异则由集体人类—AI复合体在训练设计中组织。这一点与第11章讲的「人怎样进入大模型训练」是同一回事，只是换到了画面上。

由此可以提出一个检验视觉对象是否站稳的办法，而且它对人和机器都适用：给出一段正常的物体运动，再给出一段违反常理的运动，比如一个东西被挡住后在另一处凭空出现，或者穿过了本该挡住它的东西，看系统是否对后者表现出明显更大的「意外」。婴儿研究长期使用注视时间来做这类实验。也有公开工作把「违背预期」测试搬到了机器上：一篇题为《直觉物理理解从自然视频的自监督预训练中涌现》的论文（arXiv　2502.11831）报告，在学到的表示空间里做预测的视频模型，在物体恒存、形状一致等若干直觉物理性质上表现出明显高于随机的区分，而直接预测像素的视频模型和多模态大语言模型则接近随机水平；即使只用约一周长度的独特视频训练，前一类模型也仍高于随机（作者自报）〔待核：所用模型的具体版本与各测试项的结果幅度，以正文为准〕。如果这类结果成立，它说明的是视觉显露中的对象连续性可以从大量观看中训练出来；它不说明这个对象在被托起、被敲击时会怎样，也不能当作「视觉足以产生物的意识」的本体论证明。三层话在这里要分清：对象连续性能被视觉训练出来，是有待核对的经验事实；对象稳定化取决于多路差异一致运行，是机制假说；视觉是一套完整的 SDE、但不等于物本身的全部发生，是本体论立场。

◆ 六、看得再清楚，也要伸手

回到水果摊。小舟挑中一只看起来最好的瓜，周敏没有马上付钱。她伸出手，在瓜上拍了两下，又用手掌托住瓜底掂了掂。小舟问：「看都看好了，为什么还要拍？」这个问题正好是本章留给下一章的。

视觉 SDE 让西瓜作为一个有轮廓、有位置、能被追踪的东西显露出来，而且在长期经验中把「看起来熟」这类判断组织得相当细。但视觉显露所依赖的，是光在表面形成的差异；瓜皮之下发生了什么，瓜的分量和弹性如何，手碰上去会怎样回应，视觉只能间接猜测。这里不是在给视觉打低分。在很多任务里，视觉一路已经足够：认出自家的车停在哪里，判断对面行人是在走还是在站，看出桌上的杯子倒了。一个任务需要多少路过程参与，要由任务本身决定，第28章还要细讲任务如何组织世界的发生范围。本章只主张：当任务涉及物的内部、分量、弹性和接触后果时，视觉显露的稳定并不自动延伸过去。

周敏那一拍一托，正是在这种任务里补上视觉够不着的部分。手掌一拍，瓜皮在掌下回弹，这是触觉里的差异；耳朵里传来一声闷响或脆响，这是听觉里的差异；手一托，分量压到掌心，又是触觉里的另一种差异。三路差异落在同一只瓜上，彼此印证或彼此打架。本编要讲的，正是这种印证和打架怎样组织出一个物。

要让这些进入发生，需要另一套完整的过程，一套以接触、受力和形变为差异的过程。下一章从触觉开始，而且换一个不靠眼睛的场景：摸黑找开关。

> **带走的话**
>
> 视觉是一套完整的 SDE，物在轮廓、位置、遮挡与运动差异的相互印证中显露并站稳；但同一外观可以出自不同发生条件，视觉显露的一致不能替代物在接触与行动中的一致。

> **小账** 论证落在哪一层？——本章以本体论立场（视觉是完整 SDE、不与 S 对应）和机制假说（视觉对象稳定化取决于多路差异的一致运行）为主，经验事实（共同运动、视频与特征预测模型）只作支撑与对照。　世界模型补了哪一条？——补上一条边界——视觉显露的稳定可以单独形成，但它不自动包含接触与行动后果的稳定，二者需分开检验。

22

第 22 章

触觉SDE中的接触与持续影响

这一章的问题：在完全看不见的时候，一只手怎样让一个东西在接触中显露出来，而这一次接触又怎样影响下一次？

◆ 一、摸黑找开关：一套不靠眼睛的完整发生

这是我构想的场景，但几乎每个人都经历过。夜里停电，周敏摸黑起身，想去门边按亮应急灯的开关。她的手先贴到墙面，墙是凉的、平的、微微有点粗糙；手掌往右滑，碰到一道凸起的边，那是门框；顺着门框往下摸，摸到一块比墙面略高、边缘规整的方形塑料，再往中间一按，「咔」的一下。灯亮了。

整个过程里，眼睛没有提供任何东西，可「开关」照样显露了出来，而且显露得足够准，准到可以按下去。这件小事说明：触觉自己就是一套完整的 SDE，不是视觉的补丁。

触觉有它自己的显露：一个有边、有高低、有软硬、在某个位置的东西，从一整片「墙」里浮出来。触觉有它自己的差异运行：凉与不凉、平与凸、滑得动与滑不动、按下去有回弹与没有回弹，这些差异随着手的移动一个接一个地出现。触觉还有它自己的特征纠缠：「方形边缘、略高于墙、塑料温感、中间可按、按时有段落感」缠在一起，成为「开关」。

和视觉相比，触觉有一个鲜明的不同：它的差异几乎全部依赖动作。眼睛不动也能看到一片东西，手不动则几乎什么也摸不出来。手掌静静压在墙上，几秒之后连凉都觉不出了；一旦滑动，纹理、边缘、温差立刻回来。所以在触觉里，

D=G(S,E) 这条方程格外显眼：已经摸出来的东西（门框在这里）和已有的纠缠（开关一般装在门边、齐腰高），直接决定手接下来往哪里摸、用多大劲。

这里同样要守住纪律：触觉不等于 D，也不对应「场」。触觉里显露、差异、纠缠一样不缺，它只是与视觉的差异来源不同，运行方式不同。

◆ 二、受力、阻力、形变与滑移

触觉中的差异，可以粗分为四类。

第一是受力。手压下去，压力大小、分布在哪里，这是最基本的差异。拎起西瓜时掌心的压感，按开关时指尖的压感，都属于这一类。

第二是阻力。手往前推，东西推得动还是推不动，推的时候吃不吃劲。墙推不动，门推得动，抽屉推起来先紧后松。阻力让「这里有个东西」变成「这个东西会怎样回应我」。

第三是形变。按下去会不会凹，凹多少，松手后弹不弹回来，弹得快不快。墙不形变，海绵大幅形变而缓慢恢复，熟透的桃子形变后不恢复。形变是触觉区分材料最主要的差异之一。

第四是滑移。手指在表面上滑，滑得顺不顺，有没有一下一下的卡顿；拿着的东西有没有往下溜。滑移是接触中最微妙的差异，它往往出现在要失手之前的一瞬间。人拿杯子的时候，并不是一开始就用很大力气，而是用刚好够的力，一旦指尖觉出轻微的滑移，握力会在极短时间内自动加大。生理学研究对这一点有较多报告，指出手指皮肤中对微小滑移敏感的感受器参与了这种快速调节〔待核：相关经典研究者与反应时间量级〕。

这四类差异的共同点是：它们都不是「物的属性」被读取出来，而是手与物相遇时发生的事。同一块面团，揉之前和揉之后，手感完全不同；同一块海绵，干的和湿的，阻力和形变都不一样。属性不是先在的数值，而是在特定的接触方式下显露出来的稳定态。

◆ 三、接触历史如何影响后续发生

视觉中的一瞥，可以在下一瞥时被完全替换；触觉不同，一次接触会留下延续的后果，影响后面的发生。这可以从两头说。

从物这一头说，接触改变物。一只西瓜被按过一下，按下的地方可能留下极轻的凹陷；门把手转过一次，弹簧的位置变了；面团揉了五分钟，筋已经起了一部分，下一次按下去的回应就不同。接触不是无痕的读取，它本身就是物的发生条件之一。

从手这一头说，接触改变手接下来怎样接触。周敏摸到门框之后，手的高度和方向就被定住了；她刚才摸墙时觉得凉，摸到塑料时那一点「不凉」就格外分明。前一次接触在手里留下的东西，并不是一张「墙面记录」，而是一种调好了的准备：力度调到多少，手掌压多紧，下一步往哪边去。

回到水果摊，看周敏托瓜那一下，接触历史的作用更清楚。她不是把瓜一把抱起来，而是先用手掌贴住瓜底，稍微往上一送，再停一下，让瓜的分量完全落进掌心。这几步有先后：先贴住，是为了让手和瓜之间不滑；再上送，是为了感到瓜「跟不跟手」；最后停住，才去感觉沉不沉。每一步都建立在上一步留下的接触状态之上。如果第一步没贴稳，第二步感到的就主要是滑移，而不是分量。

她托过上百只瓜，手里已经缠好了「这么大个头的瓜，大致该有多沉」的纠缠。所以她托起一只瓜，显露出来的不是一个重量数字，而是一个落差：比预想的沉，或者比预想的轻。她常说「同样大小，压手的那只水分足」。这个判断对不对，是生活经验，本书不作保证；值得注意的是它的结构：触觉的显露是相对于已有纠缠的差异，而已有纠缠来自此前一只只瓜的接触历史。

小舟学着托，结果完全不同。他两只手抱住瓜，用力过猛，瓜被抱得紧紧的，他只觉得「重」，分不出比谁重。这正是触觉 SDE 在混沌态上的样子：差异太粗，又没有可以比较的纠缠，显露只有一个笼统的「重」。要等他托过许多只、托过之后又知道切开是什么样子，触觉里的「压手」才会慢慢从「重」里分化出来，

进入介生态，再站稳成秩序态。这个过程与第2章所说的和面揉出筋同理，揉面的手感也是一次次接触叠出来的；这里只一句带过，不再展开。

把这两头合起来，就得到本章的一个机制假说：在触觉 SDE 中，一次接触的影响不止于接触那一刻，而是在一定范围、一定时间内持续地改变后续接触的发生条件。这个范围有大有小。轻轻擦过墙面，影响很快就消散了；用力推一扇门，门的状态改变了，接下来进出这扇门的每一次接触都在这次推的影响之下；拧紧一颗螺丝，影响可以持续几年。

◆ 四、影响范围与伞模型：一个尚未形式化的问题

这里有一个问题，本书称之为「伞模型」。必须先说清楚：伞模型目前只是一个待形式化的机制设想，下面写的是它要回答的问题和可能的形状，不是一个已经成立的理论，更不是本体论结论。

设想的图景是这样的：一次接触像在某一点撑开一把伞。伞柄是接触发生的那个点和那一刻；伞面是这次接触影响所及的范围，既包括空间（接触点周围多大一片物体被带着变化），也包括时间（之后多久的接触仍受它影响），还包括身体（这次接触之后，手的准备状态改变了多少）。落在伞面之下的后续事件，其发生条件被这次接触改写过；落在伞面之外的，则基本不受影响。

用一个日常例子说明这个设想。周敏按下开关的那一下，伞面很小：它只改变了这个开关的通断，几秒之后，她的手再去摸桌上的手电，几乎不受刚才那一按的影响。可是如果她摸黑时一脚踢到了椅子腿，伞面就大了：椅子被踢歪，接下来她走过去的路变了；脚趾隐隐作痛，接下来每一步都放得更轻更慢；她心里记住「椅子在那里」，后面一段时间里身体会自动绕开那一块。同一类事件（一次接触），撑开的伞可以差得很远。

如果这个设想有价值，它至少要回答四个问题，而这四个问题目前都没有回答。第一，伞面怎样界定：影响多大才算「在伞下」，用什么量来度量。第二，伞面怎样衰减：是随时间和距离平滑地减弱，还是有清楚的边界，或者会因为后续事件被突然收起。第三，多把伞怎样叠加：连续几次接触撑开的伞相互重叠时，影响

是相加、相互抵消，还是由最近的一把说了算。第四，伞的大小由什么决定：材料、力度、接触方式，还是主体的身体状态与任务。

虽然没有形式化，伞模型的设想至少可以落到可以做的实验上，这也是它被称为机制假说而不只是比喻的理由。一种做法是「探针式」的：先让手或机械手以某种方式接触物体一次，然后在不同的时间间隔、不同的空间位置做一次标准化的轻探，比较这次轻探的反馈和「没有那次先前接触」时的反馈相差多少。差异显著的区域和时段，就是那把伞的伞面的粗略轮廓。另一种做法是「预测式」的：训练一个预测接触后续的模型，分别给它「只看当前接触」和「附带若干步接触历史」两种输入，看后者在哪些材料和动作上明显更准；准得多的地方，说明接触历史的持续影响大，伞面宽。两种做法都可能得出伞面很窄、可以忽略的结论，那时就该收起这个设想。

伞模型也和三原理有一层关系，同样只是推测。一次接触撑开伞面之后，后续接触在伞下进行，如果新的触觉差异与伞下已经改写过的纠缠对不上，比如本以为椅子还在原处，脚下却空了，就会出现张力；张力积累到一定程度，迫使显露改变形状（「椅子被人挪了」），再回写到纠缠里。伞面的边界，或许正是这种张力开始频繁出现的地方。这一层关系更远离可检验，此处只记下，供以后形式化时参考。

这四个问题回答之前，「伞模型」只能作为一个提问方式使用。它的用处在于提醒世界模型研究者：触觉经验不能只按「一帧一帧的接触读数」去组织，因为每一帧都可能在后续一段范围里持续起作用。至于是否需要专门的表示结构来承载这种持续影响，还是现有的循环记忆、状态空间方法已经足够承担，需要实验来回答，而不是由本书预先断言。

◆ 五、身体、工具和机器接触反馈的差别

触觉 SDE 的发生条件，会随接触的承担者不同而变化很大。

身体直接接触时，皮肤、肌肉、关节一起参与，压力、温度、振动、关节角度同时涌来，差异极其丰富，而且与动作控制紧密连在一起。周敏的手不只是「感到」开关，它同时在调节方向和力度。

通过工具接触时，情况变了。老人用拐杖探路，拐杖尖碰到台阶边缘，手里传来的只是一根杆子上的震动和阻力；可熟练的人会觉得自己「摸到了台阶」，而不是「摸到了拐杖」。用筷子夹豆腐也一样，熟练之后，注意力落在筷子尖上，而不是手指上。工具改变了差异的来源，压缩了差异的种类，却并没有让触觉 SDE 消失；经过训练，特征纠缠会把工具的另一端纳进来，显露的对象落在工具之外。这一点与第8章所说「身体与工具改变现实发生条件」相呼应。

身体自身的状态，也是触觉发生条件的一部分，而且常被忽略。冬天手冻僵了，指尖对纹理和滑移的分辨明显变钝，这时去拧一个细小的螺丝，常常拧不住；戴上棉手套，摸黑找开关几乎做不成，因为墙和塑料的温差、边缘的细小高低全被手套吃掉了；手上沾了水或油，滑移的差异又变得格外大，拿东西要格外小心。这些例子说明，触觉 SDE 中的差异不是物单方面送来的，而是手与物共同组成的。同一个开关，对一只温热干燥的手和一只冻僵的手，显露的清晰度不同；手套一戴，原来站稳的「开关」又退回介生态，需要用别的差异补，比如更用力地按、靠「咔」声确认。

这一点对机器同样有启发：传感器表面的磨损、温漂、外覆材料的老化，都会改变接触差异的来源，等于手套戴到了机器手上。一个在新传感器上训练好的接触模型，换了一层磨损的表皮，是否还能维持原来的判断，是一个需要专门检验的问题〔待核：是否有公开工作系统报告触觉传感器老化对学得模型的影响〕。

机器接触又不同。机器人手上的触觉传感器，有的测压力分布，有的用摄像头拍摄软胶表面的形变，有的测振动。近年已有研究专门训练模型预测接触怎样随动作演化，用于灵巧手的局部操作，DexTacWAM（arXiv 2609.24976）与 TouchWorld（arXiv 2607.07287）是两项2026年的公开工作，它们的做法与边界放到第33章细看。这里只取一点：这些研究说明，机器也可以在接触差异中训练出一定的预测能力。但与人手相比，有三处差别需要单独检验，不宜预先认定：机

器的传感器覆盖面通常远小于人手；它的接触信号与动作控制的耦合方式，取决于具体系统的设计；它的接触历史能保留多久、以什么形式影响后续，是工程选择，而不是自然给定。这些差别是否构成根本限制，要按具体系统、具体任务去测。

三处差别里，第三处最值得多想一步，因为它牵涉到「谁在摸」。人摸东西，手是自己动的：往哪里滑、停在哪里、压多重，都是手跟着刚摸出来的东西随时改的，这正是前面说的 D=G(S,E)。换一种情形，别人捏着你的手指在一块布上划过去，你的手不动脑筋，只是被带着走，同一块布就难认得多。心理学里有过「主动触摸」与「被动触摸」的对照研究，报告前者辨认形状明显更准〔待核：经典研究的出处与具体结果〕。放到机器上，这就成了一个设计问题：一只机械手如果只是沿着事先编好的轨迹走一遍、把一路的接触读数记下来，它得到的更像被动触摸；如果它能按刚摸到的东西改下一步往哪里摸、用多大劲，触觉的差异运行才由它自己组织起来。两种做法哪一种在具体任务上更有用，同样要靠对照实验回答；第32章讲主动试探，讲的就是这件事在训练里的一般形式。

◆ 六、触觉不知道远处，也不知道名字

触觉的力量在于贴身。它告诉你这个东西推不推得动，按下去会怎样，拿着会不会滑。在摸黑找开关这类任务里，它几乎单独完成了整件事。

但触觉也有它的边界。它只能触及伸手可及的范围，远处的东西摸不到；它的差异要一点一点地扫出来，慢；它对整体形状的把握往往不如一瞥。更重要的是，触觉中的显露常常有歧义：手碰到一个凸起的方块，可能是开关，也可能是插座盖板，或者是贴在墙上的一块胶带。周敏按下去之前其实并不完全确定，是「咔」的那一声和随之亮起的灯，把这个显露最后站稳了。

这一声，就是下一章的入口。声音不需要伸手，也不需要光，它在事件发生的那一刻传来，告诉人「发生了什么」以及「是哪里发出的」。听觉是第三套完整的发生过程。

带走的话

触觉是依靠动作运行差异的完整 SDE，受力、阻力、形变与滑移让物在相遇中显露；接触留下持续影响，其范围如何刻画（伞模型）仍是待形式化的机制问题。

小账 论证落在哪一层？——本章以本体论立场（触觉是完整 SDE，属性是接触中的稳定态）与机制假说（接触影响的持续范围、伞模型设想）为主；经验事实只用于滑移调节、主动触摸与机器触觉研究的定位，细节标〔待核〕或指向第33章。　世界模型补了哪一条？——补上一条条件——物理世界模型若涉及接触，需要处理接触历史对后续发生的持续影响；其表示方式尚未形式化，应作为开放问题进入实验设计。

23

第 23 章

听觉SDE中的事件与物

这一章的问题：一声响只持续零点几秒，它凭什么能告诉人「发生了什么事」，又凭什么能把这件事归到某一个物身上？

◆ 一、声音首先是事件，其次才是物

看见的东西，常常是静的：一只瓜放在摊上，可以一直看着它。摸到的东西，也可以一直按着。声音不同：没有事情发生，就没有声音。一只西瓜静静躺着，它不发声；手一拍，它才「嘭」地响一下。开关不响，按下去才「咔」。所以在听觉里，最先显露出来的往往不是一个物，而是一件事：有东西被敲了，有东西掉了，有门关上了。物是在事件之后、借着事件被归属出来的。

这就是本章的本体论立场：听觉同样是一套完整的 SDE，但它的显露以事件为先。听觉有它自己的显露：一个在某个时刻、从某个方向传来、有始有终的「响动」，以及它背后那个被认出来的发声者。听觉有它自己的差异运行：音高的高低、响度的强弱、音色的明暗、起音快慢、余音长短、左右两耳之间极短的时间差和强度差，以及这一声和上一声之间的差别。听觉还有它自己的特征纠缠：「起音很快、余音短、偏高、清脆」缠在一起，成为「塑料件扣合」；「低沉、闷、有一点拖尾」缠在一起，成为「实心的大东西被拍」。

纪律在这里同样适用：听觉不对应 E，也不因为声音在物理上是振动就对应「波」。听觉在本书里是一种发生过程，不是一种物理量的名字。

◆ 二、发声、振动、时间变化与声源归属

一声响从哪里来？粗略地说，是某个物在某种作用下振动起来。同一个作用（敲一下），落在不同的物上，振动方式不同：实心的、空心的、绷紧的、松软的，各有各的响法。同一个物，受到不同的作用（敲、刮、摔、摇），声音也各不相同。所以一声响里，同时缠着「物是什么样的」和「对它做了什么」这两件事。

听觉要把这两件事分开，靠的是时间变化中的差异。一声敲击，先是一个很快的起音，然后是逐渐衰减的余音。起音更多反映「做了什么」：用指节敲、用掌心拍、用硬物碰，起音的尖锐程度不同。余音更多反映「物是什么样的」：衰减得快，说明东西软、吸收振动；衰减得慢、有持续的嗡声，说明东西硬、或者里面空。这是一个粗糙的经验划分，不是严格的物理定律，但它说明了听觉怎样在时间中运行差异。

接着是声源归属：这一声是谁发出的。声源归属有三条线索。一是方向，两只耳朵听到同一声响的时间差和强度差，可以给出大致方位，但精度有限，尤其是前后、上下容易混。二是与其他声音的关系，同时响起、同节奏起落的声音倾向于被归为同一个声源，这和视觉里的「共同命运」是同一类原理。三是与看见、摸到的事件对齐：手拍下去的那一刻正好响了一声，这一声就归给了手下的那只瓜。

第三条线索特别重要，因为它意味着声源归属常常不是听觉单独完成的。这里提出一个机制假说：声源归属的稳定，主要取决于听觉事件与动作、触觉或视觉事件在时间上的对齐程度；对齐越准、越可重复，归属越快站稳。这个假说可以检验：在实验中人为延迟声音与动作之间的时间，看归属判断何时开始动摇。心理学中关于视听同步窗口的研究，报告过人对音画错位的容忍有一个大致范围〔待核：常见报道的视听同步窗口量级与出处〕，这类结果可以作为检验的起点。

◆ 三、案例：周敏敲西瓜

还是那只西瓜，场景接着上面构想的那一段。周敏伸出手，用指节在瓜身上敲了两下，偏着头听，然后换一只再敲。小舟在旁边学，也敲，敲完一脸茫然：「都一样啊。」

周敏告诉他，熟的瓜敲起来「嘭嘭」的，有点闷又有点弹，像拍在一面鼓上；生瓜敲起来「当当」的，声音脆、发硬；熟过头的瓜敲起来「噗噗」的，闷而发哑，没有弹性。这套说法流传很广，靠不靠得住，要看品种、个头、气温，本书不作担保。我们关心的是听觉 SDE 在这里怎样发生。

小舟听到的确实是声音，但对他来说，差异还没有被组织起来。三种声音在他耳朵里都只是「敲瓜的声音」，这是听觉上的混沌态。周敏能听出区别，是因为她在多年买瓜、切瓜之后，听觉里缠好了「这种声音—切开后这种瓤」的对应。她敲的时候，D=G(S,E) 也在起作用：她不随便敲，而是敲在瓜的腰部，用同样的力度，敲两下，第二下和第一下对比。这些动作都是在控制差异：固定「做了什么」，让声音之间的差别只剩下「物是什么样的」。

小舟要学会，靠的不是背下「嘭嘭」「当当」这几个字，那只是语言上的标签。他要做的，是在同一个摊子上，用同样的手法，连续敲几只瓜，让声音之间的差异在短时间内并排出现。周敏说：「你别只听一个，你连着敲三个，比一比。」并排比较的时候，原本相同的「敲瓜声」开始分化，有的闷一点，有的脆一点。再等切开一两只，看见里面的瓤，这种分化才会和别的东西缠上，从介生态慢慢走向秩序态。

敲瓜这件事还可以看出听觉里 S=F(D,E) 的作用。同一声「嘭」，摊主听来是「这只沙了」，周敏听来是「还行」，小舟听来什么也不是。三个人耳朵接收到的振动差不多，差异运行也差不多，可各自的纠缠不同，显露出来的东西就不同。摊主的纠缠来自几千只瓜，周敏的来自几百只，小舟的几乎是空的。这也解释了为什么「教」在这里收效有限：周敏可以把她的显露说给小舟听（「这只闷」），但说出来的是语言，进入的是小舟的理念世界；要让它进入小舟的听觉 SDE，还得靠他

自己一次次敲、听、比、切开。语言可以指出该注意哪种差异，却不能替他缠好纠缠。

听觉里还有一种很容易被忽略的主动性：人会为了听而动作。周敏把瓜挪到摊子边上，偏着头，把耳朵凑近；敲完一下停一停，等余音散了再敲第二下。这些动作都是在组织差异运行，让要听的差异突出出来，让干扰的差异退下去。和第21章讲眼睛跳动、第22章讲手的滑动一样，听也是一种行动，只不过它的动作更隐蔽：转头、靠近、静下来、控制敲击的力度和节奏。一个只被动接收录音的系统，没有这一层；它听到的是别人当时怎样敲、在多吵的地方敲。这不等于说被动录音没有价值，第17章讨论的从记录中学习的路线，恰恰说明大量记录能训练出可观的能力；只是说，录音里的差异运行是由录音时的那次行动组织的，新的行动要求新的差异时，记录未必能给出。

再往前推一步，可以提出一个可检验的机制假说：在敲击类判断中，如果允许学习者自己控制敲的位置、力度和节奏，并能把每一次敲击与之后的后果（切开看）对上，其听觉判断的形成会明显快于只听别人敲击录音的学习者。这个假说对人和机器都可以设计实验：对人，可以用同一批果实或同一批材料样本分组训练；对机器，可以比较主动敲击并获得后果标签的训练，与只用被动音频记录的训练。结果也可能相反，比如被动录音量足够大时差距消失，那就说明在这一类任务上，差异的组织方式不是瓶颈。第32章讲「从被动记录到主动试探」，正是这一类问题的工程展开。

这里有一个细节值得一提：周敏一边敲，一边用另一只手托着瓜。她说托着敲，「手心能感到它在震」。这时听觉的差异和触觉的差异，在同一个事件里同时出现。这个细节下一章要展开。

◆ 四、声音与行动后果的联系

在摸黑找开关的场景里，「咔」那一声的作用很特别。周敏的手已经按下去了，触觉告诉她「按动了」，但按动的是不是开关、开关到底通没通，在灯亮之前并不确定。「咔」的一声，把一个行动的后果用事件的方式宣告出来。

日常生活中，大量声音承担着这种「后果确认」的作用。拧瓶盖时「咔哒」一声，说明封口断开了；插头插进插座，听到一下轻响，说明插到位了；锁门时钥匙转到头的那声，说明锁舌出来了；水壶里的声音从嘶嘶变成咕嘟，说明水开了。这些声音和动作一起，构成了对后果的及时回写：动作做完，声音立刻告诉你世界变了没有、变成了什么样。没有这声回响，人常常要多看一眼、多试一次。

声音和后果之间的这层联系，是学出来的，而且可以被改写。家里换了一种静音开关，按下去几乎没有声音，周敏头几天常常按完还要抬头看一眼灯，甚至按两次，结果又把灯关了。原来缠好的「按下—咔—灯亮」被拆掉了一环，动作和后果之间出现了一段空白，她只好用视觉去补。过了一阵，她的手学会了用按键的段落感来确认，触觉接过了原来声音的职责，新的纠缠站稳了。这个小事说明：哪一路过程承担后果确认，不是由感官的天性固定的，而是在具体的物、具体的行动里分工出来的，分工还会随物的改变而重组。

人造的声音也会利用这层联系。手机键盘按下去会发出模拟的「哒」声，有些电动汽车低速行驶时会播放人工提示音。前者是设计者有意给一个本来无声的动作配上后果确认，后者是为了让行人能够听到一辆本来安静的车。这些设计恰好反过来证明，人在行动中确实依赖声音来确认后果、来发现正在逼近的事件，以至于当物理上不再有声音时，人们要把它造出来。当然，人造声音也意味着声音和物的真实状态之间可以脱钩：一个「哒」声并不保证输入已经成功。这又是一种需要其他过程来核对的歧义。

这里可以看出听觉在行动中的独特位置。视觉对后果的确认往往要等结果稳定（灯亮了），触觉对后果的确认往往只在接触点（按下去了），而声音在事件发生的那一刻、从事件发生的地方传来，时间上最贴近，空间上又不必伸手。所以在许多需要快速确认的动作里，声音是一个很有效的后果信号。这是一个关于功能位置的描述，不是说听觉比视觉、触觉更重要；在另一些任务里，声音完全不起作用。

◆ 五、听觉参与对象确认与异常发现

声音还有两种作用值得单独说。

一是确认对象。眼睛看到一个杯子，不确定是玻璃还是塑料，轻轻一碰，「叮」一声，玻璃的；「嗒」一声，塑料的。看与碰的结果可能完全一样，声音却把材料分开了。这是第21章所说的「同一外观、不同发生条件」的一个解法：视觉显露相同的两个物，在听觉里可以显露为不同。

二是发现异常。一个熟悉的物，它的声音也是熟悉的；一旦声音变了，往往是物变了。洗衣机转起来声音不对，可能是里面混进了硬币；走在木地板上某一块忽然吱呀响，可能是那块松了。老司机听出车上的异响，属于同一类能力，第29章会以老彭为主例细讲，这里只一句带过。异常发现的结构是：听觉里已经缠好了一个物的正常声音（纠缠），当前的声音与之对不上（差异），于是物的显露被打回介生态，「这东西好像出问题了」。这正是三原理的一次运行：声音和已有纠缠之间出现张力，推动显露改变。

这两种作用都说明，听觉不只是给物「配个声」，它能改写物在整体中的显露。

◆ 六、遮蔽与歧义：听觉的边界

听觉也有自己的边界，而且相当明显。

第一是遮蔽。周围一吵，小的声音就被盖掉。菜市场里人声鼎沸，敲西瓜的声音很难听清；周敏常常要把瓜挪到摊子边上、凑近耳朵敲。遮蔽不仅是听不到，而且会让人以为「没有声音」，从而误判「没有发生」。

第二是归属歧义。同一时刻有几个声源，声音混在一起，很难分清谁是谁。夜里听到一声闷响，可能是楼上掉了东西，也可能是隔壁关门，单靠听觉往往定不下来。方向判断的精度有限，在有回声的房间里更差。

第三是因果歧义。一样的声音可以有不同的来源：敲熟瓜的「嘭嘭」，换一只更大的生瓜，也可能敲出相近的声音；隔着一块木板，空心和实心的声音也会走样。声音告诉你「这样的事件发生了」，却不总能告诉你「是哪种物、在怎样的条件下发生的」。

第四是一过即逝。看不清可以再看一眼，摸不准可以再摸一下，声音却不能重听；没听清的那一声，只能等事件再发生一次。所以听觉判断特别依赖短时记忆和可重复的动作：周敏敲两下，是因为第一下的声音已经消失，只能用第二下来对照。这也意味着，凡是只发生一次、不能重复的事件，比如一个东西从高处掉下来摔了，听觉只给出一次机会，判断容易出错，常需要事后用眼睛去补看现场。

第五是听觉在机器上的处境。许多世界模型的公开工作以画面为主，也有工作把声音放进统一架构共同建模，第19章看过的 Cosmos 3 就是一例，它的摘要没有交代声音怎样被处理〔待核：Cosmos 3 中声音模态的具体处理方式与训练数据，以正文为准〕。从发生学看，问题不在于有没有「声音这个输入通道」，而在于声音的差异是否和动作、接触的时间点对上，是否被用来做后果确认与异常发现。仅仅把音轨和视频并排输入，与让声音在事件中参与归属和回写，是两件不同的事。第19章已经区分过「多传感器融合」与「多种 SDE 纠缠」，这里只是把它落到听觉上。

这些边界提示了下一步：听觉单独面对一只瓜，能说出「这一声闷」，却说不全这只瓜；视觉说得出它的样子，触觉说得出它的分量。三套完整的过程各说一部分，又在同一件事上同时说话。它们说的话怎样叠在一起、对不上的时候又怎样处理，是下一章的问题。

> **带走的话**
>
> 听觉是以事件为先的完整 SDE，在时间变化中运行差异，并依靠与动作和其他事件的对齐把声音归给物；它擅长后果确认与异常发现，又受遮蔽和歧义所限。

小账 论证落在哪一层？——本章以本体论立场（听觉是完整 SDE、显露以事件为先）和机制假说（声源归属的稳定取决于跨事件的时间对齐）为主；经验事实限于视听同步窗口与 Cosmos 3 的定位，均标待核。 世界模型补了哪一条？——补上一条条件——声音若进入世界模型，关键是与动作和接触在时间上对齐、参与后果回写，而不只是多一条输入通道。

24

第 24 章

三种SDE如何形成特征纠缠聚合体

> **这一章的问题**：看、摸、听三套各自完整的过程，落在同一个东西上时，是怎样缠成「这一个物」的？对不上的时候又会发生什么？

◆ 一、命题：聚合体不是拼盘

前三章分别说明，视觉、触觉、听觉各自是一套完整的 SDE，各有显露、差异运行和特征纠缠。本章提出全编的核心命题：**一个物，在经验中不是由三份感官报告拼出来的，而是由三套完整的发生过程在互动中彼此进入对方的发生条件，形成一个「特征纠缠聚合体」，物就在这个聚合体中显露并站稳。**

「聚合体」这个词要说清楚。它不是把视觉的结果、触觉的结果、听觉的结果放进同一个盘子里，那叫拼盘；拼盘里每一样还是原来那一样，彼此不改变。聚合体的意思是：三套过程各自仍然完整，各自的显露、差异、纠缠仍在运行，但它们之间建立了新的纠缠，这种跨过程的纠缠反过来改写每一套过程内部的发生。看会因为摸过而看得不同，摸会因为听过而摸得不同。三套过程并没有合并成一套，也没有分别降格为 S、D、E 三个维度；它们是三个完整的发生者，在同一件事上相互进入。

这一命题属于本体论立场。下面的分层与归并、拆分、重建，属于机制假说，可以被检验，也可能被修正。

◆ 二、主例：陌生弹性球的握压、敲击与释放

这是我构想的场景，也是本章的主例。

林岚回长沙看周敏，带来一只她们公司测试用的球。球不大，一手能握住，表面是哑光的灰色，看不出是什么材料。她把球放在桌上，对小舟说：「你别看说明，自己弄明白它是什么样的东西。」

小舟先看。球是圆的、灰的、静止的，看起来像一只网球去掉了毛，又像一块橡皮泥捏成的团。视觉给出的显露很稳，但它对「这是什么样的东西」几乎什么也说不出来。

他伸手握住，用力一压。球在手里明显地凹下去，比他预料的软；手指感到一种均匀的、越压越大的回推力。这是第一个触觉显露：软，有弹性。几乎同一时刻，他看见自己手指间的球被挤得变了形，从指缝里鼓出来；手一松，球在一两秒内慢慢鼓回原状，而不是像网球那样一下弹回去。触觉里感到的回推和视觉里看到的复原，在同一段时间里起落。

接着他拿起一把勺子，在球上敲了一下。没有他预想的「砰」，而是一声很轻、很闷的「噗」，几乎没有余音。他换个地方再敲，还是「噗」。勺子敲下去时，手上感到勺子被吞了一下，没有反弹。

最后，林岚让他把球举到半臂高，松手。他以为球会像皮球一样弹起来，结果球落在桌上，只弹了很低的一下，「噗」一声，就停住了，稍稍扁着，过一会儿才圆回来。

小舟皱着眉说：「它看起来像个弹力球，摸起来也有弹性，可是一掉下来不弹。」林岚笑了：「对，这就是它要你弄明白的。」

这一连串动作在几分钟内完成，三套过程都在运行，各自显露了一些东西，又在几个地方对不上。我们就用它来分出跨过程纠缠的三层。

◆ 三、三层：同步出现、相关变化、相互制约

本书提出的机制假说是：三套 SDE 之间的纠缠，可以按由浅到深分为三层。

第一层，同步出现。不同过程的差异在同一时刻、同一处出现。小舟手一压，触觉的回推、视觉的变形同时开始；勺子一敲，手上的「被吞」与耳中的「噗」同

时到来。同步出现是最基本的一层，它让三套过程的显露被归到同一件事上，归到同一个物上。第23章所说的声源归属，就主要依靠这一层。

但同步只是在时间上碰在一起，并不说明这些差异之间有什么关系。电视里一个人拍桌子，同时窗外一辆车按喇叭，二者也同步，却毫无关联。所以仅有第一层，聚合体很松，容易把偶然同时发生的东西缠进来。

第二层，相关变化。一路差异变化时，另一路差异跟着有规律地变化。小舟压得越用力，手里的回推越大，看到的形变也越大；松得越快，球复原的速度却并不随之变快，而是自有节奏。他多压几次，就会发现「压力大小」和「凹陷深浅」之间有一个稳定的对应，而「松手快慢」和「复原快慢」之间没有。相关变化比同步出现深一层：它不只说「这些同时发生」，还说「这些一起变」。

这一层开始让物的「属性」显露出来。所谓「软」，不再只是手里一个笼统的感觉，而是一组触觉与视觉之间稳定的相关：多大的力对应多深的凹。所谓「慢回弹」，是视觉里的复原速度与触觉里撤力时机之间的一种不跟随的关系。属性在这里显露为跨过程的相关结构，而不是某一路单独读出的数值。

第三层，相互制约。一路过程的显露，限制了另一路过程可以显露什么；当它们不一致时，一方会被另一方修正。小舟看到的是「圆的、像弹力球」，视觉纠缠由此推出「掉下来会弹」；可是落地只弹了一点，听觉里是闷的「噗」。这时，听觉和视觉的落地事件制约了「像弹力球」这个视觉推断，迫使它改变；同时，先前触觉里的「越压越回推」又制约了「它不弹是因为它不弹性」这种过快的结论，因为它分明是有弹性的。几路显露在这里互相卡住，谁也不能单独说了算。

相互制约是聚合体真正成形的一层。只有到这一层，一个物才不只是「同时被看见、摸到、听到的东西」，也不只是「几路感觉一起变化的东西」，而是「几路过程相互约束、彼此担保的东西」。也正是在这一层，冲突才会出现，因为只有相互约束，才会有约束不上的时候。

为方便对照，把三层列成下表。

层次	跨过程关系	小舟和球的例子	不足之处
同步出现	差异在同一时刻同一处出现	压的同时既感到回推又看到变形	会把偶然同时的东西缠进来
相关变化	一路变化时另一路有规律地跟着变	压力越大凹得越深，复原速度不随松手快慢变化	只说一起变，不说谁限制谁
相互制约	一路显露限制另一路可能显露的东西，不一致时彼此修正	「像弹力球」被落地闷响与低弹修正，「有弹性」又阻止「不弹性」的结论	会产生冲突，需要冲突后的重组

需要强调：三层是一个机制假说，不是三个必经的先后阶段。在熟悉的物上，三层几乎同时成立；在陌生的物上，它们才会依次显出来。这个假说可以这样检验：在受控的人机实验中，分别破坏同步（让声音延迟）、破坏相关（让形变与力的对应随机化）、破坏制约（让视觉与触觉给出矛盾的材料线索），观察对象同一性判断在哪一种破坏下最先瓦解、瓦解后怎样恢复。如果三种破坏的效果没有可分辨的差别，这个分层就需要修正或放弃。

◆ 四、冲突推动归并、拆分与重建

聚合体的发生，不靠一路顺利地堆积，而常常靠冲突推动。用三原理的话说：一路过程显露出来的东西，和已经缠好的预期对不上，显露与纠缠之间的张力积累起来，推动差异运行改变（换个动作、换个法子去试），试出来的东西再回写进纠缠。这里说的两维、第三维，仍然是S、D、E，不是看、摸、听三种感官。小舟和这只球，前后经历了三种不同的重组。

归并。一开始，小舟手上的「软」和眼里的「慢慢鼓回来」是两件分开的事：一件是手的感觉，一件是看到的样子。多压几次之后，他发现两者总是一起出现、一起变化，于是它们被归并为同一个东西的同一种性质，他自己给它起了个名字：「慢弹」。归并的动力，是两路显露反复同步、相关，而且互相印证，分开处理反而解释不了它们为什么总在一起。

拆分。最初的「像弹力球」，是一个视觉上的整体纠缠：圆、光滑、有弹性、会弹起来，这几样缠在一起。落地那一下，「有弹性」和「会弹起来」第一次在同一个物上分开了：它有弹性，可它不弹。原来缠在一起的东西被迫拆开。小舟脱口而出的那句「看起来像弹力球，摸起来也有弹性，可是一掉下来不弹」，就是拆分正在发生时的语言痕迹。拆分的动力，是同一个旧纠缠在新事件上给出了互相矛盾的预期。

重建。拆分之后，小舟需要一个新的组织来容纳这些显露。他开始自己试：压下去马上松手，看回弹；压下去停五秒再松，看回弹；从不同高度扔，听声音、看弹跳。几轮下来，他得出一个新的说法：「它会吃掉力气，慢慢再还给你。」这个说法不是专业术语，却把触觉的回推、视觉的慢复原、听觉的闷响、落地的低弹，四样显露组织进了一个新的纠缠。林岚这时才告诉他，这是一种慢回弹的材料，能把冲击的能量吸掉一大部分，她们公司拿它做机器手指尖的垫子（这是构想场景中的设定，不对应具体产品）。重建的标志是：新的纠缠不仅解释了已有的冲突，还能对新动作给出预期，比如「从更高处扔，它也不会弹多高，但声音会大一点」，而这个预期可以马上去试。

归并、拆分、重建，在这个例子里先后出现，但在别的情形里次序会不同，也可能反复。第一编讲「稳定化、分化与重组」（第2章），讲的是一般学习中的纠缠变化；本章是把它落到三套感官过程之间。区别在于：这里的冲突，来自几套各自完整的过程互不让步。视觉有它自己的纠缠，不会因为一次低弹就放弃「圆的光滑的东西」这个判断；触觉也有它的，不会因为声音闷就否认「有回推」。正是因为三套过程各自完整、各自有站得稳的显露，它们之间的冲突才有力量。如果某一路只是另一路的附属，冲突一出现，附属的那一路直接被覆盖掉就是了，不会推动重组。

重建之后还有一个容易被忽略的结果：三套过程各自内部也变了。这一点正好检验「各自完整且共同参与」这句话是不是空话。

先看小舟的手。一开始他握球是一把攥紧，后来他学会了分两种压法：快压快松，用来感觉回推；慢压停住，用来感觉「吃力」。触觉里运行哪些差异，被聚合

体改写了，这是触觉内部的 D 在变。再看他的眼睛。起初他只看球的形状，后来他会盯着松手后的那一两秒，专门看复原的快慢；视觉的注意点从「它是什么形状」移到了「它怎样变回形状」，这是视觉内部的差异组织在变，视觉里显露出来的东西也随之多了一层。再看他的耳朵。第一次敲出「噗」时他只是意外；后来他会有意比较从不同高度落下的声音，用声音的大小去估计落地的力，听觉的纠缠里多了一条「声音大小—冲击大小」的联系。

三套过程各自改变了，但没有哪一套变成另一套的附庸，也没有哪一套被取消。视觉仍然用自己的差异显露它所能显露的，触觉、听觉也是如此；不同的是，每一套过程的发生条件里，都多了来自另外两套的纠缠。这就是「各自完整且共同参与」的具体含义：参与不是让出完整，而是让完整的过程在彼此的影响下继续发生。

反过来设想一种情况，可以看清这个区别。如果三路感觉只是三个探头，各自把读数交给一个中央处理者，中央处理者算出「慢回弹材料」，那么探头本身不会变：手还是一把攥，眼睛还是只看形状。小舟的例子里，恰恰是「探头」变了，而且是朝着能更好地与另外两路相互制约的方向变的。这一点可以作为区分融合与纠缠的又一个可观察迹象：看在冲突与重建之后，各路过程自身的差异运行方式有没有被改写。

◆ 五、反例与边界：融合不等于纠缠，三套也不必总在场

这一章的命题，容易被两种方式误读。

第一种误读，是把聚合体理解为「多传感器融合」。工程上的多传感器融合，常见做法是把不同传感器的数据在某一层拼接或加权，输出一个更准的估计。第19章已经区分过融合与多种 SDE 纠缠。本章的分层给了这个区分一个可操作的说法：如果一个系统只在第一层（同步）上把信号对齐，或只在第二层上学到各路信号之间的统计相关，却从不在某一路与其他各路矛盾时调整自己对物的显露，那么它做的是融合，还没有走到相互制约这一层。这不是说现有系统都停在融合上，具体系统做到哪一层，要看它在冲突情形下的行为。按这三层来看工程做法：多接几

路输入并按时间对齐，做到的是同步出现；让模型从数据里学到几路信号一起变，做到的是相关变化；要走到相互制约，模型得把几路的变化一起当作要预测的对象，一路预测错了能被另一路查出来。第33章「跨模态预测与纠缠重组」讲的「联合预测」，就是朝第三层去的训练办法。

第二种误读，是认为三套过程必须同时在场，物才能显露。摸黑找开关时，几乎只有触觉在工作，再加最后那一声「咔」；远远认出一个熟人，几乎只靠视觉。物的显露并不总需要三路齐备。更准确的说法是：一个物在经验中的聚合体，是在过往许多次互动中由多路过程共同建立的；建立之后，某一次显露可以只调用其中一路，其他几路以纠缠的形式留在背后。小舟以后再看到这只灰球，即使不去摸，也会「看出」它是慢弹的；这时视觉里显露出来的东西，已经包含了触觉和听觉留下的纠缠。

这又带出一个边界：聚合体建立之后，单一路的显露会变得更有力，也更容易出错。一只外观相同、材料不同的灰球放在桌上，小舟一眼看去会把它当作慢弹球，直到他再一次握压。被调用的纠缠越丰富，视觉越能「看到」看不到的东西，也越可能看错。这不是聚合体的缺陷，而是它的代价。

还要说明一条纪律：本章用小舟作主角，不意味着这一切只对孩子成立。成年人面对陌生材料，同样经历这三层和这三种重组，只是快得多，快到察觉不到。陌生的东西把过程放慢了，才让我们看清楚。

◆ 六、向物的意识推进

小舟弄明白这只球之后，发生了一件小事。林岚把球收起来，问他：「刚才那个球，你觉得它要是放在冰箱里冻一晚上，会怎样？」小舟想了想说：「可能会变硬，就不吃力气了。」他没有试过，他只是在一个已经组织好的物上做了推断。

这时，这只球对小舟来说，已经不是「一些看、摸、听到的东西」，而是一个有自己的样子、自己的性质、可以在没见过的条件下被设想的「它」。它跨越了感官，也跨越了刚才那几分钟。它还和用途连在一起：林岚说「我们拿它做手指尖的垫子」，小舟立刻明白为什么，因为它会把力气吃掉。

这就是下一章要回答的问题：当一个特征纠缠聚合体稳定到这种程度，我们能不能说，一种「物的意识」发生了？如果能，凭什么判断它发生了，又凭什么判断它失效了？

> **带走的话**
>
> 三套各自完整的 SDE 在同一事件上，经同步出现、相关变化、相互制约三层缠成特征纠缠聚合体；冲突推动归并、拆分与重建，物由此显露并站稳。

小账 论证落在哪一层？——本章以本体论立场（聚合体由三套完整过程相互进入而成，不是拼盘）和机制假说（三层分法、冲突驱动的三种重组）为主，几乎不用经验事实；主例为构想场景。　世界模型补了哪一条？——补上一条条件——世界模型中的对象若要稳定，需要在各路过程矛盾时能修正对物的显露；只做对齐与统计相关，停在融合层。

25

第 25 章

物的意识如何发生及接受检验

> **这一章的问题**：说一个人或一个系统对某个物「有了意识」，能否给出一组可以检验的判据？主观体验问题又该放在哪里谈？

◆ 一、先定判据，再谈体验

「物的意识」这个说法很容易把讨论引向一个深坑：机器到底有没有感受？它「觉得」那只球软吗？这是一个真问题，但如果一开始就从这里谈，讨论会陷入无法检验的争执，而那些可以研究、可以改进的东西反而被搁下了。

所以本章的做法是分两步。先给物的意识一组功能性判据：在不涉及「内部有没有感受」的前提下，一个主体（人或机器）是否在经验中形成了一个物，这个物能否被维持，在什么条件下失效，失效后能否重建。这组判据只看能力和行为，可以在人身上测，也可以在机器上测。然后，在单独的一节里谈主观体验问题，说明本书在这个问题上主张什么、不主张什么。两步不混。

需要先说清楚用词。本书所说的「物的意识」，在功能层面指的是：一个跨感官、跨时间稳定的对象同一性，以及这个同一性与用途、与行动可能性的共同形成。它是第24章所说的特征纠缠聚合体走到足够稳定时的状态，不是一个额外附加上去的「意识实体」。这个定义是本体论立场；下面的判据是把它变成可以检验的东西。

◆ 二、对象同一性的四条功能性判据

判据一：形成。主体能否在有限的互动中，把多路过程的显露归到同一个物上，并对这个物形成可以说出或可以用行动表现出来的组织。检验的办法是：给一个陌生的物，允许有限次数的看、摸、听和操作，然后要求主体对没做过的动作给出预期。小舟说出「冻一晚上会变硬」，就是形成的标志；如果他只能复述「它是灰的、软的、噗一声」，那还只是几路显露的清单，聚合体尚未形成。

判据二：维持。主体能否在条件变化时，仍然把它认作同一个物。条件变化包括：换一个角度看、在暗处摸、隔一天再拿出来、放在另一堆东西中间、被手挡住一部分。检验的办法是：在这些变化下测试主体是否仍能认出它，并且仍然用对它的那套预期来操作它。维持跨越两种距离：跨感官，即只用其中一路过程也能认出它（摸黑也知道这是那只慢弹球）；跨时间，即离开之后再回来仍能认出它。第21章提出的「被扰动后多快恢复跟踪」，是维持判据在视觉中的特例。

判据三：失效。主体能否察觉到这个物「不对了」。一只熟悉的物，换了材料、坏了、被调了包，主体能否在互动中发现预期与显露对不上，并且把这种对不上归到物身上，而不是忽略过去或归给别的原因。检验的办法是：悄悄换上一只外观相同、材料不同的球，看主体在第几次互动中发现，发现时用的是哪一路过程。能察觉失效，说明这个物的同一性不是一个死贴的标签，而是一个在互动中不断被担保的组织。

判据四：重建。察觉失效之后，主体能否通过新的互动，重新建立对这个物（或那个新的物）的组织，并且新组织能对后续动作给出可检验的预期。重建可以是修补（「哦，它冻过了，所以硬」），也可以是拆分（「这根本是另一只球」）。检验的办法是：看重建花了多少次互动，重建之后的预期在新的测试中准不准。

这四条判据合起来，描述的是一个对象同一性的完整生命：生出来、站得住、塌得下、立得回。少了后两条，判据就只测到了「认得出」，测不到「懂得它」。

一个系统能在千百张图里认出某种球，却在球被换掉时毫无察觉，或者察觉了也无法重建，那么在本书的意义上，它对这个物的意识还停留在很浅的一层。

关于这四条判据，还有两点需要说明。

其一，判据是分程度的，不是有或无。小舟对那只球的同一性，第一天就形成了，但维持得如何、失效时察觉得多快，要几个月后才知道。同一个主体，对不同的物，四条判据的达成程度也可以很不一样：周敏对西瓜的同一性很深，她能在敲第一下时就察觉「这只不对」；对那只慢弹球，她和小舟差不多是新手。所以与其问「它有没有物的意识」，不如问「对这一类物、在这一组条件下，它的对象同一性在四条判据上分别到了什么程度」。这样提问，才能让判据真正可比、可测、可改进。

其二，在机器的情形里，要问清楚是谁的对象同一性。按本书统一的发生单位，机器系统的训练属于集体人类—AI复合体：哪些物被拿来训练，哪些动作被允许，什么样的失效被标为失效，都由工程师和数据制度参与组织；具体使用时，一线操作人员的纠正和反馈又让它进入个人—AI复合体。因此，当一个机器人在测试中察觉到球被换了，这个能力的发生，不能被写成机器单独完成的「发现」，而是复合体在训练与使用中组织出来的能力，暂时封装在这个系统里。这并不贬低机器，而是让检验落在正确的单位上：失效时，要追问的不只是模型参数，还有当初的经验组织和评价标准。林岚说过一句话，可以作为这一点的注脚：「它认不出来的东西，多半是我们没让它碰过的东西。」这句话是构想场景里的人物台词，不是任何人的实际原话。

◆ 三、物、用途与行动可能性的共同形成

上面四条判据，偏重物本身。但人对物的意识，从来不只是「它是什么样」，还有「它能干什么」「我能拿它干什么」。

摸黑找开关的例子最能说明这一点。周敏摸到的那块方形塑料，之所以显露为「开关」，不是因为它的形状和材料，而是因为它「可以按」，按了「灯会亮」。如果那是一块形状相同、按不动的装饰板，触觉显露出来的就是另一个东西。物在

这里和它的行动可能性一起形成：可按、可拧、可握、可坐、可踩。第8章介绍过吉布森的「可供性」概念，这里同样不援用其理论体系，只取「物与行动可能性共同显露」这个方向。

西瓜也是这样。在水果摊上，瓜是一个「可以挑的东西」：可以看、可以敲、可以托；拿回家，它变成「可以切的东西」；切开之后，它成了「可以吃的东西」。同一只瓜，随着任务变化，显露出来的行动可能性不同，与之缠在一起的特征也不同：挑的时候纹路和声音重要，切的时候瓜皮的厚度和刀进去的阻力重要。

弹性球的例子则把用途也带了进来。林岚一句「做手指尖的垫子」，让小舟在已经形成的物上又接上了一条用途；这条用途反过来让「吃掉力气」这个性质变得更显眼、更站得稳。用途、行动可能性和物的性质，在这里是一起发生、一起改写的，没有哪一个先在。

由此可以补一条判据的要求：检验物的意识时，不应只问主体能否认出和预测这个物，还要问主体能否在新的任务中发现这个物的新用途，或者判断它不适合某种用途。小舟说「冻硬了就不吃力气了」，如果再追问「那冻过的还能不能做手指垫」，他若能答「不太行」，这就是物、性质与用途在他那里一起组织好的标志。这一点对机器同样适用，第28章讲任务如何组织世界的发生范围，会从任务一侧展开。

◆ 四、失效与重建：一次构想的对照

这是我构想的场景。过了几个月，林岚又回长沙，带来两只外观一模一样的灰球，一只是上次那种慢弹球，一只是普通的高弹力球，外面喷了同样的哑光灰漆。她把两只球放在桌上，问小舟：「你还记得上次那个球吗？哪个是？」

小舟看了看，两只一模一样，视觉给不出答案。他随手拿起一只，握了一下，回推很快，一松手就鼓回来了，他说：「这个不是。」又拿起另一只握，「这个是，它慢。」这是维持：隔了几个月，只用触觉一路，他认出了那个物。

然后林岚换了个法子。她把慢弹球在冰箱里冻了一晚上，第二天和另一只一起摆出来。小舟握了一下第一只，硬；握第二只，弹得快。他犹豫了：两只都不慢，那上次那个呢？他拿起硬的那只往桌上一扔，「砰」一声，比上次响，弹得也比上次高一点。他愣了愣，想起自己几个月前说过的话：「冻了会变硬。」于是他说：「这只是那个，冻过了。」

这一段里，四条判据都出现了：之前的形成，几个月后的维持，冻硬后的失效（原有的「慢」这一显露对不上了），以及借着先前的推断完成的重建。重建成功的关键，不在于他记住了哪个球，而在于他当初形成的组织里，就已经包含了「温度会改变它」这样一条可以被调用的预期。物的同一性不是靠外观的相同维持的，而是靠一个能容纳变化、并能解释变化的聚合体维持的。

同样的检验，可以原样搬到机器上，第34章会把它写成可以失败的对照实验，这里只说四条判据落到那里是哪几项指标：形成，对应「只给一路信号能否预测另一路」的跨模态一致性和「同貌异质」的辨别；维持，对应跨时间的维持；失效，对应失配检出时间；重建，对应恢复所需的交互次数和旧能力保持率。这组测试既不要求机器「说出感受」，也不以机器某次的成败直接证明或否证 SDE 本体论。

◆ 五、主观体验问题：单独论证

现在单独谈主观体验。

问题是这样的：即使一个系统完全满足上面四条判据，甚至在每一项测试中都比小舟做得好，我们能不能说它「感到」那只球软？它的内部有没有「软的感觉」这回事？这个问题与功能判据是两个问题，必须分开论证。

本书在这里的立场分三点说。

第一，功能判据的满足，不构成主观体验存在的证明。四条判据测的是能力和行为的组织，它们在逻辑上可以与「内部有感受」相容，也可以与「内部没有感受」相容。从「它能做到」直接跳到「它感到了」，就是第20章所说的「从显露相似跳到发生机制相同」的推论缺口，在意识问题上同样成立。

第二，功能判据的不满足，也不构成主观体验不存在的证明。一个刚出生的婴儿，按上述判据大概大多不满足，但我们并不因此断言婴儿没有任何感受。判据是对能力的检验，不是对体验的检验。

第三，SDE 在这个问题上能提供的，是一个提问方式，而不是答案。按本书的本体论，人的体验也是在特征纠缠中发生的，自我世界（第9章）本身就是记忆、情感、关系与价值缠在一起的发生。由此可以提出一个可以讨论的问题：主观体验是否对应于特征纠缠聚合体中某种特定的组织方式，比如与自我世界的纠缠相连的那一部分？如果是，那么问一个系统有没有体验，就变成问它的聚合体是否与某种自我世界的组织纠缠在一起。这是一个推测性的方向，本书不把它当作结论，也不认为它能用当前的任何实验判定。

在此之外，本书不就「某某系统是否有意识」作任何断言。这样做不是回避，而是因为在这个问题上，目前既没有公认的判据，也没有能区分两种答案的实验。把一个尚无法检验的问题与可检验的功能判据混在一起，只会让后者也失去检验的力量。

◆ 六、三模态必要性：研究边界，而非前提

最后要处理一个本编一路都绕着走的问题：物的意识是否必须由视觉、触觉、听觉三者共同形成？缺了其中一路，是否就不可能形成？

本书的回答是：这是一个研究边界，不是一个已经成立的前提。

先看理由。本编的论证说明了三点：三套过程各自完整，各有自己能显露而其他过程显露不了的东西；它们在互动中通过三层纠缠形成聚合体，冲突推动重组；聚合体建立后，单一路的显露可以调用其他各路留下的纠缠。这三点支持一个较弱的主张：对于涉及物的内部、分量、弹性与接触后果的任务，多路过程共同参与，比单一路更有可能形成稳定的、经得起失效与重建检验的对象同一性。

但这三点不支持一个更强的主张：三者对所有任务都不可替代。理由也有三条。第一，生活中有大量例子表明，缺少某一路感觉的人，依然能形成丰富的物的

世界，其他过程会以不同的方式承担起相应的组织；本书不在这里展开，更不作任何医学方面的判断。第二，任务不同，所需的过程不同，第28章会讲任务如何决定需要保留的差异与范围；摸黑找开关就几乎不需要视觉。第三，除了视听触，还有本编没有讨论的其他过程，比如身体自身的位置与平衡感、温度感、嗅觉与味觉，它们同样可能参与物的聚合体；本书集中讨论视听触，是为了论证的集中，不是认定只有这三路。

因此，「三模态必要性」在本书中的位置是：它划出一组需要实验回答的问题。对某一类任务，去掉某一路过程，对象同一性的形成要慢多少、维持要弱多少、失效能否察觉、重建能否完成？去掉之后，是否可以用更多的其他过程互动或更多的记录数据来补偿，补偿到什么程度？对机器系统来说，这些问题都可以设计对照实验，也都可能得出「在这类任务上不必要」的结论。本书提出视听触三种SDE 的共同互动作为核心研究路线，是认为这条路线值得被认真检验，而不是预先宣布它的胜利。

至此，第五编完成了它的论证：物不是先在地摆在那里等待被识别，而是在三套各自完整的发生过程的互动中，作为特征纠缠聚合体显露、站稳、失效、重建。下一编要追问的是：如果物的发生如此依赖具体的互动与接触历史，一个试图统一模拟全部物理世界的模型，会碰上怎样的信息负担？这是第26章「物理发生的信息负担」的问题。

> **带走的话**
>
> 物的意识先以对象同一性的形成、维持、失效与重建四条功能判据来检验，并与用途和行动可能性一起形成；主观体验另案讨论，三模态必要性是待实验回答的研究边界。

小账 论证落在哪一层？——本章以本体论立场（物的意识是稳定的特征纠缠聚合体）与机制假说（四条功能判据及其测试方法）为主；主观体验部分明确为推测性提问，不作结论；经验事实仅涉及可供性概念，出处见第8章。 世界模型补了哪一条？——补上一条检验标准——世界模型中的「物」应以形成、维持、失效、重建四条判据检验，而不只以识别准确率或画面逼真度检验；三模态是否必要须分任务实验确定。

第 六 编

统一物理模型的困境与任务边界

编序 第五编讲到，一个物在视觉、听觉、触觉三种 SDE 的共同互动中显露出同一性，也会失效、会重建。顺着这条路往下走，马上会碰到一个工程上非常现实的问题：既然物理世界里的物是这样一个个发生出来的，那能不能造一个足够大的模型，把所有材料、所有尺度、所有接触情形都学进去，一次性得到一个「统一物理模型」？本编五章就是对这个问题的分步回答。

第26章先算账：物理发生到底牵涉多少要素，组合起来有多大，压缩有没有可能，以及「信息量巨大」这句话能推出什么、推不出什么。第27章换一个角度：就算一个模型把画面、状态预测得很像，行动为什么还不一定可靠。第28章提出本编的正面主张：世界的发生范围由任务组织，任务决定保留哪些差异、多细的粒度、多大的时空范围，而且这个边界在互动中不断修订。第29章拿驾驶做一个完整样本，看人—车辆—道路复合体怎样在视觉、听觉、接触和车辆响应里专门发生。第30章收束：可复用的底座和任务专门训练怎样协作，不同任务世界之间怎样接口，为什么「共享」不等于「一个模型穷尽全部物理发生」。

本编主例是老彭开车（雨夜、异响、方向盘手感）、人手拧各种瓶盖、装修师傅的手艺。拧瓶盖这里只讲人手怎样对付不同的盖子，机器怎样学，留给第七编林岚的团队。需要提醒读者：本编不把「统一架构必然失败」当作前提，它是一个要拿证据回答的经验问题；本编要澄清的，是这个问题该怎样问才问得对。

26

第 26 章

物理发生的信息负担

这一章的问题：物理世界里一件事要发生，牵涉的东西到底有多少？「信息量太大」这句话，究竟能证明什么？

◆ 一、从一只瓶盖算起

先从一件人人做过的小事说起：拧开一只瓶盖。

矿泉水瓶的塑料盖，捏住一拧就开，开的时候「咔」一声，那是防盗环断开。辣椒酱那样的玻璃罐，铁皮盖子拧不动，人会先拿刀背在盖沿敲两下，或者放到热水下冲一冲，再垫块毛巾去拧。药瓶的儿童安全盖，要先往下压再转，光转是转不开的。冰箱里刚拿出来的果酱瓶，瓶口挂着水珠，手一滑就打空。开过一次又拧紧的瓶盖，和全新的瓶盖，劲道完全不一样。手上沾了油，或者手指刚洗过还湿着，同一只瓶盖又换了一副样子。

这几种情形里，起作用的东西各不相同：瓶盖和瓶身的材料（塑料、铁皮、玻璃）；螺纹的尺度和密封方式（防盗环、真空吸住、压转结构）；温度（热胀冷缩让金属盖松一点）；手的状态（干、湿、油、有没有力气）；工具（毛巾、刀背、开瓶器）；这只盖子以前被怎样对待过（新的、开过的、被谁拧得过紧）。把这些合起来，才是「这一次拧开」实际发生的条件。

还要加上一项最容易被当作「背景」的东西：身体。下面是我构想的场景：同一只玻璃罐，老彭手大、握得住，一拧就开；九岁的小舟手小，只能握住盖子的一半，力气使不到一处，他会换成两只手夹住盖子，再用肚子顶住瓶身。手的大小、握力、手腕能转多大角度，决定了哪些办法可用、哪些办法根本轮不上。对小舟来

说，「拧开这只罐子」和对老彭来说并不是同一件事，两人面对的是两个不同的发生条件。物理发生里的身体不是一个可以略去的常数，它本身就是参与纠缠的一方；换一个身体，组合表里就要多出整整一列。机器也一样：两指夹爪、五指灵巧手、带触觉皮肤的手，对同一只瓶盖面对的根本不是同一组可能性。

再看时间尺度。拧瓶盖是一两秒的事，可瓶盖为什么拧不动，可能是半年前封装时抽了真空，可能是上周有人拧得太紧，也可能是果酱糖分在螺纹里慢慢结晶、把盖子粘住了。眼下这一拧，背着一段看不见的接触历史。一个想要表示「拧瓶盖」的模型，不仅要表示此刻，还要在某种程度上表示此刻是怎样来的。

用本书的话说，这一次拧开是在一团特征纠缠（E）中，经过手上的力、滑与不滑、松与不松这些差异的运行（D），最后显露为「开了」或者「没开」（S）。前面各编已经讲过，三方程 S=F(D,E)、D=G(S,E)、E=H(S,D) 缺一不可。这一章要问的是另一件事：如果有人想把「拧瓶盖」这件事在一个模型里完整表示出来，他得准备多少东西?

◆ 二、组合扩张与长尾

单看每一种因素，好像都不多：材料几类，结构几种，温度几档，手的状态几种。麻烦在于它们是乘起来的，不是加起来的。假设材料有五类、结构有六种、温度分四档、手的状态有五种、使用历史分三种、工具有四种可选，那么光是这六项的组合就是五乘六乘四乘五乘三乘四，七千多种。这还只是粗分。真实世界里每一项都是连续变化的：温度不是四档，是一度一度变的；手上的湿不是有和没有，是多湿、湿在哪个手指上。

这就是「组合扩张」：每多一个相关因素，可能的情形就成倍增加。

再看装修师傅。一位做了二十年的师傅在墙上打孔装空调支架，要考虑的东西远比拧瓶盖多：墙是老房子的红砖墙、新楼的混凝土剪力墙、还是轻质的加气块或者石膏板；墙里有没有暗埋的水管和电线；钻头是冲击钻头还是开孔器，新的还是磨钝了；电钻开不开冲击档；孔要打多深，膨胀螺丝选几号；师傅当时站在梯子上

还是踩在阳台边。每一项都和别的项相互牵扯：同样的冲击档，在红砖墙上很顺，在加气块上会把孔打大，膨胀螺丝就吃不住劲。

组合扩张之外还有「长尾」。所谓长尾，是说常见的情形只占一部分，剩下大量的情形每一种都很少见，可是少见的情形加在一起并不少。师傅一年里碰到的绝大多数是普通混凝土墙，但总有那么几回：墙里藏着一根以前的铸铁管；某户人家自己改过墙，外面抹了一层厚灰，里面是空的；某栋楼的混凝土标号特别高，钻头打下去直冒火星。每一种都是第一次碰到，或者十年才碰到一回。

还有一层比数量更麻烦：这些因素不是各管各的。如果它们彼此独立，学会每一项的影响，再把影响加起来，就能对付组合。可物理发生偏偏不是这样。铁皮盖子遇热会松，这是一条；手湿会打滑，这是另一条。但用热水冲完盖子，手也湿了，盖子松了却握不住，两条加在一起，结果不是「更好拧」，而是「得先擦干手、垫上毛巾才好拧」。冲击档对红砖墙有利，新钻头对混凝土有利，可新钻头配冲击档打加气块，孔会打得又大又毛。一个因素的作用，取决于别的因素当时处在什么状态。用本书的话说，这些因素是纠缠在一起起作用的，不是并排摆着的。正因为纠缠，组合里的每一格都可能出现前面各格都没出现过的结果，这才让组合扩张真正成为负担。

这一节说的是什么层次的话？组合扩张是一个算术事实，长尾是对现实经验分布的一种描述。物理与工程中常见的数据分布有长尾性质，这是很多研究者都会同意的一般经验陈述，至于某个具体数据集的分布形状，需要具体去测。

◆ 三、有限的表示，压缩为什么可能

如果故事到此为止，结论似乎就是：物理世界的组合太多，任何有限的模型都装不下，世界模型注定做不成。但这个推论太快了。

先看人。老彭开了三十年车，他没有在脑子里存下每一种路况、每一种车况的组合。小舟拧瓶盖，也不需要先见过七千种瓶盖。人能应付组合扩张，靠的不是穷举，而是压缩。手上形成的不是一张清单，而是一种「劲道」：盖子拧不动的时候，手会自己去找角度，先加一点力，再试试反方向松一下，或者换一个握法。这

种劲道是在一次次具体拧盖中纠缠出来的，它把很多不同的情形归到了少数几种「对付办法」之下。

科学里的规律也是压缩。热胀冷缩一句话，就把「热水冲一冲铁盖就好拧」和「夏天铁轨要留缝」放到了一处。摩擦与正压力的关系，就把「垫毛巾」「手上沾油打滑」放到了一处。用本书的话说，压缩是特征纠缠的一种组织方式：许多零散的差异被缠成少数几个稳定的结（结构是显露的稳定态），以后遇到新情形，可以从这几个结出发去试。

所以压缩是可能的，而且是学习本来就在做的事。这对技术也成立：一个大模型的参数是有限的，但它面对的输入可以是无限多样的，它能处理没见过的输入，靠的就是某种压缩。这是机制层面的一般说法；至于某一个具体模型压缩了什么、压缩得好不好，要看它的训练机制和测试结果，不能一概而论。

不过压缩总要有所取舍。拧瓶盖的劲道，压缩掉了瓶盖的颜色、商标、生产日期，保留了紧不紧、滑不滑。装修师傅打孔的经验，压缩掉了墙面刷的什么漆，保留了墙体材料和里面有没有管线。压缩之所以能成功，是因为它丢掉的是对眼下的事无关紧要的差异。那么，谁来决定什么是无关紧要的？这个问题，第28章会正面回答。这里只先指出：压缩得好不好，离不开「为了做什么」。

压缩还有一个容易被忽略的来源：边做边补。装修师傅碰到一面拿不准的墙，他不会站在那里把所有可能性都想一遍。他先用小钻头在不显眼的地方试一个浅孔，听声音、看粉末的颜色、感受钻头往里走的阻力。红砖出红粉，混凝土出灰粉，加气块的粉又细又轻，钻头一下子就进去了；打到钢筋，手上会突然一顿，声音也变尖。几秒钟的试探，就把一大片可能性收成了一两种。人对付长尾，靠的不是事先装好所有答案，而是带着一套压缩好的经验，在现场用一个小动作去换回关键的差异。这说明，有限的表示加上即时的互动，可以覆盖远比表示本身大得多的情形范围。这同样只是一个机制假说：它在人身上有大量日常经验支持，用到机器身上，要看机器有没有这样的试探通道，以及试探的结果能不能进入它的判断，这是第七编第32章「从被动记录到主动试探」要讨论的事。

四、「信息量巨大」能支持什么结论

现在可以老老实实算一下这笔账了。物理发生的信息负担很大，这一点没有争议。争议在于从这一点能推出什么。

它能支持的结论，至少有三条。

第一，靠被动收集记录来穷举物理情形，是不现实的。不论是视频、传感器日志还是文字描述，记录总是有限的，组合扩张和长尾决定了总有大量情形没被记录到。这条结论比较稳，因为它只依赖于算术和经验分布的一般性质。

第二，任何有限的模型都必须压缩，而压缩一定带着取舍。一个模型在某些情形下表现好，往往是因为它保留了这些情形需要的差异；它在另一些情形下失败，往往是因为恰好丢掉了那里需要的差异。所以，一个模型的能力必须连同它的适用范围一起说明，单说「它会物理」是不够的。

第三，评估一个世界模型，不能只看它在常见情形下的平均表现，必须专门检验长尾和组合边缘。这是从前两条直接引出的方法论要求。

它不能支持的结论，同样要说清楚。

第一，它不能证明「统一架构必然失败」。信息量巨大反驳的是「穷举」，不是「统一的表示方式」。一个统一的架构，完全可能通过很好的压缩，在许多任务上都表现得不错。这是一个经验问题，要靠具体模型、具体测试来回答，不能从算账里推出来。

第二，它不能证明「大规模训练学不到可迁移的物理规律」。公开文献里已有不少报告，说某些在大量视频或交互数据上训练的模型，在没见过的情形中有一定的迁移能力（具体到哪个模型、迁移到什么程度，要按版本核查，第四编已逐一讨论过各条路线）。信息量大，只说明迁移不能靠穷举，不说明迁移不存在。

第三，它不能代替本体论论证。这一点最容易混。SDE 的本体论主张是：物理世界不是一个先在的、等着被装进模型的对象清单，物是在纠缠中经差异运行而显露的。这个主张不依赖信息量大不大。就算哪一天发现物理世界的组合其实没那么

多，SDE 的本体论立场也不会因此改变；反过来，信息量很大这个事实，也不能当作 SDE 本体论的证明。两者是两层话：前者是本体论立场，后者是关于组合与分布的经验事实，中间要靠机制假说来连接，比如「压缩的好坏取决于任务所需差异的保留」。

那么，什么样的证据才能回答「统一架构行不行」？至少要先把「统一」说清楚。它可能指同一种网络结构，可能指同一套参数，可能指同一个训练目标，也可能指同一个对外接口。同一种网络结构，分别训练出拧瓶盖和开车两套参数，这叫统一吗？一套共享的底座参数，再针对每个任务加一层专门训练，这叫统一吗？这几种意思混在一起，争论就会各说各话。按本书的口径，真正需要检验的问题是：一套参数在不经任务专门训练的情况下，能在多大的任务范围内、以多高的可靠性行动；加上专门训练以后，又提升多少；失效的时候，失效集中在哪些组合边缘。这几个问题都可以设计对照实验来回答，第七编第34章「建立可以失败的实验与评价」会讨论这类实验怎样设计。在没有这些对照之前，无论说「统一架构必然成功」还是「必然失败」，都是在拿信息量这一个事实替一个更复杂的问题下结论。

还要补一句：信息量大，对人和对机器同样成立。人也没有穷尽物理发生，老彭照样会碰到没见过的路况，装修师傅照样会打穿过一回水管。所以「人能做到、机器做不到」也不能从信息量里推出来；能推出来的只是：人和机器都得靠有取舍的压缩加上现场的试探与修正，区别在于各自的压缩是怎样纠缠出来的、试探的通道是否畅通、失败之后能不能回写。

把这三层分开，「信息负担」这个说法就有了确定的分量：它是一个严肃的工程约束，提醒人们不要指望穷举，要说明压缩的取舍与适用边界；但它不是判决书，不能直接宣判任何一条技术路线的死刑。

◆ 五、负担之外，还有一道更难的坎

算完这笔账，还剩一个问题。假设某个模型压缩得很好，它看了大量拧瓶盖的视频，能把「接下来瓶盖会怎么转、手会怎么动」预测得非常像，甚至画面以假乱真。这样它就能拧好瓶盖了吗?

直觉告诉我们未必。看过一百遍别人拧瓶盖的录像，和自己手上拧过一百只瓶盖，是两种不同的经验。录像里有显露，却没有自己手上那一下的阻力；预测了画面，却不知道自己这一拧会怎样改变瓶盖。这里的缺口不在于信息不够多，而在于预测显露与可靠行动之间隔着一道坎。下一章就专门讨论这道坎。

> **带走的话**
>
> 物理发生的信息负担很大，它足以否定「靠记录穷举世界」，并要求每个模型说明自己的压缩取舍与适用边界；但它不能单独证明统一架构必然失败，更不能代替本体论论证。

小账 论证落在哪一层？——本章以机制假说为主（压缩依赖于保留任务所需的差异），辅以经验事实（组合扩张的算术、长尾分布的一般性质）；本体论立场只用于区分层次，不由信息量推出。 世界模型补了哪一条？——补上一条边界——世界模型不能以穷举为目标，必须以有取舍的压缩为目标，并连同适用范围一起交代能力。

27

第 27 章

为何显露预测不能保证行动可靠

> **这一章的问题**：一个模型能把接下来的画面、声音或状态预测得很准，为什么它照样可能在行动时出错?

◆ 一、命题：预测显露和改变世界是两件事

先把本章的命题摆出来：对显露（S）的预测，哪怕在已有记录上做得很好，也不能保证据此采取的行动可靠。原因不在预测不够精细，而在于两件事回答的问题不同。预测回答的是「照这样下去，接下来会显露成什么样」；行动要回答的是「我这样动一下，世界会变成什么样」。前一个问题里，预测者站在一旁；后一个问题里，行动者本身成了差异运行（D）的一部分，他这一下改变了纠缠（E），世界从此走上另一条路。

用三方程来说：S=F(D,E)。一个只学会从过去的显露推出后来显露的模型，学到的是显露与显露之间的相关；而一旦行动介入，D 被换掉了，E 也被改写了，原来那条相关就不一定还成立。下面分三处缝隙来看：已有记录中的不可区分配置；新干预下的不同后果；时间累积误差与分布变化。最后看错误预测怎样被发现，并转化为新的经验。

◆ 二、看起来一样，底下不一样

雨夜，老彭开车经过一段路，路面反着光。同样是反光，可能是一层薄薄的雨水，可能是积水，可能是刚刷过的沥青，也可能是一摊洒在路上的油。从行车记录

仪的画面看，这几种情形可以几乎一样：路灯照下来，一片亮。如果后面的车都正常开过去了，记录里这几种路面的「接下来」也几乎一样：车照常往前走。

这就是「已有记录中的不可区分配置」：底下的纠缠不同，显露出来却相同；在被记录下来的那一段里，后续的显露也相同。一个只看记录的模型，没有任何理由把它们分开，因为把它们分开对预测记录毫无帮助。

可是对行动来说，它们差别巨大。老彭如果在这片反光上踩一脚急刹，薄水膜上车会稳稳停住；油污上车可能打滑；积水深的地方，车速快了方向盘会一下子发飘。三十年驾龄让老彭对这件事格外小心：雨夜看到前面路面特别亮、又是刚出过事故的路段，他会提前松油门，不在亮处踩重刹车。他并没有看出那是油，他只是知道：这种亮，可能底下藏着几种东西，而它们对刹车的反应不一样。

拧瓶盖也一样。两只看上去一模一样的玻璃罐，一只抽过真空，一只没有；从外面看分不出来，从别人轻松拧开的录像里也分不出来，因为录像里的人碰到的恰好是没抽真空的那只。只有手上去一拧，阻力才把两种配置分开。

值得注意的是，在一条通道里不可区分，换一条通道往往就能区分。雨夜那片反光，眼睛分不清，耳朵却能听出一些：车轮压过薄水膜是细细的「沙沙」声，压进积水是「哗」的一声，车身还会被轻轻拽一下。老彭开着车窗的时候，常常是先听见、后看清。也就是说，不可区分不是配置本身的属性，而是「这一组配置」相对于「这一种记录方式」的属性。一个只有画面的记录，会把听觉和接触能区分的东西压成一样；第五编讲三种 SDE 各自完整、共同参与，在这里显出了它的工程含义：增加一条通道，不只是多了一份数据，而是可能让原来混在一起的配置分开。当然，这也有边界，有些配置换了通道照样分不开，只有动手试才分得开。

这一节说的是什么层次的话？「表面相同、底下不同」是日常经验和物理常识都承认的事实；「只靠被动记录训练的模型没有动力把它们分开」是一个机制假说，它可以检验：拿一批这样成对的配置去测，看模型的内部表示是否把它们区分开。在机器学习里，这类问题有一个更一般的名字，叫可辨识性问题：如果两种不

同的底层情况产生同样的观测，那么仅凭观测就无法辨认它们。这是统计学中的一般结论，不是本书的发明。

◆ 三、新的一下，打开新的后果

记录里的车都是匀速开过那片反光的。可如果你在那里急刹、急打方向，记录里就没有答案了。这就是第二处缝隙：新干预下的不同后果。

所谓干预，就是行动者自己改变了差异运行的方式。被记录下来的，总是某些人在某些时候做过的那些动作；而一个要真正行动的系统，迟早会做出记录里没有的动作，或者在记录里没有的情形下做出某个动作。在这种时候，从记录里学到的相关性就不再可靠了。

举一个家里的例子，这是我构想的场景。周敏家的老式燃气灶，打火要先按下旋钮、停两秒、再转。小舟看妈妈做了几十次，画面上每次都是「按、转、着火」，他预测得非常准。轮到他自己，一按就转，火没着，他以为是坏了，又连着转了好几次，厨房里飘出一股燃气味。他看过的所有记录都是「做对的动作」接「着火」，里面没有「做错的动作」会怎样。他的预测在记录上是对的，他的行动却不对，因为他做出了记录之外的动作。好在周敏在旁边，关了阀门，开窗通风，再手把手让他按住旋钮数两秒。

技术上，第四编讲过的几条路线（Genie 的隐动作、V-JEPA 2 的动作条件后训练、DreamerV3 在自己的交互经验上学世界模型，见第17、18章）都在把「动作」放进预测的条件里，正是在回应这一缺口。本书不预写它们成败，只指出判断的关键：训练用的动作覆盖了多宽的干预范围？动作条件下的预测，在多远的新干预上仍然可靠？这要按版本、按公开评测去核实〔各路线在新干预上的泛化程度，待核〕。

小舟的燃气灶还提示了一层：记录里缺的不只是「别的动作」，还缺「做错之后该怎么办」。妈妈的几十次示范里，从没出现过「没打着火、闻到气味、先关阀开窗」这一段，因为妈妈从来没做错过。示范越熟练，越少留下纠错的样子；学的

人照着熟练的示范学，偏偏学不到出错之后的那一套。这是被动记录的一个结构性空白，不是多录几遍能补上的。

◆ 四、一步一步错下去

第三处缝隙是时间。即使每一步的预测都只差一点点，往前推得多了，误差会累积。

老彭年轻时跑过长途（以下是我构想的场景）。他说开夜车最怕的不是一下子的大意外，而是方向盘上那一点点偏：车子有点跑偏，每一秒都只偏一点，人困了不去修正，十几秒后车就压到了线上。这是控制的道理：每一步的小误差，如果没有被现场的反馈及时修正，会一步步叠上去。

一个世界模型如果被用来「想象」未来十秒、一分钟的走向，它在第一步的小偏差，会成为第二步的输入，第二步的偏差又成为第三步的输入。这被称为累积误差或者复合误差，在用模型做长时程推演的研究里是一个公认的难题。这里只说现象，不引任何模型的具体误差数字。

和累积误差相伴的是分布变化。模型在某一类数据上训练，用在另一类情形上，统计性质变了。老彭在长沙开了三十年车，有一年去东北过冬，第一次在压实的雪地上开，他说「感觉方向盘是空的」：转向、刹车、加速的手感全变了。长沙的雨夜经验，到了雪地上，只有一部分还用得上。更隐蔽的分布变化来自行动者自己：一个系统开始按某个模型行动，它就会把自己带进一些从前记录里少见的状态，而它在这些状态上的预测本来就弱，于是越走越偏。这一点在模仿学习的研究里早有讨论，不是新问题，但它恰好说明了本章的命题：预测得好的地方，正是过去的行动者常去的地方；新的行动者一旦行动，就离开了那里。

这里也要说清边界，免得把命题说过了头。显露预测并不总是不够用。有两种情形，它离可靠行动很近。第一种，行动者的动作基本不改变被预测的过程。天气预报（第四编已作主例讨论）就是这样：人带不带伞，不影响明天下不下雨，所以预测准了，决策也就稳了。开车时预测远处一辆车会不会变道，也有几分这个性质，至少在老彭没有逼近之前，他的动作对那辆车影响很小。第二种，行动的范围

被限定在记录已经充分覆盖的区域里。工厂流水线上的机械臂，每天对同一型号的瓶子做同一个拧盖动作，瓶盖、力矩、速度都被标准化了，过去的记录几乎就是将来的情形，预测与行动之间的缝隙被生产线本身压得很窄。

这两种例外恰好从反面印证了命题：预测足以支撑行动的地方，要么是行动不介入，要么是有人事先把世界规整成了记录覆盖得到的样子。后一种尤其值得注意。流水线的可靠，有一大半不在机械臂的模型里，而在工厂对瓶子、盖子、传送带的标准化里，是工程师替机器把世界的发生范围收窄了。这已经是在用任务组织世界，只是组织者站在机器之外。第28章会把这件事说开。

还有一个常见的反驳：只要记录足够多，总能包含各种干预，不就把缝隙填上了吗？这个想法有一半是对的。记录里的干预越多样，模型学到动作与后果之间关系的机会就越大，这正是现在许多路线努力收集交互数据、或在仿真中大量试探的原因。但它有两个限度。其一，记录里的干预是别人在别人的身体、别人的任务下做出的，换一个身体和任务，同一个动作的后果可能不同，第26章讲小舟和老彭拧同一只罐子就是如此。其二，越是危险的干预，记录越少：没有人会为了收集数据在雨夜的油污上反复急刹。长尾里最要命的那部分，往往正是记录最难覆盖的部分。所以多收集干预数据是有用的方向，但它把问题推到了「在哪里、以什么代价试探」，而不是让问题消失。

◆ 五、错了以后：怎样发现，怎样变成新经验

讲到这里，容易得出一个悲观的结论：既然预测和行动之间隔着这么多缝隙，世界模型不就靠不住了吗？其实不然。人也是一样，老彭的预测也不总是对的，但他的驾驶依然可靠。**可靠的秘密不在于预测永远正确，而在于错误能被及时发现、并且被转化**。

老彭怎样发现预测错了？靠的是多条通道之间的对不上。他以为前面那段路是普通湿路，可一踩刹车，脚下的踏板反馈和车身的减速感和他预想的不一样，方向盘上也传来一点轻微的抖动。眼睛看到的显露、脚下与手上的接触显露、身体感到的减速，三者之间出现了失配。这种失配就是信号。他立刻松一点刹车、重新踩、

方向盘回正。用本书的话说，这是一次差异运行把预测与实际之间的张力暴露出来，而张力推动他改变动作；张力过后，这段路在他心里被重新标记了，下次经过会更早减速，这就是回写，纠缠被改写了。

这里要守住本书反复强调的一条区分：控制反馈、参数训练、持久学习是三件事（第18章用导航软件立过这个分法，第35章还要落到机器人身上细讲）。老彭当场松刹车、回正方向，是控制反馈，只修正这一次；下次提早减速，是经验的持久改变；讲给同行、同行也改了开法，就进了集体经验。不能因为一个系统当场修正了动作，就说它「学会了」。

这里还藏着一个更细的问题：错误是在哪一刻被发现的。老彭那一脚刹车，若是在离危险还远的地方踩的，失配暴露得早，他有余地去改；若是到了跟前才踩，失配暴露的时候已经来不及。老司机在雨夜会有意在安全的路段轻点一下刹车，试试今天的路面，这一下不是为了停车，而是为了让失配在还来得及的时候暴露出来。这说明，发现错误的时机本身是可以安排的：与其等错误自己撞上门，不如在代价小的地方先去碰一碰。这一点放到机器上，就是第32章要讲的主动试探。

错误预测要变成新经验，至少需要三个条件。一是错误要可被觉察：有另一条通道能和预测对照，或者有一个任务结果能判定失败。二是失败的片段要被保留，并且和当时的条件一起保留，而不是只留一个「失败」的标签。三是保留下来的经验要能进入下一轮训练，改写纠缠，而不是停留在日志里。第七编第31章「围绕共同事件组织训练经验」会专门讨论成功与失败过程怎样共同进入训练。

还有一层容易被混淆的东西：评价标准。一个预测模型通常用「预测的画面或状态和真实的差多少」来评分，差得越少分越高。可是对行动来说，误差的分量并不平均。雨夜路面上，把一棵路边树的叶子预测错了几片，无关紧要；把前方地面的附着力预测错了一点，可能就是追尾和不追尾的分别。一个在整体画面误差上得分很高的模型，可能恰好在那一点关键的差异上出错，而平均分把这个错误淹没了。反过来，一个画面糊一些、却把附着力、距离、相对速度把握得很准的模型，对开车更有用。所以，「预测得准不准」必须问「对什么准」；而「对什么」，又一次指向了任务。这也是为什么本书一再强调，世界模型的成功标准要和它所服务

的行动一起定，不能只看显露复现得像不像，第16章「世界模型究竟以什么为世界」已经把这件事列为审查入口。

由此可以回头补正第26章：信息负担之所以构成真正的困境，并不只因为情形太多，更因为在被动记录里，有一些对行动至关重要的差异根本没有显露。行动的可靠性要靠干预来打开这些差异，靠多通道的失配来发现错误，靠回写来改写纠缠。那么，面对几乎无穷的可能差异，究竟该打开哪些、保留哪些？下一章给出答案：由任务来组织。

> **带走的话**
>
> 预测显露回答「接下来会怎样」，行动可靠要回答「我这样做会怎样」；两者之间隔着不可区分配置、新干预和累积误差三道缝隙。可靠不靠预测永远正确，而靠错误能被发现、被保留、被回写成新的纠缠。

小账 论证落在哪一层？——本章以机制假说为主（被动记录不提供区分底层配置的动力；行动带来分布变化），用经验事实作支撑（可辨识性、累积误差、模仿学习中的分布偏移都是公开文献中的一般结论；各技术路线的具体表现标〔待核〕）；本体论立场体现在「行动者是差异运行的一部分」。 世界模型补了哪一条？——补上一条条件——世界模型的可靠性必须在新干预下检验，并具备失配觉察、失败保留与回写三个通道。

28

第 28 章

任务如何组织世界的发生范围

> **这一章的问题**：面对几乎无穷的物理差异，一个学习者凭什么知道哪些该留、哪些可以不管？

◆ 一、命题：世界的发生范围由任务组织

前两章留下了两个悬着的问题。第26章说，压缩之所以可能，是因为丢掉了无关紧要的差异，可是谁来决定什么无关紧要？第27章说，行动的可靠要靠干预去打开那些在记录里没有显露的差异，可是可以打开的差异无穷无尽，该打开哪些？

本章的回答是：由任务来组织。任务决定需要保留哪些差异、要多细的粒度、在多大的时空范围内关心。任务在这里不是一个外加的「用途」，贴在一个已经完成的世界模型上；任务本身参与了世界的发生。对一个正在做事的复合体来说，什么算一个对象、对象有哪些属性、对象之间有什么关系，都是在任务中显露出来的。

这是本书本体论立场在物理世界模型上的落点。发现学的设想是：世界先在那里，对象和属性都已经备好，模型只要把它们抄下来，任务只是从抄好的清单里挑几项用。发生学的设想是：对象、属性和关系在特征纠缠中经差异运行而显露，而任务正是组织这种纠缠与差异运行的主要力量之一。注意这里说的是「主要力量之一」：好奇、偶然的碰撞、身体的结构、别人教的东西，也都参与组织；本章只把任务拎出来讲，因为在工程上它最可控。

◆ 二、三样东西由任务决定：差异、粒度、时空范围

先说差异。装修师傅进门，第一眼看墙，他看的是什么？他不看墙纸的花色，不看墙上挂的照片，他看墙面有没有空鼓、墙角有没有开关插座、哪里可能走着水管电线。同一面墙，女主人看的是颜色配不配沙发，孩子看的是能不能贴贴纸。三个人对着同一面墙，显露给他们的差异完全不同。不是墙变了，而是三个人带着三种任务进入这面墙，三种任务组织了三种不同的纠缠，于是显露出三面不同的「墙」。

再说粒度。还是那位师傅。要在墙上打孔装空调支架，他关心的是厘米级：孔的位置偏一两厘米没关系，只要避开管线、找到实墙。要铺地砖，他关心的是毫米级：两块砖之间的缝差一毫米，铺到房间那头就能看出来是歪的。要判断这面墙能不能拆，他关心的又是另一种粒度：不是长度，而是承重不承重，是整面墙在整栋楼里的位置。同一个师傅、同一套房子，任务一换，需要辨认的差异就换了一种细度。粒度太粗，该分开的分不开；粒度太细，被无关的细节淹没，反而做不成事。

再说时空范围。拧瓶盖的任务，时间范围是几秒，空间范围是一只手和一只瓶。但如果任务是「把这瓶果酱打开、吃完、下回还能盖严」，时间范围就拉长了：要关心螺纹里会不会留果酱、盖子拧回去时用多大的劲，否则下回又会拧不开。老彭开车，眼前这个路口的时间范围是几秒；整趟行程的时间范围是半小时，要关心油量、乘客要去的地方、哪条路这个点会堵；对于这辆车，他关心的时间范围是几个月，那一声异响是现在就得去修，还是可以再跑一阵。

这三样东西怎样定下来？不是靠事先想清楚，而是靠做的时候碰出来的。这是我构想的场景：小舟第一次帮妈妈拧一罐蜂蜜，他使劲握住盖子往一个方向转，没开；再使劲，还是没开。周敏让他看一下盖沿，原来上面印着一个箭头，要先往下按再转。他这才知道，对这一类盖子，「往下按」是一个必须保留的差异，而他原先的纠缠里根本没有这一项。下一回再碰到儿童安全盖，他的手会先试着往下压一压。差异的取舍、粒度的选择，都是在一次次「没成」里被修出来的。任务提供的，是判断「成没成」的那把尺子；有了这把尺子，失败才会把注意力推向原先不

在范围里的差异。没有任务，就无所谓成败，差异也就没有轻重之分，学习者面对的只是一片平均的、没有起伏的显露。

这也解释了第26章的那个疑问：压缩为什么可能，又靠什么取舍。取舍的根据，不是某种从天上掉下来的「重要性」，而是任务的成败把差异分出了轻重。对拧瓶盖来说，颜色的变化从来不会让盖子拧不开，于是颜色被压缩掉了；手湿的变化常常让盖子拧不开，于是手湿被保留下来，而且保留得很细。

用三方程来说：任务改变了 D 的运行方式（关心哪些差异、在什么粒度上比较），也改变了 E 的组织（哪些特征被缠到一起），于是 S 显露成不同的样子。这三样不是任务从外面「选择」的，而是在做事的过程中一起形成的。

◆ 三、任务相关的对象与可行动关系

任务不只决定看哪些差异，还决定什么算「一个对象」。

瓶盖和瓶子是一个东西还是两个东西？对喝水的人来说，拿起来就是「一瓶水」，一个东西；对要拧开它的人来说，瓶盖和瓶身必须是两个东西，一个要握住不动，一个要转。对回收垃圾的人来说，瓶盖和瓶身又要分开，因为塑料种类不同。哪里是一个对象的边界，取决于你要对它做什么。

更进一步，任务决定对象之间有什么「可行动关系」。对拧瓶盖的人，毛巾和瓶盖之间有一种关系：毛巾能增大摩擦。这种关系在平常并不显露，一块毛巾就是毛巾；只有在「拧不开」的时候，它才作为「可以垫在手上的东西」显露出来。装修师傅的工具箱里有一卷胶带，平时用来缠电线；打孔时他撕一段贴在墙上要打的位置，这样钻头起钻时不打滑、瓷砖不崩边。胶带和钻头之间的这层关系，是在「瓷砖上打孔」这个任务里才发生的。

放到机器上看，这一点同样清楚。一个工程团队给机器人做感知，首先要回答的问题之一就是：要把画面切成哪些东西？林岚曾跟周敏聊过她们组里的一次争论：给一台整理桌面的机器人标注数据，一位同事主张把「杯子」和「杯子把手」分开标，因为机器人要抓的是把手；另一位同事说没必要，杯子整体标出来就够

了，抓的位置让抓取模块自己找。争到最后，他们发现答案取决于下游任务：如果机器人只是把杯子推到一边，整体就够；如果要端着装满水的杯子走，把手就必须是一个单独的、可以被稳稳握住的对象。标注体系看起来是在「描述世界里有什么」，实际上已经替任务划好了对象的边界。这是我构想的场景，但标注规范随任务而定，是机器学习工程中普遍的做法。这层划分一旦写进数据，模型学到的世界就带着它；任务换了，原来的划分可能就不够用，甚至会妨碍学习。所以，在技术系统里，任务组织对象这件事往往不是发生在模型训练的时候，而是更早，发生在采什么数据、怎样标注、用什么当奖励的时候，第三编第11章讲「数据选择、目标设置与评价反馈」时已经指出了人在这里的参与。

第五编第25章讲「物、用途与行动可能性的共同形成」，这里可以看到它在具体任务中的样子：物不是先有了一张完整的属性清单，再从里面挑出用途；物的哪些方面显露出来，和这一刻能对它做什么是一起形成的。

◆ 四、相关性会变：任务内部的转移

任务组织世界的发生范围，但这个范围不是一次定死的。即使任务不变，什么相关也会在做的过程中变化。

这是我构想的场景：老彭载一位乘客去火车站。出发时，相关的是路线、路况、时间。开到半路，下起大雨，路面的附着力、前车的刹车灯、积水的路段一下子成了最相关的东西，路线反倒退到了后面。再开一阵，车里传出一声从没听过的「咯噔」，那一刻，相关的又变成了：这声音从哪里来，跟踩刹车有没有关系，方向盘的手感有没有变。任务始终是「把乘客安全准时送到」，可在这条路上，哪些差异要留意，变了三回。

这种变化的推动力，用本书的话说，正是三原理：某两维之间出现矛盾，推动第三维改变。车里那声「咯噔」，是听觉上一个新的差异与老彭对这辆车的熟悉纠缠（这车平时什么声音都有数）之间出现了对不上，张力积累起来，推动他改变显露：他开始把注意放在车身底下、放在脚下的踏板反馈上，放在手里方向盘的细微抖动上。这一改，就把原先不在范围里的东西拉进了范围。等他确认那声音跟刹车

无关、只是后备箱里一只没放稳的箱子，张力消退，范围又收回来，这次经历被回写进他对「这辆车会发出什么声音」的纠缠里。

这对技术系统意味着什么？如果任务的相关范围在系统设计时就被一次性写死，那么系统在遇到范围外的差异时，要么根本感觉不到，要么感觉到了却不知道往哪里放。可靠的系统需要一种「范围可以扩张」的机制：觉察到范围外的差异和范围内的预期发生冲突时，能把注意与处理资源移过去。这是一个机制假说，在工程上它有多种可能的实现方式，比如异常检测触发更细的处理、由更通用的模块接手、或者交还给人。哪一种更好，要按任务去测。

这里顺带要分清「任务」和工程上的「目标函数」。在技术系统里，任务常常被写成一个可计算的分数：到达终点得分，碰撞扣分，耗时扣分。这样写是必要的，否则无法训练。但分数只是任务的一个压缩。老彭心里的「把乘客安全准时送到」，包含着分数里没写的东西：不要急刹让后座的老人摔着，不要为了快一分钟去抢一个黄灯，乘客晕车了要开慢些。分数写得越窄，系统就越可能找到一条分数很高、却违背任务本意的路，机器学习研究里把这类现象称为「奖励投机」或「规格博弈」，已有大量公开讨论。用本书的话说，目标函数是任务在某一时刻的一个显露，它会随着复合体对任务理解的加深而修订；把它当作任务本身，就又回到了发现学：以为任务是先在的、一次写清的。

◆ 五、边界：初始任务边界在互动中不断修订

说「任务组织世界」，有一个容易滑过去的误解：以为任务是事先清清楚楚定义好的，像一张需求说明书，世界的范围就按说明书裁出来。实际上，任务本身的边界也是在互动中发生、不断修订的。

装修就是最好的例子。一户人家请师傅来「换个卫生间的水龙头」。师傅一拧开旧龙头，发现墙里的水管接头锈透了，换了龙头也会漏。任务从「换龙头」变成了「换一段管子」。可换管子要开墙，开墙就要动瓷砖，瓷砖是十几年前的款式，买不到一样的了。于是又要和主人商量：是补几块不一样的，还是干脆把这面墙重新贴。一个二十分钟的活，变成了两天的活。老师傅的本事，不只在于把一件定好

的活干好，更在于在干活的过程中及时发现任务的边界不对，并且知道该怎样改它、什么时候该停下来跟人商量。

这说明两件事。第一，任务的边界不是封闭的说明书，而是一种初始的组织，它在与世界的互动中会被差异冲破，需要重新划定。第二，划定任务边界本身往往不是一个人或一个系统独自完成的，而是在复合体里完成的：师傅和主人一起决定「这面墙重贴」。这和本书的发生单位口径一致：具体使用讲个人—AI复合体，任务的划定也发生在复合体之中，而不是由机器单方面决定。

还要防止另一种过头。说任务组织世界，不等于说世界是任务随意捏出来的。墙里的锈管子不因为师傅只想换龙头就不存在，路面的油污不因为老彭只想准时到站就不滑。现实有它的阻力，第二编第8章讲过，现实阻力正是检验现实发生的力量。任务决定的是「哪些差异被组织进这一次的发生」，不是「世界里有什么」；那些没被组织进来的东西，会以失败、意外、冲突的方式重新闯进来，逼着任务的边界修订。任务组织与现实阻力，是同一个过程的两面。

最后还有一个反方向的风险：任务组织得太窄。一个只为「拧这一种瓶盖」练出来的手，碰到别的盖子就手足无措；一个只为某条线路练出来的司机，换个城市就晕头转向。任务把差异分出轻重，也就有了把某些差异压得太死的危险：被判为「无关」的差异，一旦在新情形里变得要紧，学习者可能根本察觉不到。好的师傅手上干一样活，眼睛会多看几样：手上做的是眼前这件事，眼睛还留意着旁边那些暂时不相干、但可能突然相干的东西。用本书的话说，这是给任务边界之外留一道缝，让范围外的差异仍有机会进入。这道缝留多宽，本身也要按任务权衡：开车时留得宽些，因为长尾代价是人命；在流水线上留得窄些，因为环境已被标准化。这同样是一个可以通过对照实验检验的设计问题，而不是一个可以一劳永逸回答的哲学问题。

◆ 六、向下一章：驾驶是一个专门发生的世界

把这一章合起来：世界模型不必以穷举整个物理世界为目标，它可以、也应当围绕具体任务，保留任务所需的差异、粒度和时空范围，形成任务相关的对象和可

行动关系；同时它必须具备让范围随相关性变化而调整、让初始边界在互动中修订的机制。这就是「任务组织世界的发生范围」的完整含义。

这个说法要经得起一个真实而复杂的任务的检验。驾驶正是这样一个任务：它关乎生命安全，牵涉视觉、听觉、接触与车辆响应，有着极长的长尾，又有大量的公开技术积累。下一章就把驾驶当成一个完整的样本，看一个专门的世界怎样在人—车辆—道路复合体里发生。

> 带走的话
>
> 任务不是贴在世界模型外面的用途，它组织了要保留的差异、粒度、时空范围和可行动关系；任务边界在互动中被现实阻力冲破、在复合体中重新划定，世界模型的可靠封装由此才有了范围。

> **小账** 论证落在哪一层？——本章以本体论立场为主（对象、属性与关系在任务组织的纠缠中显露），机制假说为辅（相关范围的扩张需要冲突觉察与资源转移机制）；经验事实主要来自日常生活。 世界模型补了哪一条？——补上一条条件——世界模型的发生范围应由任务组织，且须带有范围随相关性变化而扩张、边界在复合体中修订的机制。

29

第 29 章

驾驶世界模型的专门发生

> **这一章的问题**：开车这件事，究竟是在什么样的世界里发生的？一个驾驶世界模型，要在哪些地方专门长出来？

◆ 一、命题：驾驶的发生单位是人—车辆—道路复合体

说到「驾驶世界模型」，人们脑子里常常浮现的是一个画面：车顶上有一双眼睛，看清路上的一切，然后决定怎么开。这个画面把驾驶写成了「一个观察者面对一条路」。本章的命题是：驾驶不是一个观察者对着一个世界，而是人、车辆、道路三者结成的复合体里专门发生的一个世界。这个世界里，车不只是被开的工具，它的转向手感、刹车脚感、发动机的声音、轮胎的抓地，都是这个世界显露的一部分；路也不只是被看的背景，它的坡度、湿滑、坑洼、车流、信号，时时刻刻通过车身传回来。人在中间，用眼、耳、手、脚、屁股底下的座椅和整个身体，与车和路纠缠在一起。

对于自动驾驶系统，复合体的构成换了一个样子：传感器和计算单元取代了人的感官与判断的一部分，但车辆和道路依然在，背后设计、标注、测试、监管它的人也依然在。按本书的发生单位口径，驾驶系统的训练是一个集体人类—AI复合体的事情：工程师选数据、定目标、写评价，路上千千万万司机的行为成了数据，交通法规和道路设计成了约束。它不是一台机器独自在路上「发现」驾驶世界。

◆ 二、老彭的驾驶世界：雨夜、异响、方向盘

先看人怎样开车。下面这个雨夜是我构想的场景，但其中每个细节，开过车的人都不陌生。

晚上十点，老彭拉完最后一单往家开。雨下得不小，雨刮器刮一下，前面的尾灯就从一团红晕里清楚一下，又糊掉。他在看什么？他在看前车尾灯的距离变化，看路面的反光哪里亮得不正常，看路边有没有打伞冲出来的人。这是视觉。

他也在听。雨点打车顶的声音告诉他雨在变大还是变小；轮胎压过积水的「哗」声让他知道前面那段低洼；旁边车道一辆大货车超过去时，风压和水声一起砸过来，他手上不自觉地握紧方向盘。突然，底盘传来一声闷闷的「咚」，像是压到了什么东西，又不太像。他放慢车速，侧耳听，同时感觉方向盘有没有往一边拽。这是听觉，也是接触。

方向盘的手感在雨夜最要紧。干路上，方向盘有一种「沉」，转多少车走多少；路面一湿，那种沉会变轻一点，老彭说「方向发飘」。他不会去想「路面附着系数下降了多少」，他的手直接告诉他：现在不能猛打方向。脚下也一样，刹车踏板踩下去的那一段行程、车头往下点的那一下，告诉他车能停得多快。

这些感觉不是各自分开传来、再在脑子里拼起来的。雨声变大的同时，他已经松了一点油门；方向一发飘，他的眼睛自然拉远了跟车距离。第五编讲三种 SDE 各自完整又共同参与，在老彭身上，就是这样一种合在一起的「车感」。三十年的驾驶让这团纠缠稳定成了秩序态，大部分时候他根本意识不到自己在处理这么多东西；只有雨夜、异响这种时候，平常隐在底下的东西才被一一拉到前面来。

这团纠缠是怎样训练出来的？不是从书上读来的。老彭学车时，交规和操作要领都背过，但真正的车感是后来一公里一公里磨出来的。新手开车，眼睛死盯着前方，耳朵里什么也听不进，方向盘握得死紧，手上感觉不到路面传来的细微变化；这时三种感官几乎各管各的，甚至互相干扰，听见后面按喇叭，手上一慌就打了方向。开到一两年，眼睛学会了往远处看、隔几秒扫一眼后视镜，手松下来了，开始

能从方向盘上读出路面；再往后，声音也进来了，发动机转速一高一低，他不看转速表就知道该换挡。这是一个三种 SDE 从并排、到相互制约、到归并成一团车感的过程，第一编第2章讲的「稳定化、分化与重组」，在这里有一个完整的样子。

关键是，这个训练离不开车辆响应。每一次他打方向、踩刹车、松油门，车都以它的方式回应：转多少、停多快、车身怎么晃。手和脚发出的动作，与眼、耳、身体收到的回应，一次次对上或者对不上，对不上的地方逼着他调整。没有这种「我动、车应、我再动」的往返，看再多行车录像也练不出车感，这也是第四编用看录像学游泳和下水对照时讲过的道理。驾驶的车感，本质上是人和这辆车之间一段长期的相互训练：车在训练人，人也在适应车，老彭换了新车，头一个星期总觉得刹车「太灵」，就是旧纠缠在新车上失配，需要重新磨合。

还要注意：老彭的驾驶世界是专门发生的。他的车感是这辆车、这几条路、这种城市里练出来的。第27章说过，他去东北开雪地，觉得「方向盘是空的」。车感不是一个装在他身上、拿到哪里都能用的通用物理知识，它是在一个具体任务、具体复合体里长成的。

◆ 三、机器的驾驶世界：公开机制里能看到什么

再看自动驾驶系统。这里要守住本书的纪律：判断落实到具体版本和公开机制，不预先断言「驾驶 AI 全部是纯视觉」。

不同公司的路线差别很大。Waymo 在 2025 年 12 月的博文里公开了它的基础模型：摄像头、激光雷达、毫米波雷达随时间融合的编码器，借助 Gemini 训练、专管少见复杂情形的驾驶视觉语言模型，二者共同送入「世界解码器」；仿真里有强化学习的内循环，真实运行日志驱动的外循环。这些第19章已经逐项转述过，这里不再重复。本章只补两处第19章没有展开的细节：一是博文说车上的轨迹在执行前还要经过一个独立的校验层；二是训练时先把大模型适配成驾驶者、仿真器、评判器三种角色的「教师」，再蒸馏出车上运行的较小「学生」模型。内部的具体结构、数据规模与评测数字，博文没有完全公开，本书不作推断〔待核〕。

另一些厂商以摄像头为主。不同路线的取舍各有理由，也各有公开争论，本书不在这里裁判孰优孰劣。

听觉在机器驾驶里的位置，公开资料比视觉少得多，但并非没有。Waymo 2017 年 7 月的一篇博文写过：它当时的新一代车辆装了自行设计的音频系统，能在比上一代远约一倍的距离上听到警笛，团队还在教软件判断警笛从哪个方向来，因为司机往往先听见、后看见；那一轮测试是和亚利桑那州钱德勒市的警察与消防部门一起做的。这是公司自述，后来的版本怎样，本书不推断。至于像老彭那样从底盘异响判断自家车况，在公开的自动驾驶机制描述中并不常见，车辆状态更多是通过车上的各类电子传感器和总线信号获得〔各厂商是否使用车内声学监测车况，待核〕。

警笛这个例子正好落在本书的口径上。警笛之所以要靠耳朵，是因为它常常在视线之外：消防车还在拐角后面，声音已经到了。这时听觉显露的不是一个看得见的物，而是一个正在逼近的事件，第23章说「声音首先是事件，其次才是物」，在驾驶里有了一个实打实的用处。可是只听见还不够，还要把它和看得见的东西对上：声音从哪边来，路口哪辆车要让，自己该靠边还是停在原地。Waymo 那篇博文里说车辆在警笛从后面来时可能靠边、从前方来时在路口让行，这已经是听觉显露与视觉、与本车动作之间的一种相互制约，而不只是多了一个「有警笛」的标签。它做到了多深，要看在声音被遮挡、被高楼反射、与喇叭声混在一起时的表现，这些公开材料没有给出。

把这些公开机制放进本书的框架里看，第19章已经说过两点：它不是「纯视觉」，也不是纯被动记录；多传感器融合不等于多种 SDE 纠缠，公开材料不足以判断它属于哪一种。本章只再补一点：外循环从真实日志里挑问题、仿真检验、再部署的流程，在结构上和第27章讲的「错误被发现、保留、回写」一致；但这是工程组织上的相似，不等于证明了 SDE 本体论。

◆ 四、身体感觉与机器信号：对应的边界

人有车感，机器有信号。能不能把两者一一对上，比如老彭的「方向发飘」就等于转向系统里的某个力矩读数，「刹车有点软」就等于制动压力的某种曲线?

部分可以。汽车工程里，转向手感、制动脚感本来就是用力矩、行程、压力这些量来调校的，人觉得「发飘」，往往对应着某些可测的变化。这是有工程基础的。

但对应有边界，至少有三层。第一，人的身体感觉分类是在任务中长出来的，不是按传感器的通道分的。老彭说「车不对劲」，可能同时来自声音、方向盘的微振、座椅传来的抖动和他对这辆车平时状态的记忆，是一个跨通道的判断；把它拆成几个信号分别对应，会丢掉它们之间的相互制约。第二，同一个身体感觉，在不同情境下指向不同的东西。方向发飘，雨天可能是路滑，晴天可能是胎压不足，高速上可能是侧风。人分得开，是因为他同时知道天在下雨、刚补过胎、这段桥面风大；机器信号若只对应「发飘」这一个特征，就分不开。第三，有些机器信号人根本没有对应的感觉，比如激光雷达测到的百米外精确距离、毫米波雷达穿过雨雾测到的速度。机器的驾驶世界在这些地方比人宽，不必也不应削足适履去模仿人的车感。

还要反过来看人的局限，免得把老彭写成驾驶的标准答案。人的车感同样有失效的时候：开夜车犯困，感官之间的纠缠会松掉，眼睛看着路，手却没跟上；心里有事、看了一眼手机，注意的范围就塌缩到一个点上；车感太熟，反而会让人对一些新的差异视而不见，老司机有时比新手更容易在熟悉的路段放松警惕。人的驾驶世界也是在任务中专门发生的，也有它的盲区和失配。把人车感当作机器要追的唯一样板，和把机器信号当作取代人的全部理由，犯的是同一个错：把某一个复合体里发生的驾驶世界当成了驾驶世界本身。更有意义的比较，是在同样的任务和情境里，看两种复合体各自在哪里稳、在哪里会失效、失效后怎样恢复，这需要公开、可比的统计，而不是印象。

所以，「人的车感」和「机器的信号体系」是两个各自发生的驾驶世界，它们在某些地方可以对译，在某些地方各有对方没有的东西。对一个驾驶世界模型来说，有意义的问题不是「能不能复制老彭的车感」，而是：在这个复合体里，哪些差异对驾驶安全是必须保留的，它通过哪些通道、以什么组合被保留下来，以及这些通道之间冲突时怎样处理。

◆ 五、陌生情境：扩展与重新判断

驾驶的长尾极长。一棵被风吹倒横在路上的树，一个在路中间指挥交通的交警、手势和信号灯相反，一个突然从车底滚出来的球，一段刚刚塌陷的路面。每一种都很少见，加起来却天天发生。

老彭碰到陌生情境怎么办？他先慢下来，给自己时间；然后用手边的一切去重新判断：交警手势和灯不一致，听交警的；路面有个说不清的东西，宁可绕开；车里传来从没听过的声音，靠边停下检查。这一套不是针对每一种陌生情形分别学的，而是一种通用的「陌生应对」：降速、扩大注意范围、找最保守的选项、必要时求助。用第28章的话说，这是让任务的相关范围在冲突出现时扩张，并让初始边界得到修订。

机器系统怎样做到这一点？前面提到的 Waymo 公开描述里，用一个借助 Gemini 训练的驾驶视觉语言模块专门处理少见而复杂的语义情形，这可以看作是一种把更宽的集体经验接入驾驶任务的尝试；车上的校验层，可以看作把「重新判断」交给更保守的规则。还有一条通向人的路：Waymo　2026 年 2 月的一篇博文说，它的远程协助人员不远程驾驶车辆，也不持续盯着每辆车，而是在车辆自动驾驶系统主动发出请求时提供建议，系统可以采纳，也可以不采纳。按本书的说法，这是在复合体里给「拿不准」安排了一个正式的出口，求助由车发起，决定仍留在车上。这些安排是否足够，不能凭架构图判断，只能靠在陌生情境里的公开表现和统计来判断。

陌生情境里还有一类特别容易被忽略：道路上的其他人。复合体里的「道路」不只是沥青和标线，还包括路上别的司机、行人、骑电动车的人，以及他们之间约

定俗成的默契。老彭在长沙开车，知道哪些路口电动车会闯灯，知道前车打了双闪往往是要临时停车，知道对面车闪一下远光多半是在说「你先走」。这些不是交规写的，是一个城市的司机们相互训练出来的集体经验。换一个城市，默契就变了。驾驶世界模型若只学「物」的运动，而不学这层人与人之间的相互预期，就会在这些地方显得笨拙或者危险。这层集体经验怎样进入机器的训练，是一个值得专门研究的问题，它连接着第三编讲的集体语言世界与本编讲的物理世界。

这里也要说清一个边界：陌生情境的应对，不能被理解为「再多收集一些长尾数据」就能彻底解决。第26章已经算过，长尾是收集不完的。可靠性来自两件事的结合：一是在任务范围内把常见情形做得非常稳；二是在范围之外有一套觉察冲突、收缩行动、请求支援的机制。前者靠专门训练，后者靠任务边界的设计与复合体的协作。

◆ 六、向下一章

驾驶世界是专门发生的：它在人—车辆—道路复合体里，经视觉、听觉、接触与车辆响应的共同训练而形成，带着它特有的差异、粒度、时空范围和陌生应对机制。可是驾驶系统里用到的很多东西，并不是为驾驶专门从零造的：识别行人的能力、理解交通标志文字的能力、关于雨天路滑的常识，都来自更广的共同底座。专门的世界与共享的底座怎样协作？共享到什么程度，才不至于变成「一个模型穷尽全部物理发生」的老路？这是下一章的问题。

> **带走的话**
>
> 驾驶世界不是车顶上的一双眼睛看到的世界，而是人—车辆—道路复合体中专门发生的世界；人有车感，机器有信号体系，两者可以部分对译，但各自发生、各有对方没有的东西，可靠来自范围内的稳定加上范围外的觉察与求助。

小账 论证落在哪一层？——本章以经验事实（Waymo 2025-12 博文的公开机制，其余标〔待核〕）与机制假说（身体感觉分类与机器信号的对应边界、陌生应对机制）并重；本体论立场体现在复合体作为发生单位。　世界模型补了哪一条？——补上一条条件——专门任务世界模型的评价，要连同复合体构成、通道组合、冲突处理与范围外应对一起交代，而不能只看感知精度。

30

第 30 章

专用模型的协作与共享

这一章的问题：如果世界模型要围绕具体任务专门发生，那不同任务之间还能共享什么？共享到什么程度，才不变回「一个模型包揽一切」？

◆ 一、命题：共享底座，专门发生，接口协作

前四章走到这里，容易被误读成一句口号：「不要统一模型，要专用模型」。这不是本编的结论。本编的结论更细一些：可复用的底座与任务专门训练应当协作；底座提供的是许多任务都用得上的纠缠，专门训练在具体任务中把这些纠缠组织成可靠的能力；不同任务世界之间通过接口相互调用、相互校验。共享是真实的，也是有价值的；但共享一个底座，不等于一个模型穷尽了全部物理发生。

先说清「统一」与「共享」的分别。统一，在本编的意思里，是指一个模型、一套参数，被期待在不经专门训练的情况下覆盖所有物理任务。共享，是指多个任务世界从同一个底座出发，继承它的纠缠，然后各自专门发生。前者是一个强断言，第26章说过它是一个要由证据回答的经验问题；后者是一个结构安排，人类社会和工程实践里都有大量的先例。

◆ 二、人怎样共享：手艺的底座与专门的活

装修师傅是看清这件事的好例子。一位老师傅会的东西可以分两层看。底下一层是很多活都用得上的：手稳，眼准，知道力往哪里使；看得懂墙体、看得出水平；知道材料遇水会胀、遇冷会缩；会用电钻、角磨机、水平尺。上面一层是专门

的：贴瓷砖有贴瓷砖的手法，砂浆调多稀、怎样敲实、留多大的缝；装水电有装水电的规矩，管子怎么走、接头怎么做、打压试多久；做木工有木工的讲究。

一个学徒进门，先学底下那一层，然后跟着师傅一样一样学专门的活。底下那层让他学专门的活快得多：手稳的人学贴瓷砖，比手抖的人快；懂材料的人学防水，比不懂的人快。但底下那层不会自动变成任何一门专门手艺。手再稳，没有贴过几百平方米的瓷砖，也贴不平一面墙。专门的手艺是在专门的活里，带着它特有的差异、粒度和失败，一次次磨出来的。

拧瓶盖也有这两层。底下一层是手的抓握、旋转、用力的一般能力，小舟从会拿勺子开始就在练；上面一层是对各种瓶盖的专门对付：安全盖要压、真空罐要敲、湿手要垫毛巾。没有底下那层，专门的对付无从谈起；只有底下那层，一碰到安全盖照样打不开。

用本书的话说，底座是在大量不同任务中反复出现、反复被训练的那部分纠缠，它稳定成了秩序态，可以被很多任务调用；专门能力是底座之上，在具体任务的差异运行中进一步组织起来的纠缠，它带着这个任务的边界。两层不是简单地叠在一起：学了贴瓷砖，师傅的手也会更稳，专门训练会回写到底座里去。这是三方程在两层之间的循环。

◆ 三、机器怎样共享：基础模型与专门训练

技术上，「大规模预训练的共享底座＋针对任务的后训练」已经是一种常见的做法。第三编讲的大语言模型是一例：预训练形成一个集体语言世界，再经指令微调、反馈训练等适配到具体用途。物理世界模型的路线里也有类似的结构。按公开描述，V-JEPA 2 先在大量视频上做自监督的特征预测，再用较少量的机器人交互数据做动作条件的后训练，用于规划；Waymo 2025 年 12 月的博文描述了大模型作教师、蒸馏出车上运行的较小模型，并接入一个借助 Gemini 训练的视觉语言模块；Cosmos 3 按其摘要把自己定位为给具身智能体用的通用底座（第19章已述），下游怎样适配，摘要没有细说〔待核〕。这些都是「共享底座＋专门训练」的不同形态。

这类做法的经验依据是什么？公开文献里有不少报告，说在大规模数据上预训练的底座，能让下游任务用更少的专门数据达到较好的表现。这是机器学习领域广泛报告的现象，但它的程度高度依赖任务、底座和数据，不能一概而论〔具体到物理操作与驾驶任务上的迁移幅度，需按各路线公开评测核实，待核〕。

放到本书的框架里，有两点要看清。其一，底座本身是集体人类—AI复合体的产物：它的数据来自千千万万人拍下的视频、写下的文字、开过的车，它的目标与评价来自工程团队的选择。它不是机器独自对物理世界的「发现」。其二，底座继承的是判断，不只是数据。第一编第5章讲「封装所继承的判断」：一个底座在训练时被判定为「好」的标准，比如画面预测的误差小、特征预测的误差小，会被带进所有用它的任务里。如果某个任务需要的差异恰好在底座的训练标准里不被看重，比如第27章说的那片反光底下的附着力，那么底座不但帮不上忙，还可能把这一差异压得更死，专门训练就要花额外的力气把它重新打开。

共享还有一个反方向：专门训练的经验回流到底座。人类手艺里这是常事。一位老师傅在某个老小区里摸索出一个办法，能判断老式预制板墙里哪里是空心的，他教给了徒弟，徒弟又在别的工地上传开，慢慢就成了这一行的常识。专门活里磨出来的东西，一部分会沉淀为下一代人的底座。机器体系里也可以有类似的回流：各个专门任务里发现的失败片段、难例，被收集起来进入底座的再训练。Waymo博文描述的外循环，就有从真实运行日志中挑出问题、用于后续训练的结构。但回流要有筛选：并不是每个专门任务里的经验都适合进入底座，有些差异只在某个任务里要紧，带进底座反而会干扰别的任务。哪些经验该留在专门层、哪些该沉入底座，本身就是一个需要设计和检验的问题，它决定了共享体系是越用越好，还是越用越乱。

◆ 四、不同任务世界之间的接口

共享还有另一种形态，不是从同一个底座长出来，而是不同的专门能力之间相互调用。这需要接口。

装修现场是一个多工种协作的地方。水电师傅先走管线，泥瓦师傅再贴砖，木工最后装柜子。他们之间的接口是什么？是一些约定俗成的东西：水电师傅在墙上用记号笔画出管线走向，拍照留给后面的人；泥瓦师傅贴砖时在插座位置留好开孔；木工装柜子前先问清楚哪里有管子不能打钉。这些接口不要求每个师傅都会别人的手艺，只要求在交接的地方说清楚：我这里有什么、你那里不能做什么。接口一旦出错，就会出事：没拍照的管线，被后来的人一钉子打穿，这是装修行里最常见的事故之一。

不同任务世界之间的接口有几个特点。第一，它是有选择的：只传递对方任务需要的差异，不传递全部。水电师傅不会告诉木工他用了哪个牌子的电线，只会告诉他线走在哪里。第二，它是有边界说明的：「这面墙可以打孔，但离地三十公分以下不行」。第三，它需要校验：后来的人不能全信记号笔，打孔前还要用探测仪扫一遍。这第三点尤其重要：接口不是把前一个任务世界的结论直接搬进后一个，而是一个相互校验的地方。

再看一个跨行的接口，仍是我构想的场景。老彭那辆车底盘的异响，后来响得越来越频繁，他开去修理厂。修车师傅问他：什么时候响？他说过减速带的时候响，冷车的时候响得更厉害，左边比右边明显，踩刹车时倒不怎么响。这几句话，就是驾驶世界和修车世界之间的接口。老彭不懂悬挂结构，修车师傅也没开过这辆车跑那些路；但老彭把自己车感里那团纠缠，翻译成了几个修车世界能用的差异：触发条件、温度、方位、与刹车无关。师傅听完，心里已经把范围缩到左前悬挂的几个部件上，再把车架起来一看一摇，很快找到一只老化的胶套。接口在这里做了三件事：它把一个世界里的跨通道感受压成了另一个世界能处理的少数差异；它传递了边界（「刹车时不响」排除了一大片可能）；它也接受校验，师傅并不全信老彭的描述，还要亲手摇一摇。设想将来车上的系统能把运行中记录到的振动和声音异常整理成类似的描述，交给维修方，这就是两个任务世界之间一个机器参与的接口；它好不好，同样要看它是否保留了下游需要的差异、是否带着边界、是否接受校验〔车载系统向维修方提供此类声学或振动诊断信息的现有公开实践，待核〕。

技术上，驾驶系统内部的感知、预测、规划、控制之间，机器人系统里的视觉模块、抓取模块、力控模块之间，都有这样的接口。有的系统采用模块化接口，有的采用端到端训练、把接口隐在网络内部，各有利弊，公开争论仍在持续。本书不偏向某一种，只提出一个发生学的检验：接口处是否保留了下游任务需要的差异？是否带着上游能力的适用边界？下游发现上游的结论不对时，能否把这一失配回写给上游？这三点，可以作为评价任何多任务系统协作质量的问题。

◆ 五、为什么共享不等于一个模型穷尽全部物理发生

现在可以正面回答本章最要紧的问题。共享底座、共享接口都是真实而有价值的，但它们不等于一个模型穷尽了全部物理发生。理由有三层，要分开说。

第一层是机制假说。底座之所以能被共享，是因为它保留了很多任务共同需要的差异；可是共同的东西，恰恰是去掉了各任务专门差异之后剩下的东西。装修师傅的「手稳」可以带到每种活里，但贴瓷砖需要的那一毫米、水电需要的打压试漏，都不在「手稳」里。底座越是追求覆盖面宽，越要在每个任务的专门差异上有所舍弃。这不是说底座做不大，而是说它做得再大，各个任务的专门训练与边界检验依然需要。这个假说可以被检验：如果有一天，某个底座在不经任何专门训练的情况下，在大量专门任务上的可靠性都达到了专门训练的水平，这个假说就要修正。

第二层是经验事实。到目前为止的公开报告中，物理操作与驾驶等高要求任务，普遍仍在底座之上进行任务专门的数据收集、后训练与评估〔这一概括须按最新公开资料复核，待核〕。这是现状，不是定律；现状会变，本书只把它当作现状。

第三层是本体论立场。即使某一天，一个模型在极大范围的任务上都表现可靠，从 SDE 的立场看，它也没有「穷尽物理发生」，因为物理发生本来就不是一个可以被穷尽的既成清单。新的材料会被造出来，新的工具会改变接触的方式，新的任务会组织出新的对象和关系。装修行业二十年前没有的墙板、胶粘剂、智能开关，今天都要重新学。世界在继续发生，任何模型的覆盖范围都是一个此时此地的封装，带着它的适用边界，需要在新的发生中被检验、解封、重训。这一层话不依

赖任何模型的成败，它是本书对世界的基本看法；正因为如此，它也不能拿某个模型的成败来证明或推翻。

还要预先回应一个反例。设想某个大型通用模型，用很少的示范就能完成许多以前要专门训练的操作任务，似乎「专门训练」正在变得多余。对这种情形，本章的说法并不需要回避，而是要细看：少量示范本身就是一次很轻的专门训练，或者至少是一次在上下文中的任务组织；任务依然在组织发生范围，只是组织的代价变小了。这是工程上的巨大进步，值得认真对待。但还要问三个问题：在少量示范下达到的成功率，在新干预、长时程和长尾情形下还剩多少？出错时它能不能觉察、能不能回写？它的适用边界有没有被交代？如果这三个问题都得到了好的回答，那说明底座与专门训练之间的分工在移动，底座承担得更多了；它仍然不说明物理发生被穷尽了。本书一直要求区分暂时语境变化、外部记忆、控制反馈、参数训练与持久学习，这里正好用得上：少量示范带来的往往是暂时语境变化，它要转成持久能力，还要经过检验与封装。

这三层分清了，「统一物理模型的困境」就有了确切的意思：**困境不在于统一的架构注定做不出来，而在于把「一个模型穷尽全部物理发生」当作目标**。这个目标背后，是把世界当作先在的、可一次性封装的对象。一旦放下这个目标，共享底座、专门训练、接口协作就各得其所：底座提供广泛的起点，专门训练形成可靠的能力，接口让不同任务世界相互调用与校验，而整个体系通过持续评估与回写，随着世界的继续发生而继续发生。

◆ 六、向第七编

本编从信息负担出发，经过预测与行动之间的缝隙，落到任务对世界发生范围的组织，用驾驶做了样本，最后说清了共享与专门的协作。到这里，「为什么要围绕任务」已经讲完；还没讲的是「具体怎样训练、怎样验证」：经验怎样围绕共同事件组织，怎样从被动记录走向主动试探，跨模态预测怎样推动纠缠重组，怎样设计可以失败的实验，可靠性怎样形成并持续封装。这些是第七编「发生式世界模型的训练与验证」的任务。

带走的话

共享底座和任务专门训练不是对立的两条路，而是同一个发生体系的两层；共享让起点更高，专门训练与接口校验让能力可靠。共享不等于穷尽，因为物理世界本来不是可以一次封装的清单。

小账　论证落在哪一层？——本章三层并用且明确分开——机制假说（底座覆盖越宽、越需舍弃专门差异）、经验事实（共享底座＋后训练的公开做法，迁移幅度标〔待核〕）、本体论立场（物理发生不可穷尽）。　世界模型补了哪一条？——补上一条结构——世界模型体系＝可复用底座＋任务专门训练＋带边界说明与失配回写的接口，并以持续评估代替「一次封装」。

第 七 编

发生式世界模型的训练与验证

编序　第六编说到：物理发生的信息负担太大，显露预测又不能保证行动可靠，所以要让任务来组织世界发生的范围。读者有权追问：那到底怎么训练？怎么知道这话对不对？本编是全书的「出路」部分，也是最该接受检验的部分。

五章依次接力。第31章讲训练经验的单位是「共同事件」：看见的、听到的、手上感到的、做了什么、结果怎样，按时间对齐成一件事。第32章讲系统要在自己拿不准的地方主动去试，并给出组织试探的六步法、九步法。第33章讲多种感官放在一起，要紧的不是多几路输入，而是改变「要预测什么」，以及预测失配时怎样找原因。第34章把本书的核心主张写成可以失败的对照实验，事先写明什么结果算本书判负。第35章讲能力怎样从「成功率」走到「在哪些条件下可靠」，以及封装后怎样持续评估、回写和解封。

两个主例：林岚团队训练机械手拧瓶盖、捏豆腐；厨房里做菜试咸淡。第六编讲的是人手怎样对付各种瓶盖，本编接着问：一只没有这份手上劲道的机械手，要怎样学出来。各处会标明哪些是本体论立场、哪些是可检验的机制假说、哪些是公开文献里的经验事实。

31

第 31 章

围绕共同事件组织训练经验

> **这一章的问题**：如果学习是特征纠缠的建立，那么喂给机器的训练经验，应当以什么为一个单位?

◆ 一、命题：训练经验的单位是「共同事件」，不是分开的数据流

先说本体论立场。按照本书的母题，学习是特征纠缠（E）的建立与训练；一块块经验缠到一起，靠的不是它们被放进同一个硬盘，而是它们在同一件事里相互约束：这一眼看到的，和这一下手上感到的、这一声听到的、这一动作造成的后果，是同一件事的不同侧面。由此推出一条训练上的主张：训练经验的基本单位，应当是一次「共同事件」——在一段时间内，观察、动作、接触、声音与后果按发生次序对齐，构成一件可以被回看、被比较、被重做的事。

这条主张本身是本体论立场的延伸，但它会在工程上落成一个可检验的机制假说：同样的数据量下，按共同事件组织的训练，比把各路数据分开收集、分开训练再拼在一起，更容易形成跨感官的对象同一性。这个假说第34章要正式拿去接受对照实验，这里先把它说清楚。

用林岚的例子。这是我构想的场景：林岚所在的团队要训练一只多指机械手拧开各种瓶盖。第26章已经讲过人怎样对付五花八门的瓶盖：防盗环、真空罐、安全盖、湿手打滑，人手靠的是一次次拧出来的劲道，再加上现场的一个小试探。那一章没有回答的是：一只没有这份劲道的机械手，它的劲道从哪里来？本编从这里接着讲，不再重复瓶盖的种类，只讲机器的经验该怎样收、怎样试、怎样验。最省事的做法是分头收：摄像头录一批拧瓶盖的视频，力传感器另录一批读数，麦克风又

录一批声音，各自打标签。这样收来的东西，数据量可以很大，但「这一帧画面里瓶盖刚开始转」和「这一刻指尖摩擦力突然下降」和「这一声轻轻的咔哒」之间的对应关系，并不在数据里。对应关系不在，模型就只能从统计上去猜它们有没有关系；而本书认为，恰恰是这种对应关系，构成了「这个瓶盖」作为一个对象的同一性。

◆ 二、机制：怎样把一件事按时间对齐

把各路信号对齐成一件事，并不只是给每条数据打上同一个时间戳那么简单。至少有四层工作。

第一层是采样节奏不同。摄像头每秒几十帧，触觉传感器往往每秒成百上千次，声音更密，关节位置和电机电流又是另一套节奏。对齐不是把快的降到和慢的一样——那会把滑移开始那几毫秒的变化抹掉——而是保留各自的节奏，同时记录它们在同一条时间轴上的位置。

第二层是延迟。不同设备从物理变化发生到信号被记录下来，各有延迟，而且延迟会漂移。林岚的团队如果发现视频里瓶盖开始转动的时刻，总比指尖切向力下降晚几十毫秒，就要先弄清这是物理上本来如此（手指先打滑、瓶盖后转），还是摄像头本身慢了。前一种是事件的结构，应当保留；后一种是记录的毛病，应当校正。把两者混淆，模型学到的「先后」就是假的。

第三层是事件的边界。一件事从哪里开始、到哪里结束？对拧瓶盖来说，可以从手指接近瓶盖开始，到瓶盖完全松脱或任务被放弃为止；中间还有几个自然的分界点：初次接触、建立握持、开始转动、打滑、换手位、松脱。这些分界点不必全靠人工标注，许多可以从信号本身的突变里找出来，但需要人来确认哪些分界对任务有意义。这里体现的是差异运行（D）的组织：一件事内部，哪些差异是要紧的差异，决定了事件被切成哪几段。

第四层最容易漏掉：记录者自己也在变。做菜的人连尝五六口，舌头对咸味会变得迟钝，后面那几口尝出来「刚好」，其实已经偏咸了。机器人的触觉皮肤也一样，用久了表面磨损、弹性变化，同样的压力读出来的数会慢慢漂。这意味着「传

感器」不是一个站在事件外面的公正记录员，它本身就在事件里。工程上的对策是定期用标准物（比如一块已知硬度的标定块）重新标定，并把标定记录也作为情境写进事件。理论上的意思更深一层：显露（S）总是在某一个具体的纠缠状态里显露，记录工具的状态是这纠缠的一部分，不能被假装成零。

◆ 三、一件事的记录长什么样

说「对齐成一件事」容易，研究者要落手，得知道一条事件记录里至少该有哪些东西。下面这张表是本书提出的一个最小清单，不是现成标准，供设计数据管线时参照。

记录项	内容	为什么不能少
时间轴	各路信号在统一时钟上的位置及估计延迟	先后关系是事件结构的一部分
观察	视觉、声音、触觉、本体感受（关节角、电机电流）等原始或压缩信号	同一件事的不同侧面
动作	实际下发的指令与实际执行的轨迹（二者常不一致）	区分「想做什么」与「做成了什么」
后果	任务层结果（拧开／未拧开／捏碎）与物理层结果（形变、位移）	让显露有站稳或失效的依据
分段	接触、握持、转动、打滑、松脱等分界点及其来源（自动／人工）	标出要紧的差异
缺失与来源	哪路信号缺、哪路是仿真值、哪路是推测值	让模型知道自己看见了什么
情境	物体编号、材料、温度湿度、手上是否沾水等可记录条件	为后面的适用边界留证据

表里有两项最容易被省掉。一是「实际执行的轨迹」。机械手接到「顺时针转二十度」的指令，实际转了多少，取决于当时的阻力；如果只记指令，模型会以为每次都转了二十度，阻力的信息就丢了。二是「情境」。同一只瓶子，手上沾了水和没沾水是两件不同的事；今天拧开了，明天湿度大了没拧开，若没有记录情境，

这两次就只能被当成噪声。第35章讲适用条件时，依靠的正是这一栏日积月累的记录。

这张表也提醒一件常被忽略的事：事件怎么切、哪一栏算要紧、哪些情境值得记，都是人定的。林岚和她的同事决定「打滑」算一个分界点，这个决定本身来自无数人拧瓶盖的经验，来自机器人学多年积累的接触力学知识。训练的发生单位不是孤零零的一台机器，而是工程师、数据、工具与模型共同构成的集体人类—AI复合体。说清这一点，不是为了谦虚，而是为了在出错时知道去哪里找原因：模型学偏了，有时是参数的问题，有时是当初切事件的人把一处要紧的差异切掉了。

◆ 四、经验来源与缺失信号

共同事件要从哪里来？常见来源大致有四种：人通过遥操作带着机械手去做；从人做事的视频里学；在仿真里大量生成；让机器自己去做并记录。四种来源各有缺口，而且缺口的位置不同。

人类视频里没有力。看一段人拧瓶盖的视频，看得到手形、看得到瓶盖转了，看不到指尖用了多大劲、什么时候开始打滑。遥操作数据里有机械手自己的力和位置，但操作者手上那一瞬的感觉——「有点涩，要再加点劲」——通常没有被记录，除非遥操作设备本身有力反馈并被一同记下。仿真里什么都能读出来，但接触、摩擦和软物体形变的仿真精度有限，读出来的力未必是真的力。机器自己做的数据最完整，但成本高、慢，而且一开始做得很差。

本章主张的做法是：把缺失当作事件的一部分标记出来，而不是悄悄补齐。一条来自人类视频的拧瓶盖事件，应当明确标注「此事件无接触信号」；一条仿真事件，应当标注「接触信号为仿真值，可信区间待定」。模型在训练时要知道自己哪里是看见了、哪里是没看见。这样做的理由仍然回到三问里的那一问：显露出来的东西靠什么站稳？如果把缺失的力用猜测值补上而不加标记，模型学到的「站稳」，有一部分是站在猜测上的。

四种来源混在一起训练时，还有一个比例问题。人类视频便宜而多，机器自采贵而少；如果按条数平均混合，模型的「世界」会被视频的样子主导，接触那一面

被稀释。一个可检验的做法是：在评估中分别报告模型在「有接触信号的事件」和「无接触信号的事件」上的表现，看两者是否此消彼长。这样至少能看出，增加便宜数据是在扩展能力，还是在冲淡能力。

◆ 五、成功与失败要一起进入训练

许多训练数据只收成功的示范：只有拧开的过程，没有拧不开的过程。原因可以理解——失败的示范「教坏了」怎么办？但按本书的思路，失败过程里有成功过程不能提供的东西：边界。

林岚的团队会遇到几种典型的失败：指尖打滑，瓶盖纹丝不动；握得太紧，薄塑料瓶被捏变形，里面的水溢出来；斜着拧上去，螺纹咬错；换成捏豆腐的任务，则是一下捏碎。每一种失败都在告诉系统：显露（S）在哪里不再站得住了。打滑那一刻，视觉上几乎看不出变化，触觉上切向力的曲线却变了形；捏碎豆腐之前，形变在加速，阻力反而先变小。只看成功的示范，模型学到的是一条窄窄的通道；把失败也收进来，它才有机会学到通道两边的墙在哪里。

这里要分清两件事。失败进入训练，不等于失败被当作要模仿的目标。工程上可以把失败事件用于学习后果预测（「这样做会捏碎」），而不是用于学习动作（「要这样做」）。本书在这里只提出机制假说：含失败事件的训练，在失配恢复上应优于只含成功事件的训练。这条也进第34章的指标。

失败数据怎样收，也有讲究。让机械手故意去捏碎一百块豆腐，代价不大；让它故意把一百只装满热水的瓶子捏变形，就不行了。所以失败的采集要分级：便宜且安全的失败可以主动制造；昂贵或危险的失败，只能从自然发生的失败里收，或者在仿真里制造并标明来源。还有一种很有价值的「准失败」——快要打滑但被及时纠正过来的那几次。它们离墙最近，又没有撞上去，最能告诉模型墙的位置。林岚的团队如果在每次事件里都自动检出「切向力与法向力之比接近打滑阈值」的片段并加以保留，就是在专门收集这类经验。

◆ 六、任务目标怎样引导经验采集

共同事件不可能无限地收。收什么、在哪里多收，要由任务来引导。这一点第28章已经说过：任务决定需要保留的差异、粒度和时空范围。这里把它落到采集上。

用做菜试咸淡来看。一个人炒一盘青菜，不会每一秒都尝。他在几个要紧的时刻尝：放盐之后、出锅之前，或者加了酱油、发现颜色比平时深的时候。他的「采样」是被任务引导的——做一盘咸淡合适的菜——也是被预期与现状之间的差距引导的。颜色深了，他心里有一个预测「可能偏咸」，于是在这个时刻多尝一口。这是共同事件组织的一个日常版本：看见的颜色、闻到的气味、尝到的味道、刚才放了几勺盐，在一次「试咸淡」里被对齐成一件事。

机器人那边，任务引导的采集意味着：拧瓶盖时，初次接触到开始转动这一段，以及打滑前后，是信息最密的地方，要用高采样率、要多收样本；手臂在空中移动接近瓶子那一段，可以收得稀一些。捏豆腐时，从接触到形变加速的那几百毫秒最要紧。任务越清楚，采集就越有重点。

但任务引导也不能做成百分之百。这里有一个可操作的建议：在采集预算里留出固定的一小份，比如一成到两成，用于「不按任务重点」的采集——换不常见的瓶子，换奇怪的握法，在手臂空中移动的那一段也偶尔高采样。这一份数据短期内看不出用处，它的作用是给后来的任务留余地：等到有一天任务从「拧瓶盖」扩展成「拧瓶盖之后把瓶子递给人」，原来被当作无关的那一段手臂运动就变得要紧了。做菜的人也有类似的习惯：偶尔在不该尝的时候尝一口，结果发现汤底本身就带咸，原来一直放的盐都多了。这一口不是任务计划里的，却改写了以后的计划。把这一份比例定多少，本书不给答案，只建议把它当作一个实验变量，在第34章的对照里一起测。

◆ 七、边界与向下一章推进

这一章的主张有几条边界要说在前面。

第一，「共同事件」不是什么神奇的数据格式。它对工程界并不陌生，许多机器人数据集本来就同步记录多路信号〔待核：具体哪些公开机器人数据集同时含视觉、触觉、声音与动作并统一时间轴〕。本书的主张不在于「要同步」，而在于把同步上升为训练经验的单位，并把缺失、失败、事件边界当作事件的组成部分。

第二，共同事件组织得再好，它仍然是被动记录。系统只能从已经发生的事里学；而第27章已经说过，已有记录中存在不可区分的配置——两个瓶盖看起来一样，录像里的拧法也一样，一个松一个紧，要等真拧下去才知道。被动记录给不出这个区分。

第三，任务引导采集有代价：任务定义得太窄，收来的经验就会把世界剪得太小，换一个瓶盖材料就不认识了。这个代价要在后面的评估里显式地量出来。

第四，事件与事件之间并不互相独立。第22章讲过，接触历史会影响后续发生，第26章也提到果酱瓶的盖子会被螺纹里的糖分粘住：同一只瓶盖，第一次拧时螺纹里有糖渍，拧过一次再拧上，第二次就松多了；一块豆腐被捏过一下，表面已经有了细小裂纹，第二次轻轻一碰就碎。如果把每次事件当作互不相干的样本随机打乱来训练，这种「上一件事留下的痕迹」就会被当作噪声抹平。所以事件记录里最好还有一栏「前史」：这个物体此前经历过哪些事件。这一栏不难记，却常常没人记。它是对象同一性跨越时间的那一半——同一只瓶盖之所以是同一只，不只因为它此刻在各个感官里对得上，还因为它带着自己的经历。

所以下一步顺理成章：系统不能只等经验送上门，要在自己拿不准的地方主动去试。这就是第32章的内容。

> **带走的话**
>
> 训练经验应以「共同事件」为单位——观察、动作、接触、声音与后果按发生次序对齐，缺失要标明，失败要收进来，采集由任务引导。对应关系本身，就是对象同一性的原材料。

> **小账** 论证落在哪一层？——本章以机制假说为主（共同事件组织优于拼接、含失败事件利于失配恢复），以本体论立场为出发点（纠缠靠同一件事中的相互约束建立）。 世界模型补了哪一条？——补上「训练经验单位」这一条件——世界模型的发生，需要按共同事件组织的、标明缺失的、含失败的经验。

32

第 32 章

从被动记录到主动试探

> **这一章的问题**：机器什么时候该停下来不再看录像，而是自己伸手去试一下？试什么、按什么次序试？

◆ 一、命题：区分发生配置，要靠干预

第27章说过，已有记录中有些配置是不可区分的，那里的例子是两只外观一样、一只抽过真空的玻璃罐。放到林岚的机械手面前，问题一模一样：两只同款矿泉水瓶，一只瓶盖拧得松，一只拧得死，录像拍得再清楚也分不出来，只有上手拧一下。人是凭手上那一下阻力分开的；本章要问的是，机器怎样学会挑这一下、在哪里挑。这不是数据不够多的问题——把同样的录像再收一万段，还是分不出来——而是被动记录本身的边界：它只记得已经发生的事，而这两个配置的差别，要在一个尚未发生的动作下才显露。

本体论上，这对应全书的一个基本看法：显露（S）不是物体自带、等人来拍的一张照片，而是在特征纠缠（E）中、经过差异运行（D）才显露出来的。没有「拧」这个差异运行，「松」和「紧」就没有机会显露。由此得到本章的命题：发生式世界模型的训练，必须包括主动试探——系统在自己拿不准的地方，选择能区分不同发生配置的动作去做，并把后果收回来。

这条命题有一个日常版本，每个人都熟。做菜试咸淡，从来不是看一眼就算。汤的颜色、放盐的勺数，只能让你猜个大概；真正分出「偏淡」「刚好」「偏咸」的，是舀一小勺尝一下。尝，就是干预。它的代价很小（一小勺），却能区分锅里那几种看起来一样的状态。

◆ 二、机制：用预测分歧选择行动

问题是：什么时候尝？尝哪里？对机器来说，就是：在千万种可能的动作里，选哪一个去试？

一个有用的机制是「预测分歧」。让系统对同一个动作的后果做几种预测——可以是几个独立训练的预测器，也可以是同一个模型给出的几种可能后果。如果几种预测大体一致，说明系统对这个动作「心里有数」，去试的收获不大；如果几种预测分歧很大——一种说瓶盖会转，一种说手指会打滑，一种说瓶身会被捏扁——说明这正是系统拿不准的地方，去试一下最能增长见识。用分歧来选择探索方向，在强化学习和世界模型的研究里已有公开工作：2019 年有研究用一组预测器之间的分歧作为自我驱动探索的信号；2020 年的 Plan2Explore 把这一做法放进世界模型，让智能体在内部想象中寻找分歧大的地方，再去真实环境里探索；本书在这里不是提出新算法，而是给这一做法一个发生学的读法：分歧之处，正是纠缠尚未定形的地方，也就是混沌态与介生态交界的地方。

但分歧不等于值得试。分歧大的地方有两种：一种是「试了就能分清」，比如这只瓶盖松还是紧；另一种是「试多少次也分不清」，因为结果本来就有很大的随机性，比如豆腐碎开时裂纹往哪个方向走。前一种叫「可消除的不确定」，后一种叫「固有的随机」，机器学习里分别称为认知不确定性与偶然不确定性。用分歧选动作时，要尽量把两者分开，否则系统会被那些天生随机的地方吸住，一遍遍地去捏豆腐看裂纹，白白花掉探索预算。工程上常用模型集成、贝叶斯神经网络等办法估计前一种，并让模型直接输出结果的散布来表示后一种；这些办法在接触密集的操作任务上分得有多准，还要具体去测。本书只强调：这个区分要进入试探的选择规则，而不能事后再补。

◆ 三、通过干预区分发生配置

主动试探最有价值的地方，是设计「区分性动作」：一个动作，在配置甲下会产生一种后果，在配置乙下会产生另一种后果。

林岚的团队可以这样做（这是我构想的场景）。机械手面前有一只没见过的瓶子，瓶盖是光滑的金属盖，系统对它的摩擦系数和拧紧程度都拿不准。直接用平常的力度去拧，可能有三种结果：拧开、打滑、把瓶子带着一起转。三种结果混在一起，分不清是哪一个因素在起作用。更好的办法是先做两个小试探：第一，轻轻压住瓶盖，沿切向慢慢加力，直到出现微小滑动，记下此时切向力与法向力之比——这粗略告诉系统摩擦有多大；第二，一只手扶稳瓶身，另一只手以很小的扭矩试转，看瓶盖动不动——这告诉系统拧得多紧。两个试探各自隔离了一个因素。之后再正式去拧，失败的话，原因已经缩小到很窄的范围。这和第26章装修师傅先在不显眼处打个浅孔、听声看粉是同一个道理，区别只在于：师傅的试探是多年手艺自动带出来的，机器的试探要被明确设计、记录和计账。

这背后有一个一般原则：好的试探动作，是让两个或几个可能的发生配置在后果上拉开距离的动作。同一个原则在厨房里也成立。有经验的人试咸淡时，不只是尝汤，有时还会单独尝一口菜叶——汤咸而菜淡，说明盐还没吃进去，要再等一等；汤和菜都咸，才是真的放多了。尝汤和尝菜，是两个区分性动作。

区分性动作还有一个常被忽视的问题：什么时候该停。试探可以无穷无尽——再轻压一次，摩擦估计会更准一点；再试转一次，松紧判断会更稳一点。但任务不需要无穷的精确。拧瓶盖只要知道「够不够我用中等力度拧开」，不需要把摩擦系数测到小数点后三位。所以停止规则应当由任务来定：当剩下的几种可能配置在「该采取什么动作」上已经不再分歧时，就不必再试了——哪怕它们在物理参数上仍有差别。做菜的人尝到「差不多，再少许一点盐」就停，不会去追究这锅汤的含盐量究竟是千分之八还是千分之九。这条停止规则也是可以检验的：如果它定得太早，任务失败率会上升；定得太晚，交互成本会上升。在两者之间找平衡点，本身就是一项需要实验来确定的工作，而且对不同任务、不同对象，平衡点很可能不一样。

◆ 四、六步法与九步法：组织训练的两个方法框架

以下两个框架是本书提出的机制假说，用来组织主动试探的训练，不是已被证实的工程规律。它们的价值在于给出可以检验的组织方式，若实验不支持，应当修改或放弃。

六步法对应六路径，即三维的六种次序：D→E→S、E→D→S、D→S→E、S→D→E、E→S→D、S→E→D。本书把每一种次序读成一类训练动作。D→E→S是「先动手」：换一种动作（差异），看它怎样改变系统内部已有的纠缠，再看显露出什么——林岚让机械手换一种握法去拧，看模型对这只瓶盖的「理解」怎样被改写。E→D→S 是「以旧推新」：从已经建立的纠缠出发（这像一只常见的塑料瓶），推出应当怎样动作，再看显露是否符合。D→S→E 是「先看后果再回写」：做一个动作，直接记录后果，再把后果写回纠缠，这是最常见的试错。S→D→E 是「从异常追因」：先看到一个反常的显露（瓶身意外变形），追问是哪一处差异引起的，再重组纠缠。E→S→D 是「先想象再动」：从纠缠出发在内部生成预期后果，再据此规划动作，这与 Dreamer 一类方法在内部想象中学习策略的做法有可比之处（第18章讲过其公开机制）。S→E→D 是「先认再做」：从眼前的显露认出它属于哪类对象，调出相应的纠缠，再决定动作。

六步法的主张是：一个训练过程如果长期只走其中一两条路径——比如只走D→S→E 的试错，或只走 E→S→D 的内部想象——它形成的纠缠会有系统性的缝隙；有意识地让六种次序都得到训练，应当比只用其中一两种更有利于对象同一性的形成和失配后的恢复。这是一个可以检验的假说：可以统计一段训练中六类动作各占多少比例，再看比例与第34章那些指标之间的关系。

九步法对应三维九宫。九宫不是本章自拟的分层，而是 SDE 的定式：S 维三格，S1 规范（对比、变化、分布）、S2 模态（粒子、波、场）、S3 价值（真、善、美）；D 维三格，D1 意义目标（创造、自由、幸福）、D2 路径组织（六步、九步）、D3 优化约束（误差、冗余、损耗）；E 维三格，E1 三界（理念、现实、自我）、E2 信息三模态（符号、逻辑、数学）、E3 能量三状态（内能、动能、势

能）。九格按编号连成三个平行的三角：S1·D1·E1、S2·D2·E2、S3·D3·E3。六步法与九步法自己就落在 D2 这一格里，它们是差异怎样组织成路径的办法。要提醒一句：S2 的粒子、波、场说的是显露的样态，不与看、听、摸一一对应，第五编已经立过这条界线。

本书组织训练，以 S1·D1·E1 三角为定式。这三格是三合一的，有它们自己的三方程、六路径、三原理。E1 回答这一轮训练扎在哪一界的土里：理念界是符号与推演，现实界是接触的后果，自我界是这台机器自己的身体与它在一个具体复合体里的使用历史。D1 回答这一轮差异朝哪里运行：创造，是发生原来没有的区分，比如第一次把「看着一样、软硬不同」的两块豆腐分开；自由，是这份区分能在更多情境里被调用，而不是只在训练台上灵；幸福，是任务对承接它的人真正完成、可以放心托付。S1 回答显露怎样接受检验：对比，同一外观、不同后果分不分得开；变化，干预前后的差异有没有被预先说出来；分布，在多少对象、多少情境里站得住。三格之间也按三方程互相改写：S1＝F(D1,E1)，显露站不站得稳，取决于差异朝哪里跑、扎在哪一界；D1＝G(S1,E1)，方向从检验结果与三界的土里调出来；E1＝H(S1,D1)，哪一界的纠缠被加固，由检验与方向共同决定。按三原理，当 S1 与 E1 对不上——规范说「这是北豆腐那一类」，现实界的后果却是碎了——两者的张力就推动 D1 改向：从「照旧完成」转向「创造一个新区分」。

九步法的用法，是把九格当作一张训练与诊断的检查单，次序从定式三角起：第一到第三步过 S1·D1·E1，问这一轮在哪一界发生、朝哪一个意义目标运行、显露经不经得起对比、变化、分布三种检验；第四到第六步过 S2·D2·E2，问这次失配显露为一个离散的事件、一段连续的起伏还是一片持续的影响，本轮走了六路中的哪几路，学到的东西能不能写成日志、规则或参数这样可传递的形式；第七到第九步过 S3·D3·E3，问预测与后果是否相符、动作是否不伤人不毁物、结果是否被人愿意接受，误差、冗余、损耗各花在哪里，已积累的纠缠储备、正在运行的互动和尚未调用的可迁移能力各有多少。每一轮之后看九格里哪几格得到了改写，哪几格没有动。九步法的假说是：按九格检查、先补定式三角里的空格再安排其余试探，应当比不分格地增加数据更省数据。

要让这两个框架可被检验，最关键的是给每一步一个操作定义，让不同的人看同一段训练日志，能把同一个训练动作归到同一类。本书建议的判法是看「这一步从哪里起手、改动落在哪里」。起手于改变动作参数、改动最先体现在执行轨迹上的，记为 D 起手；起手于调用已有的对象类别或已有模型的内部状态的，记为 E 起手；起手于一个新观察到的显露、特别是一个未被预测到的显露的，记为 S 起手。第二步、第三步依此类推。这样的判法当然会有模糊地带，所以建议两人独立标注同一批日志，报告一致率；一致率太低，说明操作定义还要改，而不是框架已被证实或推翻。

下面用捏豆腐走一轮，看两个框架怎样配合（这是我构想的场景）。林岚的系统已经能稳稳夹起北豆腐，换成嫩豆腐，第一次就捏碎了。先过定式三角：E1，事情发生在现实界，后果是碎了，可这个后果还没写回纠缠；S1，对比这一项没过，系统把嫩豆腐当成了「北豆腐那一类」，调出的是北豆腐的力度；D1，这一轮需要的是创造一个新区分，而不是把旧力度用得更熟。缝隙定位在 S1 与 E1 的张力上。于是下一轮的试探从 S 起手：拿那次捏碎的显露做起点，走 S→D→E，回放事件，找出形变开始加速的那一刻对应的力值，再在嫩豆腐上做一组从极轻到稍轻的按压（D），把「同样外观、不同软硬」写进纠缠（E）。接着走 E→S→D：用新纠缠预测下一块嫩豆腐在某一力度下的形变，再据预测去夹。若预测与实际吻合，S1 的对比一项算过；再换几块不同产地的嫩豆腐去夹，看分布一项站不站得住。后面六步顺手过一遍：每捏碎一块就是一份 D3 的损耗，所以按压从极轻起步；新区分若能迁到捏鸡蛋、托草莓，就是 E3 里多出来的一份势能。整个过程留下的日志，正好可以拿来统计六种次序的比例，以及九格中哪几格被改写了几次。

六步法管「次序」，九步法管「格位」；前者回答下一步该从哪一维起手，后者回答这一轮在九宫里该补哪一格，而且先补 S1·D1·E1 这一角。两者同属 D2，都不是要求每次训练机械地走满六步、九步，而是提供一套可以被记录、被统计、被检验的组织语言。

◆ 五、探索成本、行动边界与数据效率

主动试探不是免费的。林岚的机械手每试一次都要时间，会磨损指尖，可能摔瓶子，可能捏碎豆腐。真实世界里的探索成本，远高于在录像上多训练一轮。所以主动试探必须有边界。

第一类边界是安全边界：有些试探永远不该做，比如对装着热水的瓶子用大力试捏。这些边界应当在试探之前由人设定，并在系统里成为硬约束，不交给模型自己去「学」。第二类边界是成本边界：一次试探带来的预期信息量，应当和它的成本相称。用分歧选动作时，可以把成本写进选择规则：同样分歧大小的两个动作，选便宜的那个。第三类边界是可逆性：优先试那些做了还能撤回的动作——轻压一下可以松开，捏碎了就回不去。做菜的人也懂这一点：盐宁可少放、尝了再加，因为淡了能补，咸了难救。

数据效率是主动试探最该交出的成绩单。如果主动试探真如本书所说能区分被动记录区分不了的配置，那么在同样的真实交互次数下，它应当更快达到同样的任务能力，或者在同样能力下需要更少的交互。这是可以测的。反过来，如果实验发现：主动试探的数据效率与随机采集差不多，或者只在很窄的任务上有优势，那么本章的主张就要收缩。

衡量数据效率，要先说清楚「一次交互」算什么。一次完整的拧瓶盖尝试算一次？还是一次轻压试探也算一次？本书建议按真实世界的耗时和磨损来折算，而不是按次数：一次两秒的轻压与一次二十秒的完整拧开，代价不同。否则主动试探把一次大试拆成许多小试，就会在「次数」上显得很省，实际并不省。另外，仿真里的试探与真机上的试探要分开计账；在仿真里试一万次几乎不花钱，但仿真本身的失配会带来另一种成本，这一点第33章讲失配归因时还会碰到。

◆ 六、边界与向下一章推进

本章要防止两种误读。一种是把「主动」理解成「机器自己决定一切」。主动试探的选择规则、安全边界、成本函数，都是人参与设定的；训练的发生单位仍是

集体人类—AI复合体，机器的「主动」是在这个复合体里分到的一份。另一种误读是以为现有世界模型都没有互动。并非如此：强化学习路线里的世界模型本来就在与环境互动中学习，V-JEPA 2 也有用机器人交互数据做动作条件后训练的公开工作（第18章已讲）。本章的主张不是「互动从无到有」，而是互动要被组织成区分发生配置的试探，并按六种次序、九个格子来安排与检查。

还要承认一个反例：并非所有任务都需要主动试探。如果任务面对的对象高度标准化——流水线上同一型号、同一批次的瓶子，拧紧力矩由机器统一设定——那么被动记录里几乎没有不可区分的配置，多试反而浪费。主动试探的价值，随对象与情境的多样性上升而上升。这本身就是一个可检验的推论：在对象越标准化的任务上，主动试探相对被动采集的优势应当越小；若实验发现两者的优势与多样性无关，说明本章对「为什么要试」的解释有问题。

主动试探收回来的，不只是更多数据，而是多路信号在同一个干预下的共同变化。下一章要追问：多路信号放进同一个模型，究竟改变了什么——是多了输入，还是改变了要预测的东西？

> **带走的话**
>
> 被动记录分不清的配置，要靠干预来分；用预测分歧选择试探，用六步法安排次序、九步法检查格位，并让探索服从安全、成本与可逆的边界。

小账 论证落在哪一层？——本章以机制假说为主（预测分歧引导试探、六步法与九步法的训练组织均为本书提出的假说），本体论立场为依据（显露在差异运行中发生），经验事实只作对照。 世界模型补了哪一条？——补上「主动干预」这一条件——世界模型要能区分被动记录中不可区分的发生配置，并给出组织试探的可检验框架。

33

第 33 章

跨模态预测与纠缠重组

> **这一章的问题**：给模型多接几路感官信号，和让模型学会「看、听、摸是同一件事」，差别在哪里?

◆ 一、命题：关键不在增加输入，而在改变预测对象

先把两种做法分开。第一种：模型原本只看画面，现在给它再接上声音和触觉，作为额外输入，要预测的东西不变——比如仍然只预测下一帧画面，或只预测下一个动作。第二种：要预测的东西本身变了——模型要同时预测下一刻的画面会怎样、声音会怎样、指尖的力会怎样，而且三者要彼此对得上。

本书的机制假说是：第一种做法主要改善的是「多一路线索」，第二种做法才迫使不同感官在模型内部缠在一起，形成指向同一对象的特征纠缠。理由并不玄：如果触觉只是输入，模型完全可以学会在多数情况下忽略它，只在画面信息不够时偶尔借用一下；可如果触觉的变化本身是要被预测的对象，模型就必须在内部建立「这种画面变化对应那种力的变化」的关系，否则预测就会错。**预测对象决定了误差从哪里来，误差从哪里来，决定了纠缠往哪里长**。

这条假说与第五编的立场一致但不等同。第五编说，视觉、听觉、触觉各自是完整的 SDE，三者在互动中形成特征纠缠聚合体；本章说的是，在工程训练中，「联合预测」是促成这种聚合的一种具体手段。用第24章的三层来说：同步采集只保证了「同步出现」；只把多路当输入，模型最多学到「相关变化」；要让一路的显露去限制另一路可能显露的东西，也就是走到「相互制约」，预测对象本身就得

包含各路的变化，并且要求它们彼此对得上。手段可以被实验检验，立场不因某次实验的成败而直接被证明或推翻——这是三层论证纪律的要求。

◆ 二、联合预测：让三路变化彼此约束

用林岚的拧瓶盖来具体化（这是我构想的场景）。设想她的团队训练一个模型，给它过去半秒的画面、声音、指尖触觉与关节状态，再给它接下来准备执行的动作，要它预测接下来两百毫秒里三件事：瓶盖在画面里转了多少；会不会出现螺纹脱开那一声轻响；五个指尖的法向力与切向力怎样变化。

这样训练出来的模型，内部被迫学会几种对应：切向力突然下降而画面上瓶盖不动，意味着打滑；切向力平稳下降而瓶盖开始转，意味着拧开了；在拧开前的某一刻，往往伴随一声轻响和力的小跳变〔这一声学—力学对应在不同瓶盖上是否稳定，需实验确认〕。这些对应不是工程师一条条写进去的，而是在联合预测的误差压力下长出来的。

联合预测还有一个附带的好处：它让模型可以用一路去「审」另一路。如果画面预测说瓶盖转了，触觉预测却说手指在打滑，两个预测互相矛盾，说明模型对眼下这件事拿不准——这正好接上第32章的预测分歧，可以用来选择下一步试探。用三原理的话说：两维之间的矛盾积累张力，推动第三维改变——这里是视觉与触觉两路预测的矛盾，推动系统采取新的差异运行（换个动作去试），进而回写纠缠。

联合预测落到工程上，马上会碰到三个具体选择，每个选择都会影响纠缠往哪里长。

第一，预测到什么粒度。可以预测原始像素与原始力值，也可以预测压缩后的特征。第18章讲过，V-JEPA 一路的思路是在特征空间里做预测，而不去逐像素重建画面。这个选择在联合预测里同样要做：如果要求逐像素重建画面，而触觉只要求预测几个力值，模型的大部分力气就会花在画面的纹理、光照这些和拧瓶盖关系不大的细节上；如果三路都在特征空间里预测，就需要另外检查这些特征有没有把要紧的差异（比如打滑开始的那一瞬）压掉。哪种粒度更好，本书不预设，只提醒：粒度是由任务决定该保留哪些差异的，这一点与第28章一致。

第二，各路误差怎样加权。三路信号的数值尺度不同、可预测程度不同，若简单相加，往往是最大最吵的那一路主导训练，最要紧却最细微的那一路被淹没。拧瓶盖时，打滑前后切向力的变化可能只有一点点，却比整幅画面的变化更能说明发生了什么。加权方式可以随训练动态调整，比如让近期预测得最差的那一路暂时获得更大权重；但无论怎样调，都要把权重的变化记下来，否则事后无法判断模型的进步来自哪一路。

第三，预测多远。只预测下一瞬间，模型容易学成「照抄当前状态」，因为下一瞬间和这一瞬间差别很小；预测得太远，误差累积，又学不到东西。对拧瓶盖来说，本书建议至少包含一个跨越「事件分界点」的预测窗口——比如从「建立握持」预测到「开始转动」之后——因为跨界点的变化，正是不同感官的对应关系最集中显露的地方。

捏豆腐的任务把这三点放在一起看得很清楚。豆腐被捏的过程中，画面上的形变一开始很小，触觉上的阻力却已经在变；临近碎裂时，画面上的形变突然加速，而阻力反而先松了一下。要让模型学到「阻力先松是碎裂的前兆」，就得用特征而不是像素来描述形变，给触觉误差足够的权重，并让预测窗口跨过碎裂这个分界点。三个选择任何一个做错，这条对应就可能学不出来。

◆ 三、失配归因：错了，是哪里错了

联合预测必然会出错。真正考验一个训练体系的，不是错得少，而是错了之后能不能找到原因并改对地方。本书把这叫做「失配归因」。

一次预测失配，可能源自几个不同的地方。一是信号本身出了问题：触觉传感器沾了水，读数漂了。二是对象认错了：系统把这只金属盖当成了它见过的塑料盖，调出了错的纠缠。三是对象没认错，但这只瓶盖有一处它没见过的特性：螺纹里有糖渍。四是动作没按指令执行：电机在阻力下没转到位。五是事件本身固有的随机：同一个动作，结果本来就会有一点散布。

这五种原因要求五种不同的调整。信号问题要标定或剔除，不该去改模型；对象认错要改关于这个对象的纠缠，让系统学会区分；新特性要扩展纠缠，把「有糖

渍的瓶盖」作为一种新配置收进来；执行偏差要改控制或把实际轨迹记准；固有随机则什么都不该改，只需要让预测表达出散布。把五种原因混成一种，用同一个梯度去更新全部参数，就可能出现「为了解释一次传感器漂移，把关于瓶盖的理解改歪了」的情况。

做菜试咸淡里也有同样的结构。尝了一口觉得淡，可能是真淡，也可能是刚喝了一口咸汤、舌头麻了，也可能是盐还没化开。老手不会一觉得淡就加盐，他会先想一想：刚才吃了什么？盐放了多久？必要时漱口再尝，或者搅一搅过一会儿再尝。这就是失配归因：先问错在哪一层，再决定改哪一层。

工程上，失配归因可以部分借助前两章的工作：事件记录里有情境与缺失标注，可以先排除信号问题；按九格分开检查（S1 的对比、E1 的现实后果、D3 的误差），可以看失配集中在哪一格；多路联合预测，可以看是哪一路先偏离。本书提出的机制假说是：带显式失配归因的更新，应当比不加区分的整体更新，在失配恢复速度和对已有能力的保持上都更好。「对已有能力的保持」要特别测：学会了有糖渍的瓶盖，原来的干净瓶盖还拧得好不好？

◆ 四、表示调整：归并、拆分与重建

找到原因之后，纠缠的调整大致有三种方向：归并、拆分、重建。这三种重组第24章用陌生弹性球已经讲透，这里只看它们落到机械手上是什么样：两种瓶盖在看、听、摸三方面的变化规律几乎一样，就并成一类，互借经验；嫩豆腐和北豆腐原来算一类，捏碎了才知道得拆开；原来按「材质」分瓶盖，后来发现真正决定拧法的是「螺纹圈数和密封圈的松紧」，就得整个换一种分法。

这三种调整，在模型内部未必对应清清楚楚的模块操作。本书不预设现有神经网络内部存在可直接读出的「对象」单元；归并、拆分、重建在这里是对行为与预测模式变化的描述，至于在参数层面怎样实现，是一个开放问题〔需要可解释性研究支持，暂无定论〕。可以检验的是行为层面的后果：归并之后，两类对象之间的迁移应当变好；拆分之后，原来混淆的两类在预测上应当分开；重建之后，一批原来分散的失败应当一起消失。

重组有一个现实的风险：改了这里，坏了那里。系统为了学会嫩豆腐，把「豆腐」这一类拆成两类，结果原来夹北豆腐的力度也变得犹犹豫豫，夹起来更慢、更容易掉。这在机器学习里被看作「灾难性遗忘」的一种温和形式；这个问题早在上世纪八十年代末就有人指出，后来的缓解办法主要有回放旧经验、对旧任务要紧的参数加约束等。从发生学上看，它说明纠缠的重组不是在空白处添一块，而是在一张已经缠好的网上动刀，每一刀都可能牵动别处。工程上可行的办法有几种：保留一部分旧事件，在学新东西时一起回放；每次重组后用一套固定的旧任务检查集复测；对改动幅度设上限，宁可分几轮小改。不管用哪一种，原则是同一个：重组之后，要同时报告「新的学会了多少」和「旧的丢了多少」，只报前者等于只看了一半账。

声音在这里的位置值得单独说一句。在拧瓶盖和捏豆腐这两个任务里，声音提供的信息未必多，但它常常在恰好的时刻出现：拧开前的轻响，豆腐碎裂时极轻的一声。这些声音对重组的价值，不在于它们单独能告诉系统什么，而在于它们给视觉与触觉的对应提供了第三个锚点。是否真的如此，要看实验：去掉声音一路，归并与拆分的准确性是否下降。若不下降，就应当如实承认，在这类任务上声音可以省去——本书不预设三种感官对所有任务都不可替代。

◆ 五、局部研究的启示：DexTacWAM 与 TouchWorld

下面谈两项公开的局部研究。它们是经验事实层的对话对象，不是本书主张的证明；以下内容依据其 arXiv 公开摘要，全文细节以原文为准。

DexTacWAM（arXiv 2609.24976）是一个面向灵巧操作的视觉—触觉「世界—动作模型」。按摘要的说法，它把接触的演化作为被预测的世界状态的一部分，在视频扩散世界模型中联合预测视觉与触觉的未来状态，并输出动作；每根手指的触觉信号先分别编码，再经一个考虑手指与姿态的压缩器合成触觉隐表示，注入视频模型。作者在一台 22 自由度的双手机器人上做了六个接触密集的灵巧操作任务，摘要报告它在每个任务上都得分最高，平均 70.6，最强基线为 38.0；并报

告：保持同样的触觉特征与动作模块，只去掉「触觉世界建模」这一部分，四个任务的平均成绩从 74.7 降到 26.6〔数字依 arXiv 摘要，评分口径与实验细节待核〕。

TouchWorld（arXiv 2607.07287）提出一个「预测—反应」式的触觉基础模型。摘要写道，视觉与语言能提供语义与几何上的引导，却难以可靠揭示力、滑移、接触稳定性这类隐藏的接触状态。它采用分层策略：高层用视觉—语言做子任务规划并预测触觉子目标；中层是视觉—触觉的目标条件策略，生成名义动作片段；底层用最近的触觉与本体感受反馈做高频的残差修正。摘要报告，在六个长时程、接触密集的灵巧操作任务上，干净设置下成功率为 65.0%，有人为扰动时为 53.7%，分别比最强基线高 15.7 和 18.5 个百分点〔数字依 arXiv 摘要，已核；实验设置细节待核〕。

这两项工作对本章有三点启示，也有清楚的边界。

启示之一：两者都把触觉放进了「要预测的东西」里，而不只是当作输入。DexTacWAM 预测接触演化，TouchWorld 预测触觉子目标。DexTacWAM 摘要里那个消融——同样的触觉特征，只是不再让模型预测触觉——成绩明显下降，与本章「改变预测对象比增加输入更要紧」的假说方向一致。但要说清楚：方向一致不等于证明，那是特定平台、特定任务上的结果，本书不能把它推成一般结论。

启示之二：TouchWorld 把触觉同时用作「预测的参照」和「快速反馈信号」，并把慢的任务推理与快的接触修正分在不同层。这对应本书反复强调的一个区分：控制反馈与参数训练不是一回事。底层的残差修正是在执行中即时纠偏，不等于模型学到了新东西；它会不会进一步变成持久学习，取决于这些修正经验是否被收回去更新模型；摘要里没有提到这一步〔全文是否有此回写机制待核〕。第35章还要专门讲这几层的分别。

启示之三：两项工作都是局部的——限于特定的手、特定的任务集、视觉与触觉两路（声音在摘要中未见作为主要信号）。它们说明「接触演化预测与局部操作」这条路走得通，并不说明三种感官的联合预测已经被做成，更不说明对象同一

性已经在跨任务的意义上形成。本书在这里能说的，只是「这些局部结果与本书的机制假说不冲突，并为第34章的实验设计提供了可借鉴的消融方式」。

◆ 六、边界与向下一章推进

本章的主张有三处可能失败。第一，可能联合预测并不比「多路输入、单一预测」更好，只是更贵——在某些任务上，视觉加一点触觉输入就够了。第二，可能失配归因在实践中做不准，五种原因经常互相缠在一起，强行区分反而引入新的错误。第三，可能归并、拆分、重建在行为层面看不出清楚的模式，只是一团连续的参数漂移。

这三处都不能靠论证来排除，只能靠实验。下一章就把本编到此为止的几条机制假说——共同事件组织、主动试探、联合预测——放进一组对照实验，事先写好什么结果算本书的预期被推翻。

> 带走的话
>
> 多接几路感官只是增加线索；把三路的变化一起当作要预测的对象，才迫使它们缠成同一件事。错了之后先归因、再分层调整，纠缠才会往对的方向重组。

> **小账** 论证落在哪一层？——本章以机制假说为主（联合预测促成跨感官纠缠、显式失配归因优于整体更新），以经验事实为对照（DexTacWAM、TouchWorld 的公开摘要），并标明二者不构成本体论证明。 世界模型补了哪一条？——补上「预测对象」与「失配归因」两条条件——世界模型要以跨感官变化为预测对象，并能把失配归到正确的层次再调整。

34

第 34 章

建立可以失败的实验与评价

> **这一章的问题**：本书关于世界模型发生的主张，怎样写成一组如果错了就会被看出来的实验？

◆ 一、命题：一个出路，要先说清它怎样会被证明走不通

前三章提出了几条机制假说：按共同事件组织训练优于分开收集再拼接；主动试探优于被动记录；联合预测优于单纯增加输入；显式失配归因优于整体更新。把它们收成一句话，就是本书在工程层面最核心的主张：在具体任务中，让视觉、听觉、触觉三种 SDE 通过主动互动形成纠缠，比把它们拼接或被动融合，更能形成对象同一性、更能迁移、更能从失配中恢复。

这句话属于机制假说，不属于本体论。本体论立场——学习是特征纠缠的建立与训练——不会因为一次实验的结果而被直接证明或推翻；但机制假说必须可以。**如果一个主张无论实验出什么结果都能自圆其说，它就不是在描述世界，而是在保护自己**。所以本章要做的事情很朴素：把对照组设计好，把要控制的变量列出来，把指标定义清楚，并且在实验之前写下来：什么结果算本书的预期判负。

◆ 二、三组对照：拼接、被动融合、主动互动

实验以林岚团队的两个任务为载体（这是我构想的实验方案，供研究者参照改写）：一是拧开多种瓶盖，二是夹起并移动多种软硬不同的豆腐块。两个任务都要求接触，都有看起来一样、实际不同的配置，适合检验对象同一性。

第一组，拼接组。视觉、听觉、触觉三路各自训练一个编码器或预测器，分别用各自的数据（不要求同一事件），训练完后把三路的输出拼在一起，送给一个策略模块学习动作。三路之间的对应只在最后的策略学习阶段才可能被学到。

第二组，被动融合组。数据按共同事件同步采集（第31章），三路信号作为同一个模型的输入，模型做联合预测（第33章），但所有数据都来自事先采集的示范与随机动作，训练过程中模型不能选择自己去试什么。

第三组，主动互动组。与被动融合组同样的模型结构与联合预测目标，但训练中的一部分真实交互由模型按预测分歧自行选择（第32章），并带显式失配归因的更新。

三组之间的差别要一层一层地加：拼接组到被动融合组，加的是「共同事件＋联合预测」；被动融合组到主动互动组，加的是「主动选择试探＋失配归因」。这样，如果主动互动组胜出，可以分辨胜出来自哪一层；如果只有被动融合组胜过拼接组、主动互动组并不更好，也能清楚地看出本书哪一半主张站住了、哪一半没站住。

必要时还应加两个辅助组：一个是「只有视觉」的单模态组，作为所有比较的底线；一个是「主动互动但不做联合预测」的组，用来检验主动试探的效果是否依赖于联合预测。辅助组多一个，成本就多一份，研究者可以按条件取舍，但至少要保留单模态底线。

把三组与要检验的假说对照起来，可以列成下表，便于实施者核对每一组究竟多给了什么、要回答什么。

组别	数据组织	预测对象	试探方式	与前一组相比要回答的问题
单模态底线	仅视觉事件	画面与动作	被动	多感官是否有任何增益
拼接组	三路分开采集	各路各自预测，末端拼接	被动	多路信息本身带来多少
被动融合组	共同事件，同步对齐	三路联合预测	被动	共同事件与联合预测是否促成对象同一性

组别	数据组织	预测对象	试探方式	与前一组相比要回答的问题
主动互动组	共同事件，同步对齐	三路联合预测	按分歧主动选择，带失配归因	主动试探与归因是否进一步改善迁移与恢复
辅助组	共同事件	不做联合预测	主动	主动试探的效果是否依赖联合预测

◆ 三、控制：数据量、算力与先验

对照实验最容易出的毛病，是让某一组悄悄多吃了东西。三样东西必须控制。

其一，数据量。主动互动组的长处据称是数据效率，所以必须在真实交互预算相等的条件下比较。按第32章的建议，预算按真实世界的耗时与磨损折算，而不是按条数。拼接组的三路数据可能来自不同来源，总量也要折算到同一预算。此外，建议在几个不同预算水平上各做一遍——比如预算的四分之一、二分之一、全部——画出能力随预算增长的曲线。只比一个点，很容易被偶然的结果误导；比曲线，才能看出优势是在小预算时更明显、还是在大预算时消失。

其二，算力。联合预测与主动选择都要额外的计算。如果主动互动组比其他组多用了几倍的训练计算，它的胜出就可能只是算力的胜出。所以要报告每组的训练计算量，并在计算量相当的条件下至少做一次比较；做不到严格相等时，要给出计算量—能力的曲线供读者自己判断。

其三，先验。三组如果从同一个预训练底座出发，底座里已经带着的视觉知识对三组是公平的；但如果某一组额外用了物理仿真的先验、人工设计的接触模型或专门调过的超参数，就要么三组都给，要么都不给。最容易被忽略的是人的先验：研究者在主动互动组上花了更多时间调试，这本身就是一种不对等的投入。建议由事先固定的调参预算约束各组，并在报告里写明。

◆ 四、三类指标：对象同一性、迁移与失配恢复

任务成功率当然要报，但只报成功率，看不出本书关心的东西。本书建议三类专门指标，每一类都给出可操作的测法。

对象同一性。这是第25章四条功能性判据在实验里的落地，名目照搬：形成、维持、失效、重建。形成：给一只没见过的瓶盖，允许有限次数的看、摸、听和操作，再要它对一个没做过的动作给出预期，比如只给触觉、让它预测画面上瓶盖会转多少；预测得准，说明几路在模型内部已归到同一个对象上。维持：一只瓶盖被拧了一半、手松开、过一会儿再握上去，或者只剩触觉一路，模型能否认出这还是那只拧了一半的瓶盖，而不是一只新瓶盖。失效：悄悄换上一块外观相同、更嫩的豆腐，看模型在第几次接触时察觉不对，用的是哪一路。重建：察觉之后，要多少次互动才建立起对这块新豆腐的可用预期，新预期在后续测试中准不准。第25章强调过，四条判据是分程度的，不是有或无，所以每一条都要报一个程度，而不是打一个勾。这里还要强调，这些是功能性判据，测的是行为与预测的表现；它们不回答机器有没有主观体验，那个问题第25章已单独划出，不混入这里。

迁移。在训练中没见过的对象上测：新材质的瓶盖、新尺寸的瓶子、新品牌的豆腐。迁移要分近迁移和远迁移：近迁移是同类对象的新个体，远迁移是换任务——比如从拧瓶盖迁到拧开果酱罐。本书预期主动互动组在远迁移上的优势比在近迁移上更明显，因为远迁移更依赖对象层面而不是信号层面的纠缠。这个「优势随迁移距离而变」的预期，本身也是可以判负的。

失配恢复。在任务中途人为制造失配：给瓶盖抹上一点油，悄悄把豆腐换成更嫩的一块，让一根手指的触觉传感器暂时失灵。然后测三件事：模型多快发现不对（失配检出时间）；发现之后需要多少次交互才恢复到原来的成功率（恢复交互数）；恢复之后，原来的对象和任务还做得好不好（旧能力保持率）。第三项最容易被漏掉，却最能看出纠缠是被重组了，还是被冲乱了。失配恢复与对象同一性里的「失效」「重建」两条有重叠，分工是：那两条看模型对这个物的认识怎样塌、怎样立回来；这里看的是任务能力怎样恢复、旧本事有没有被连带冲掉。

指标之间也要防止互相替代。一个模型可能任务成功率很高，对象同一性测试却很差——它找到了一条捷径，比如永远用同一套「先轻后重」的力度序列，对哪种瓶盖都管用，却从未真正把瓶盖认成对象。这种情况不应被算作失败，也不应被算作支持，而应如实报告为「任务能力形成，但未见对象同一性」。它恰好说明任务成功与对象同一性是两回事；反过来，如果对象同一性的四条判据得分高，任务成功率却不高，也要追问是策略模块没跟上，还是测法本身测到了别的东西。

还要警惕测试集被训练「看见」。迁移测试用的新瓶盖、新豆腐，在整个调试过程中都不能被任何一组接触过，最好由不参与训练的人事先封存，实验结束时一次性开封。做菜比赛里评委要盲尝，道理一样：厨师知道评委喜欢什么口味，就会不自觉地往那边做，结果比的就不再是手艺。

判负条件之外，还应写下竞争解释各自预言什么，这样实验结果出来时才知道它更偏向哪一边。最主要的竞争解释是「规模解释」：差别不在组织方式，只在数据与计算的总量，给拼接组足够多的数据，它也会学出同样的对应。规模解释预言：随着预算增大，三组的差距会缩小乃至消失。本书的解释预言：在被动记录分不清的配置上（同貌异质的豆腐、松紧不同的同款瓶盖），差距不会随被动数据增加而消失，因为那些区分只能在干预中显露。两种预言在第三节所说的「能力—预算曲线」上会画出不同的形状。如果曲线显示差距随预算稳定收敛到零，规模解释占上风，本书应当承认。

◆ 五、什么结果算本书预期判负

以下几条事先写定，实验做完不得改。

第一条，在同等数据预算、同等计算量、同等先验的条件下，如果主动互动组在对象同一性四条判据上均未显著优于被动融合组，且被动融合组也未显著优于拼接组，那么本书「三种 SDE 经由主动互动形成纠缠更利于对象同一性」这一机制假说在该任务上判负。

第二条，如果主动互动组只在近迁移上占优，在远迁移上与其他组无差别或更差，那么本书关于「对象层面的纠缠支持远迁移」的预期判负。

第三条，如果主动互动组在失配恢复上不比被动融合组快，或者恢复速度快但旧能力保持率明显更低（学了新的、丢了旧的），那么本书「显式失配归因有利于纠缠重组」的预期判负。

第四条，如果去掉某一路感官（比如声音）后，三类指标都没有可测的下降，那么在该任务上「这一路参与纠缠是必要的」不成立。这一条判负不伤及本书的总立场——本书本来就不预设三种感官对所有任务都不可替代——但必须如实记录，并据此修订「哪些任务需要哪几路」的判断。

第五条，如果三组的差异在不同随机种子之间的波动大于组间差异，那么本次实验不能支持任何一方，结论应写为「未能分辨」，而不是「趋势支持」。

「显著」的统计标准、重复次数、效应量的最低门槛，应当在实验前登记，并对两个任务分别报告〔具体统计方案由实施者按领域惯例确定，本书不预设数值〕。还要说一句：两个任务上结果不一致，是很可能出现的情况，比如拧瓶盖支持、捏豆腐不支持。这时候不能挑支持的那个报，而要两个都报，并把不一致本身当作需要解释的发现。

◆ 六、不支持预期时的修正程序

实验判负，不等于本书就此作废，但也绝不能随手打个补丁说「其实我们本来就是这个意思」。本书建议一个固定的修正程序，分四步。

第一步，先查实验本身。核对数据预算、计算量、先验是否真的对等；核对事件对齐与缺失标注是否做对；核对主动选择的规则是否按设计执行。实验执行有误，先修实验，不动假说。

第二步，定位判负发生在哪一格。先用定式三角 S1·D1·E1 的语言问：是 E1 的现实后果没有写回纠缠（各路信号对不上，后果落不到对象上），是 S1 的对比、变化、分布没过（对得上，却认不出同一个对象，或换个情境就站不住），还是 D1 的方向没有发生（认得出，却没有长出新的区分，任务因此连不起来）？再看另两个三角里有没有连带的空格。定位越准，后面的修正越有针对性。

第三步，在机制假说的层面修正，并且要让修正后的假说比原来的更容易被检验，而不是更难。例如：如果发现主动互动只在对象多样性高的条件下占优，就把假说收窄为「对象多样性超过某个程度时，主动互动占优」，并给出这个程度的测法，再做一轮实验。如果修正之后假说变得什么都能解释，那就是在逃避，应当宁可承认原假说失败。

第四步，区分哪些修正触及本体论表述，哪些不触及。多数判负只会改动机制假说——比如联合预测的粒度、主动试探的比例。只有当一系列不同任务、不同实现、不同团队的实验都反复显示「跨感官的对应关系对对象同一性的形成没有作用」时，才需要回头检视本书关于特征纠缠如何建立的表述。这一层的修正不应由单次实验驱动，但也不能永远推说「还需更多实验」：本书建议，在第40章所说的研究纲领中，把这种累积性证据的审查定期做一次，把结果写进公开记录。

做菜的人改方子，走的也是这个程序。一道菜按新方子做出来不好吃，先看是不是火候没按方子来（查实验），再看是咸了、淡了还是腥了（定位层次），然后只改那一样、下次再做一遍（修正可检验），而不会一次失败就断言「这道菜根本不能这么做」，也不会把每次失败都推给「今天的菜不新鲜」。

◆ 七、边界与向下一章推进

这一章写的是实验设计的骨架，不是一个已经跑完的实验；它所有的结论都悬置到实验之后。它的价值，在于让本书的出路可以被检查：研究者照着做，可能得到支持的结果，也可能得到不支持的结果，两种结果都有用。

但即使实验支持了本书的预期，也只是说明在这两个任务上，按这种方式训练出来的能力更好。「更好」离「可靠」还有一段路：可靠意味着知道它在什么条件下好、在什么条件下会坏、坏了怎么办。下一章就讲这段路。

带走的话

本书的出路要写成可以失败的实验——拼接、被动融合、主动互动三组对照，数据、算力、先验对等，以对象同一性、迁移、失配恢复为指标，并事先写定判负条件与修正程序。

小账 论证落在哪一层？——本章以机制假说的检验设计为主，明确区分机制假说（可判负）与本体论立场（不由单次实验判定）。 世界模型补了哪一条？——补上「可检验性」这一条件——世界模型发生的主张，须给出对照组、控制变量、功能性指标与事先登记的判负标准。

35

第 35 章

可靠性形成与持续封装

这一章的问题：一个机器能力，什么时候可以说「靠得住」？靠得住之后，又怎样知道它什么时候不再靠得住？

◆ 一、命题：可靠不是成功率高，而是知道在什么条件下成功

假设林岚团队的拧瓶盖系统在测试中成功率达到了很高的水平。这能不能说它可靠了？还不能。一个百分比只是把许多次尝试平均起来的结果，它不说明哪些情况下会失败。如果那少数失败恰好集中在某种瓶子上——比如装了碳酸饮料、瓶内有压力的瓶子——而明天的任务恰好全是这种瓶子，那么昨天的高成功率对今天几乎没有意义。

本章的命题是：可靠性的形成，是从「任务成功率」走到「适用条件」——知道这项能力在哪些条件下可以放心调用，在哪些条件下需要小心，在哪些条件下不该调用。这对应第5章讲的「从纠缠形成到能力封装」。那里给封装下的定义是：一组特征纠缠经过充分训练和检验，作为整体能力固定下来，可以在一定条件下直接调用，同时带着适用边界，失效后要解封重训。第5章里周敏让小舟在车库出口下车推行，能力没变，变的是封装的「接口说明」。本章要做的，就是把这份接口说明在机器上写实：它继承了哪些判断，能在什么范围里运行。

做菜的人对此再熟悉不过。一个人说「我炒青菜从来不咸不淡」，这话靠得住的前提是：在他自己家的灶上、用他常买的那个牌子的盐、炒他常炒的那几种菜。换到朋友家，灶火更旺，盐颗粒更细，同样一勺就更咸，他的「从来不咸不淡」立

刻打折。真正的老手会说：「在我家灶上，我心里有数；换个地方，我得先尝一口。」这句话就是一份带适用条件的封装说明。

◆ 二、从成功率到适用条件

怎样从成功率走到适用条件？靠的正是第31章那张事件记录表里的「情境」一栏，以及第34章的迁移和失配测试。

具体做法是把每一次尝试按情境分组：瓶盖材质、瓶内是否有压力、手上是否潮湿、瓶身软硬、环境温度，等等。然后看成功率在各组之间怎样变化。分组之后，常常会看到一幅和总平均很不一样的图：大多数组都很高，个别组明显偏低，还有一些组样本太少，根本说不准。这三种情况要写进说明书的不同位置：高的那些组成为「可放心调用」的条件；偏低的组成为「已知的薄弱区」；样本太少的组成为「未经检验区」——这一类最要紧，因为最大的风险往往不在已知的薄弱处，而在根本没测过的地方。

这里要防止一种常见的倒错：把「没在测试中失败过」当作「在这个条件下可靠」。没测过和测过没出错是两回事。适用条件的说明书里，「未经检验」必须是一个明确的类别，而不能被默认归到「可靠」那一边。

一份适用条件说明书，至少应当回答五个问题。这项能力做什么（任务的定义，包括什么算成功、什么算失败）；在哪些条件下测过、各自结果如何；已知在哪些条件下表现差、差到什么程度；哪些条件从未测过；出现什么信号时应当停止调用。说明书不必长，但这五项缺一项，使用者就会在那一项上替说明书「默认」一个答案，而默认的答案往往是乐观的。

以捏豆腐为例（这是我构想的场景）。林岚团队的说明书可能这样写：本能力用于夹起并平移方块豆腐，成功指无可见裂纹且位移误差在规定范围内；在北豆腐、南豆腐、冷藏后取出十分钟内的条件下测过，成功率均高；在嫩豆腐上测过，成功率明显偏低，主要失败为夹起瞬间碎裂；在冻豆腐、油豆腐上未测；当首次接触时阻力上升曲线比训练分布平缓得多，或画面中豆腐表面已有裂纹时，应改用托底铲取而不是夹取，或交给人。这样一份说明书，比一个总成功率有用得多。

还要说明：说明书由谁来写、由谁来签。它不是模型自己生成的，而是工程师根据测试记录整理、由使用方审核的；其中每一条判断，都带着团队的经验与取舍。封装所继承的判断（第5章），在这里就是一份份具体的签字。训练的发生单位是集体人类—AI复合体，可靠性的形成同样如此——它不是机器单方面的属性，而是机器、记录、工程师与使用者之间一份持续更新的约定。

◆ 三、反馈控制、情境记忆与长期训练的分层

系统在运行中遇到不顺利的情况，会用几种不同的方式应对。全书反复要求区分五件事：暂时语境变化、外部记忆、控制反馈、参数训练、持久学习。第18章用导航软件把这个分法立住了，这里把它落到机器人任务上。

控制反馈是最快的一层。拧瓶盖时手指开始打滑，底层控制在几毫秒到几十毫秒内加大握力。第33章说的 TouchWorld 底层残差修正，就属于这一层。它是即时纠偏，任务一结束就不留痕迹，模型本身什么也没学到。

情境记忆是中间一层，大致对应「暂时语境变化」与「外部记忆」。系统在今天这批瓶子上试了几次，发现它们的盖子都偏紧，于是在本批次内把起始力度调高一档。这个调整存在于本次任务的上下文或一个外部记录里，换一批瓶子就应当清掉。它让系统在一段时间里「记得」，但没有改变系统的参数。

参数训练是慢的一层：把积累的事件拿回去更新模型参数。持久学习则是在参数训练之上的更高要求：更新之后，新能力在一段时间后、在不同情境下仍然稳定，并且没有损坏原有能力。

分层的意义在于：每一层的改变都要放在自己该在的地方，不能越层。如果把控制反馈的每一次纠偏都直接写进参数，系统会被大量偶然的扰动冲得摇摆不定；如果把本批次的「瓶盖偏紧」写进参数，换一批松瓶盖时就会用力过猛。反过来，如果一个反复出现的问题始终只在控制层被纠偏、从来没进入参数训练，系统就永远在「补救」而从未「学会」。何时让下层的经验上升到上层，是持续封装要管的核心问题。本书提出的一个可检验规则是：同一类失配在多个独立批次、多种情境

下反复出现，才把它送入参数训练；只在单一批次出现的，留在情境记忆。这条规则的阈值怎么定，应当用第34章的方法去实验确定。

分层还有一个时间上的含义。控制反馈在毫秒尺度上起作用，情境记忆在分钟到天的尺度上，参数训练在周到月的尺度上。三个尺度错开，系统才能既灵敏又稳定：快的一层吸收偶然扰动，慢的一层只吸收反复出现的规律。人的学习也有类似的分层。一个刚学颠勺的人，菜要飞出锅时手腕一收，是即时纠偏；这一晚上他记得「这口锅比家里的重，起势要早」，是情境记忆；颠上几个月，换哪口锅手上都有数，才算持久学习。机器上的分层未必与人的分层机制相同，本书不作这种推论；这里借用的只是一个结构上的提醒：不同快慢的改变，要放在不同的地方。

◆ 四、识别封装边界

封装边界不能只靠事先写的说明书，还要能在运行中被识别：系统自己要能察觉「我现在可能已经走出了适用范围」。

可用的线索有几种。一是预测误差：模型对接下来的画面、声音、力的联合预测，误差突然变大且持续不降，说明眼前的对象可能不是它认识的那一类。二是预测分歧：第32章用来选择试探的那个信号，在运行中也可以用来报警——多个预测器开始各说各话，说明系统拿不准了。三是情境比对：当前情境的各项记录（材质、温湿度等）落在训练与测试从未覆盖的区域。三条线索中任何一条亮起，系统都不应当继续按原样执行，而应当降速、改用更保守的动作、转为试探模式，或者把任务交给人。

这一节对应第25章「对象同一性的失效与重建」。对象同一性失效，在功能上就表现为这些线索：系统对这件东西的看、听、摸对不上了，或者对接下来会发生什么拿不准了。能察觉自己的同一性正在失效，是可靠性的一个必要部分；只会成功、不会察觉失败的系统，平时成功率再高，也不能叫可靠。

运行中识别边界，还有一个容易被忽视的难处：误报与漏报的权衡。报警线定得太敏感，系统动不动就降速、交给人，实际上没法用；定得太迟钝，真走出边界了还在照常执行。这条线没有一劳永逸的定法，但可以把它变成可以检查的东西：

记录每一次报警之后实际发生了什么——报了警、交给人，人一看其实没问题，算一次误报；没报警、结果出了事，算一次漏报。两种错误的代价不同：在捏豆腐上漏报，碎一块豆腐；在搬运热汤上漏报，可能烫伤人。所以报警线应当按任务的风险分别设定，风险越高越宁可误报。本书不对任何具体应用的风险等级作政策表态，只指出：这条线的位置，是封装说明书里应当写明并能被复核的一项。

还有一种边界识别要靠「对象同一性」本身来做。系统握住一只瓶盖，看、听、摸三方面一开始对得上，拧到一半，触觉突然显示瓶盖在动、画面上却显示它没动——原来是瓶盖外圈的塑料圈断了，空转。这时候，三路信号指向的已经不再是同一个「可以拧开的瓶盖」。能察觉这种「同一性在任务中途裂开」，比察觉「这是一只没见过的瓶子」更难，因为开始时一切正常。第34章的失配恢复指标中，「失配检出时间」测的就是这种能力。

◆ 五、持续评估、回写、解封与协作处理

封装不是一劳永逸。林岚团队的系统上线以后，至少有四件事要持续做。

持续评估。定期用一套固定的检查任务复测，看适用条件里的各组成功率有没有悄悄下降。下降的原因可能在机器（指尖磨损），可能在世界（市面上换了新款瓶盖），也可能在模型（某次更新带来的副作用）。不复测，就只能等出了事才知道。

持续评估还要防一种慢性的偏移：检查任务本身老化。一套固定的检查任务用久了，团队会不自觉地围着它优化，检查成绩年年好看，真实世界里的新瓶盖却越来越多地拧不开。所以检查集要分两部分：一部分长期固定，用来比较历次版本；另一部分定期从最近的真实运行事件里抽取更新，用来跟上世界的变化。两部分的成绩出现背离——固定集越来越好、更新集越来越差——就是一个明确的信号：能力在向检查集收缩，而不是在向世界扩展。

回写。运行中积累的事件——尤其是失败与「准失败」——按第三节的分层规则，该进情境记忆的进情境记忆，该进参数训练的进参数训练；同时把新发现的薄

弱区、新检验过的条件写回说明书。说明书要像病历一样有更新历史，而不是一纸永久不变的声明。

解封。当发现某一项能力在某类条件下系统性地失效，就要把这部分「解封」：在说明书里把它从「可放心调用」移到「已知薄弱」或「暂停使用」，并重新进入训练。解封不是失败的标志，而是封装诚实的标志。第5章已讲过失效、解封与重新训练；这里补充一条操作上的要求：解封应当有明确的触发条件（比如某组成功率连续若干次复测低于某值），而不是靠人临时判断，否则解封会因为「不想承认」而一拖再拖。

协作处理。有些失效靠系统自己解决不了，要交给人或别的系统。拧不开的瓶盖，交给人；捏豆腐时遇到从未见过的食材，暂停并请求示范。这对应第30章「专用模型的协作与共享」：一个任务世界的封装边界，往往正是它与其他任务世界、其他复合体之间的接口。把「交出去」设计成正式的流程，而不是失败后的兜底，是可靠性的一部分。被交出去的那些事件，还要连同人的处理过程一起记下来，回到第31章的事件库——人怎样拧开了那只拧不开的瓶盖，是下一轮训练最好的材料之一。

用做菜收个尾。一个在家做了二十年饭的人，到了一家陌生的厨房，他的「封装」并没有失效，但他会自动进入一种更小心的状态：第一道菜少放盐、多尝几次；火候先小后大；有不认识的调料先闻一闻、蘸一点尝尝。做完一两道菜，他对这间厨房「心里有数」了，小心的状态才慢慢退去。这一整套——知道自己的手艺在哪里可靠、察觉环境变了、降速试探、更新判断、必要时问主人「你家这酱油咸不咸」——就是本章所说的持续封装在人身上的样子。机器能力要做到可靠，大致也要走过这样一条路，只是每一步都要写成可以记录、可以检验的规程。

◆ 六、边界与本编收束

本章的主张有两处需要警惕。一是适用条件不可能穷举，说明书总会有没写到的地方；所以第四节的「运行中识别边界」不是锦上添花，而是对说明书必然不完

整的补救。二是分层规则与解封触发条件里的各种阈值，目前没有普适的值，本书也不预设；它们应当针对具体任务、通过实验确定，并随经验更新。

至此，第七编走完了一条线：以共同事件组织经验，以主动试探区分配置，以联合预测促成纠缠、以失配归因引导重组，以可以失败的实验检验这些主张，以适用条件与持续封装形成可靠性。这条线回应了全书判词的后半句——世界模型的出路，不在一次封装一个全能的世界，而在具体任务的互动里，先让一个物站稳；也给判决五「任务组织世界，封装要能失败」交了一份可以照着做、也可以照着驳的方案。

但这条线讲的都是机器人在物理任务中的学习。第三编讲过的那个集体人类—AI语言世界，怎样在每一个具体的人手里继续发生？一个具体的人和 AI 一起备课、写作、做研究时，也有共同事件、主动试探、失配与封装吗？这是第八编要接着回答的问题。

> 带走的话
>
> 可靠不是成功率高，而是知道在什么条件下成功、能在运行中察觉自己走出了边界；封装要持续评估、分层回写、按条件解封，并把交给别人处理也设计成正式流程。

> **小账** 论证落在哪一层？——本章以机制假说为主（分层上升规则、运行中边界识别、解封触发条件），兼有工程规程性的建议；经验事实仅作对照。　世界模型补了哪一条？——补上「可靠封装」这一条件——世界模型的能力须附带适用条件、能在运行中识别边界，并经持续评估、回写与解封保持可靠。

第 八 编

从集体世界到具体复合体的学习

编序 前七编先把学习重新理解为一块块经验缠到一起（特征纠缠，E）的建立与训练，再看三个世界怎样发生，然后进入集体人类—AI复合体语言世界，最后转向物理世界模型，追问对象同一性与可靠行动从哪里来。到这里，书的视线要从「集体」回到「具体」。一个公共模型训练完成以后，它并不停在服务器里等人来取用；它要进入千千万万个具体的人的备课、作业、写作、实验与日常决定之中。每一次进入，都是一个「个人—AI复合体」在发生。

本编五章接力完成这次回转。第36章问：集体训练出来的那份相对同一性，怎样成为某一个人继续发生的条件，而不变成压在每个人头上的统一答案。第37章把个人与AI的往来拆开看，严格区分暂时语境变化、外部记忆、控制反馈、参数训练与持久学习这五件事，说明一次流畅对话为什么不能证明能力已经长成。第38章落到教育：周敏和AI一起备课，小舟用AI学写作文，看学习怎样从「知识呈现」转向「纠缠训练」，又怎样用迁移和重组来检验。第39章落到科研：一个小小的研究小组怎样在概念、模型推演与现实实验之间往返，让异常变成新的对象与新的方法。第40章收束全书的理论工作，把本体论命题、机制假说与工程证据分三层列出，提出特征纠缠的形式化方向和可以动手推进的研究计划。

读这一编，可以一直带着那三问：这里纠缠的是什么？差异是怎样运行的？显露出来的东西靠什么站稳、又在哪里失效？

36

第 36 章

集体同一性怎样成为个人发生条件

> **这一章的问题**：一个由千万人共同训练出来的公共模型，怎样成为某一个具体的人继续学习与发生的条件，而不是把同一个答案压给所有人？

◆ 一、继承下来的是条件，不是结论

先把本章的命题说清楚。第13章「集体性的语言世界如何发生」已经讲过，大模型的训练让无数人的表达习惯、推演约束与知识沉淀进入共同训练，形成一个具有相对同一性的语言世界。这里的「相对同一性」，白话说就是：不管谁来问，它大体讲同一种话，守同一套推理规矩，知道差不多同样多的事。它是集体的，因为它来自许多人；它是相对的，因为它内部容纳差异、并且会随再训练而改变。

本章的立场是一条本体论主张：**当一个具体的人开始使用这个公共模型时，他继承的不是一套现成的结论，而是一组发生条件**。所谓发生条件，是指那些让新的显露（S）有可能出现、让新的差异运行（D）有可能展开的已有纠缠。公共模型能把物理概念讲顺、能把一段话改得通顺、能列出一道题的几种解法，这些能力本身已经是集体复合体发生出来的封装。个人拿到它，就像一个新厨子走进一间别人收拾好的厨房：刀磨好了，调料摆好了，火候的常识也写在墙上，可今天这桌菜给谁吃、谁怕辣、谁的老人牙口不好，厨房不会替他知道。菜是在这间厨房里，由这个厨子和这桌客人一起做出来的。

这一判断有它的反面。另一种看法认为，公共模型就是一个装好了正确知识的容器，个人使用只是取用，取得越多越好，取得越准越好。按这种看法，个人只是终端，集体同一性向下灌注，个人的差异是噪声。本书在第一编已经说明，这正是

「知识进入大脑」那幅传统图景的翻版，只是把大脑换成了服务器。若接受它，个人—AI复合体这个概念就多余了，学习也就只剩下一件事：把公共答案搬到自己这里来。

◆ 二、公共能力怎样接上具体经历

那么继承来的条件怎样变成个人的发生？下面给出一个机制假说。说它是假说，是因为它可以在使用记录与学习结果上被支持或推翻。

假说的要点有三层。第一层，具体的人带着自己的问题进来。问题不是一串字，而是他已有纠缠里的一个张力点：他手头有一件事办不成，或者两样东西对不上。这个张力点决定了他会从公共能力里拉出哪一部分，也决定了他会对模型回答里的哪一句起疑。第二层，模型的回应是在公共纠缠上展开的差异运行，它给出的显露总带着「平均人」的样子：用最常见的例子，取最稳妥的说法，照顾最多数的读者。第三层，个人把这份显露拿回自己的处境去碰：和自己的经历对照，和自己的价值掂量，和自己面对的那些具体的人比一比。碰得上的部分被他吸收，碰不上的部分形成新的张力，推动他追问、改写或者干脆放下。这正是三原理所说的过程：两维发生矛盾，产生推动第三维改变的动力。在这里，公共显露（S）与个人已有纠缠（E）对不上，矛盾推动个人重新组织他的差异运行（D）：他换一种问法，换一个例子，换一条路走。

这个假说可以导出几条能检验的推论。如果它对，那么同一个公共模型、同一个问题，在不同的人手里会走出明显不同的后续追问序列；追问越多地引入个人处境的人，最后形成的方案与模型首轮回答之间的差距越大；而只接受首轮回答的人，其结果更接近彼此。反过来，如果实际记录显示人们的后续追问大同小异、最终方案也趋于一致，那么至少在那个场景里，「个人经历进入」的作用比本书设想的弱，需要重新考虑。

这里还要把三种进入分开。一种是具体经历的进入：这个人遇到过什么，失败过什么。一种是具体问题的进入：他此刻要解决的是哪件事，限制条件是什么。一种是价值的进入：他在乎什么，愿意为了什么牺牲效率或稳妥。三者都改变他对公

共显露的取舍，但方式不同。经历提供对照的材料，问题划定相关的范围，价值决定冲突时站哪一边。一个人若只带问题不带经历，他得到的多半是可用的通用方案；若带了经历却不带价值，他会在许多合理选项面前犹豫不决。

◆ 三、周敏备一节浮力课

下面是我构想的场景。

周敏在长沙一所中学教物理，下周要上八年级的浮力。这不是第9章讲过的那节课：那是她教书第二年，一个男生问「铁船为什么能浮」，她临场让学生用铝箔糖纸折船，第一次接住了教案外的问题。那节课改的是她的教师自我，这里不再重讲。十几年后再备浮力，她手边多了一个AI。她打开一个通用的大模型，输入：「帮我设计一节四十分钟的浮力入门课。」几秒钟后，一份规整的教案出来了：导入用阿基米德洗澡的故事，新课讲浮力的定义与方向，再做弹簧测力计称重的演示实验，归纳阿基米德原理，最后是课堂练习。环节齐全，时间分配合理，语言也准确。

这份教案就是集体同一性的样子。它浓缩了大量公开教案、教材与教学经验里最常出现的组织方式。周敏承认它是好的起点，她自己第一年教书时写的教案还未必有这么齐整。

可是她读到「阿基米德洗澡」那一段就停住了。她教了十几年书，知道这个故事学生在小学就听过，一讲，后排就有人接话「我知道，尤里卡」，课堂的好奇心反而先泄了气。更要紧的是，她去年带过的一个班，有好几个学生把「浮力」理解成「水往上推的力，东西越轻推得越多」，期末考到铁块沉底时浮力是否为零，错了一大片。这些是她的经历，模型不知道。

于是她追问：「我的学生容易以为沉下去的东西不受浮力，能不能把导入改成直接冲着这个误解来？」模型给出几种方案，其中一个是让学生先预测：把一个铁块用测力计吊着慢慢放进水里，读数会不会变。周敏觉得这条路对了，但又改了一处：她不用铁块，用学生从家里带来的东西，土豆、钥匙、橡皮、一小瓶矿泉水，让他们先猜哪个会沉、沉下去的东西读数会不会变小。她这样改，是因为她在乎一

件事：她要学生把物理和自己书包里、厨房里的东西缠在一起，而不是和课本插图缠在一起。这是她的价值，模型的首轮回答不会自动替她选。

碰撞也不只朝一个方向。模型在列常见误解时还提到一条：不少学生以为物体浸得越深，受到的浮力越大。周敏愣了一下，她没把这一条当回事过，回想起来，去年确有学生在作业里这样写，她当时只当是粗心。这一回是公共能力补上了她经历里的一个盲点。于是她在实验里加了一步：把吊着的土豆在水里上下移动，让学生盯着读数看。个人经历修正公共方案，公共积累也修正个人经验，两股力在同一份教案里互相改写。

到这一步，一份教案已经从公共的样子长成了周敏的样子。注意这里发生的事情：公共模型提供了教案的骨架、实验的常规做法、误解的常见类型，这些都是继承来的条件；周敏的经历让她识别出公共方案在她班上会失效的地方，她的价值让她在几个可行方案里挑出一个。最终的那份教案，既不是模型单独写的，也不是周敏凭空想的，它是在这个个人—AI复合体里发生出来的。

还可以再往下看一步。这节课上完，周敏发现学生对「土豆为什么沉、苹果为什么浮」吵得最凶。她把这个现象记在本子上，下次备「密度」时，她会带着这个记录去问模型。她的问题因此又变了。公共能力没有变，变的是她拿来碰它的那一面。

◆ 四、同一性训练的边界：别让它变成单一答案

现在要处理一个真实的风险。集体同一性是好东西，它让人与人之间有共同语言，让一个刚入行的老师很快拿到合格的起点。但同一性训练有一种滑落方式：从「提供共同条件」滑向「强制单一答案」。

滑落可能从三处发生。第一处在训练这一端。第15章「集体语言世界的边界与修正」讨论过，数据选择和评价标准会留下偏向。如果评价反馈一再奖励某种最稳妥、最常见的回答样式，模型的显露就会越来越窄，同一个问题来回只给出几种标准写法。第二处在使用这一端。人如果只拿首轮回答，不追问、不对照，那么千千万万个周敏会上出几乎一模一样的浮力课。单看每一节课都不差，合起来却是一种

贫瘠：各地课堂原本不同的经验，在一代人之间被压平了。第三处在制度这一端。一旦有人把「和模型标准答案的一致程度」当作评判标准，个人的差异就会从资源变成扣分项。

这里要避免两种过头的说法。一种说法是公共模型必然导致趋同。这不是本书的结论，它是一个可以检验的经验问题，答案可能随场景、随使用方式而不同，不宜预写。另一种说法是个人差异总是好的、公共同一性总是压迫的。这也不对。一个刚学做菜的人若完全不看菜谱，全凭「我的个性」放盐，得到的多半不是创造，而是一锅咸汤。共同条件的价值恰恰在于让个人的差异有一个可以站稳的地方。

所以本章给出的边界是这样的：同一性训练应当训练的是共同的差异辨认能力和共同的推演约束，而不是共同的结论。用SDE的话说，它稳定的应当是E的一部分，即那些让人们能互相理解、能互相检验的纠缠，而不应当把S锁死在一种显露上。判断一个公共模型、一种使用方式有没有越过这条边界，可以问三件事：它是否在被追问时给出真正不同的方案，而不只是同一方案的换词；它是否能把使用者提供的具体处境吸收进回答，而不是把处境当作装饰；它是否在不同方案之间说明取舍理由，让人自己选。三件事都做到，同一性是条件；三件事都做不到，同一性就在变成强制。

还有一层边界要说明。本章说的「个人发生」，并不意味着每个个人—AI复合体都会长出更好的东西。周敏的经验让她改对了导入，换一个经验更少或判断有偏的人，也可能把一份合格的公共教案改坏。个人进入带来的是差异的可能，不担保差异的质量。质量要靠下一步的检验，靠课堂上学生的反应，靠第38章要讲的迁移与重组。

◆ 五、个人的发生怎样回到集体

前面说的都是从集体到个人这一个方向。其实还有反方向的一条路，它更慢、更间接，却决定了集体同一性会不会继续发生。

周敏把「土豆、钥匙、橡皮」的导入在教研组里讲了，隔壁班的老师也试了，效果不错，有人把它写进了一篇教学心得，发在一个公开的教师论坛上。几年以

后，这样的心得文章如果被收进某个模型的训练数据，它就以极其稀薄的方式进入了下一轮集体训练。周敏的一次经历，最后成了千万份样本中的一份。这条路径就是第11章「人怎样进入大模型训练」所说的历史中介：人的活动先在文本、制度和工具里沉淀，再被训练吸收。

另一条路更直接。许多公开的大模型服务会收集使用者的反馈，例如对回答点赞或点踩、在几个回答中挑一个更好的；这些信号在符合各家公开政策与使用者授权的前提下，可能被用于后续的评价与再训练。各家怎样使用、默认是否开启、怎样退出，做法并不相同，要以各产品当时公开的说明为准。这里不对任何一家的做法下判断，只说结构：个人与AI的一次往来，有可能经过筛选、汇总，回写到集体复合体的纠缠里。

这一回写有两面。好的一面是，集体同一性不是一次定死的，它在吸收个人发生的过程中继续改变，第13章所说的「集体同一性怎样容纳差异」在这里有了一个具体通道。坏的一面是，回写也会放大第四节说的趋同：如果大家点赞的总是最稳妥的回答，回写就会把显露进一步收窄。所以，个人在使用中保留的那些「碰不上」、那些追问与改写，不仅关乎他自己，也关乎集体世界下一轮会长成什么样子。一个只会取用的使用者群体，最后得到的是一个越来越像平均数的公共模型；一个在使用中留下差异的群体，才给集体同一性留下重新组织的材料。

需要声明，这一段是对结构的描述，不是对某种回写效果大小的断言。个人反馈在再训练中究竟占多大分量、能否保留差异而不被平均掉，都是需要具体研究的经验问题。

◆ 六、从「继承」走向「继续」

本章开头的问题，到这里有了回答：集体同一性成为个人发生条件，靠的是个人带着经历、问题与价值去碰公共显露，在碰不上的地方产生张力，进而重组自己的差异运行。公共模型因此不是终点，而是起点。

可是这一切还只是在一次使用里看。周敏这一次备课改好了，下一次她再打开模型，模型还记不记得她班上学生的那个误解？如果记得，是怎样记得的？是这一

次对话里暂时记住了，还是她自己存了一份笔记，还是模型的参数真的因此改了？而周敏本人呢，她这次备课是否让她的教学能力有了持久的变化？这些问题如果不分开，就会把「它好像懂我了」误当成「它学会了」，也会把「我用它做成了一件事」误当成「我学会了这件事」。下一章专门把这几件事拆开。

> **带走的话**
>
> 从公共模型那里继承来的是发生条件，不是现成答案；个人的经历、问题与价值去碰公共显露，碰不上的地方才是个人发生的起点。同一性训练应当稳定共同的辨认与推演，而不是锁定单一结论。

小账 论证落在哪一层？——本章以本体论立场为主（继承的是条件而非结论），第二节给出可检验的机制假说及其推论，第四节划出经验问题的边界（趋同与否不预写）。 世界模型补了哪一条？——补上一条条件：集体训练形成的世界，只有在具体复合体的处境里被碰撞、被修订，才继续发生；不经具体任务碰撞的「全能同一性」不等于任何一个人的可用世界。

37

第 37 章

个人与AI如何继续发生语言世界

> **这一章的问题**：一个人和AI来来回回地问、答、改，到底是什么在变？哪一种变化才算得上学习?

◆ 一、五件事，不能混成一件

本章先立一个区分，它是全书论证纪律里最容易被日常经验冲垮的一条。第15章讲集体语言世界的修正时，已经拿「湘菜红烧肉放不放酱油」的纠正把其中几层过了一遍；这里换到个人这一侧，把五件事逐一摆开。人和AI聊得顺了，常会说「它懂我了」「它学会了」；自己借AI办成了一件事，也常会说「我学会了」。这些话在日常里无妨，在研究里却会把五件不同的事混成一件。

第一件是暂时语境变化。同一个模型，前面说过的话会影响后面的回答。你告诉它「我是小学四年级学生」，它后面的用词就浅了；你给它看一段范文，它仿写就像了。可是这些变化只活在当前这段对话的上下文里，模型的参数一点没动；对话关掉，重开一个窗口，它又回到原样。打个比方，这像在一块小黑板上写了几行提示，讲课的人看着提示讲，课一完黑板就擦了。

第二件是外部记忆。有些系统会把你说过的事存下来，下次对话时再读进上下文；有些人自己建一个笔记文档，每次开头贴给模型。这比暂时语境走得远，跨了对话，但它仍然在参数之外：存下来的是模型之外的一份记录，模型每次是「重新读到」它，而不是「已经变成」它。各家产品的记忆功能怎样存、何时写，做法不一；一个设计得好的记忆，应当让使用者自己看得到、改得了、删得掉。好比一个厨子案板旁边钉着一张顾客口味卡，卡在，他就照着做；卡掉了，他就不知道。

第三件是控制反馈。你说「太长了，砍一半」「这句不对，重写」，模型立刻改；机器人抓东西滑了，传感器读数一变，控制器马上加力。控制反馈是在一个闭环里根据偏差修正下一步动作，它让当下的结果变好，却不必改变系统内部的长期组织。就像开车时手不停地微调方向盘，车走直了，手的技术并不因为这一次微调就长进。

第四件是参数训练。模型的权重通过梯度更新真的改了：预训练、微调、根据人类偏好做的强化训练，都属于这一类。参数一改，以后所有对话都带着这次改变。对绝大多数公开服务来说，这种训练是由开发方在离线流程中成批完成的，不是某一个用户在对话中实时触发的。有的平台允许开发者拿自己的数据做微调，得到一个专门的版本，但那也是一项单独提交、离线跑完的训练任务，不是聊着聊着参数就变了。

第五件是持久学习。这是本书意义上的学习：一种能力稳定地形成了，换了情境还能用，过一段时间还在，不靠当场提示也能调出来，并且失效时能被识别。用本书的话说，就是特征纠缠（E）建立起来并站稳，使得在新的差异运行（D）中能再次显露（S）出合格的结果。

这五件事的关系要说准。**前四件都是机制，第五件是结果**。持久学习可以借助参数训练实现，也可以借助外部记忆加上稳定的使用习惯在复合体层面实现；反过来，参数训练并不自动等于持久学习，参数改了，能力可能没长，甚至可能让原有能力退化，第1章讲过的遗忘在模型这里同样存在。暂时语境变化和控制反馈本身都不构成持久学习，但它们可以是持久学习的材料。

◆ 二、语言世界在复合体里怎样继续发生

有了这个区分，再来看本章的命题：个人与AI的往来，确实在继续发生语言世界，但发生在哪一层、持续多久，要分开说。

提问是第一步。一个问题不是空白的请求，它把提问人已有纠缠中的一个张力点交给对方。回应是第二步，模型在集体纠缠上展开一次差异运行，给出一个显露。纠正是第三步，人发现显露与自己处境碰不上，指出来。共同表达是第四步，

几轮之后，一段话、一个方案、一种说法出现了，它不属于提问前的人，也不属于训练好的模型，而属于这一次复合体的发生。

这四步在暂时语境层面很充分：一段对话之内，语言确实在长，说法越来越贴合，概念越来越精确。它在外部记忆层面能延续：人把这次得到的好说法记下来，下次带进来。它在参数层面通常不直接发生，个人的一次对话不会马上改写公共模型。它在人这一侧的持久学习层面，则要看人自己有没有经历足够的差异运行，把新的说法缠进自己的纠缠里。

四步之中，纠正最值得多看一眼。纠正的时候，人必须把「哪里不对」说出来，而说出来本身就是一次差异运行：原先只是隐约觉得别扭，现在要找到别扭在哪个词、哪个前提、哪个例子上。很多人有这样的经验，向别人解释为什么一个答案不对，解释到一半，自己先想明白了。所以在复合体里，纠正不只是对模型的修理，它往往也是人这一侧纠缠重组最集中的时刻。一段只有提问和接受、没有纠正的对话，语言世界多半只在模型那一侧暂时铺开，人的这一侧没怎么动。

这里有一个常被忽略的不对称。在复合体中，模型那一侧的变化大多是暂时的或外挂的，而人这一侧的变化可以是持久的，前提是人确实动了手、动了脑。于是一个简单的推论是：在个人—AI复合体里，持久学习最可靠的落点常常在人身上，或在人有意维护的外部记忆上，而不在模型参数上。这是一个机制假说，可以被检验：比较两组人，一组让AI直接给出终稿，一组只让AI提问、指错，人自己改；隔两周不借助AI独立完成同类任务，看哪一组更好。本书预期后一组更好，但这要用数据说话，并且结果可能随任务难度、学习者水平而变化。

◆ 三、小舟的一篇《我的小狗》

下面是我构想的场景。

周小舟九岁，读四年级，老师布置作文《我家的小动物》。他家有一条小狗叫豆豆。他写了一段：「我家有一只小狗，它很可爱，它每天都很开心，我很喜欢它。」写完就卡住了，三行，离要求的四百字差得远。

周敏允许他用家里的AI学习工具，但定了规矩：不许让它替你写，只能让它问你问题。小舟输入：「我要写我的小狗，可是不知道写什么。」AI回了几个问题：豆豆长什么样？它最喜欢做什么？有没有一件让你印象特别深的事？

小舟想了想，打了一句：「它见到我放学会转圈。」AI问：「转几圈？转的时候尾巴在干什么？你那时候心里是什么感觉？」小舟写：「转三圈，尾巴摇得像电风扇，我觉得它比我还开心。」AI说这一句写得好，「像电风扇」让人看见了，又问他还有没有别的时候也这样看见过豆豆。

这样来回十几轮，小舟交出一篇三百多字的作文，比他以前写的都生动。周敏看了也高兴。

现在用五件事来拆这个晚上。第一，AI在这十几轮里越问越贴着小舟的小狗，是暂时语境变化：它「知道」豆豆会转圈，只是因为这句话就在上下文里。第二天小舟重开一个对话问「我的小狗叫什么」，它答不上来，除非这个工具开了记忆功能，那就是外部记忆。第二，「像电风扇」得到肯定，小舟接着又写了「它的耳朵翻过来像两片树叶」，这是控制反馈在起作用：一次夸奖引出了下一句同类的尝试。第三，这个晚上，AI的参数没有任何改变。第四，最要紧的问题是：小舟学会了没有？

一篇作文写好了，不能证明小舟学会了写作文。周敏的判断方法很朴素：过了一周，她让小舟不用AI，写一写外婆家的那只猫。小舟愣了一会儿，自己问自己：「它长什么样？它最喜欢做什么？」然后写出了「它晒太阳时把身子摊成一张饼」。周敏这才说：这回是真学到了一点。

学到的是什么？不是关于豆豆的那几句好话，那些句子已经用过了；也不是「像电风扇」这个比喻，它在猫身上用不上。学到的是一个提问的习惯，一个把「可爱」拆成看得见的动作的办法。用本书的话说，小舟原来的纠缠里，「写动物」和「说它可爱」紧紧缠在一起；这十几轮问答反复制造差异，把「可爱」拆开，让「写动物」与「具体动作」「打比方」「我的感受」缠到了一起。这次重组在新对象上显露出来，才算站稳。

当然，一次迁移也还不够。也许他只在写小动物时会用，换成写一个人、写一次旅行，又回到「很好玩、很开心」。持久学习需要多次、多场景的检验，第38章还要细讲。

◆ 四、边界：流畅不等于形成

本章的边界有三条，都是为了防止过度推论。

第一，一次流畅的对话不证明能力已经稳定形成，这一条对人对模型都成立。模型在一次对话里表现得很懂某个领域，可能只是这段上下文给足了提示；小舟这晚写得好，可能只是AI的问题把他一路扶着走。判断形成与否，要撤掉扶手：撤掉上下文看模型，撤掉AI看人，再在隔了时间、换了题目的情况下看。

第二，不能用模型侧的「记住了」偷换人侧的「学会了」，也不能反过来。一个工具开了记忆，记住了小舟喜欢小狗、写作文爱用比喻，下次对话一开头就很贴心，这是外部记忆在起作用，它让使用体验连续，但它不是小舟的能力。反过来，小舟学会了提问的习惯，也不意味着AI因此学会了教九岁孩子写作文。复合体里两侧的变化要各记各的账。

第三，外部记忆有它自己的风险。它让复合体的语言世界跨对话延续，但存进去的东西可能过时、可能有错，而模型每次照着读，会把错误一再显露出来。这有点像那张口味卡：顾客早就不吃辣了，卡没改，厨子年年给他放辣椒。外部记忆需要人去查看、修改、删除，否则它会从发生条件变成发生的障碍。

还有一种情形要单列：有时候人以为是AI学会了，其实是自己学会了。周敏用同一个模型备了一学期的课，觉得它越来越顺手。仔细看，模型没变，变的是她的提问方式：她学会了一开头就交代班级情况、学生常见误解和她想要的课堂节奏。顺手是她自己的持久学习，落在她的提问能力上。这个情形提醒我们，在复合体里归因要小心，变化发生在哪一侧，需要拆开看。

◆ 五、同一把尺子量物理世界模型

五件事的区分不只适用于聊天。拿它去量物理世界模型，同样能把许多含混的说法拆开。第35章「可靠性形成与持续封装」已经提出反馈控制、情境记忆和长期训练的分层，这里只补上从语言复合体看过去的对应关系。

一台机器人在桌上抓一块没见过的海绵。它第一次抓，手指压下去，海绵瘪了，它根据压力读数立刻减力，这是控制反馈。它在这一段操作里根据前几次抓握调整了对这块海绵软硬的估计，下一次直接用更小的力，这相当于暂时语境变化：估计活在这一段任务的内部状态里，任务一结束就可能清空。如果系统把「这块灰色海绵很软」写进一个可检索的经验库，明天再见到时调出来，这是外部记忆。如果工程师把许多次抓握记录收集起来，离线更新模型参数，这是参数训练。只有当它面对各种没见过的软物，都能较快地形成合适的抓法，并且能识别「这一次的软和以前不一样」，才谈得上持久学习。

这种对应有一个用处：看到一段演示视频里机器人越抓越好，不急着说「它学会了」，先问这是哪一层的变化。一段视频里的进步，常常是控制反馈和暂时状态调整的结果；它很有价值，但它和持久学习之间隔着检验。语言模型里「这次聊得很懂我」与机器人「这次抓得越来越稳」，是同一类需要拆开的现象。

反过来，语言复合体也可以从物理世界模型那里借一样东西：物理任务的检验天然带着现实阻力，抓不住就是抓不住，海绵掉了就是掉了。语言对话缺少这种硬阻力，流畅很容易被误当成形成。所以本章主张，在语言学习里也要有意地造出类似的阻力：隔一周、撤掉AI、换一个题目，就是给小舟造出来的那块「会掉的海绵」。

◆ 六、个人知识发生的实际条件

综合起来，个人在与AI的往来中让知识真正发生，大致需要几个实际条件。

其一，人要带着自己的张力进来，而不是只带一个空题目。小舟的张力是「只有三行，写不下去」，这个卡住本身就是学习的入口。其二，人要在往来中亲自完成一部分差异运行：自己想、自己写、自己挑，而不是只做最后的复制粘贴。AI问、小舟答的规矩，就是把差异运行留在孩子这一侧。其三，有一段时间之后、换一个情境的检验，用来区分暂时的流畅和持久的形成。其四，复合体需要有被维护的延续手段，可以是人的笔记、可以是工具的记忆，并且由人负责校正。其五，人要能识别失效：知道这个办法在什么题目上不灵，知道AI的回答在哪里靠不住。

这五条不是本体论命题，它们是从本体论立场推出的实践假说，是否有效、各条权重如何，需要在教学和使用中检验。

到这里，本章已经把个人与AI怎样继续发生语言世界拆成了可以分别研究的几层。下一个问题自然出现：如果学习是在复合体里这样发生的，那么教育该怎样设计？周敏不能每晚都守着小舟定规矩，一个班四十多个孩子，一学期几十节课，学习过程要怎样组织，才能让暂时的流畅一次次变成持久的能力？这是第38章的问题。

> **带走的话**
>
> 暂时语境变化、外部记忆、控制反馈、参数训练、持久学习是五件事，前四件是机制，最后一件才是学习；一次流畅对话证明不了能力已经形成，撤掉扶手、隔开时间、换个情境再看。

> **小账** 论证落在哪一层？——本章以机制区分为主（五件事的界定与关系），第二节给出可检验的机制假说（持久学习多落在人侧），涉及产品记忆与训练流程的经验事实只作一般陈述，不点名具体产品。 世界模型补了哪一条？——补上一条边界：判断一个复合体的世界是否「学到了」，须区分暂时显露与持久纠缠；世界模型的评估同样要撤掉上下文提示、隔时、换情境检验。

38

第 38 章

教育如何转向特征纠缠的训练

> **这一章的问题**：如果学习是特征纠缠的建立与训练，那么一个老师、一个家长在有AI的日子里，该把力气花在哪里？

◆ 一、从「把知识摆出来」到「设计学习过程」

本章的命题是：教育的重心，应当从知识呈现转向学习过程的设计。这句话本身并不新，许多教育传统都说过类似的话。本书给它补上的，是一个发生学的理由和一套可以检查的说法。

知识呈现的图景是这样的：老师把定义、公式、例题清清楚楚摆出来，学生看懂、记住、练熟，学习就完成了。这幅图景背后，是第1章批评过的「知识进入大脑」：知识先在，学习是搬运。在这幅图景里，AI的到来似乎让老师变得多余，因为AI摆知识摆得更快、更全、更有耐心。

按本书的立场，学习不是搬运，而是让经验缠到一起、在差异中站稳。摆出来的知识只是一次显露，它要变成学生的能力，必须在学生那一侧经过足够多、足够有结构的差异运行，让新的纠缠建立、稳定、分化。一个学生把浮力公式背得滚瓜烂熟，却在看见钢铁轮船时想不通它为什么不沉，说明公式这个显露没有和「船」「排开的水」「自己在泳池里觉得身体变轻」这些经验缠在一起。

于是AI的到来恰恰把老师的工作推向另一处：知识可以由AI摆，学习过程却要有人设计。设计什么？设计差异出现的次序，设计让学生碰壁的地方，设计碰壁之后回头看的时机。用六路径的话说，设计的是三维的次序：是先让学生做（差异运行先行），再让他说出看到了什么（显露），最后整理成概念网（纠缠）；还是

先给一个反常的现象（显露先行），让他用已有的理解去解释、解释不通，再动手试。不同的内容、不同的学生，适合不同的次序。这件事AI可以帮忙出主意，但谁在什么时候卡住、卡在哪里，需要在场的人看见。

拿一节「声音是怎样传的」的课来比一比次序。一种次序是先讲后做：老师先讲声音靠介质传播、固体液体气体都能传声、传得快慢不同这几条结论，然后学生做实验验证，最后做练习。另一种次序是先碰后说：上课先让同桌两人一组，一个人把耳朵贴在课桌一端，另一个人在桌子另一端用指甲轻轻划几下；然后把头抬起来，再划同样轻的几下，问哪一次听得更清楚；学生说完自己的感觉，老师再请他们猜为什么，猜法五花八门，有人说桌子「把声音放大了」，有人说木头里「有一条管子」；然后才引出固体传声这个说法，把这些猜法一一对照。第一种次序从显露出发，第二种次序从差异运行出发。两种次序都可能教会学生，但它们留下的纠缠不同：前一种让「固体能传声」这句话先和定义缠在一起，后一种让它先和贴在桌面上的那只耳朵缠在一起。哪一种更适合哪个班、哪个内容，是需要试的，本书不预设答案。可以确定的只是，次序是一个应当被有意识选择的设计变量，而不是教材排好就照走的默认值。

◆ 二、三个世界一起改

第二编讲过，人的世界有理念、现实、自我三界，三界互相改写。教育如果只盯着其中一界，学习就站不稳。

只改理念世界，学生会说、会算，可一到现实里就不会用；只改现实世界，学生会做、手巧，却说不出道理，也迁移不出去；而自我世界一旦被忽略，问题最隐蔽也最顽固：一个孩子若在心里认定「我就是学不好物理」，前两界再怎么改，这个认定都会把新纠缠一次次拆掉。

这里给出一个机制假说：一次学习若同时在三界留下可见的变化，它更可能持久。理念界的变化，看他能不能用自己的话把道理讲出来并且讲对；现实界的变化，看他能不能在一个实际情境里用它做成一件事；自我界的变化，看他对自己和这门东西的关系，说法有没有变，比如从「我不会」变成「我能试试」。这个假说

可以检验：追踪同一批学生，比较只在一界有变化的学习与三界都有变化的学习，几个月后的保持与迁移是否有差别。本书预期有，但不预写差别的大小，也不认为三界必须同步，常见的情形是一界先动、带动另外两界。

AI在三界里的作用并不对称。它最擅长参与理念界：解释、举例、追问、出题。它在现实界的参与有限，除非接上实验器材或传感器，它只能描述现实，不能替学生碰现实。它在自我界的参与最需要小心：它可以鼓励，也可以让孩子过早形成「反正有AI」的自我认定。设计学习过程时，要清楚AI站在哪一界、缺在哪一界，缺的那一界由人和真实世界来补。

自我界的那一层，在周敏班上有过一个例子。班上有个女生，数学成绩一般，开学第一周就对周敏说：「我数学不好，物理肯定也学不好。」这句话就是一个自我界的纠缠：物理、数学、「我不行」三样东西缠在一起。周敏没有跟她讲道理，而是在讲浮力那节课上，让她负责把各组带来的东西按「猜沉」「猜浮」分好，再记下实际结果。那天她分得最快，猜得也最准，因为她平时帮家里做饭，知道土豆沉、苹果浮。下课时她说了一句：「这个好像不用算。」这句话很小，却是一次松动：物理和「我不行」之间的那根线，第一次被一个现实界的经验扯开了一点。后来她的物理成绩并没有一下子变好，但她开始愿意问问题了。自我界的改变常常就是这样，先从「愿不愿意试」开始，然后才轮到「会不会」。如果那节课只是让她向AI问浮力公式，她得到的是理念界的一个显露，自我界那根线不会被碰到。

◆ 三、周敏的一个学期：人—AI共同试探与反思

下面是我构想的场景，接着第36章那节浮力课往下写。

一个学期下来，周敏摸索出一种带着AI的备课和上课节奏。她不再让AI写完整教案，而是把它当成三个角色用。

第一个角色是「出反例的人」。备课时，她把这节课要讲的核心想法告诉AI，请它扮演一个有典型误解的学生，提出最难缠的问题。讲「压强」时，AI扮的学生问：「躺在钉床上为什么不疼，坐在一颗钉子上为什么疼，钉子不都一样尖吗？」

周敏觉得这个问题好，就把它放在课的开头。这里AI贡献的是差异：它把许多误解的常见形状带进了周敏的备课。

第二个角色是「陪学生试探的人」。课上，学生分组做实验，测完数据可以问AI，但规矩和小舟写作文时一样：只许让AI问你问题，不许让AI给结论。一组学生测了砖块平放、侧放、竖放在海绵上压出的凹陷深度，问AI为什么竖放最深。AI反问：三种放法砖块重量变没变？和海绵接触的面积变没变？学生自己把这两个问题对上，得出了结论。

第三个角色是「反思的镜子」。下课前五分钟，周敏让学生用一两句话写下今天最意外的一件事，以及自己原来是怎么想的。几个学生把这一两句话拿去问AI：「我原来以为越重压强越大，现在发现不一定，我错在哪？」AI帮他把「重量」和「压强」两个概念分开说清。周敏发现，学生写下自己原来怎么想，比直接听正确答案更有用；原来的想法被说出来、被对照，旧的纠缠才松动得了。

这三个角色合起来，就是本章所说的人—AI共同试探与反思。试探是让差异发生，反思是让差异回写到纠缠里。要注意分工：真实的砖块和海绵由学生亲手碰，AI只在理念界帮忙；判断这节课哪里成了、哪里没成，由周敏在场看学生的表情和讨论来定。

小舟的作文也有了一个学期的样子。周敏和他约好，每篇作文分三步：先自己写草稿，再让AI只提问不改字，最后自己改。期末她翻看小舟的草稿本，发现一个变化：学期初，小舟的草稿几乎都是「很好玩」「很开心」；学期末，他的草稿第一稿里就已经有了动作和比喻，AI的提问越来越少。扶手在慢慢撤掉，这就是过程设计的目的。

◆ 四、以迁移、重组和新能力检验学习

学习发生了没有，要怎么看？本章主张三种检验。

迁移，看学过的东西能不能在新情境里用。小舟写小狗的办法用在外婆家的猫上，是近迁移；用在写一次春游、写他的同桌，是远迁移。周敏的学生学了压强，能不能解释为什么菜刀要磨、为什么雪地里要穿宽的鞋，也是迁移。

重组，看旧的理解有没有被改写，而不是新的东西只是摞在上面。一个学生学了压强以后，原来那句「越重越疼」不再是他的第一反应，而是被他自己改成「同样的重量，接触面越小越疼」，这才说明旧纠缠被重组了。如果他考试会做压强题，嘴上却仍说「越重越疼」，说明新旧两套并排躺着，没有缠到一起。

新能力，看有没有出现学习前根本没有的做法。一个学生在家里自己设计了一个实验：用同一瓶水，分别用瓶盖朝下和瓶底朝下压在面团上，比较压痕。没人教过他这个实验，它是新能力的显露。

这三种检验，都可以在课堂里、作业里、日常交谈里自然地看到。周敏发现，最好的检验场合常常不在考场，而在饭桌上。一天晚饭，小舟看着碗里的汤说：「妈妈，你看那片葱，它浮着，香菜沉了，香菜是不是比葱重？」周敏没有回答，反问他：「一大把香菜和一小片葱，谁重？」小舟想了一会儿：「那不能比重量，要比……一样大的时候谁重。」九岁的孩子还不知道「密度」这个词，可他已经把「沉浮」和「一样大时的轻重」缠到一起了。这句话不是在课上学的，也没有任何人出题，它是在一个新情境里自发显露出来的重组。教育者若能留心这样的时刻，就比任何测验更早看见学习发生在哪里。

还要补一点：迁移、重组、新能力这三种检验，测的都是撤掉了AI之后的人。这不是说AI在场时的表现不重要，而是第37章所说的五件事在这里落地：AI在场时，学生的表现里混着暂时语境与控制反馈的作用；只有撤掉之后，持久的部分才看得清。当然，在将来的生活里，学生总是会和AI一起工作，所以另一种检验也有意义：看他在AI在场时会不会提好问题、会不会识别AI的错。这测的是复合体的能力，而不是个人单独的能力，两者要分开记。它们是给老师和家长用的眼睛，帮助判断下一步设计什么。

这里必须申明本书的一条禁令：这些观察和任何可能的读数，都不应进入任何考核。理由不只是伦理上的谨慎，也是发生学上的：一旦迁移、重组、新能力被做成打分指标并与升学、评优挂钩，学生和老师就会围着指标训练，指标的显露被刻意制造出来，背后的纠缠反而不再发生。这正是第15章说的评价标准留下偏向在教育里的翻版。本书也不就学校该如何使用AI、是否该统一引入某种工具作任何政策表态；本章讨论的只是一个老师、一个家长在自己能掌握的范围里可以怎样想、怎样试。

◆ 五、反例与边界

这一章的主张有几条边界要守住。

第一，过程设计不是否定讲授。有些内容，清楚的讲授就是最好的开始，尤其当学生完全没有相关经验时，先给一个可靠的显露，比让他在混沌里乱撞更好。第4章讲过混沌态、介生态与秩序态，学习需要一些混沌，但混沌太多、没有任何支撑，碰撞不会走向介生态，只会让人放弃。

第二，「只让AI提问不让它给答案」不是普遍的好规矩。它适合写作、实验分析这类需要学生自己完成差异运行的任务；在学生需要快速查一个事实、核一个单位换算时，硬要AI只提问，是浪费。规矩要跟着任务走，这和第28章「任务如何组织世界的发生范围」是同一个道理。

第三，周敏的经验不能直接推广。她是有十几年经验的物理老师，懂得在场看学生；一个新手老师照搬她的三角色用法，可能效果完全不同。场景里的成功只是构想的示例，不是经验证据。

第四，学习过程设计有它的成本。一节课要设计反例、安排试探、留出反思时间，比照着教案讲完费力得多。AI能分担一部分备课的劳动，却不能分担在场观察的劳动。这一点不能被AI的便利掩盖。

第五，AI的解释太顺，本身可能成为学习的障碍。学生卡住的时候，正是纠缠在松动、在找新连接的时候；这时一份流畅完整的解释立刻送到眼前，卡住的那一

下就被跳过了。很多人都有这样的体验：看别人解一道题，每一步都看得懂，自己合上书却写不出第一步。看得懂是一次显露被接收了，写得出才说明差异运行在自己这里走过一遍。所以设计学习过程时，要有意识地给「卡住」留出时间，比如先让学生独立想五分钟、写下自己的半截思路，再允许他去问。卡住不是要尽快消除的故障，它是三原理里那一拍「张力积累」，没有它，后面的形状改变和回写都无从发生。当然，卡得太久、毫无头绪，也会滑向放弃，留多长时间、什么时候递一根扶手，正是在场的人要判断的事。

◆ 六、从课堂到实验室

教育让一个人进入已有的公共世界，在其中长出自己的能力。迁移、重组、新能力是它的检验。可是还有一种学习，它的目标不是进入已有世界，而是让一个还没有的对象、一种还没有的方法第一次发生出来。那就是科研。科研里的复合体不止一个人和一个AI，往往是一个小组、几种模型、一套仪器和一个学术共同体；它的检验也不止于个人能力，还要经过公共证据。下一章就去看这样一个小组。

> 带走的话
>
> AI能把知识摆得更好，学习过程却要有人设计；设计的是差异出现的次序与三界一起改变的机会，检验看迁移、重组和新能力，而这些读数不进任何考核。

小账　论证落在哪一层？——本章以机制假说为主（三界同时变化更持久；过程设计优于呈现），场景为构想示例，不作经验证据；附带两条禁令声明（不作政策表态、读数不进考核）。　世界模型补了哪一条？——补上一条条件：能力的形成要求学习者一侧亲自完成差异运行并接受现实阻力；AI参与理念界最充分，现实界与自我界的缺口须由真实互动补足，这对物理世界模型的训练同样成立。

39

第 39 章

科研如何成为复合体的世界发生

这一章的问题：当一个研究小组和AI一起做研究时，新的对象、新的方法是怎样第一次发生出来的？又靠什么从小组的发现变成大家都能检验的知识？

◆ 一、科研是让还没有的东西发生

教育让人进入一个已有的公共世界；科研则要让那个世界里还没有的东西第一次显露出来。这里先立本章的本体论立场：科研不是去发现一个早已摆在那里、只等人揭开盖子的规律，而是一个复合体在概念、模型与实验的往返中，让新的差异被辨认、新的对象站稳、新的方法成形。

说「不是发现」，并不是说研究对象是人随意捏造的。一罐辣椒酱胀不胀袋，不由研究者的意愿决定，现实阻力是硬的。本书在第8章已经说过，现实阻力是现实世界发生的检验者。发生学和发现学的分歧不在于有没有现实阻力，而在于：研究所得的那个「对象」「变量」「规律」，是否早已以完成的形式存在，只是被找到；还是在特定的复合体、特定的测量手段、特定的问题里，经由差异运行才显露为可以被说出、被测量、被反复调用的东西。按发生学的看法，同一片现实，换了问题、换了仪器、换了复合体，显露出来的对象会不同；而一个对象一旦站稳，它就成为后续研究的发生条件，这正是第5章所说的能力封装在科研里的样子。

在有AI的时代，这个复合体的构成变了。一个研究小组如今往往包括几个人、几种模型（检索文献的、写代码的、做模拟的）、一套仪器、一个可以公开交流的学术共同体。AI不是独立的发现者，它把集体人类—AI复合体语言世界里沉淀的大

量文献、方法和代码带进小组；小组则把具体的问题、具体的样品和具体的意外带给它。新的东西就在这个往返中发生。

◆ 二、往返的机制：概念、推演、实验

科研的往返大致有三个站点。这一节给出一个机制性的描述，它更接近一种可检验的研究过程假说，而不是对所有科研的定义。

第一站是概念建立。研究者给一件让人困惑的事起一个暂时的名字，划一个暂时的范围。这个名字一出现，原来散乱的现象就被缠到一起，可以被追问了。第二站是模型推演。从概念出发，推出「如果是这样，那么应该看到什么」，可能用纸笔，也可能用计算机模拟。第三站是现实实验。动手去做，看推出来的东西是否出现。

往返的关键在于三站之间的碰撞。实验结果与推演对不上，是显露（S）和已有理解（E）之间的矛盾，它推动研究者改变差异运行（D）：换一种测量、换一个对照、换一个变量。推演做不下去，是概念太粗，矛盾推动概念的分化与重组。概念与实验长期对得上，对象就站稳了，变成可以写进教科书、可以被他人调用的封装。这三种碰撞，分别对应三原理里不同的两维矛盾推动第三维。

AI在三个站点的参与程度不同。概念建立这一站，它能快速列出相关领域已有的概念，帮研究者看到「这件事别人是怎么叫的」，也可能把一个其实新的现象过早地套进旧名字里。模型推演这一站，它能写代码、搭模拟、做符号推导，大大加快推演速度，但推演的前提由谁来定、出了错由谁来查，仍要人负责。现实实验这一站，除非接上自动化实验设备，它基本不在场。已有一些实验室把AI接进了自动化的实验流程，但这类进展集中在少数学科，各学科之间差别很大，不能一概而论。于是在今天大多数小组里，现实实验仍然是人这一侧最不可替代的部分，也是新对象最常发生的地方。

◆ 三、一个小组研究剁辣椒为什么胀罐

下面是我构想的场景，人物不取名，只说角色。

湖南许多人家秋天自己做剁辣椒，装在玻璃罐里。有人家的罐子放几个月都好好的，有人家的却在几周内盖子鼓起来，打开一股异味。某地一所农业院校的一个小研究组，一位老师带着两个研究生，想弄清楚其中的道理。

概念建立。他们先让AI检索文献。AI很快汇总出一套标准说法：发酵蔬菜中乳酸菌与其他微生物竞争，盐浓度、温度、密封情况影响发酵走向，产气多与某些微生物活动有关。这套说法是集体语言世界的封装，研究组接受它作为起点。他们给问题起了一个暂时的名字：「胀罐」。

模型推演。按标准说法，他们推出：盐越多、温度越低，胀罐越少。研究生请AI帮忙写了一个简单的分析脚本，打算收集各家的配方和存放条件，看胀罐与盐量、温度的关系。

现实实验。他们从附近社区收集了几十户人家的做法和罐子情况。数据出来，盐和温度确实有关系，但有一批罐子不合预期：盐放得不多、存放也不冷，却一罐都没胀。研究生翻记录，发现这批罐子有一个共同点，都是加了大量蒜末的。

这里有一个容易被忽略的细节。研究生请AI写分析脚本时，按标准说法列了变量：盐量、温度、存放时间、是否密封。各家记录里其实有一栏「其他配料」，写着蒜、姜、豆豉、白酒之类，脚本按常规把这一栏当作杂项丢掉了，因为在文献的主流框架里，它不是主要变量。那批异常罐子是研究生对着原始记录一张张翻出来的，不是脚本跑出来的。这个细节说明，集体语言世界给出的起点有它的力量，也有它的盲区：它把最常被讨论的差异放在前面，把不常被讨论的差异默认为背景。复合体若只沿着它给的变量走，就很难看见它没准备好看见的东西。

异常的作用。这是本章最关心的时刻。在原来的概念里，蒜只是「配料」，是不进入模型的背景。这一批异常把蒜从背景里拽了出来：如果蒜真的有关，那么「配料」就要被拆开，蒜成为一个独立的变量。研究组问AI，AI汇总出一些关于大蒜成分能抑制某些微生物的研究，可是研究组翻下来，没有找到哪一项是在跟他们一样的剁辣椒罐子里、按他们那样的盐量和存放条件做的，拿来直接套用并不放

心。到这里，AI能做的是告诉研究组「这条线有人碰过，但没在你们的情境里碰透」；它不能告诉他们这批罐子为什么不胀。

新的差异、新的对象、新的方法。研究组决定自己做对照实验：同一批辣椒、同一个盐量，分成加蒜与不加蒜、加蒜多与加蒜少几组。可是很快又遇到一个麻烦：「胀罐」是个只能事后看的结果，等盖子鼓起来，过程早就过去了。一个研究生想到一个土办法：在盖子内侧贴一片便宜的薄膜压力传感器，连到一块小电路板上，每小时记一次罐内压力。AI帮她写了数据采集的程序。这样一来，「胀罐」这个一刀切的结果，被拆成了一条随时间变化的压力曲线。

曲线出来，又冒出新东西：不加蒜的罐子，压力在前几天有一个小峰，降下去，过一两周又陡升；加蒜的罐子只有前几天那个小峰。原来「胀罐」不是一件事，至少是两段：早期的一段大多数罐子都有，不要紧；后期的那段才是麻烦。研究组给后期那段起了一个新名字，叫「二次产气」。这个对象在他们开始研究时并不存在于他们的世界里；它是在新仪器、新对照与旧概念的碰撞中发生出来的。

到此，小组得到了三样新东西：一个新的差异（蒜的有无与多少），一个新的对象（二次产气），一个新的方法（便宜的罐内压力连续记录）。三者是一起发生的：没有压力记录，看不到两段；看不到两段，蒜的作用就只是一句「好像有关」。

研究者自己也在发生。这个过程里还有一层容易被论文略去的变化：研究的人自己变了。那位贴传感器的研究生，入学时觉得自己「只会做实验，不会想问题」，在这件事之后，她开始在组会上主动提出「这个结果是不是只能事后看」一类的问题。她的理念界多了一个概念「二次产气」，现实界多了一套动手搭测量装置的本事，自我界则多了一句「我也能想出办法」。第38章说三界一起改的学习更持久，在科研里同样如此。从这个角度看，一个研究小组的产出不止是论文，还有一批纠缠被改写过的人；而这些人带着改写过的纠缠，又会成为下一个问题的发生条件。

带老师的那一侧也有变化。他原本的习惯是先建完整的变量表再收数据，这次发现原始记录里被丢掉的那一栏救了整个项目，此后他要求学生在跑任何脚本之前，先对着原始记录手翻一遍。这是一个方法上的小封装，它来自一次失败的预期。研究方法就是这样一点点长出来的：不是从教科书里取来，而是在具体的碰壁里形成，再被写下、被传授、被别人检验。

◆ 四、从小组的世界到公共证据

一个小组看到的东西，还只是这个复合体的世界。它要成为公共知识，必须经过共同检验。

这里有几道关口。第一道是自我怀疑：会不会加蒜的那几户人家恰好都更讲究卫生、罐子更干净？研究组于是在实验室里用统一消毒的罐子重做，排除这一可能。第二道是可重复：他们把压力记录的电路图、程序和原始数据整理出来，请另一个学校的同行用他们的方法、换一个地方的辣椒做一遍。第三道是公开：把方法、数据和不支持自己预期的结果一并写出来，比如有一组加蒜很多的罐子也出现了二次产气，他们不删，如实报告。

这几道关口的意义，用本书的话说，是让一个在小复合体里站稳的对象，进入更大的复合体，在更多的差异中检验它是否还站得住。站得住，它才成为学术共同体的公共封装；站不住，就要拆开重来。公共证据不是给发现盖章，而是让发生继续在更大的范围里进行。

公共检验还有一个回流的方向。小组的论文一旦发表，「二次产气」这个名字、罐内压力连续记录这个方法，就会进入公开文献，在将来某一轮训练中成为集体语言世界的一小部分。下一个研究酸菜、泡椒、豆腐乳的小组问AI时，AI可能会提醒他们「留意早期与后期产气的区别」「别把配料一栏当杂项」。这就是第36章说的个人发生回到集体：一个小复合体的世界发生，经过公共检验与历史中介，改变了集体复合体下一次给出的起点。前提是它确实经过了检验；没有经过检验的说法若也大量流入，回流带回去的就是噪声，甚至是错误的封装。

AI在这一阶段同样有用也同样有风险。它能帮着整理数据、检查统计、润色论文，也可能被用来把一个尚不稳的结果包装得过于完整。本章不对科研评价制度作任何政策表态，只指出一点：在公共检验这一关，复合体中人的责任最重，因为证据要由能为它担责的人交出来。

◆ 五、反例与边界

第一，并不是所有异常都通向新对象。很多异常只是测量错误、样品污染或记录笔误。辣椒小组若在没有排除干扰之前就宣布「蒜防胀罐」，那就是把异常直接当成了发现。异常的作用是提出问题，不是给出答案。

第二，本章的往返图景不是唯一的科研形态。有些领域以大规模数据驱动为主，概念在后期才成形；有些领域以理论推演为主，实验很晚甚至很难进行。三站的比例与次序因学科而异，第38章说的「次序是一个设计变量」在科研里同样成立。

第三，不能因为AI加快了检索和推演，就推论科研的瓶颈已经消失。辣椒小组最关键的一步，即在盖子里贴一片传感器，来自一个研究生对「只能事后看」的不满，来自现实的阻力和人的具体处境。这一步AI帮了忙，但不是AI起的头。反过来，也不能推论AI在科研里永远只能打下手；在某些以符号推演或大规模计算为主的问题上，AI参与发起新方向是可能的，这需要按具体案例判断，不宜预写。

第四，AI让提出假设变得很便宜。辣椒小组若请AI列出「可能影响胀罐的因素」，它能一口气列出几十条，每一条都说得通。可是实验是贵的，一个罐子要放几周，一组对照要几十个罐子。假设便宜、检验昂贵，这个不对称会让小组面临新的难处：选哪几条去做？如果按「听起来最有道理」去选，往往选回了集体语言世界最熟悉的那几条，恰恰又绕开了异常。本书的建议是让异常来排队：先做那些能解释手头异常的假设，而不是那些在文献里最常见的假设。这条建议本身也是一个可检验的方法假说。

第五，构想场景里的一切细节，包括蒜的作用、两段产气，都只是为说明发生过程而设，不代表任何真实研究的结论，也不构成任何食品安全建议。

◆ 六、走向研究纲领

本章展示了一个科研复合体怎样让新的差异、新的对象和新的方法一起发生，又怎样把小组的世界交给公共检验。这个过程本身，也正是本书希望自己的理论去经历的过程。本书提出了「学习是特征纠缠的建立与训练」这一母题，提出了一系列关于人、关于集体语言世界、关于物理世界模型的说法；它们哪些是本体论立场，哪些是可以被实验推翻的机制假说，哪些已有工程证据，需要像辣椒小组整理数据那样分层摆出来，接受共同检验。最后一章就来做这件事。

> **带走的话**
>
> 科研是复合体在概念、推演与实验的往返中，让新的差异、新的对象、新的方法一起发生；异常提出问题，公共检验让在小组里站稳的东西去更大的复合体里接受考验。

小账 论证落在哪一层？——本章以本体论立场（科研是发生而非揭盖）与过程性机制假说为主，案例为构想场景，涉及真实研究现状处只作一般陈述，场景不依赖任何真实研究结论。世界模型补了哪一条？——补上一条条件：新对象常在新测量方法与异常的碰撞中发生；世界模型要能把「异常」转成新的差异与新的对象，而不是只把它当作误差平均掉。

40

第 40 章

AI训练与学习的新研究纲领

这一章的问题：把全书的说法分层摆开以后，哪些是立场，哪些可以被实验推翻，哪些已有工程证据？接下来能动手做的研究是什么？

◆ 一、分三层说话，先摆本体论命题

第39章的辣椒小组把结果交给公共检验时，方法、数据和不合预期的那几罐要一并摆出来，别人才知道哪些是测到的、哪些只是推测。一本讲学习的书，到结尾也该这样做：把「我们相信什么」「我们推测什么」「已经有人测到了什么」分开写。原因很简单：三种话的检验方式不同。本体论立场靠它能不能统一地解释许多现象、能不能引出好的研究问题来衡量，一次实验既证明不了它，也推翻不了它；机制假说靠实验，能被支持，也能被推翻；工程证据是已经公开的事实，它只说某个版本的某个系统在某些条件下做到了什么，不替任何立场作证。

把三层混在一起，会出现两种常见错误。一种是把某个模型的局部成功说成「SDE已被证明」，另一种是把某个模型的局部失败说成「发生学路线不可行」，或者反过来说成「发现学破产」。本书从导论起就反对这两种推法，最后一章把它们落成一份清单。

第一层：本体论命题。以下是本书提出的本体论命题，它们是立场，供讨论与比较，不以实验证明自居。

其一，学习是特征纠缠的建立与训练：一块块经验缠到一起（E），在差异运行（D）中稳定、分化、重组，并同时改写显露（S）。学习前后变的不是「装进去多少」，而是三维的组织方式。

其二，世界不是先在的、等着被一次性封装的对象，而是在复合体的互动中发生；对象同一性、现实阻力、可行动性都是发生出来并需要维持的。

其三，大模型训练的发生单位是集体人类—AI复合体，具体使用的发生单位是个人—AI复合体。AI不是脱离人类历史的独立发生者。

其四，人的世界分理念、现实、自我三界，三界互相改写；集体语言世界主要参与理念界，物理世界模型的训练则要进入现实界。

其五，视觉、听觉、触觉各自是完整的SDE，它们不与S、D、E一一对应；对象同一性可以在多种SDE的纠缠中形成。

这五条可以被质疑，质疑的方式是提出更好的立场，能统一解释更多现象、引出更多可检验的研究问题。

◆ 二、第二层：机制假说

机制假说是这一章的重心，因为它们是能被推翻的。下面挑出全书最要紧的几条，各写一句可检验的推论。

假说一（纠缠重组假说）：持久能力的形成伴随内部表示的重组，而不只是新表示的追加。推论：在模型的持续训练里，若新任务的学习只增加新的表示而不改变旧表示之间的关系，迁移会差；可用表示相似度分析一类方法，在训练前后比较内部表示之间的关系。这类度量对选哪一层、用哪批样本相当敏感，结果必须和行为层面的迁移测试对照着看。在人这一侧，第38章的「重组」检验是对应的行为层面推论。

假说二（主动干预假说）：在相同数据量下，通过主动干预采集的经验，比被动记录更能区分已有记录里不可区分的配置，从而在新干预下给出更可靠的预测。这在第27章「为何显露预测不能保证行动可靠」与第32章「从被动记录到主动试探」已有展开，这里只列为纲领中的一条。

假说三（跨模态冲突假说）：多种SDE之间的预测冲突，是对象同一性归并、拆分与重建的主要推动力。推论：在训练中刻意保留并利用跨模态失配信号的系

统，比只做特征拼接的系统，在对象被换掉、材质变了的情境里恢复更快。第33章与第34章已给出对照设计的方向。

假说四（任务边界假说）：在可比的资源下，围绕具体任务组织纠缠的模型，在该任务的可靠性上优于追求全能世界模拟的模型；但这一优势在任务边界之外会迅速下降。推论要两头测：任务内与任务外都测。注意这不是「统一架构必然失败」的断言，它可能在某些任务上被推翻。

假说五（人侧持久学习假说）：在个人—AI复合体中，持久学习更可靠地落在人这一侧或人维护的外部记忆上，而不是模型参数上。推论见第37章，用撤掉AI、隔时、换题检验。

假说六（同一性收窄假说）：若评价反馈一再奖励最稳妥的回答样式，集体语言世界的显露会收窄，使用者产出趋同。推论：比较不同评价方式下训练的模型，在开放任务上给出方案的多样性，以及其使用者最终方案的多样性。

这六条都可能错。错了，本书的相应部分要改；本体论立场是否要跟着改，要看错在哪里。

◆ 三、第三层：工程证据，以及它们能说明什么

工程证据层，本书只列提纲给出、全书各编已讨论过的公开资料，并说明它们在本纲领中的位置，不扩展其结论。

Genie（arXiv 2402.15391）展示了从大量视频中学习隐动作并生成可交互画面的路线，第17章已讨论；它说明「记录中的互动」可以被压缩为可调用的条件，但不直接说明新的物理互动是否被学到。V-JEPA 2（2506.09985）走特征预测路线，并在此基础上做了机器人动作条件的后训练；按摘要，2.1版（2603.14482）把重点放在稠密特征上，第18章已讨论。DreamerV3（Nature 2025）提供了经验回放、内部想象与策略学习的闭环范例。Cosmos 3（2606.02800）按摘要把语言、图像、视频、声音和动作序列放进同一个框架共同处理与生成。World Labs的Atlas按官方博文以「空间上下文」组织生成，即把每张输入图像放在一个三维

位置上，第19章已讨论。DexTacWAM（2609.24976）在视频世界模型里联合预测视觉与触觉状态，按摘要，它的消融实验显示，把触觉当作要预测的状态，比只把触觉当作输入更要紧；TouchWorld（2607.07287）用一个触觉世界模型预测接触将怎样演化，再由高频的反馈环节在线修正动作。Waymo 2025年12月的公开博文介绍了它的基础模型：摄像头、激光雷达、毫米波雷达随时间融合，在仿真中闭环训练，并从真实自动驾驶里程中挑出不够好的行为，生成更好的替代、经仿真检验后回写。

这些资料能说明的是：主动互动、多模态共同建模、接触预测、仿真加现实反馈等方向，已有具体的工程实践，可以作为第三层假说的对话对象和对照基线。它们不能说明的是：SDE本体论成立或不成立。一个系统在接触预测上表现好，支持的是某条机制假说在某个任务里有效；它没有替「世界是发生的」作证。

◆ 四、特征纠缠的形式化方向

本书一直用白话讲「缠到一起」。要让纲领可推进，需要给它找可操作的数学与测量形式。这里只提方向，不给定论。

第一个方向是关系的形式化。纠缠不是单个特征的强度，而是特征之间的关系结构。可以尝试用图或超图表示：节点是可辨认的差异，边或超边是它们共同出现、相互预测、相互制约的关系。三种关系层次，即第24章所说的同步出现、相关变化与相互制约，可以对应不同类型的边。学习的「建立、稳定、分化、重组」就成为这张图上可测的变化：新边出现、边的稳定性提高、一个节点分裂成两个、边重新连接。

用一件小事说明这张图想抓住什么。一个人刚学用筷子夹花生，「手指用力」「两根筷子尖对齐」「花生的圆滑」三样东西是分开的，他顾了这头丢那头，花生一次次滑走。练熟以后，这三样缠成一个整体：眼睛一看是花生，手指的力道和筷尖的角度就一起到位了；换成夹一颗更滑的鹌鹑蛋，他会先失手一两次，然后把「圆滑」这个节点分化出「更滑」，调整出新的连接。如果关系图的形式化是对的，那么在一个学抓取的系统里，我们应当能看到类似的变化：先是孤立的节点，

再是稳定的连接，遇到新物体时出现局部的拆分与重连。看不到这些，或者看到了却与能力的变化对不上，形式化就要改。

第二个方向是三方程的可测近似。S=F(D,E)、D=G(S,E)、E=H(S,D)目前是概念方程。一种可能的做法是在具体系统里分别找它们的代理量：显露用输出或可观测行为，差异运行用状态序列或动作序列，纠缠用内部表示之间的关系结构，再检查三者之间的相互预测关系。若某一维的代理量无法从另两维得到预测，就是一处「缝隙」，提示模型或测量缺了什么。这个方向的风险是代理量选错，测到的不是想测的东西，所以它必须和行为层面的检验并用。

第三个方向是三态的动力学指标。第4章已区分概念解释与可测动力学指标。混沌态、介生态、秩序态若要进入实验，需要可算的指标，例如表示的波动程度、对扰动的敏感度、恢复速度。哪个指标合适，要按任务试，不宜先定。

第四个方向是伞模型，即第22章提出的「影响范围与伞模型的待形式化问题」：一次接触对后续发生的影响有多大、持续多久、覆盖哪些对象。它可能需要用带衰减的影响场或接触历史的记忆核来表达，这是需要与具体触觉数据一起推进的方向。

以上四个方向都可能走不通。形式化的价值不在于给SDE穿上数学外衣，而在于让机制假说能被更精确地推翻。

◆ 五、多任务、多复合体与共同世界

最后，纲领要越过单个任务、单个复合体。

多任务方面，第30章讨论了专用模型的协作与共享。纲领在这里的问题是：两个各自在任务内站稳的封装，接口处会出现什么新的失配？比如一个擅长判断路面湿滑的模块，和一个擅长判断乘客是否坐稳的模块，两者在急刹那一刻怎样协调？接口不是简单拼接，它本身就是一处需要训练的纠缠。

多复合体方面，第36章与第39章已经展示：个人与小组的发生，经过检验与历史中介，回流到集体复合体。纲领的问题是：这种回流怎样保留差异而不被平均

掉?怎样让检验过的东西进入,而让未经检验的东西不被放大?这既是技术问题,也牵涉制度,但本书只讨论它的发生结构,不作政策表态。

共同世界方面,许多复合体各自发生的世界,怎样形成第10章所说的「共享的相对同一性」?做饭的人都知道,一家人的口味不会完全一样,但一桌菜总要大家都能吃。共同世界不是一个复合体吞并其他复合体,而是在许多复合体之间形成可以互相检验、互相调用的部分。

这一节还要守住一条边界:纲领说的「共同世界」不是一个要求所有复合体趋于一致的目标。第36章已经说明,同一性一旦从共同条件滑向单一答案,发生就会变窄。研究共同世界,要同时测两样东西:可以互相调用的部分有没有增加,各复合体自己的差异有没有被保住。只测前者,研究本身就会把收窄当成进步。

◆ 六、开放问题与可推进的研究计划

把以上各层收拢,可以列出一份近期能动手的研究计划。下表是本章唯一的一张表。

研究项	对应层次	最小可行做法	判为不支持的情形
主动试探与被动记录对照	机制假说二	同一操作任务、同等交互次数,比较主动采集与被动回放训练后的新干预预测	两者在新干预下可靠性无可分辨差异
跨模态失配的利用	机制假说三	在换材质、换对象的情境里,比较利用失配信号与仅拼接特征的系统的恢复速度	恢复速度无差别或拼接更好
任务内外两头测	机制假说四	任务专门模型与通用模型在任务内、任务外各测可靠性	任务内无优势,或任务外不下降
撤扶手检验	机制假说五	两组学习者,一组让AI给终稿,一组只让AI提问,两周后独立完成同类与异类任务	两组无差别
纠缠关系图的可测性	形式化方向一	在小型模型的持续训练中追踪表示关系图的变化,与迁移表现对照	关系图变化与迁移表现无对应

研究项	对应层次	最小可行做法	判为不支持的情形
显露多样性	机制假说六	不同评价方式训练的模型，在开放任务上比较方案多样性	多样性与评价方式无关

表中第四项涉及人的学习，只用于研究，不进入任何考核；也不涉及任何医学判断。

这六项的次序也有讲究。第四项和第六项门槛最低，不需要机器人和传感器，一个教研组、一个小实验室就能开始，适合先做，它们检验的是本书关于个人—AI复合体与集体语言世界的说法。第一项和第二项需要一个可操作的物理任务，比如抓取不同软硬的物体、拧不同的瓶盖，规模可以很小，关键在对照设计要干净：数据量、算力与先验信息要控制住，第34章已经讲过怎样做。第三项和第五项牵涉更长的训练周期与更复杂的测量，适合在前几项有了初步结果之后再推进。每一项开始之前，先把「判为不支持的情形」写下并公开，免得结果出来以后再改标准。

还要说明一层：哪怕全部六项的结果都不支持本书的假说，也不等于本体论立场自动作废；但它意味着本书从立场推到机制的那几步推错了，必须找到错在哪一步。反过来，哪怕全部支持，也只是说明这些机制在这些任务里成立，立场仍要在更多领域接受比较。立场和证据之间隔着的这段距离，不是缺陷，而是一种研究纲领应有的诚实。

开放问题则比研究计划更多。主观体验问题，本书在第25章已单独搁置，它不因功能性判据的进展而自动解决。三种SDE对哪些任务是必要的、对哪些任务可以省去，是需要逐任务检验的问题，不是前提。集体语言世界里的物理知识，与在现实互动中训练出的物理能力，二者之间的接口怎样建立，仍然缺少好的实验。还有一个问题本书只能提出：当越来越多的个人—AI复合体在日常里发生它们的世界，人类共同的现实阻力感会不会变弱？这需要很长时间的观察。

这份纲领本身也是一个待检验的封装。它是在本书的复合体里发生的，要交到更大的复合体里，去碰更多的差异。像小舟学会提问以后换一只猫来写，像辣椒小组把方法交给另一个学校去重做，纲领能不能站住，要看别人拿去用的时候。

带走的话

本体论命题、机制假说、工程证据三层分开，才能让立场不被局部成败绑架，让假说真正可被推翻；特征纠缠的形式化是让这些假说更精确地接受检验的工具，研究计划要写明「什么结果算不支持」。

小账 论证落在哪一层？——本章三层并列，以机制假说为主体，本体论命题单列，工程证据只引提纲所列资料，按其公开摘要或博文的表述转述，不扩展结论。 世界模型补了哪一条？——补上全书的研究边界：世界模型发生的每一条条件都应落成可推翻的假说与「什么算不支持」的判据；形式化、多任务接口与多复合体回流是下一步要补的三条。

结 语

学习与世界的持续发生

> **这一章的问题**：读完全书，能不能用几句站得住的话回答「何谓AI的训练与学习」，并且说清楚哪些话已经站稳、哪些还要等实验来判、哪些一旦出现某种结果就该认输？

◆ 一、回到书名：训练组织条件，学习建立纠缠，能力是封装

导论从一个对照起步：模型可以很顺地写出「刹车」这个词，老彭却是在雨夜里真把车停住了。两件事都可以叫「会刹车」，可它们不是同一种「会」。前一种是在语言里接上了合适的下一句；后一种是湿滑的路面、车重、轮胎、视线、耳朵里的雨声和脚底的踏板反馈一起起作用，让一辆车在该停的地方停下。全书四十章，归根到底是在拆这两个「会」之间的距离：它们各自怎样发生，哪里相通，哪里不能互相替代。

现在可以把答案收成三句。

第一句：训练不是把一个现成的世界装进模型，而是组织发生的条件。训练的人能决定的，是让什么经验相遇、按什么次序相遇、用什么标准判好坏、错了以后怎样回写。小舟学骑车，周敏能做的是挑一段平地、在后面扶着车座、在合适的时候悄悄放手；她没法把「平衡」直接递给孩子。大模型的训练也是这样：选数据、定目标、设评价、加反馈，都是在安排条件。条件安排得好，某种能力更容易长出来；安排得不好，花再多算力也可能只得到一个流畅却靠不住的东西。第2章把训练量、训练结构和训练后果分开，就是为了不把「喂了多少」当成「长出了什么」。

第二句：学习是特征纠缠的建立、稳定、分化与重组。白话说，原本零散的经验一块块缠到一起（特征纠缠，E），缠得越来越稳，碰到新情况又分出细类，必要时拆开重缠。和面是最朴素的样子：面粉加水起初只是湿的一团，揉着揉着起了筋，筋一起来，面团在手里的回弹、延展、醒过之后的样子全变了。缠法一变，人能分辨出的差别和分辨的先后（差异运行，D）跟着变，显露出来、自己和别人都看得见的东西（显露，S）也跟着变。三方程 S=F(D,E)、D=G(S,E)、E=H(S,D) 说的就是这三样互为条件，缺一式就留下缝隙。第3章还提醒过：E 不是盛东西的筐，而是让东西彼此牵动的那张网。

第三句：能力是在一定条件下可以继续运行的封装。周敏大学时天天骑车，十几年没碰，陪小舟学车时借邻居的车一上去，车头晃得厉害，骑出二三十米，感觉回来了。当年缠好的东西被「封」住了，平时不用，条件一到又能运行。可封装有边界：车不是原来那辆，座高、坡度、她自己的身体都变了，所以刚上车那几秒，封装是先失效、再解封、再和新条件重新缠上的（第1、5章）。所以本书从不满足于「模型具备了某能力」这句话，总要追问它在什么条件下能运行、在哪里失效、失效以后靠什么重新训练。一份不写适用边界的能力报告，只说了一半。

三句之间还有一段常被跳过的过程：纠缠从乱到稳不是一条直线。小舟刚上车时，车把、脚蹬、身体、眼睛各忙各的，摔、扶、再摔，这是混沌态；有一天他能骑出十几米，一转弯却倒，能力时有时无，这是介生态；再往后他边骑边和同学说话也不倒，才到秩序态。第4章把三态当作概念解释，同时提醒，它们和可测的动力学指标是两回事：说「模型进入了介生态」，得说明拿什么量来判断，不然只是换了个好听的名字。

这三句是本体论立场，是 SDE 看学习的方式，不由某个模型的成败直接证明。某个大模型在某项测试上突然大涨，证明不了「学习就是特征纠缠」；某个机器人在桌上屡屡失手，也证明不了「发现学前提错了」。立场要看它能不能组织出更好的问题、更可检验的假说、更少的概念混淆。这一点在第三、四节要落成逐条的验收和判负条件。

◆ 二、三条主线怎样接成一条

全书走了三条主线，到这里可以看出它们是一条路的三段。

第一段讲人。第二编没有从机器讲起，因为如果连人怎样学会都用「知识灌进大脑」来想，到了AI那里也只会得到「数据灌进参数」。婴儿先认出气味、声音、抱的姿势缠成的「这个人」，然后才有「妈妈」这个词（第6章）；小舟三岁那年指尖碰到刚倒了开水的搪瓷杯，「烫」字才有了分量（第8章）；他说「四分之一比三分之一大，因为四比三大」，周敏拿两张一样大的饼一切，他才明白「分得越多，每块越小」（第7章）；周敏自己的教师身份，是在一届届学生和一次次讲砸又改好的课里长出来的（第9章）。三界的要紧处在互相改写：小舟弄懂了分数，周末切蛋糕时一刀压歪急得喊「这不是二分之一」，第二天又在学校替同桌分零食（第10章）。这一段给后两段立了两把尺。一把是：学习的结果是世界变了，而不只是输出变了；判断一个系统学到了什么，要看它能分辨的差别、能稳稳维持的对象、能采取的行动有没有变。另一把是：现实世界不是看出来的，而是视觉、听觉、触觉一起参与，再加上现实的阻力才站稳的。杯子不会因为你不信就不烫，理念可以自洽地错，现实总会在某个地方把它顶回来（第8章）。这把尺后来在第六、七编一直在用：机器在桌面上的每一次失手，都是现实在顶它。

第二段讲集体语言世界和个人。大模型的训练发生的是一个有相对同一性、有集体性的「人类—AI复合体语言世界」：进训练的文本里沉淀着无数人的说话与做事，选数据、定目标、打分的又是一群具体的人和一套制度（第11章）。它像一个县的方言，没有哪个人能代表，却确实有自己的样子，又容得下各家各户的差异（第13章）；它也像被千万人改过的菜谱，带着被反复校过的好处，也带着流传太广的误区（第15章）。第14章又把三种检验分开：说得流畅、判断可靠、能力能迁移，是三回事。「你真行」这句话，可以是夸奖，可以是讽刺，也可以是无奈，模型在大多数场合能接对，说明它在语言里缠出了相当细的差别；可接得对，不等于它对说话人此刻的处境有可靠的判断，更不等于能挪到一个没见过的情境里去。语言里也确实沉淀了大量物理知识，「杯子放桌边会掉下去」「冬天水管会冻裂」，

可这些话的来源是人在现实里做事、吃亏，模型拿到的是吃亏以后留下的话，不是吃亏本身。这句话是第二段通向第三段的门。

第八编再把这个集体世界接到具体的人身上：周敏拿AI备浮力课，小舟让AI只提问不代笔写作文，一个小组和AI一起追剁辣椒罐子的胀罐。这三个场景各自拆开了一件事。周敏的备课拆开的是「继承」：她从公共模型拿到的是教案的骨架、实验的常规做法、误解的常见类型，可她班上学生把「沉下去的东西不受浮力」当真，这一点只有她知道；她让学生从书包里、厨房里拿土豆和钥匙来，这个选择出自她在乎什么。小舟的作文拆开的是「学会」：作文写好了不等于学会，扶手撤掉以后还写得出，才算。辣椒小组拆开的是「发现」：分析脚本照着文献的主流变量，把「其他配料」一栏当杂项丢掉了，那批加了大量蒜末却一罐没胀的罐子，是研究生对着原始记录一张张翻出来的；后来一片贴在盖子里的压力传感器，把「胀罐」拆成了两段，「二次产气」这个对象才第一次有了名字（第39章）。三件事合起来说明：集体语言世界给的起点有力量，也有它没准备好看见的东西，看见那些东西的，常常是具体的人和具体的阻力。

集体世界和个人世界是连续的：前者是后者的发生条件，后者的使用、反馈与再训练又可能回写前者。这种连续有工程上的通道（上下文、检索、反馈数据、再训练），也有制度上的通道（谁的反馈被采纳、按什么标准）。连续不是自动的好事：通道设计得好，个人的新经验能进入共同基础；设计得不好，共同基础会一层层把个人的差异抹平。如果每个老师都拿同一个模型备出同一节课，那不是共同基础的胜利，而是个人的发生被吞掉了。

第三段讲物理世界模型。写出「刹车」和停住车之间的距离，到了物理世界，变成「生成一段像样的视频」和「在真实桌面上拧开一个没见过的瓶盖」之间的距离。第四编先问清各路世界模型「以什么为世界」，按公开机制谈 Genie、V-JEPA 2、DreamerV3、Cosmos 3、Atlas，不给整个领域贴标签；第五编提出视觉、听觉、触觉各自是完整的 SDE，物在三者的互动里站稳；第六编承认物理发生的信息负担太重，显露预测保证不了行动可靠，出路是让任务来划定世界的范围；第七编把这些变成可以做、也可以失败的训练与实验。

这一段里有两处值得再说一句。一处是第20章指出的推理习惯：从「显露相似」直接跳到「发生机制相同」。生成的视频里杯子掉下去碎了，看着对，不等于模型里已经有了「玻璃杯从这个高度落到这种地面会碎」的发生条件。天气预报和导航软件常常很准，准得有用，可准是靠什么条件、在哪些情境下会突然不准，从「平时准」里读不出来。另一处是驾驶。人—车辆—道路复合体里，视觉、听觉、接触与车辆响应一起被训练（第29章）。本书没有预写「驾驶AI都是纯视觉」；Waymo 2025年12月的公开博文描述的，是摄像头、激光雷达、毫米波雷达随时间融合，再加仿真里的闭环训练和来自真实里程的反馈回写。本书要问的是，这种融合在什么程度上构成了多种 SDE 的纠缠，而不只是多路输入的拼接；公开资料还不足以回答，这本身就是一个研究问题。林岚的机械手拧瓶盖、捏豆腐，做菜时一边尝一边调咸淡，都是第七编里「试探、失配、重组」的样子：一次拧滑、一下捏碎，正是下一轮训练的材料。

三段接起来，是同一个判断走了三遍：不管是孩子、是千万人和一个模型组成的复合体，还是一台要拧瓶盖的机器，学习都不是把现成的世界搬进来，而是在具体的相遇里让纠缠建立、站稳，并且知道它在哪里会失效。

◆ 三、五条判决逐条验收

全书的结论可以压成五条判决。下面逐条验收：哪些地方站住了，哪些地方要收窄，最弱的一处在哪里。

判决一 · 学习即纠缠。站住的部分：它把训练和学习分开了，把「只看输出变化」的学习定义补成了三维，并且用同一幅图解释了学会、学错、遗忘和重新学会（第1—5章）。小舟一紧张就低头看前轮，紧张、低头、车把发僵缠得很牢，那是缠歪了的稳定；对的封装和错的封装形成方式一样，区别只在碰到现实时站不站得住。这些主要是区分与澄清，在本书框架里理由清楚、没有发现内部矛盾。要收窄的部分：本书从没说「参数就是 E」，只说参数连接可以被解释为特征纠缠，这种解释有边界（第12章）；三态（混沌、介生、秩序）目前是概念解释，不是测量结

果（第4章）。最弱处：E 在AI这一侧能不能落成可测的量。第40章的纠缠重组假说和关系图方向，都还没有结果。

判决二 · 三界共生。站住的部分：人的学习同时改动理念、现实、自我三个世界，这一点有大量日常经验可对照，第二编的五章和第38章那个说「这个好像不用算」的女生，都是同一幅图。它还给AI研究立了一条有用的限制：AI在三界里的参与不对称，最擅长理念界，现实界要靠接上器材和传感器，自我界最需要小心。要收窄的部分：三界是一个分析框架，不是三个分开的仓库；三界也不必同步，常见的是一界先动、带动另两界（第38章）。最弱处：「三界同时有变化的学习更持久」到现在只是一条机制假说，自我界的变化又最难观察，眼下多半只能靠学习者自己的说法。这些观察只能用来帮人设计下一步，不能进任何考核。

判决三 · 复合体是发生单位。站住的部分：训练里人的参与有公开的工程事实可查，选数据、写规范、做比较、设目标都是人做的（第11章）；使用中五件事的区分，即暂时语境变化、外部记忆、控制反馈、参数训练、持久学习（第37章），让「它学会了」「我学会了」这类话有了可以拆开核对的办法。小舟那篇写豆豆的作文，十几轮问答里AI越问越贴切，是暂时语境；这一晚模型参数一点没动；他一周后不用AI写出外婆家的猫「把身子摊成一张饼」，才算学到了一点。归因也要拆开：周敏用同一个模型备了一学期课，觉得它越来越顺手，仔细看，模型没变，变的是她的提问方式，她学会了一开头就交代班级情况、学生常见误解和想要的课堂节奏。顺手是她自己的持久学习。复合体里两侧的变化，要各记各的账。要收窄的部分：说复合体是发生单位，不等于说AI内部没有自己的机制可研究；第36章讲个人反馈可能回写集体，只描述了结构，没有断言回写分量有多大。最弱处：回流通道。个人在使用里留下的差异，经过筛选、汇总、再训练之后，是被保留还是被平均掉，现在没有好的证据。

判决四 · 物在三种 SDE 的纠缠中站稳。站住的部分：显露预测不保证行动可靠，这是一条推理上的缺口，有日常的反例撑着。雨夜路面一片反光，可能是薄水、积水或一摊油，行车记录仪拍下来几乎一样，一脚急刹下去后果却完全不同；两只看着一样的玻璃罐，抽没抽过真空，只有手一拧才分得开（第27章）。「物的

意识」先落成一组功能性判据，即对象同一性怎样形成、维持、失效、重建（第25章），也站得住。要收窄的部分：视觉、听觉、触觉不与S、D、E一一对应，各自都是完整的SDE；本书不主张三种感官对所有任务都不可替代；主观体验是另一个问题，功能判据不代它回答。最弱处：「三种SDE在互动里联合预测，比把特征拼起来更能形成可迁移的对象同一性」，这是第33、34章的核心预期，对照实验还没有做；第22章的伞模型也还没有形式化。第24章那只陌生弹性球给了这条判决最好的样子：看着像弹力球，从半臂高松手，它却「噗」一声只弹了很低一下就停住，看和摸、听的预期打了架，小舟最后自己给出一个说法：「它会吃掉力气，慢慢再还给你。」冲突推着他把这只球重新认了一遍。可这是一个人在构想场景里做到的事，机器能不能在类似冲突里做出类似的重组，只能等实验。

判决五 · 任务组织世界，封装要能失败。站住的部分：能力必须带着适用边界说话，这一条从第5章的封装一直贯到第35章的持续评估、回写与解封。拧瓶盖要保留的是材质、螺纹松紧、手指打滑，不是瓶子的颜色；装修师傅打孔在乎厘米、铺砖在乎毫米（第28章）。任务一换，要保留的差异和粒度就跟着换。要收窄的部分：本书不主张统一架构必然失败；可复用的底座加任务专门训练、专用模型之间靠接口协作（第30章），是本书偏向的路线，不是唯一的路线。最弱处：「任务专门模型在任务内更可靠，出了边界会很快下降」需要两头测，现在没有干净的对照；任务的初始边界怎样划，第28章也只说了它会在做的过程中被修订，没有给出划法。

五条放在一起看，有一个共同的样子：站得最稳的，都是「这两件事不是一件事」「这一步推理跳了一级」这类区分；最弱的，都是要把区分变成测量、把测量做成对照的地方。这不奇怪。一本书能在书桌上做完的，是把话说清楚；要在实验室、教室、马路上做完的，是看清楚的话对不对。前一种工作本书尽力做了，后一种工作，本书只能把门打开、把判负的规矩先写在门口。也要说明，五条判决之间是一环扣一环的：判决一说学习是什么，判决二说学习的人怎样随世界一起变，判决三说AI时代学习的单位是谁，判决四说物理世界里「学会一个物」意味着什么，

判决五说这样的能力怎样带着边界交给别人用。任何一环被推翻，后面几环都要跟着检查，不能只改一处了事。

◆ 四、推翻条件逐条报账

第40章把机制假说列成六条。这里沿用同样的编号，逐条写清楚：出现什么结果就该认输，认输之后书里哪一部分要改。

假说一（纠缠重组假说）：持久能力的形成伴随内部表示的重组，而不只是新表示的追加。判负条件：在持续训练里，用多种方法、多个模型都找不到「旧表示之间关系被改写」与「迁移变好」之间的对应，或者只追加新表示的系统迁移得一样好。认输之后：「E」在AI这一侧只能算比喻，不能当研究变量；第一编的本体论立场可以保留，但「用 SDE 解释AI训练」的工程价值要大打折扣。

假说二（主动干预假说）：同样的数据量下，主动干预采来的经验比被动记录更能分辨记录里分不出的配置，在新干预下预测更可靠。判负条件：数据量、算力、先验信息都控制住以后，被动记录训练在新干预预测、陌生物体的同一性维持、出错后的恢复速度上持续不差于主动训练，而且在多种任务、多种物体上重复出现。认输之后：第七编第32章的训练组织方案要改写。

假说三（跨模态冲突假说）：多种 SDE 之间的预测冲突，是对象同一性归并、拆分与重建的主要推动力。判负条件：在设计好的冲突情境里，比如第24章那只看着像弹力球、松手却只「噗」地弹一下的球，系统的对象表示既不归并也不拆分，只是被某一路主导、其余被平均掉，而任务表现并不因此变差；或者利用失配信号的系统，在换材质、换对象时恢复得并不比只拼接特征的系统快。认输之后：「冲突推动重组」至少不是这类系统形成对象的必经机制，第五编第24章和第七编第33章要收窄。

假说四（任务边界假说）：在可比资源下，围绕具体任务组织纠缠的模型在该任务上更可靠，出了任务边界优势会很快下降。判负条件：任务内没有优势，或者任务外并不下降；尤其是，一个统一架构在多种任务上同时达到专门模型的可靠性，并且能清楚报告自己在每种任务上的适用条件和失效方式。认输之后：第六编

偏向专用模型协作的判断要修正。但要说明，统一架构即便成功，也得交出这份边界报告，否则它的成功还不算可靠。

假说五（人侧持久学习假说）：在个人—AI复合体里，持久学习更可靠地落在人这一侧，或人自己维护的外部记忆上，而不在模型参数上。判负条件：两组学习者，一组让AI给终稿，一组只让AI提问、自己改，隔两周不借AI独立完成同类和异类任务，两组没有差别。认输之后：第37、38章关于「把差异运行留在学习者这一侧」的设计主张要重新评估。

假说六（同一性收窄假说）：评价反馈若一再奖励最稳妥的回答样式，集体语言世界的显露会收窄，使用者的产出趋同。判负条件：不同评价方式训练出的模型，在开放任务上给出的方案多样性没有差别，使用者最终方案的多样性也与评价方式无关。认输之后：第36章对同一性滑向单一答案的担心要降级。这一条还有反方向的输法：如果长期观察显示，不论通道怎样设计，人在和同一类模型的持续协作里都显著趋同，那么第八编「同一性是条件而非强制」的乐观判断应当收回。

六条之外，还有立场本身的账。上面说过，SDE 立场不由某个模型的成败直接证明，那它是不是永远不会输？不是。立场也会输，只是输法不同：如果它长期组织不出比别的立场更好的问题，如果按它设计的实验一个接一个被判负、而它只能靠改口保住自己，如果它的术语越用越含混、越来越要堆叠才说得下去，它在实际上就被放弃了。本书接受这种输法。六条假说各自独立判负，合起来也是对立场的考验：如果大多数都输了，「学习是特征纠缠的建立与训练」这个母题就该认真重审。每一项研究开始之前，都先把判负条件写下来、公开出来，免得结果出来以后再改标准。

◆ 五、收窄声明与尚未解决的问题

有两句话要单独写出来，因为它们最容易被读成本书的前提：本书不主张统一架构必然失败，不主张三种感官对所有任务都不可替代；这两条是研究边界，不是前提。

前一条的意思是，本书批评的是一种推理习惯，即把世界当作先在的、可以一次封装的对象，从「显露相似」直接跳到「发生机制相同」（第20章），而不是批评某一种网络结构。一个统一的大模型如果能在多种任务上交出可靠的能力和清楚的边界报告，本书乐于承认它。后一条的意思是，三种感官哪一种对哪个任务必要、哪一种可以省，要一个任务一个任务去试。摸黑找开关时，视觉退场，触觉和身体记忆照样能把一件物维持住；挑西瓜时，看、敲、托三路互相印证，老练的人会更准。哪个任务属于哪种情形，不是在书桌上能定的。

为什么要把这两条单独拿出来说？因为一本批评「发现学前提」的书，很容易反过来犯同一个毛病：把自己偏爱的路线当成先在的真理，再去现实里找证据。如果本书预先认定统一架构一定失败，那么任何一个统一模型的进步都会被解释成「表面成功」，任何一次失手都会被当成「果然如此」，立场就再也碰不到现实的阻力了。同样，如果预先认定三种感官缺一不可，那么一个只靠视觉就把某件事做得很好的系统，也只会被说成「迟早要出问题」。这两种说法都把检验关在了门外。本书宁可把它们写成研究边界：边界的意思是，我们在这里停下，等实验来说话。第34章要求「可以失败的实验」，结语对自己也提同样的要求。

还有几个问题，本书提出了，没有答案，甚至还说不清要怎样检验。一是特征纠缠的形式化：三方程现在是概念方程，F、G、H 是什么样的函数，何时能写成可算的形式，没有回答；第22章的伞模型，即一次接触的影响能延续多久、波及多远，也还只有描述。二是复杂性的可测指标：混沌态、介生态、秩序态在训练过程里各对应什么可观测量，没有定论（第4章）。三是身体感觉和机器信号的对应：老彭手上的「方向盘发飘」，能在多大程度上对应到车上某种传感器信号的组合，要一项一项做对照（第29章）。四是主观体验：功能判据可以研究，「这个系统有没有感受」是另一个问题，本书明确不拿功能判据代答。五是集体同一性的回流：谁的反馈被采纳、按什么标准、个人的新经验怎样进入共同基础而不被抹平。这牵涉制度，本书只指出它是发生条件的一部分，不做政策表态。

把账分清楚，不是为了显得谦虚，而是让后来的人知道往哪里使劲：站住的部分要守住，待验的假说要去做实验，没解决的问题要先想清楚怎样才能做实验。

六、往后走的四个方向

第40章给了一份研究纲领。在结语里，它可以收成四个方向。每一个既是对AI的要求，也是对人自己的要求。

第一，可持续学习。今天多数大模型的参数在训练结束时就定住了，后面的「学习」多半是语境变化和外部记忆。这是一种工程选择，定住的模型更容易评估。可从发生学看，一种能力如果只能在出厂那一刻形成，它的封装就很难跟上世界的变化。要解决的不是「让模型一直在改」，而是「在哪里改、改了以后旧能力会不会坏、怎样知道改对了」这三件事都有着落。神经网络学新任务时旧任务表现可能明显下降，工程界叫它「灾难性遗忘」（第1、15章），这仍是要啃的硬骨头。周敏十几年后重新上车很快找回平衡，同时并没有因此忘掉怎样走路；人的学习在建立新纠缠时大体保得住旧纠缠，这是值得机器借鉴的地方，至于怎样借鉴，要靠具体机制去试，本书不预写答案。

第二，可识别失配。可靠的系统不是不出错，而是知道自己什么时候走到了边界外面。老彭碰到陌生情境，先慢下来：交警的手势和信号灯对不上，听交警的；路上有个说不清的东西，宁可绕开；车里有从没听过的声音，靠边停下查（第29章）。对AI来说，就是在预测分歧变大、感官之间互相矛盾、行动后果偏离预期时，说出「这里我没把握」，而不是用一段流畅的输出把它盖过去。语言模型的幻觉和机器在陌生物体上的失手，在这个方向上是同一个问题。

第三，可重建纠缠。识别出失配以后，要能拆开旧的缠法重新缠。小舟看了一周游泳教学视频，动作讲得头头是道，下了水手一划身体只是晃；几天以后才找到手能「抓住」一点水的感觉（第17章）。对机器来说，就是能在局部重组对象表示，不必推倒重来，并把一次失败转成新的训练经验，回写到能力的适用边界里（第5、35章）。

第四，可与其他复合体协作。没有哪个复合体能穷尽世界。周敏的备课复合体、小舟的写作复合体、林岚团队的机器人训练复合体、第39章那个研究胀罐的小组，各自形成的能力都要通过接口和别的复合体相连。协作不是把所有东西装进一

个大模型，而是让不同的封装能互相调用、互相检验，在对方失效的时候接手（第30、39章）。

四个方向有个次序：可持续学习是底子，可识别失配是触发，可重建纠缠是动作，与别的复合体协作是自己重建不了时的出路。

这四个方向对人一样适用，教育在这里有一份特别的责任（第38章）。如果学习是特征纠缠的建立与训练，学校要设计的就不只是把知识摆出来，而是让理念、现实、自我一起变的学习过程；判断学会没有，看迁移、重组和新能力，不看复述。小舟喝汤时看着浮着的葱和沉下去的香菜，自己说出「要比一样大的时候谁重」，那是在饭桌上冒出来的重组，比任何测验都早。人和AI一起试探、一起反思，可以是这个过程里很有力的一环，前提是它不替代那一次真正「被烫一下」，也不替代孩子在没有AI时自己写出的那个开头。一个能持续学、知道自己哪里不懂、肯拆了重来、能和别人一起干的人，正是教育想要的人。这些观察是给老师和家长看的眼睛，不是打分的尺子，不进任何考核。

◆ 七、随身三问

全书每一章都可以回头问三个问题：这里纠缠的是什么？差异是怎样运行的？显露出来的东西靠什么站稳、又在哪里失效？

这三问不只是给研究者的。下一次你用AI查一件事、改一段文字，也可以拿它们问一问。这一次聊得很顺，纠缠的是什么：是这段对话里的上下文，是工具替你存的一份记忆，还是你自己脑子里真缠上了新东西？差异是怎样运行的：是你在追问、在纠正、在对照自己的处境，还是只在接收？显露出来的那个好答案，靠什么站稳：换一个窗口、隔一个星期、换一道题，它还在不在？在哪里会失效：它碰到你班上那几个学生、你家那条路、你手里那只拧不开的瓶盖时，还灵不灵？

对一个世界模型，同样三问。它生成的视频很像，纠缠的是画面之间的关系，还是动作和后果之间的关系？它的差异是从录像里取来的，还是从自己的试探里长出来的？它说杯子会碎，靠什么站稳，换一种地面、换一个高度它还对不对，错了它自己知不知道？

对一个孩子，也是这三问。小舟这次作文写得好，缠上的是AI一路递过来的问题，还是他自己把「可爱」拆成了看得见的动作？差异是AI在推着走，还是他自己在问自己「它长什么样」？换一只猫、换一次春游，那个办法还在不在？

三问问不出现成的答案，但能把「看上去会了」和「真的会了」分开。分开了，下一步该做什么，往往就清楚了。

◆ 八、判词

把全书压成一句话，就是：

> **训练不是把一个现成的世界装进机器，而是组织发生的条件；学习，是特征纠缠的建立与训练——世界模型的出路，不在一次封装一个全能的世界，而在具体任务的互动里，先让一个物站稳。**

这句话前一半回答书名里的问题，后一半回答副题里的困境与出路。困境不在算力不够，而在把世界当成先在的、可以一次装进去的东西；出路是从具体任务出发，让一个物先在互动里站稳，再一个一个往外扩，并且每一步都写好它会在哪里失败。

◆ 九、回到雨夜

再回到导论那个下雨的夜晚。

模型写出「刹车」，背后是千百万人在语言里留下的关于刹车的经验，经过集体人类—AI复合体的训练，缠成了一个能在语言里调用得当的东西。这不是假的，也不是小事。有了它，一个从没开过车的人几分钟就能问清雨天刹车的要领；周敏也能拿它做起点，备出一节冲着学生误解去的课。

老彭停住车，背后是三十年里脚底、手心、耳朵、眼睛一起缠出来的东西，再加上那一晚那一条老街的具体条件。他的右脚离开油门，先重后稳地压下去，脚底一阵细密的顿挫，轮胎碾过积水的嘶嘶声变了调，方向盘在手里轻轻一沉，他微微

回正。那一刻没有哪一个封装是「完成了」的：路面有多湿、前面那团黑影往哪边走、车上哪一处的声音不太对，都在当场重新进入他的判断。车停住了，离那个推车的人还有一两个车身。他摇下窗，什么也没说，等那人过去。

两者之间的距离，本书没有说能靠一个更大的模型一步跨过去，也没有说它永远跨不过去。本书说的是：这段距离要靠一次次围绕具体任务的互动来走，在每一处看清纠缠的是什么、差异怎样运行、显露出来的东西靠什么站稳、又在哪里失效。

AI的训练与学习，不是一件发生在机器里、与人无关的事。训练时，集体的人在场；使用时，具体的人在场；在物理世界里行动时，世界本身以它的阻力在场。学习不是把世界搬进模型，而是模型、人和世界一起，在一次次相遇里继续发生。这本书只是这场发生里的一段记录，它的许多判断要由后来的实验、后来的使用、后来的失败和重建来检验。哪些判断被推翻了，那也是学习，一张新的网又在缠起来。

雨还在下。老彭把车重新开起来，速度比刚才又慢了一点。

> **带走的话**
>
> 训练组织发生的条件，学习是特征纠缠的建立、稳定、分化与重组，能力是有边界、会失效、能重建的封装；五条判决各有站住的地方，也各有最弱的一处，六条假说都写明了怎样算输。世界不是被一次装进模型的，而是在人、AI与世界的持续互动里，一边被检验，一边继续发生。

参 考 书 目

本书目收录正文论证中实际用到的研究、技术报告、官方博文与著作，以及少量与本书问题直接相关、读者可据以追溯源头的经典文献；正文只作线索性提及而无公开出处可核的材料（如各类产品的当期条款、二手报道）不收。全目分五类，类内外文文献按第一作者姓氏字母排列，同一作者按年份排列；编号全书连续。条目大体依 APA 格式著录，外文保留原题；作者八人及以上者列前三人后加 et al.，机构署名者以机构为作者；有 DOI 或 arXiv 号者一并列出。2026 年新出的预印本与博文，按 2026 年 10 月可见版本著录，日后若有改版或正式发表，以作者最新版本为准。条目末尾标〔待核〕者，表示该项著录信息（多为出版年份）尚待与原书版权页核对。

一、SDE 发生学与德麦国际相关著作

1. 王德生. (年份〔待核〕). 《SIO 哲学导论》. 德麦国际出版社. ISBN 978-1-970820-01-0.
2. 王德生. (2026). 《AI时代教育的认知变革——腾出的那一处，怎样才被占住》. 德麦国际出版社. ISBN 979-8-90690-096-8.
3. 王德生. (2026). 《可判性的到来——发现还是发生？》. 德麦国际出版社. ISBN 979-8-90690-092-0.
4. 王德生. (2026). 《让每个人都有自己的思想和知识工厂》. 德麦国际出版社. ISBN 979-8-90690-129-3.
5. 王德生, 胡志英. (2026). 《SDE 语义发生三原理》. 德麦国际出版社. ISBN 979-8-90690-644-1.

二、学习、发展与认知科学

1. Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. Cognition, 20(3), 191–208. https://doi.org/10.1016/0010-0277(85)90008-3

2. Bregman, A. S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press.
3. Bushnell, I. W. R., Sai, F., & Mullin, J. T. (1989). Neonatal recognition of the mother's face. British Journal of Developmental Psychology, 7(1), 3–15. https://doi.org/10.1111/j.2044-835X.1989.tb00784.x
4. Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181–204. https://doi.org/10.1017/S0140525X12000477
5. DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. Science, 208(4448), 1174–1176. https://doi.org/10.1126/science.7375928
6. DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. Infant Behavior and Development, 9(2), 133–150. https://doi.org/10.1016/0163-6383(86)90025-1
7. Ebbinghaus, H. (1913). Memory: A Contribution to Experimental Psychology (H. A. Ruger & C. E. Bussenius, Trans.). Teachers College, Columbia University. (Original work published 1885) https://doi.org/10.1037/10011-000
8. Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870), 429–433. https://doi.org/10.1038/415429a
9. Fernald, A., Taeschner, T., Dunn, J., Papoušek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. Journal of Child Language, 16(3), 477–501. https://doi.org/10.1017/S0305000900010679
10. Frank, M. C. (2023). Bridging the data gap between children and large language models. Trends in Cognitive Sciences, 27(11), 990–992. https://doi.org/10.1016/j.tics.2023.08.007
11. Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Houghton Mifflin.
12. Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology, 56(5), 872–876. https://doi.org/10.1037/h0040546
13. Henderson, J. M. (2003). Human gaze control during real-world scene perception. Trends in Cognitive Sciences, 7(11), 498–504. https://doi.org/10.1016/j.tics.2003.09.006

14. Johansson, R. S., & Flanagan, J. R. (2009). Coding and use of tactile signals from the fingertips in object manipulation tasks. Nature Reviews Neuroscience, 10(5), 345–359. https://doi.org/10.1038/nrn2621
15. Johansson, R. S., & Westling, G. (1984). Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects. Experimental Brain Research, 56(3), 550–564. https://doi.org/10.1007/BF00237997
16. Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. Cognitive Psychology, 15(4), 483–524. https://doi.org/10.1016/0010-0285(83)90017-8
17. Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: Evidence of perceptual narrowing. Psychological Science, 18(12), 1084–1089. https://doi.org/10.1111/j.1467-9280.2007.02029.x
18. Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. Nature Reviews Neuroscience, 5(11), 831–843. https://doi.org/10.1038/nrn1533
19. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253. https://doi.org/10.1017/S0140525X16001837
20. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. Nature Reviews Neuroscience, 21(6), 335–346. https://doi.org/10.1038/s41583-020-0277-3
21. Macfarlane, A. (1975). Olfaction in the development of social preferences in the human neonate. In Parent-Infant Interaction (Ciba Foundation Symposium 33, pp. 103–117). https://doi.org/10.1002/9780470720158.ch7
22. McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746–748. https://doi.org/10.1038/264746a0
23. Ni, Y., & Zhou, Y.-D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. Educational Psychologist, 40(1), 27–52. https://doi.org/10.1207/s15326985ep4001_3
24. Pascalis, O., de Haan, M., & Nelson, C. A. (2002). Is face processing species-specific during the first year of life? Science, 296(5571), 1321–1323. https://doi.org/10.1126/science.1070223
25. Piaget, J. (1952). The Origins of Intelligence in Children (M. Cook, Trans.). International Universities Press.

26. Piaget, J. (1954). The Construction of Reality in the Child (M. Cook, Trans.). Basic Books.
27. Spelke, E. S. (1990). Principles of object perception. Cognitive Science, 14(1), 29–56. https://doi.org/10.1207/s15516709cog1401_3
28. Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. Developmental Science, 10(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x
29. Thelen, E., & Smith, L. B. (1994). A Dynamic Systems Approach to the Development of Cognition and Action. MIT Press.
30. Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences, 28(5), 675–691. https://doi.org/10.1017/S0140525X05000129
31. Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. Attention, Perception, & Psychophysics, 72(4), 871–884. https://doi.org/10.3758/APP.72.4.871
32. Vygotsky, L. S. (1978). Mind in Society: The Development of Higher Psychological Processes (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.
33. Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: Sleep-dependent motor skill learning. Neuron, 35(1), 205–211. https://doi.org/10.1016/S0896-6273(02)00746-8
34. Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development, 7(1), 49–63. https://doi.org/10.1016/S0163-6383(84)80022-3
35. Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. Science, 269(5232), 1880–1882. https://doi.org/10.1126/science.7569931

三、机器学习与语言模型

1. Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
2. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185–5198). https://doi.org/10.18653/v1/2020.acl-main.463

3. Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems 33. arXiv: 2005.14165.
4. Dziri, N., Lu, X., Sclar, M., et al. (2023). Faith and fate: Limits of transformers on compositionality. In Advances in Neural Information Processing Systems 36. arXiv: 2305.18654.
5. Elhage, N., Hume, T., Olsson, C., et al. (2022). Toy models of superposition. Transformer Circuits Thread. arXiv:2209.10652.
6. Gerstgrasser, M., Schaeffer, R., Dey, A., et al. (2024). Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. arXiv: 2404.01413.
7. Gurnee, W., & Tegmark, M. (2024). Language models represent space and time. In International Conference on Learning Representations (ICLR 2024). arXiv: 2310.02207.
8. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv:1503.02531.
9. Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training compute-optimal large language models. In Advances in Neural Information Processing Systems 35. arXiv:2203.15556.
10. Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In International Conference on Learning Representations (ICLR 2020). arXiv:1904.09751.
11. Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. In Proceedings of the 41st International Conference on Machine Learning. arXiv:2405.07987.
12. Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730
13. Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361.
14. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114
15. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems 30. arXiv:1612.01474.

16. McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation, 24, 109–165. https://doi.org/10.1016/S0079-7421(08)60536-8
17. Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2025). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In International Conference on Learning Representations (ICLR 2025). arXiv:2410.05229.
18. Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. In International Conference on Learning Representations (ICLR 2023). arXiv:2301.05217.
19. Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35. arXiv:2203.02155.
20. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012
21. Petrov, A., La Malfa, E., Torr, P. H. S., & Bibi, A. (2023). Language model tokenizers introduce unfairness between languages. In Advances in Neural Information Processing Systems 36. arXiv:2305.15425.
22. Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. arXiv:2201.02177.
23. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536. https://doi.org/10.1038/323533a0
24. Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? In Advances in Neural Information Processing Systems 36. arXiv:2304.15004.
25. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. arXiv:1508.07909.
26. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y
27. Sutton, R. S. (2019, March 13). The Bitter Lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

28. Templeton, A., Conerly, T., Marcus, J., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread. https://transformer-circuits.pub/2024/scaling-monosemanticity/
29. Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4593–4601). arXiv:1905.05950.
30. Touvron, H., Lavril, T., Izacard, G., et al. (2023a). LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.
31. Touvron, H., Martin, L., Stone, K., et al. (2023b). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.
32. Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2023). Steering Language Models with Activation Engineering. arXiv:2308.10248.
33. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30. arXiv:1706.03762.
34. Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-consistency improves chain of thought reasoning in language models. In International Conference on Learning Representations (ICLR 2023). arXiv:2203.11171.
35. Wei, J., Tay, Y., Bommasani, R., et al. (2022a). Emergent abilities of large language models. Transactions on Machine Learning Research. arXiv:2206.07682.
36. Wei, J., Wang, X., Schuurmans, D., et al. (2022b). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems 35. arXiv:2201.11903.

四、世界模型与机器人、自动驾驶

1. Assran, M., Bardes, A., Fan, D., et al. (2025). V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv:2506.09985.
2. Assran, M., Duval, Q., Misra, I., et al. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). arXiv: 2301.08243.
3. Bardes, A., Garrido, Q., Ponce, J., et al. (2024). Revisiting Feature Prediction for Learning Visual Representations from Video. arXiv:2404.08471.

4. Bruce, J., Dennis, M., Edwards, A., et al. (2024). Genie: Generative interactive environments. In Proceedings of the 41st International Conference on Machine Learning. arXiv:2402.15391.
5. Chang, H., Zhang, H., Jiang, L., Liu, C., & Freeman, W. T. (2022). MaskGIT: Masked generative image transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). arXiv:2202.04200.
6. Damen, D., Doughty, H., Farinella, G. M., et al. (2022). Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. International Journal of Computer Vision, 130(1), 33–55. https://doi.org/10.1007/s11263-021-01531-2
7. Garrido, Q., Ballas, N., Assran, M., et al. (2025). Intuitive Physics Understanding Emerges from Self-Supervised Pretraining on Natural Videos. arXiv:2502.11831.
8. Goyal, R., Kahou, S. E., Michalski, V., et al. (2017). The "something something" video database for learning and evaluating visual common sense. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 5843–5851). https://doi.org/10.1109/ICCV.2017.622
9. Ha, D., & Schmidhuber, J. (2018). World Models. arXiv:1803.10122.
10. Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2020). Dream to control: Learning behaviors by latent imagination. In International Conference on Learning Representations (ICLR 2020). arXiv:1912.01603.
11. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2019). Learning latent dynamics for planning from pixels. In Proceedings of the 36th International Conference on Machine Learning. arXiv:1811.04551.
12. Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2025). Mastering diverse control tasks through world models. Nature, 640(8059), 647–653. https://doi.org/10.1038/s41586-025-08744-2 (Preprint: arXiv:2301.04104)
13. Hu, A., Russell, L., Yeo, H., et al. (2023). GAIA-1: A Generative World Model for Autonomous Driving. arXiv:2309.17080.
14. Khazatsky, A., Pertsch, K., Nair, S., et al. (2024). DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. arXiv:2403.12945.
15. Lam, R., Sanchez-Gonzalez, A., Willson, M., et al. (2023). Learning skillful medium-range global weather forecasting. Science, 382(6677), 1416–1421. https://doi.org/10.1126/science.adi2336 (Preprint: arXiv:2212.12794)
16. LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence (Version 0.9.2). OpenReview. https://openreview.net/forum?id=BZ5a1r-kVsf

17. Mur-Labadia, L., Muckley, M., Bar, A., et al. (2026). V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning. arXiv:2603.14482.
18. NVIDIA. (2025). Cosmos World Foundation Model Platform for Physical AI. arXiv: 2501.03575.
19. NVIDIA. (2026). Cosmos 3: Omnimodal World Models for Physical AI. arXiv: 2606.02800.
20. Parker-Holder, J., & Fruchter, S. (2025, August 5). Genie 3: A New Frontier for World Models. Google DeepMind Blog. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/
21. Pathak, D., Gandhi, D., & Gupta, A. (2019). Self-supervised exploration via disagreement. In Proceedings of the 36th International Conference on Machine Learning. arXiv:1906.04161.
22. Runway. (2025, December 11). Introducing Runway GWM-1. Runway Research. https://runway.com/research/introducing-runway-gwm-1
23. Schrittwieser, J., Antonoglou, I., Hubert, T., et al. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. Nature, 588(7839), 604–609. https://doi.org/10.1038/s41586-020-03051-4
24. Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., & Pathak, D. (2020). Planning to explore via self-supervised world models. In Proceedings of the 37th International Conference on Machine Learning. arXiv:2005.05960.
25. Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. ACM SIGART Bulletin, 2(4), 160–163. https://doi.org/10.1145/122344.122377
26. Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.
27. Waymo AI Team. (2025, December 9). Demonstrably Safe AI for Autonomous Driving. Waymo Blog. https://waymo.com/blog/2025/12/demonstrably-safe-ai-for-autonomous-driving/
28. World Labs Team. (2026, September 1). Atlas: A World Model for Spatial Intelligence. World Labs Blog. https://www.worldlabs.ai/blog/atlas
29. Yuan, H., Wang, Z., Shao, B., Lu, H., Darrell, T., Lourentzou, I., & Zhan, W. (2026). DexTacWAM: A Visuo-Tactile World-Action Model for Dexterous Manipulation. arXiv:2609.24976.
30. Zhou, J., Hong, F., Li, Y., et al. (2026). TouchWorld: A Predictive and Reactive Tactile Foundation Model for Dexterous Manipulation. arXiv:2607.07287.

五、哲学与科学方法

1. Clark, A. (1997). Being There: Putting Brain, Body, and World Together Again. MIT Press.
2. Craik, K. J. W. (1943). The Nature of Explanation. Cambridge University Press.
3. Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6
4. Kuhn, T. S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.
5. Lakatos, I. (1978). The Methodology of Scientific Research Programmes (J. Worrall & G. Currie, Eds.). Cambridge University Press.
6. Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. University of Chicago Press.
7. Merleau-Ponty, M. (1962). Phenomenology of Perception (C. Smith, Trans.). Routledge & Kegan Paul. (Original work published 1945)
8. Polanyi, M. (1966). The Tacit Dimension. Doubleday.
9. Popper, K. R. (1959). The Logic of Scientific Discovery. Hutchinson.
10. Scheffer, M., Bascompte, J., Brock, W. A., et al. (2009). Early-warning signals for critical transitions. Nature, 461(7260), 53–59. https://doi.org/10.1038/nature08227
11. Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424. https://doi.org/10.1017/S0140525X00005756
12. Varela, F. J., Thompson, E., & Rosch, E. (1991). The Embodied Mind: Cognitive Science and Human Experience. MIT Press.

附 录

附录五件，供读完全书或中途回查时使用：术语不熟查附录一，想知道本书在什么结果下认输查附录二，想对照技术系统查附录三，想找某个例子在哪一章查附录四，想知道这本书怎样写成、证据核到哪一步查附录五。附录只整理正文已有的内容，不增加新的论证与新的数据；凡与正文有出入，以正文为准。

附录一　术语速查

下表按全书论证的先后排列。「首见」指这个词在正文里第一次出现的位置；若首见处只是提到、正式立义在后面，两处都写。

词	本书里的意思	首见
显露（S）	某种东西在一个过程中露出来、被认出来、可以被打交道。不读作「结构」：结构是显露的稳定态，是那些反复露出来、站得住的东西	导论；第3章立义
差异运行（D）	前后、左右、这样与那样之间的不同，在过程中一个接一个出现、被区分、被承接；一个差异引出下一个，所以也叫差异序列	导论；第3章立义
特征纠缠（E）	一块块经验、一个个可区分的方面缠到一起、彼此牵动。它不是背景，不是仓库，也不是训练数据集，而是在相遇中被缠出、又回头塑造显露与差异的组织；可从范围、紧度、弹性三方面描述	导论；第3章立义
母题	学习是特征纠缠的建立与训练；它一旦发生，显露和差异运行也跟着改变	导论
三方程	S=F(D,E)、D=G(S,E)、E=H(S,D)：每一维都由另外两维决定。F、G、H 目前是概念方程，不是可代入数字的函数	导论；第3章立名

词	本书里的意思	首见
缝隙	三方程缺了任何一条，解释里就留下一道缝隙；如显露不再回写纠缠、差异运行不受调节、显露漂浮于差异与纠缠之外	导论；第3章
六路径	三维走完一轮的六种次序：D→E→S、E→D→S、D→S→E、S→D→E、E→S→D、S→E→D。不是两维之间的单箭头	第3章
三原理	某两维之间发生矛盾，产生推动第三维改变的动力；分三拍：张力积累、形状改变、回写。推动哪一维，取决于矛盾发生在哪两维之间	第3章
混沌态、介生态、秩序态	三态。混沌态：差异大量运行，纠缠松散，显露不稳定；秩序态：纠缠稳定，差异有固定节奏，显露可反复出现；介生态：局部已成形而未连成整体、新旧组织并存竞争的生成状态。三态不是单向阶梯，秩序态可以退回	第一编编序；第4章立义
概念解释与可测动力学指标	三态等说法是概念解释；说「进入了介生态」必须说明拿什么量来判断，否则只是换了个名字	第4章
训练量、训练结构、训练后果	训练量是相遇发生了多少次；训练结构是相遇以什么方式组织；训练后果分在原条件下能否重复、相近条件下能否迁移、条件变化时能否发现「不一样了」三层	第一编编序；第2章
封装	一组特征纠缠在充分训练和检验之后，作为整体能力被固定下来，可在一定条件下直接调用而不必每次重组内部细节；同时带有适用边界，边界外可能失效	导论；第5章立义
解封	封装失效后，被封好的整体暂时打开，内部组织重新暴露在注意之下、接受新差异的冲击，进而重新训练、重新封装。封装不是黑箱，打不开的封装是风险	导论；第5章
适用边界	一项能力在哪些条件下可以放心调用、哪些需要小心、哪些不该调用。能力报告不写适用边界，只说了一半	导论；第35章展开
三界	理念世界、现实世界、自我世界。三者不是三个房间，而是同一张纠缠之网的三个方向，彼此改写	导论；第二编展开
理念世界	符号、概念、逻辑与推演经训练而稳定下来的显露；它的稳定靠符号、推演和操作反复相互约束，不是挂在天上的先在之物	导论；第7章
现实世界	哪里是边界、哪里能走、什么能抓、什么碰不得，在身体与周围的反复互动和阻力中发生	导论；第8章

词	本书里的意思	首见
现实阻力	世界不因人的意愿而改变的那一面，是纠缠能否站稳的检验者：理念可以自洽地错，现实会在某处把它顶回来	第1章；第8章展开
可行动性	物对一个身体显露为可以抓、可以走、可以推的样子；随身体与工具改变	第二编编序；第8章
自我世界、我之同一性	记忆、情感、关系、价值缠成的一团，在行动历史中显露；「我」的同一性经历形成、维持、受冲击与重建	导论；第9章
发现学	以为世界、对象与规律先在，等着被找到，越发现越接近完整	导论
发生学	对象与规律在 E 中经 D 显露为 S，显露之后仍以稳定态继续接受互动检验。不主张世界随人心意而变	导论
集体人类—AI复合体	大模型训练的发生单位：人类的历史活动、数据与制度、工程与评价，和模型一道。AI 不写成脱离人类历史的独立发生者	导论；第11章展开
个人—AI复合体	具体的人提问、追问、纠正时，与模型一道构成的发生单位；在这里发生的东西既不全属于人，也不全属于模型	导论；第八编展开
相对同一性	同一个模型面对不同的人大体守着同一套表达习惯与推理约束；它不是平均值，而是能容纳差异的共同部分	导论；第13章
五件事	暂时语境变化、外部记忆、控制反馈、参数训练、持久学习。前四件是机制，第五件是结果；一次流畅对话最多说明前一两件发生了	导论；第37章
持久学习	一种能力稳定形成：换了情境还能用，过一段时间还在，不靠当场提示也能调出，失效时能被识别	导论；第37章立义
三层论证	本体论立场（SDE 的主张，不由某次实验直接证明或推翻）、机制假说（可被实验支持或推翻）、经验事实（公开文献与技术报告里的东西）。模型的局部成败不能直接当作本体论证明	导论
随身三问	这里纠缠的是什么？差异是怎样运行的？显露出来的东西靠什么站稳、又在哪里失效？	导论
显露相似与发生机制相同	从「显露相似」直接跳到「发生机制相同」是一个推论缺口，视频生成与语言生成里都成立	第12章；第20章展开
统一架构与全能世界模拟	统一架构是工程事实（一个框架处理多种模态与任务）；全能世界模拟是把世界一次性封装的设想。二者不是一回事，本书不主张统一架构必然失败	导论；第16章

词	本书里的意思	首见
记录中的互动与新互动	录像里的互动真实发生过，但它是别人选的、已经封闭、可能混淆；新互动由系统自己发起，能问「换一个动作会怎样」	第17章
多传感器融合与多种 SDE 纠缠	融合是在事先定好的对象下对齐、加权、合并多路信号；多种 SDE 纠缠是各自完整的感官过程在共同事件中相互制约，冲突能推动对象的归并、拆分或重建	第四编编序；第19章
对象同一性	一个物在跨感官、跨时间的经验里被认作同一个，并能在失效时被察觉、被重建	导论；第25章
物的意识	功能层面指跨感官、跨时间稳定的对象同一性，及其与用途、行动可能性的共同形成；是特征纠缠聚合体足够稳定时的状态。主观体验问题单独论证，不混入	导论；第25章立义
四条功能性判据	形成、维持、失效、重建：生出来、站得住、塌得下、立得回。判据分程度，不是有或无	导论；第25章
特征纠缠聚合体	视觉、触觉、听觉三套各自完整的 SDE 在互动中彼此进入对方的发生条件，物在其中显露并站稳；不是三份感官报告的拼盘	第五编编序；第24章
同步出现、相关变化、相互制约	聚合体的三层纠缠：差异在同一时刻同一处出现；一路变化另一路有规律地跟着变；各路之间彼此约束，冲突时推动重组	第24章
伞模型	一次接触像撑开一把伞：伞面是这次接触在空间、时间与身体上的影响范围，伞下的后续事件被它改写过。这是尚未形式化的工作说法，属机制设想，伞面怎样界定、衰减、叠加、由什么决定四问都未回答	第五编编序；第22章
任务组织世界的发生范围	任务决定需要保留哪些差异、多细的粒度、多大的时空范围；初始边界在互动中不断修订	第六编编序；第28章
人—车辆—道路复合体	驾驶的发生单位：视觉、听觉、接触与车辆响应在其中一起被训练	导论；第29章
共同事件	训练经验的基本单位：一段时间内的观察、动作、接触、声音与后果，按时间对齐成一件事	导论；第31章
主动试探与预测分歧	系统在自己拿不准的地方选择能区分不同发生配置的动作去做；拿不准的标志是对同一动作的几种后果预测分歧大	导论；第32章

词	本书里的意思	首见
六步法、九步法与九宫	六步法按六路径把训练动作分为六类次序；九步法按 SDE 三维九宫（S1 规范、S2 模态、S3 价值；D1 意义目标、D2 路径组织、D3 优化约束；E1 三界、E2 信息三模态、E3 能量三状态）逐格检查，次序从 S1・D1・E1 定式三角起，再过 S2・D2・E2、S3・D3・E3；六步与九步同属 D2	第七编编序；第32章
联合预测	不是多接几路输入，而是改变预测对象：同时预测下一刻的画面、声音、触觉，且三者要彼此对得上	第33章
失配与失配归因	失配是预期与显露对不上；失配归因是追问错在哪里：信号坏了、对象认错、对象有新特性、动作没执行到位，还是事件本有的随机	第2章；第33章立义
拼接、被动融合、主动互动	第34章三组对照：三路分开训练末端拼接；按共同事件联合预测但只用事先采集的数据；在此基础上按预测分歧主动选择试探并做失配归因	第34章
迁移与失配恢复	迁移分近迁移（同类新个体）与远迁移（换任务）；失配恢复测失配检出时间、恢复交互数、旧能力保持率三项	第31章；第34章立义
四个方向	可持续学习、可识别失配、可重建纠缠、可与其他复合体协作：底子、触发、动作、出路	结语

附录二　判决卡与推翻条件

全书收成五条判决。下面每条一张卡：一句话、主场在哪几章、凭什么站住、收窄到哪里、出现什么结果就要判负或收窄、本书自认它最弱在哪里。推翻条件一律从正文第34章、第40章与结语中整理，并注明出处；它们是事先写定的，实验做完不得改。

判决卡之前，先重申全书的收窄声明：本书不主张统一架构必然失败，不主张三种感官对所有任务都不可替代；这两条是研究边界，不是前提。

判决一 · 学习即纠缠

项	内容
一句话	学习是特征纠缠（E）的建立、稳定、分化与重组，同时改写显露（S）与差异运行（D）；知识不是被搬进脑子或参数里的货物
主场	第一编（第1—5章）
站住的根据	小舟学骑车、和面揉出筋、多年不骑重新上车三个主例显示，学会前后变的是三维的组织方式，而不是「装进去多少」；三方程给出「三问都有正面回答才算学习发生」的检查办法，并能按缺口分辨半截子的学习（第3章）；结语把「训练与学习应当区分」「学习同时改写三维」列入已论证的第一本账
收窄处	这是本体论立场，不由某个模型的成败直接证明（结语第一节）；把大模型的参数连接看作特征纠缠，只是候选载体，对应程度属机制假说，不是已知事实（第3章、第12章）；三态是概念解释，不等于可测动力学指标（第4章）
推翻条件	若在多种模型上都找不到任何可测量的内部量，能对应纠缠的建立、稳定、分化、重组，并预测能力的出现与失效，则 E 在 AI 一侧只是比喻，「用 SDE 解释 AI 训练」的工程价值要大打折扣（结语假说五，对应第40章假说一与形式化方向一）；若在小型模型的持续训练中，表示关系图的变化与迁移表现无对应，形式化方向一判为不支持（第40章第六节表）；若持续训练中新任务只追加新表示、不改旧表示之间的关系，迁移却并不变差，纠缠重组假说不成立（第40章假说一）。若全书各条机制假说大多数都输了，母题本身应当被认真重审（结语第五节末）

项	内容
本书自认最弱处	F、G、H 是什么样的函数，什么时候能写成可计算形式，本书没有回答（结语第三本账一）；表示相似度一类度量对选层、选样本相当敏感，必须和行为层的迁移测试对照着看（第40章）

判决二 · 三界共生

项	内容
一句话	人在学习中同时发生理念、现实、自我三个世界，主体随世界一起发生
主场	第二编（第6—10章）
站住的根据	婴儿认妈妈显出前语言经验、身体活动与社会交往在概念之前；小舟被烫一下，「烫」字才有了分量；学分数显出符号与操作缠在一起抽象才有落脚处；周敏的教师自我在一届届学生与承诺中长成；切蛋糕一例显出三界互相改写（结语第二节）。结语把「主体与世界一起发生」列入已论证的第一本账
收窄处	「主体与世界共同发生」是本体论立场；「三股线共同纠缠出对象」「三条改写通道」都是机制假说，用日常场景说明，不是被实验测量过的模型（第6章、第10章边界节）；「共同世界」是相对同一性，不等于多数人共享的就是对的（第10章）
推翻条件	本体论层面，按结语给出的输法判：若这一立场长期组织不出比别的立场更好的问题，按它设计的实验一个接一个判负而只能靠改口保住自己，术语越用越含混，就在实际上被放弃（结语第五节末）。机制层面：只给声音、不给抱与喂养，或只给面孔照片，婴儿是否同样快地建立对某个人的偏好，是三股线假说的可检验推论，本书不预设结论（第6章第六节）；在人这一侧，撤扶手检验（一组让 AI 给终稿、一组只让 AI 提问，两周后独立完成同类与异类任务）若两组无差别，第38章「以迁移、重组和新能力检验学习」的设想要重新评估（第40章第六节表，机制假说五）
本书自认最弱处	这条判决的经验支撑主要来自构想的日常场景，没有本书自己的对照数据；一次概念学习在多长时间里、以什么形式影响行动，一次现实失败怎样可被观察地改变自我描述，本书只提出问题，交给教育学与心理学的研究对照（第10章第六节）

判决三 · 复合体是发生单位

项	内容
一句话	大模型训练发生的是集体人类—AI复合体的语言世界；具体人的使用、反馈与再训练，让它继续进入个人—AI复合体
主场	第三编（第11—15章）、第八编（第36—40章）

项	内容
站住的根据	写下文本的人、挑数据的人、定目标的人、打分的人和制度都进入训练（第11章）；方言与菜谱两例显出集体同一性既统一又容纳差异（第13章）；Waymo 公开的外部学习回路，经人类团队、评判器、仿真测试与安全审查才回写新一版系统，是这一发生结构的公开实例（第19章）；五件事的区分使「模型学会了」这句话重新有了意义（第37章）。结语把「训练讲集体复合体、使用讲个人复合体」与「五件事不可混用」列入已论证的第一本账
收窄处	「集体」不是「全体」，「参与」不是「同意」（第11章）；参数连接不是特征纠缠的物证（第12章）；集体同一性只是条件，不能变成单一答案的强制；连续性不是自动的好事，通道设计得不好，共同基础会一层层抹平个人差异（结语第三节）
推翻条件	若长期观察显示，人在与同一类模型的持续协作中表达与判断显著趋同、新问题与新对象越来越少，且各种通道设计都无法扭转，本书对「连续性是条件而非强制」的判断应当收回，第八编的教育与科研设想要重新评估（结语假说六）；不同评价方式训练出的模型，若在开放任务上方案多样性与评价方式无关，同一性收窄假说判为不支持（第40章假说六及第六节表）；撤扶手检验若两组无差别，「持久学习更可靠地落在人这一侧或人维护的外部记忆上」判为不支持（第40章假说五）
本书自认最弱处	谁的反馈被采纳、按什么标准、个人的新经验怎样进入共同基础而不被抹平，涉及制度设计，本书只指出它们是发生条件的一部分，不作政策表态，也没有给出答案（结语第三本账五）

判决四 · 物在三种 SDE 的纠缠中站稳

项	内容
一句话	视觉、听觉、触觉各自是完整的 SDE；对象同一性在三者的互动中形成、维持、失效与重建；显露预测不保证行动可靠
主场	第五编（第21—25章）、第六编（第26—30章）
站住的根据	挑西瓜时看、敲、托三路各有完整判断又互相纠正；摸黑找开关显出视觉退场后触觉与身体记忆怎样维持一个物；陌生弹性球分出同步出现、相关变化、相互制约三层，冲突推动归并、拆分与重建（第21—24章）；四条功能性判据把「物的意识」变成可检验的东西（第25章）；外观一样、一只抽过真空的玻璃罐，雨夜路面上薄水、积水与油污的反光，说明记录中不可区分的配置在新干预下后果不同（第27章）

项	内容
收窄处	三种感官不与 S、D、E 一一对应，也不与粒子、波、场一一对应；「三模态必要性」是研究边界，不是前提，对某类任务完全可能得出「不必要」（第25章第六节）；功能性判据不回答主观体验问题（第25章第五节）；多传感器融合不等于多种 SDE 纠缠，Waymo 博文提供的信息不足以判断它属于哪一种（第19章、第29章）
推翻条件	同等数据、计算与先验下，若主动互动组在对象同一性三项测法上均未显著优于被动融合组，且被动融合组也未显著优于拼接组，「三种 SDE 经由主动互动形成纠缠更利于对象同一性」在该任务上判负（第34章判负第一条）；若去掉某一路感官后三类指标都无可测下降，该任务上「这一路参与纠缠是必要的」不成立，须如实记录并修订「哪些任务需要哪几路」（第34章判负第四条）；在接触丰富、外观不足以区分配置的任务上，若纯视觉或「视觉＋简单拼接」在控制条件下达到相同的迁移与恢复水平，至少在这些任务上「共同互动」不必要，若在第24章所构想的冲突情境里都找不到优势，这条假说整体判负（结语假说二）；若冲突情境中系统的对象表示被某一路主导、其余被平均掉，且不损害任务表现，「冲突推动重组」不是必经机制（结语假说三，对应第40章假说三）。关于「显露预测不保证行动可靠」：若能力—预算曲线显示组间差距随预算稳定收敛到零，规模解释占上风，本书应当承认（第34章第四节末）
本书自认最弱处	三种感官的联合预测还没有被做成；现有局部研究只覆盖视觉与触觉两路，声音在其摘要中未见作为主要信号（第33章）；老彭的「方向盘发飘」能在多大程度上对应到机器的传感器信号，需要逐项对照实验（结语第三本账三）

判决五 · 任务组织世界，封装要能失败

项	内容
一句话	世界模型的出路是围绕具体任务组织必要的纠缠，以可检验的能力与适用边界形成封装，并为它写好会失败的条件
主场	第四编（第16—20章）、第七编（第31—35章）
站住的根据	各路世界模型以不同东西为「世界」，成功标准不同（第16章）；从「显露相似」跳到「发生机制相同」是推论缺口（第20章）；装修师傅看墙一例显出任务决定差异、粒度与时空范围（第28章）；成功率不等于可靠，可靠要落到适用条件（第35章）；第34章把核心主张写成三组对照、三类指标和事先写定的判负条件
收窄处	不主张统一架构必然失败；统一架构如果成功，也应当交出一份边界报告（结语假说四）；GraphCast 这类数据驱动天气预报是正面反例：在记录极丰富、过程相对稳定、检验天天发生的任务里，封闭记录可以走得很远，限制不在「用记录」本身（第20章）；现有路线已有发生取向，不给整个领域贴标签（第20章第五节）

项	内容
推翻条件	若一个统一架构在多种任务上同时达到专门模型的可靠性，并能清楚报告自己在每种任务上的适用条件与失效方式，本书关于专用模型协作更合适的判断应当修正（结语假说四）；任务专门模型与通用模型在任务内、任务外两头测，若任务内无优势或任务外不下降，任务边界假说判为不支持（第40章第六节表）；同等预算下，若被动记录训练在新干预下的后果预测、陌生物体同一性维持、出错后恢复速度上持续不差于主动互动训练，且在多种任务、多种物体上重复出现，主动干预假说应当放弃，第七编的训练组织方案相应改写（结语假说一，对应第40章假说二）；若主动互动组只在近迁移上占优，「对象层纠缠支持远迁移」判负（第34章判负第二条）；若主动互动组失配恢复不更快，或恢复快而旧能力保持率明显更低，「显式失配归因有利于纠缠重组」判负（第34章判负第三条）；若组间差异小于不同随机种子之间的波动，结论只能写「未能分辨」，不得写「趋势支持」（第34章判负第五条）
本书自认最弱处	第34章写的是实验设计的骨架，不是跑完的实验，所有结论悬置到实验之后；统计标准、重复次数、效应量门槛留给实施者事先登记，本书不预设数值（第34章）；六步法、九步法是本书提出的训练组织假说，按九格检查、先补 S1・D1・E1 定式三角是否更省数据，尚待实验（第32章）

本书自认的最弱处

把五张卡合起来看，本书最可能先从下面几处破。

第一，特征纠缠在机器一侧的可测性没有证成。三方程是概念方程，F、G、H 没有形式，伞模型、三态的可测指标也都在待形式化之列。若一直找不到能对应纠缠变化、又能预测能力出现与失效的内部量，全书对 AI 训练的解释会退成一套好用的比喻（结语第三本账一、二，假说五）。

第二，全书没有自己的实验数据。第七编的出路部分最该接受检验，也正是迄今只有设计、没有结果的部分；第34章的对照、第40章的六项研究计划，都要等别人照着做出来才能结算。

第三，三种感官的共同互动证据最薄，尤其是听觉。已核对的局部研究只覆盖视觉与触觉，且限于特定的手、特定的任务集；它们与本书机制假说方向一致，但方向一致不等于证明（第33章）。

第四，人一侧的论证多靠构想场景。三界共生、复合体的连续性、教育与科研的设想，都用日常场景说明，长期效应需要很长时间的观察，本书只能提出问题（第10章、第40章第六节末）。

第五，集体同一性的治理、身体感觉与机器信号的对应、主观体验问题，本书明确留作未解，不用功能判据代答，也不作政策表态（结语第三本账）。

附录三　世界模型技术路线速查

下表只收正文讨论过的系统。「公开机制」一栏只取正文已经写下、并经质检对过原始材料的内容，不增加新的数字；数字凡出现，均为作者自报。「发生学读法」是本书的解释，属于本体论立场或机制假说，不是对模型内部的测量。

系统	公开机制一句话	本书在哪一章讨论	本书的发生学读法一句话
Genie（arXiv 2402.15391）	在无动作标注的平台游戏录像上，用视频分词器压缩画面、由隐动作模型推断帧间动作并压成很小的离散集合，再由动态模型以过去画面与隐动作为条件逐帧生成可操作的二维世界	导论；第17章；第40章	「记录中的互动」可以被压缩为可调用的条件；动作空间的粒度由人事先替模型定下，决定了哪些差异被抹平；它不直接说明新的物理互动被学到
Genie 3	据公开报道可按文字提示实时生成可交互环境、在一段时间内保持前后一致、可用文字插入事件；数据与结构未公开〔待核〕	第17章	只有二手报道，按「演示不是机制、不替公司补全、不从沉默推出否定」三条规矩，判断悬置
Runway 世界模型（GWM-1）	公开页面称建立在其视频生成模型之上，自回归逐帧实时生成，可用相机位姿、事件、机器人位姿和语音控制，分为环境、角色、机器人三个后训练版本；参数、数据与动作条件的接入方式未公开	第16章；第17章	产品化系统往往只公开功能与演示；能说的只到公开所及，其互动是否来自新互动无从判断
V-JEPA 2 与 V-JEPA 2-AC（arXiv 2506.09985）	在大规模互联网视频上「遮住一部分，预测被遮住部分的表示」做无动作预训练；再用少量已录好的机器人视频冻结编码器、后训练动作条件预测器，以目标图片在表示空间里规划，零样本部署到机械臂	导论；第18章；第40章	训练目标从「显露得像」移向「对行动有用」，是一种向任务靠拢的发生取向；执行后看新画面再规划属于反馈控制，就公开所及不是在新环境互动中持续学习〔在线参数更新待核〕；以目标图片定任务，对摆放类任务自然，对力道类、时机类任务需要别的信号

系统	公开机制一句话	本书在哪一章讨论	本书的发生学读法一句话
V-JEPA 2.1（arXiv 2603.14482）	按摘要把重点放在稠密特征：可见与被遮位置都参与训练信号，在编码器多个中间层施加自监督，并用多模态分词器同时训练图像与视频	第18章；第40章	仍属特征预测路线的推进；作者自报的提升本书照录，不作独立评价
DreamerV3（《自然》2025）	学一个环境模型，在紧凑的内部状态上「想象未来」训练策略；经验来自智能体自己的行动，新策略回到环境产生新经验，循环往复；一套固定超参数应对多种任务	导论；第18章；第20章	差异从「读出」转向「做出」，想象的错误要回到环境里暴露并纠正，是 E=H(S,D) 被走通的一例；但它的「水」是游戏与仿真，在真实机器人上的表现不在摘要范围内
Cosmos 3（arXiv 2606.02800）	按摘要称「全模态世界模型」，用统一的混合 Transformer 架构同时处理与生成语言、图像、视频、声音和动作序列，面向具身智能体，代码、权重与合成数据集开放发布	导论；第19章；第23章；第40章	「统一架构」的清楚实例；声音与动作是并列生成，还是在同一事件里相互约束，摘要层面答不了〔待核〕；合成数据扩大的是记录中的差异，不能替代来自真实世界的新差异
World Labs Atlas	按官方博文，从零预训练、原生处理文字、图像、视频与三维；把输入编码进「空间上下文」，每张图像都放在一个三维位置上，输出新视角图像、深度与显式三维结果，也可从仿真机器人传感器视角输出彩色图与深度	导论；第19章；第40章	为对象同一性的空间一侧提供了工程承载；但空间上的同一只是对象同一性的一部分，机器人从中得到的是一个视觉—空间训练世界，有多少能带回真实手感要靠真机实验
Waymo 基础模型（2025年12月博文）	传感器融合编码器把摄像头、激光雷达、毫米波雷达随时间融合，另有驾驶视觉语言模型处理罕见复杂情形，共同送入世界解码器；内部回路在闭环仿真中训练，外部回路从真实里程挑出欠佳行为，生成替代、仿真检验、安全确认后再部署	导论；第16章；第19章；第29章；第40章	说明实际运行的驾驶系统不是纯视觉、也不是纯被动记录；外部回路是集体人类—AI复合体发生结构的公开实例；融合是否构成多种 SDE 纠缠，博文不足以判断；车辆自身状态与接触信息的位置〔待核〕

系统	公开机制一句话	本书在哪一章讨论	本书的发生学读法一句话
DexTacWAM（arXiv 2609.24976）	按摘要，视觉—触觉「世界—动作模型」：逐指编码触觉、压缩后注入视频扩散世界模型，联合预测未来的画面与触觉状态并输出动作；消融显示只去掉触觉世界建模，成绩明显下降	导论；第22章；第33章；第40章	把触觉当作要预测的状态而非只当输入，与「改变预测对象比增加输入更要紧」方向一致；但这是特定平台、特定任务的局部结果，方向一致不等于证明
TouchWorld（arXiv 2607.07287）	按摘要，「预测—反应」式触觉基础模型：高层视觉—语言规划并预测触觉子目标，中层视觉—触觉策略生成名义动作，底层用触觉与本体感受做高频残差修正	导论；第22章；第33章；第35章；第40章	慢的任务推理与快的接触修正分层，对应「控制反馈不等于参数训练」；修正经验是否收回去更新模型，摘要未提〔待核〕
GraphCast（arXiv 2212.12794）	直接用再分析资料训练、做十天以内的全球中期预报，作者报告在大多数检验目标上优于当时的业务确定性预报	第20章	本书正面对待的反例：在记录极丰富、过程相对稳定、检验天天发生的任务里，封闭记录可以走得很远；限制不在用记录本身，而在任务是否要求记录之外的新区分；罕见极端天气下的表现〔待核〕
直觉物理研究（arXiv 2502.11831）	在表示空间里做预测的视频模型，在物体恒存、形状一致等「违背预期」测试上明显高于随机，直接预测像素的模型与多模态大语言模型接近随机	第21章	说明视觉显露中的对象连续性可以从大量观看中训练出来；不说明这个对象被托起、被敲击时会怎样，不能当作「视觉足以产生物的意识」的证明

附录四　全书主例与人物

人物

全书陪读人物四位，均为构想，所在场景都是「我构想的场景」，不对应真实个人与真实公司。一次性出场的路人不取名。

姓名	设定	在书里主要做什么
周敏	41岁，长沙中学物理老师，长沙郊区长大；大学时天天骑车，工作后十几年没骑；小舟的妈妈	教育、讲解、追问；多年不骑重新上车，敲西瓜，炖红烧肉时辨出模型的错，和 AI 一起备课
周小舟	9岁，小学四年级，周敏的儿子；骑车只在小区内，正在学骑车、学游泳	学习与身体经验：学骑车、学「热」、学分数、看录像学游泳、握弹性球、用 AI 学写作文
老彭	58岁，网约车司机，驾龄三十年，讲长沙话；周敏的邻居	驾驶、听车、手感：雨夜停车，听异响，方向盘发飘，陌生情境先慢下来
林岚	29岁，深圳一家机器人公司的算法工程师，周敏以前的学生	工程现场与模型训练：用大模型查资料改代码，带团队训练机械手拧瓶盖、捏豆腐

各编主例一览

本书内同一个例子只当一次主例，别处可以一句带过；下表列主例所在的章。

编	主例	所在章
导论	模型写出「刹车」这个词，与老彭在雨夜真把车停住	导论
第一编	小舟学骑车：摔、扶、放手	第1章，第3—5章续用
第一编	和面揉出筋	第2章
第一编	周敏多年不骑，重新上车	第1章、第5章
第二编	婴儿认妈妈；戴口罩的妈妈	第6章
第二编	小舟学分数	第7章，第10章再看一遍

编	主例	所在章
第二编	小舟学「热」，被烫了一下	第8章
第二编	周敏的教师自我	第9章
第三编	一份菜谱的一生：外婆的红烧肉被千万人改	第11章、第13章
第三编	方言与家乡话，湖南话里的「冇」	第12章、第13章
第三编	「你真行」这句话	第14章
第三编	周敏炖红烧肉时查询模型、辨出错误	第15章
第四编	天气预报的几种做法	第16章
第四编	看录像学游泳与真下水	第17章
第四编	导航软件「重新规划路线」	第18章
第四编	气象站的多传感器与泳池里的多感官	第19章
第五编	挑西瓜：看、敲、托	第21章、第23章
第五编	摸黑找开关	第22章
第五编	陌生弹性球的握压、敲击与释放	第24章、第25章
第六编	一只瓶盖、一只抽过真空的玻璃罐：人手怎样对付不同的盖子	第26章、第27章
第六编	老彭的雨夜：路面反光底下不一样	第27章
第六编	装修师傅看墙、打孔、铺砖	第26章、第28章、第30章
第六编	老彭开车：雨夜、异响、方向盘手感	第29章
第七编	林岚团队的机械手学拧瓶盖、捏豆腐	第31—35章
第七编	做菜试咸淡	第32章、第35章
第八编	周敏和 AI 一起备一节浮力课，一个学期的备课节奏	第36章、第38章
第八编	小舟用 AI 写《我家的小动物》	第37章
第八编	一个研究小组研究剁辣椒为什么胀罐	第39章
第八编	学用筷子夹花生（形式化方向的小例）	第40章
结语	回收全书主例，不引新主例	结语

附录五　全书说明

生产方式。本书由王德生立题、拟定提纲并对关键问题作出裁定。写作阶段由Claude 统领十个智能体，按提纲分编并行执笔；五个智能体交叉质检，逐章核对术语、三层论证纪律、主例分配与技术事实；再由五个智能体终审统稿。执笔的智能体不给自己的稿子评分。全书是这样一个人—AI复合体的产物，书中凡称「我」，都在这个意义上理解。

三层论证纪律。全书每章把三种话分开说：本体论立场、机制假说、经验事实。每章末的小账写明论证主要落在哪一层。模型的局部成功或失败，不直接当作本体论的证明；从技术事实走到本体论结论，中间必须经过机制假说一层，并把推论的每一步摆出来。本书也不预写「驾驶 AI 全部是纯视觉」「现有世界模型都没有互动」「统一架构必然失败」「三种感官对所有任务均不可替代」这类结论。

技术资料核验到哪一步。本书讨论的技术系统，以截至 2026 年 10 月可见的版本为准：arXiv 论文以摘要与正文所见为准，企业系统以官方页面与博文为准，质检时逐条打开公开材料核对过。2026 年的几项预印本与博文（V-JEPA 2.1、Cosmos 3、World Labs Atlas、DexTacWAM、TouchWorld）按当时所见的摘要与博文转述，若原始资料日后改版，以原文为准。只有二手报道的系统（如 Genie 3），去掉具体参数与日期，只作一般陈述。书中引用的数字均为作者或公司自报，本书照录，不作独立评价。本书不编论文结论，不编原话，不编数据。

〔待核〕的含义。正文中以六角括号标出的〔待核〕，表示这一处的细节在成书前还没有拿到足以下判断的原始材料，或原始材料只到摘要一层；它不影响所在段落的论证，但读者不应把它当作已核实的事实引用。全书剩余的待核处见正文灰色小字，成书前再核一遍。

人物场景为构想。周敏、周小舟、老彭、林岚以及书中所有家庭、工程团队、科研小组的场景，均为构想，用来把道理摆到眼前，不是实录，也不对应任何真实

个人与真实公司。第34章的对照实验是构想的实验方案，供研究者参照改写，不是已经跑完的实验。

不构成专业建议。本书讨论 AI 的训练与学习、世界模型与驾驶、教育与科研，仅用于思想、研究设计与概念澄清，不构成工程、医疗或政策建议；涉及感官缺失之处不作任何医学判断；书中提出的各项检验与读数只用于研究，不进入任何考核。

版本与修订方式。这是本书的第一版。全书把可以被推翻的地方都写在了附录二：若以后有实验结果落在判负条件里，相应部分应当修改；若有更好的立场能统一解释更多现象、引出更多可检验的问题，本书应当让路。勘误与修订将在后续版本中说明修订了哪里、依据是什么。