

判断的品味

AI 时代，如何沿着布鲁姆的梯子练出评价力

竹小竹8167

费曼式写法 · 由多个 AI 代理撰写与互相审校

导读

AI 时代，如何沿着布鲁姆的梯子练出评价力

先做一个十秒钟的小实验。

打开你常用的 AI，让它「写一段关于拖延的话」。它会给你一段通顺、正确、毫无记忆点的文字。再让它写第二遍、第三遍。现在问自己一个问题：**这三段里哪段最好？为什么？**

如果你发现自己答不上来——只能说「都还行」或者「这段读着顺一点」——恭喜你，你刚刚亲手摸到了这个时代最值钱的那个缺口。AI 三秒钟能生成的东西，你可能三分钟都判断不了。生成和判断之间，裂开了一道口子，而且越裂越大。

这本书就是来填这道口子的。

一个过时的假设

我们从小到大接受的教育，默认「产出」是稀缺能力：考试考你写不写得出来，公司奖你做得出来不。「判断好坏」被视为附属品——你会做，自然就会评。这个假设在 AI 时代失效了：当任何产出都能三十秒生成，「做得出来」不再是瓶颈，「知道哪个好」才是。位置上，人从「动手的人」变成了「验收的人」，而验收这件事，偏偏没法外包——AI 可以给你一百个选项，「就它了」这三个字的责任永远在人身上。

这本书的地图

我们把「评价能力」这门看似玄学的本事，拆成可以逐个练习的部件。拆解的梯子借用教育学的经典框架——布鲁姆目标分类学：记忆、理解、应用、分析、评价、创造，六级层层垫高。评价站在第五级，它评不准，病根常常在下面几层；而 AI 恰好从塔基往上承包，越到塔尖越需要人守。

顺着这个结构，全书十章：先看清楚评价这个动作为什么值钱（第一、二章），再把它拆成标准、证据两道工事逐个攻克（第三、四、五章），然后处理最像天赋的部分——品味怎么喂出来、判断怎么写成理由、怎么知道自己什么时候在瞎判断（第六、七、八章），最后落到日常训练法和人机分工（第九、十章）。

每一章都尽量回答「那我明天具体做什么」。如果读完某章你说不出下一步动作，跳过它，别客气。

读法建议

可以从头顺着读——章节是递进设计。也可以从中间进：手里正好有一份 AI 生成的东西要验收，直接读第五章（检查）；总被人问「好在哪」却答不上来，直接读第七章（把判断写成理由）。每章独立成篇。

最后一个小要求：读的时候，手边最好放着一份你最近要判断的东西——一篇稿子、一个方案、一份代码。这本书不是用来看的，是用来对照着用的。判断能力长在判断的动作里，不长在阅读里。

第一章 · 最值钱的那句话

想象两个场景。

第一个：老板让你写一份竞品分析报告。你打开 AI，输入一句话，三十秒后，一份结构工整、数据翔实、措辞专业的报告躺在屏幕上。你又让它写了三个不同风格的版本，挑了挑，把其中一份发给了老板。

第二个：三个下属各自交上来一份方案，AI 也生成了一份。四份东西摆在你面前，措辞都很漂亮，结构都很工整。老板问你：「你觉得哪份靠谱？」

第一个场景里，AI 是主角，你更像一个收发员。第二个场景里，AI 帮不上你——它生成的那一份，恰恰也是需要被判断的一份。这两件事中间，隔着这本书要讲的东西：**生成已经不值钱了，判断才值钱。**

一件悄悄发生的倒挂

过去几十年，职场奖励的是「能产出」的人：能写出漂亮 PPT 的，能写出顺滑代码的，能做出完整方案的。「产出」门槛高、耗时长，所以值钱。而「评价」——说这页 PPT 哪里不对、这段代码有没有坑、这个方案能不能落地——是少数资深者的事，藏在幕后，不太被单独提起。

AI 把这件事倒了过来。产出的成本几乎降到零：任何人都能在一分钟内拿到一篇及格线以上的文章、一段能跑的代码、一份像样的方案。当供给无限，稀缺的就不再是供给本身。真正稀缺的变成了：**面前摆着十份都「看起来不错」的东西，谁能指出哪份真的好、好在哪里、哪份埋着雷。**

这就是评价能力——对一个东西的质量、价值、可信度做出判断，并且说得出依据的能力。注意后半句：「说得出依据」。随口一句「我觉得不错」不算评价，那只是口味反应。评价是一种可以拿出来给别人检验的判断。

大白话：以前值钱的是「会做的人」，现在值钱的是「识货的人」。AI 让「会做」变得人人可得，于是「识货」成了分水岭。

为什么偏偏是「评价」

你可能会问：人的本事那么多——共情、领导、创意——为什么单拎「评价」出来说？

因为在一个 AI 深度参与的工作流里，评价是**所有其他环节的总开关**。你给 AI 出一个题，它给你十个答案，选哪个、改哪里、要不要推翻重来，每一步都是评价。你让 AI 写代码，它写得飞快，但「这段代码敢不敢上线」是你的判断。你让 AI 起草合同，它引经据典，但「这一条会不会坑了我们」是你的责任。

换句话说，AI 越强，人的位置越往链条的末端挪：从「动手做的人」变成「验收的人」。而验收这个位置有个残酷的特点——**它没法外包**。AI 可以帮你生成一百个选项，但「就它了」这三个字的责任，永远落在人身上。出问题的时候，没有人会接受「这是 AI 选的」作为解释。

好消息：它是一门可以拆解的手艺

「评价能力」听起来很玄，像某种天赋——有人就是「有眼光」，有人就是不行。这本书要做的就是拆掉这个神秘感。

我们用的拆解工具，是教育心理学里一个经典框架：布鲁姆目标分类学（Bloom's Taxonomy）。它是 1956 年由教育学家本杰明·布鲁姆带领的一群学者提出的，把人的认知活动从低到高分成几个层级——从最简单的「记住」，一直到最高级的「创造」。2001 年，安德森和克拉斯沃尔等学者把它修订成今天我们常用的六级版本：**记忆、理解、应用、分析、评价、创造**。

这个分类学本来是老师设计教学目标用的：教一门课，不能只让学生背下来（记忆），得让他能讲明白（理解）、能上手用（应用）、能拆开看结构（分析）、能判断好坏（评价）、最后能自己做出新东西（创造）。它像一架梯子，每一级都踩在下面几级之上。

对我们来说，它恰好回答了两个关键问题：

- **评价能力建立在什么之上？** 评价站在梯子第五级，下面垫着记忆、理解、应用、分析四层。一个人评不了，往往不是「没眼光」，而是下面的某一级缺了功课。这给了「怎么补」的地图。
- **AI 时代人的位置在哪里？** AI 承包的主要是梯子下面几层——查资料、做总结、套模板、拆结构都又快又好。越往上走，尤其是「评价」这一级，越需要人来守。这不是安慰人的话，后面几章会讲清楚为什么。

这本书的走法

全书围绕一个问题推进：**当 AI 什么都能生成，人凭什么说它好不好——以及这个「凭什么」怎么练出来。**

我们会先花一章把布鲁姆的梯子讲透（第二章），然后钻进「评价」这一级内部，看它到底由哪些动作构成：标准从哪来（第四章）、证据怎么核（第五章）、眼光怎么喂出来（第六章）、判断怎么写成能站得住的理由（第七章）。再往后是更细的功夫：知道自己什么时候在瞎判断（第八章），以及把评价练成日常肌肉的训练法（第九章）。最后一章回到 AI：人机怎么分工，人守住的那个位置到底长什么样。

有一个提醒放在开头：这本书里没有任何一章会说「你要培养批判性思维」「你要提升审美」这种正确的废话。每一章都试图回答一个更硬的问题——**具体做什么动作，评价能力才会长出来。**如果某一章读完你说不出「那我明天该干嘛」，那一章就是失败的，你完全可以跳过它。

好，上梯子。

第二章·布鲁姆的梯子

上一章立了个flag：评价能力不是天赋，是可以拆解的手艺。这一章把拆解的工具本身讲清楚——布鲁姆目标分类学到底分了哪几层，层与层之间是什么关系。别把它当教育学名词背，我们用一个贯穿全章的例子来走一遍：**学做一道菜，比如番茄炒蛋。**

六级梯子，一级一个动作

第一级：记忆——背下菜谱

你记住「番茄炒蛋要先炒蛋、盛出、再炒番茄、混合、加糖加盐」。这是记忆：把信息原样存进来，能原样取出来。它不问懂不懂，只问在不在。

第二级：理解——说得为什么

有人问你「为什么蛋要先盛出来？」你能答：因为蛋老得快，一直留在锅里会柴；番茄要炒出汁，耗时更长。这是理解：你能用自己的话解释一件事，知道步骤背后的道理，而不只是复读步骤。检验理解的办法很简单——换一个问法还答得上，就是懂了；只能背原文，就是没懂。

第三级：应用——换个厨房也能做出来

你去朋友家，灶火比家里猛，番茄特别酸。你调小了火，多放了一勺糖。这是应用：把懂的道理用到一个新情境里。背菜谱的人在陌生厨房会翻车，懂原理的人会自动调整。

第四级：分析——拆开看哪步出了问题

你做出来一盘蛋是蛋、番茄是番茄、汤水汪汪的「番茄炒蛋」。你能拆开诊断：蛋没炒到位就混进去了？番茄出汁太多没收干？顺序错了？这是分析：把一个整体拆成部件，看清各部件之间的关系和因果。分析不问「好不好」，只问「它由什么构成、哪里连着哪里」。

第五级：评价——这盘菜到底行不行

现在，桌上摆着三盘番茄炒蛋，一盘你做的，一盘朋友做的，一盘 AI 菜谱指导你室友做的。有人问：哪盘好？凭什么？

注意这一刻发生了什么变化。前四级功夫都派上了用场——你记得标准做法（记忆），懂得每步的道理（理解），能自己实操（应用），能拆解这盘菜的得失（分析）——但第五级要你做的是一件新事：**给出一个站得住的裁决。**

「这盘好」不能只是口味反应。你得说：按「蛋要嫩滑、汁要收浓、酸甜平衡」这个标准（这是**标准**），这盘蛋老了、汁稀了、但酸度合适（这是**证据**），所以整体七分，输在火候。这就是**评价**：依据标准、对照证据，对事物的质量做出可辩护的判断。它是本书的主角，第三到第八章会把它拆得更细。

第六级：创造——发明你自己的版本

最高一级是**创造**：在掌握所有下面层级之后，做出新东西——你发明了「番茄炒蛋配烤馒头片蘸汁吃」，或者针对健身人群发明了低油高蛋白版本。创造不是凭空来，它踩在下面五级之上。

梯子结构里藏着三个关键事实

第一，高层功夫建立在低层之上，跳级会塌。一个从没做过菜、也没看懂菜谱的人，没资格评价一盘番茄炒蛋——他的话不是判断，是噪音。布鲁姆的梯子解释了生活中大量「评价失灵」：有人分析头头是道却从没动手应用过，评出来的东西落不了地；有人连基本事实都记错，评价自然建在沙子上。**一个人评不准，病根常常在梯子下面某一级。**这就是为什么本书说评价能力是「可以练的」——它不是一颗单独的果子，它是整架梯子的长势。

第二，层级越高，越没有标准答案，越需要人来负责。「菜谱是什么」有标准答案；「这盘菜值七分还是八分」没有。越往上走，问题越开放，判断的分量越重。评价和创造这两级，本质上是「在没有标准答案的地方做决定并承担责任」——这正是机器最难替人干的事。

第三，AI 正在从塔基往上承包。查菜谱？AI 秒回（记忆）。讲原理？它讲得比多数厨师清楚（理解）。给你按手头的食材出做法？可以（应用）。拆解你失败的那盘菜？给个照片它能分

析个八九不离十（分析）。梯子下面四层，AI 已经是廉价而优秀的供应商。但到第五级——「这盘菜该不该端给客人、这套方案该不该拍板、这段代码该不该合并」——事情变了。

大白话

大白话：AI 能替你做卷子，不能替你担分数。梯子越往上，「对不对」越变成「你敢不敢签字」。签字这件事，外包不出去。

一个常见的误会：评价就是「会挑刺」

很多人对「评价能力」的印象是：眼光毒、嘴巴利、一眼看出毛病。这是把评价矮化成了挑刺。挑刺只是评价里最便宜的部分——找缺点永远比定标准容易，说「这不行」永远比说清「怎样算行」省力。

真正的评价包含三件挑刺给不了的东西：**标准**（用什么尺子量）、**证据**（量的结果是什么）、**权衡**（几项指标冲突时怎么取舍）。一个只会挑刺的食客会砸掉所有餐馆；一个会评价的食客能告诉你「这家火候顶级、调味保守，适合带长辈来，不适合约会」。后者的话能指导行动，前者的不能。

这三件东西——标准、证据、权衡——就是接下来几章要逐个攻克工事。先让它们的样子看清楚：评价不是一句话，是一个结构。

本章小结

布鲁姆的梯子把认知分成六级：记忆、理解、应用、分析、评价、创造。梯子三个性质决定了这本书的写法：高层踩低层（所以要练就要从塔基查起）；越往上越没标准答案（所以评价需要人来负责）；AI 从塔基往上承包（所以人的价值集中在评价这一级）。下一章，我们把镜头怼到「评价」这一级内部，看它到底由哪几个动作构成。

第三章·评价不是「我觉得」

上一章结尾说了：评价不是一句话，是一个结构。这一章把这个结构拆开。拆完你会发现，「评价」这个动作内部其实有两道工序，它们用的肌肉不一样，练法也不一样——很多人评价能力弱，其实是卡在其中一道上，却一直在练另一道。

先立个定义，不然全书没法谈

本书给「评价」下一个可操作的定义：

评价 = 拿一把明确的标准，去对照一份可查的证据，得出一个可以辩护的结论。

三个部件，缺一不可。**标准**回答「用什么量」；**证据**回答「量的对象是什么、量出来的结果是什么」；**可辩护**回答「别人不服时，你能不能把这个判断的来路摆出来给他检验」。

对照这个定义看日常语言，会发现我们嘴里大量「评价」其实不合格：

- 「这个设计挺高级的。」——没有标准。什么叫高级？对谁而言高级？
- 「我觉得这篇写得不好。」——没有证据。哪一段、哪一句、为什么？
- 「反正就是不行，你不懂。」——拒绝辩护。判断的来路被封死了。

这些话不是评价，是意见——一种不需要依据、也不接受检验的个人反应。意见没有错，人人都有意见；但意见不解决问题。当老板说「四份方案你挑一个」，他要的不是你的意见，是你的评价：能摆出来给全组人检验、最后能为之负责的那个判断。

大白话

大白话：意见是「我喜欢」，评价是「按这把尺子量，它三米二」。前者没法争论，后者可以——而可以争论恰恰是它的价值，因为错了能被发现。

评价的两道工序：检查与评判

2001 年修订布鲁姆分类学的那批学者（安德森、克拉斯沃尔等）做了一件很细的事：他们把「评价」这一级又拆成了两个子动作。

第一个叫检查（checking）：发现一件东西内部的矛盾和错误。它问的是「这东西自己跟自己打架吗？它说的事实是真的吗？它的逻辑链条断没断？」

第二个叫评判（critiquing）：拿一件东西去对照外部的标准，判断它好不好、合不合用。它问的是「按这个目的、这个标准、这个受众，它够格吗？」

用体检打比方最清楚。检查是看化验单本身有没有问题：数值自相矛盾吗？样本是不是拿错了？评判是拿着化验单对照健康标准：这个血压对这个年龄的人来说危险吗？需不需要干预？

再用工作场景过一遍。同事交来一份市场分析：

- **检查**：里面的数据是真的吗？引用来源存在吗？第三章的结论和第一章的数据一致吗？「增长 300%」是从哪个基数算的？
- **评判**：就算数据全对，这份分析回答了我们真正的问题吗？对我们这个阶段的决策有用吗？深度配得上它要支撑的那个决定吗？

看出来了吗？这两道工序的失败方式完全不同。**检查不过关的东西是「错的」，评判不过关的东西是「不对题的」**。一份报告可以数据全真、逻辑全顺（检查满分），却答非所问、没法用（评判零分）；反过来，一份洞察极准的报告里可能埋着一个致命的错误数据。高手两道关都过，而且清楚地知道自己此刻在做哪一道。

为什么这个拆分在 AI 时代特别要命

因为 AI 的输出有一个非常独特的失败模式：**它极擅长通过「评判」的第一印象，又极擅长埋「检查」才能发现的雷。**

AI 生成的东西，往往结构漂亮、措辞专业、看起来很切题——在评判这道工序的初筛里容易拿高分。但同一篇输出里，可能有一个不存在的引用、一个编造的数字、一处前后矛盾的论述——这些只有耐着性子做检查才抓得到。人类作者写错东西时往往「看起来也有点勉

强」，错误和水平大体匹配；AI 不一样，它的语言质量和事实质量是**脱钩**的。文笔的可信度不再能代理内容的可信度。

这意味着面对 AI 输出，老编辑的直觉（「读着顺，应该没问题」）会系统性地失灵。你必须把「检查」从后台直觉升级成一道独立的、刻意的、有流程的工序。第五章会专门讲检查的功夫怎么练。

还有一个隐藏前提：你得先知道「这是要拿来干嘛的」

评判工序里藏着一个经常被跳过的步骤：**明确目的**。同一份方案，用来「探索方向」还是「明天上线」，评价标准天差地别。同一段代码，原型阶段「能跑就行」是优点，生产环境「能跑就行」是事故。

很多评价争执的根源，是两个人在用不同的目的量同一件东西，却以为在争同一个问题。一个说「这方案太糙」，另一个说「这是用来试水的，糙是对的」——其实他们都对，只是目的没对齐。所以成熟的评价者在开口前会先问一句：**这东西是拿来干嘛的？**把这个答案写下来，后面的评判才有靶子。

本章小结

评价 = 标准 × 证据 × 可辩护，三者缺一就是意见。评价内部有两道工序：检查（对内部，找错）和评判（对外部，定好坏），失败模式完全不同。AI 输出的麻烦在于文笔质量与事实质量脱钩，让检查从直觉变成了必须刻意执行的工序。而一切评判的前提是先把目的说清楚。接下来两章，一章攻标准，一章攻证据。

第四章 · 先给自己造一把尺子

第三章把评价拆成了「标准 × 证据 × 可辩护」。这一章攻第一个部件：标准。它是三件里最反直觉的一个——多数人以为自己缺的是「眼光」，其实缺的是一把写下来、能递给别人用的尺子。

一个实验：不给尺子，人人都在乱量

找五个人，分别让 AI 写一篇「产品介绍」，然后请这五个人各自评价哪篇好。你大概率会得到五种答案：有人说「最有文采的」好，有人说「最短最清楚的」好，有人说「数据最多的」好，有人说「最有创意的」好。他们不是在争论哪篇好，他们是在用五把不同的尺子量同一堆东西，只是谁都没把自己的尺子亮出来。

这个实验揭示了一个朴素但常被无视的事实：**没有标准，就没有评价，只有口味碰撞**。而口味碰撞是不可调和的——因为争论双方根本不在量同一个东西。前面说的「先对齐目的」是第一步；目的定了，下一步就是把目的翻译成可操作的尺子。

尺子的学名：rubric

教育界给这种尺子起了个名字：rubric（评分细则）。它是一张表，横轴是「评价的维度」，纵轴是「每个维度的好坏等级」，格子里写清楚每个等级长什么样。

比如评价一篇文案的 rubric 可能是：

- **准确性**：有没有事实错误？（好 = 零错误且来源可查；差 = 有一处以上硬伤）
- **切题度**：是否回答了目标读者真正的问题？（好 = 每段都指向读者的一个真实疑问；差 = 自说自话）
- **行动性**：读完知不知道下一步干嘛？（好 = 有明确的下一步；差 = 读完只觉得「有点道理」）

注意 rubric 的关键设计：**每一级都要写出可观察的样子**，而不是「好/中/差」三个空洞的圆圈。「准确性：优秀」不是标准，「准确性：所有数字都能指出来源」才是。前者是把判断重新藏回黑箱，后者才是把判断摊开在阳光下。

大白话：rubric 就是把「什么叫好」从一种感觉，翻译成一张可以两个人一起核对的清单。尺子一旦写下来，争论就从「谁的品味好」变成「哪个格子该打勾」——后者是能解决的。

标准从哪来？三条路

「尺子要自己造」听着吓人：我又不是专家，凭什么定标准？其实标准有三个现成的来源，由浅入深：

第一条路：抄。几乎所有成熟领域都有现成的公开标准。代码圈有代码评审清单（能不能读懂、有没有测试、边界情况处理了吗）；写作圈有「奥威尔六条」（能用短词就不用长词、能删的字就删，等等）；招人有岗位说明书。新手的第一把尺子不需要原创，抄一份业内现成的，先用起来。用的时候你会碰撞、会不满，这个不满就是你迭代尺子的起点。

第二条路：拆。找你公认最好的两三个样本，拆它们「好在哪」。拿三篇你认为顶尖的同行报告，逐段标注：这段为什么有力？这段如果删掉会怎样？拆上十几个样本，你会发现重复出现的特征——那就是你所在领域的事实标准。这个过程本身就是第六章「识货是喂出来的」要讲的核心功夫。

第三条路：写。在动手做一件事之前，先把「做成什么样算合格」写下来。这在工程上叫验收标准，在管理上叫目标对齐，在日常生活里就是「点菜前先说想吃辣还是想吃淡」。事前写标准有一个神奇的功效：它防止你事后挪动球门。人有一种本能，看到一个东西的成品后，会不自觉地拿它反过来定义「好」——先写尺子就是为了防这个。

AI 时代的标准学：出题的能力

这里出现一个 AI 时代特有的转折。当 AI 负责生成、你负责验收时，**你写给 AI 的任务说明，本质上就是你给它的评分标准。**「帮我写个方案」是无标准的任务，你只能拿到无标准的输出，然后用「感觉」去验收——两头都靠运气。而「给 X 类客户写一份一页纸方案，必须回答他们的三个顾虑，语气克制不用感叹号」是一段带尺子的任务，验收时你有章可循。

所以这本书反复强调：评价能力不只是「事后挑货」，它前置为「出题的能力」。**会把「好」说清楚的人，才能把「要」说清楚。**这两个是同一门手艺的正反手。这就是为什么评价能力

是 AI 时代的元技能——它同时决定了你拿到的东西的质量上限，和你验收时的可靠程度。

一个诚实的提醒：尺子永远不完美

写下标准不会让你从此客观公正。rubric 有三类已知毛病：它会漏（真正重要的东西没被写进去）；它会偏（写的人把自己的偏好当成了普遍标准）；它会僵（领域变了尺子没变）。所以标准要当成「版本 0.1」来用：写下来，用它评，评完复盘「这把尺子这次漏了什么」，然后改。标准是消耗品，不是传家宝。

但「不完美的尺子」和「没有尺子」之间隔着一条鸿沟：前者可以改，后者连「哪里错了」都无从谈起。这就是评价能力的第一个硬功夫——**敢于把「什么叫好」写下来，并接受它会被证伪。**

本章小结

标准是评价的第一个部件，没有它就只有口味碰撞。可操作的标准叫 rubric：维度 × 等级，每级写出可观察的样子。标准有三个来源：抄现成的、拆好样本、事前自己写。在 AI 时代，给 AI 的任务说明就是评分标准，评价能力前置成了出题能力。尺子永远不完美，但不完美的尺子可以迭代，没有尺子无从谈起。下一章攻第二个部件：证据——以及为什么在 AI 面前，核查证据从好习惯变成了保命技能。

第五章 · 证据这关：AI 最会一本正经地错

评价的第二个部件是证据。第三章说过，「检查」这道工序——核实事实、抓自相矛盾——在面对 AI 输出时从好习惯变成了保命技能。这一章讲为什么，以及具体怎么查。

AI 的错误长什么样

先建立一个关键认知：**AI 的错误和人类作者的错误，长相完全不同。**

人类写错东西，通常有迹可循：他不熟的地方写得含糊，他编造的地方笔调发飘。一个有经验的读者能从字里行间嗅出「这段作者心里没底」。文字的可信度和作者的把握程度大致是挂钩的——这是一个进化了很久的社交信号系统。

AI 没有这个系统。它生成一段文字时，「确定的事实」和「合理的猜测」用的是同一种笔调、同一种自信。它能用引用学术论文的权威口吻，报出一本不存在的书、一个不存在的页码、一个差两倍的数字。这在业内叫幻觉（hallucination）：模型生成的听起来可信但与事实不符的内容。重点不是「它会错」——人也会错——重点是**它错的时候和它对的时候一模一样地自信**。你赖以以为生的那个「这段写得这么顺，应该没问题」的直觉，在 AI 面前是个坏仪表。

大白话

大白话：人类撒谎会脸红，AI 不会。所以你那套「看脸色识谎」的功夫在 AI 身上全部作废，只能用笨办法：一条条查。

检查的三种靶子

面对一份 AI 输出（或任何人的产出），检查工序瞄准三类靶子，难度递增：

靶子一：事实性硬伤

数字、日期、人名、引文、法条、API 参数——一切「有唯一正确答案」的东西。这是最容易查也最容易被 AI 搞砸的一类。查法没有捷径：**凡是可以核查的，一律核查**。数字回到原

始来源，引文去搜原文，参数去翻官方文档。一个实用的心法是「风险加权」：这个数字错了代价多大？报告里的市场规模错一位数可能没人发现，合同里的违约金错一位数会出大事。检查力度跟着代价走，不跟着「看起来可疑程度」走——因为 AI 的可疑程度不报错。

靶子二：内部一致性

东西自己跟自己打架：第三页说「用户增长强劲」，第七页的数据显示月活在下滑；开头假设「预算有限」，结尾推荐最烧钱的方案。这类错误 AI 常犯，因为它是逐段生成的，写第七页时并不「记得」第三页承诺过什么。查法是把长文档里的**所有断言和数字列成一张清单，两两对照**。这很枯燥，但它是抓这类错误唯一可靠的姿势——靠通读的「整体感觉」几乎必漏，因为 AI 每一段单独读起来都很顺。

靶子三：前提与语境

最难的一类：每句话单看都对，合在一起对不上你所在的真实处境。AI 给的方案默认了「团队有十个人」，你们只有三个；它引用的最佳实践来自一个烧得起钱的大厂；它说「一般建议这样」，而你面对恰恰是不一般的那个 case。这类错误没法用搜索引擎核查，只能靠你自己对处境的掌握——这恰恰是 AI 给不了、只能靠人的部分。查法是每读一条建议就问一句：**这条建议背后的默认前提是什么？这个前提在我这里成立吗？**

检查的流程化：把「仔细点」变成一张单子

「你要仔细一点」是最没用的建议——仔细不是态度问题，是流程问题。航空业早就想明白了：飞行员技术再高，起飞前也照清单逐项打勾，不靠「我今天状态好」。这个思路落到评价上，就是给你常验收的东西各做一张**检查清单**（checklist）。

举几个例子：

- **验收一篇 AI 写的文章**：每个数字有来源吗？每个引文搜得到吗？有没有「正确的废话」段落（删掉它全文意思不变）？结论和论据之间有没有跳步？
- **验收一段 AI 写的代码**：边界条件处理了吗？它调用的函数/库真实存在吗、参数对得上吗？错误路径会怎么死？
- **验收一份 AI 做的方案**：它默认的前提列出来了吗？最贵的那个假设是什么？如果数据是编的，结论还成立吗？

清单的价值在于它把检查从「依赖当天的专注度」变成「依赖一个结构」。注意力会波动，清单不会。而且清单是积累的：每被 AI 坑一次，就把那个坑变成清单上的一项。被坑三次之后，你的清单就是你个人的「AI 错误图谱」——这是花钱都买不来的资产。

一个心态转变：从「读者」变成「质检员」

检查功夫真正难的不是技术，是姿势。我们读东西的天然姿势是「读者」：顺着作者的思路走，被说服，点头。而检查要求你切换到「质检员」：不顺着走，处处反问「凭什么是这样」。这个切换很耗能，所以人会偷懒——尤其是当文本读起来舒服的时候，而 AI 恰恰最擅长生产读起来舒服的文本。

有一个简单的触发动作能帮你切换：**先找一处错，再开始读**。心理学上有个现象，一旦你在一份文档里抓到过一处错误，你对整份文档的信任校准会立刻重置——你会从「默认它对」切到「默认它可能错」。所以拿到 AI 输出的第一件事，不是通读，是抽查：随机挑两个事实去核，核完再决定这份东西值得你投入多少信任。

本章小结

AI 的错误特征是「错得自信」，人类的直觉仪表在它面前失灵，检查必须升级为独立工序。检查瞄准三类靶子：事实硬伤（逐条溯源）、内部一致性（断言清单两两对照）、前提语境（问「这条建议默认了什么」）。把检查流程化成清单，每被坑一次就加一项。姿势上，先抽查找一处错再通读，强制自己从读者切到质检员。下一章换个方向：证据查完了，东西「对」了——但「对」不等于「好」。判断「好不好」需要另一种资本：见识。

第六章 · 识货是喂出来的

上一章结尾埋了个问题：东西查完了，事实都对——但「对」不等于「好」。两份报告都没有硬伤，为什么行家扫一眼就说「这份是高手写的，那份是新手攒的」？这一章讲评价能力里最像「玄学」、其实也最有规律的部分：见识，或者说，品味。

为什么新手评不了，行家一眼看穿

先看一个常见现象。一个刚入行的人看两份设计稿，觉得「都挺好的」；总监扫一眼说「这份不行」。新手的感受是总监在装——明明差不多。但三个月后，这个新手再回头看当初觉得「挺好」的那份，自己也看不下去了。

这中间发生了什么？不是他学会了什么口诀，而是**他在这三个月里看了几百份设计稿，脑子里攒出了一个样本库**。所谓「眼光」，本质是一个参照系：你脑子里存着几百个例子，新的东西一进来，自动跟它们比对——这个比平均水平好，那个比顶尖水准差一截。没有样本库的人，眼前这份稿子就是他世界的全部，他没有坐标系可以安放它，只能说「挺好」。

大白话

大白话：识货不是一门知识，是一个数据库。数据库是拿眼睛一条条喂出来的，没有捷径，但有喂法。

品味的科学名称：内化的分布

用更精确的话说：行家对一个领域「好」的判断，其实是他脑子里内化了这个领域作品的质量分布——绝大多数作品长什么样（平均水准），顶尖的长什么样（天花板），糟糕的长什么样（地板）。新手没有见过分布，只能看到手里的一个点；一个点没有好坏可言，好坏是分布里的位置。

这解释了很多事：

- **为什么「我觉得挺好」常常是错的**：你的样本太少，随便一个及格线上的东西都会超出你的期望。

- **为什么行家会「苛刻」**：他不是苛刻，是他的分布里见过真正好的，天花板抬高了他的及格线。
- **为什么品味可以迁移有限**：美食家评得了餐厅评不了代码，因为他的分布是喂在菜上的。品味是领域特异的——这是第八章要展开的重要警告。

怎么喂：样本库的三条投喂纪律

既然眼光是数据库，问题就变成工程问题：怎么高效地建这个库。三条纪律：

纪律一：要看就看好的，别在平庸里泡着。这是反直觉的——很多人觉得「我水平低，先看点入门的」。错了。样本库的价值锚点在天花板：你得先知道顶尖的东西长什么样，才知道「好」的方向在哪。学写作就去读公认最好的那些作者，学设计就去看得奖作品，学代码就去读优秀的开源项目。**你的品味上限，取决于你接触过的最好的东西。**天天刷平庸内容的人，分布被喂歪了——他会把平庸当成平均水准。

纪律二：看好的也要看烂的，而且要知道烂的烂在哪。只看好样本有个副作用：你会觉得好是「理所当然的」，说不出它做对了什么。对照样本才显影。找一份公认糟糕的东西和一份公认优秀的，逐段对比：同样的信息，高手是怎么处理的？这一步是把模糊的「好」逼成具体的差异——它同时为第四章的「拆样本造尺子」供弹药。

纪律三：带着问题看，别裸看。「多看好东西」是废话，关键是怎么看。裸看是消费，带着问题看才是采集。每看一个好样本，问三个问题：它哪里比同类强？这个强法我能拆出来吗（是什么操作造成的）？这个操作能搬到我的场景吗？三个问题问完，这个样本才真正入库——从「看过」变成「存了结构」。

AI 时代的一个新风险：样本库被污染

这里有一个这个时代特有的陷阱。过去，你能接触到的「样本」大多经过人工筛选——出版有编辑，上架有采购，传播有口碑。你看到的东西整体质量高于平均。而现在，AI 生成的内容正在灌满信息环境，它们的水准集中在「像模像样的平庸」。如果你的样本库主要被这类内容喂养，你的分布会整体下移：**你会慢慢把「AI 水准」当成「好」的标准。**

这是评价能力在 AI 时代的第一场保卫战：守住你的输入质量。具体做法不复杂——刻意保留高质量的人类作品来源（经典的书、顶尖的专栏、获奖的作品），定期回到这些锚点重新

校准。品味像视力，长期在昏暗的光线下看东西，会忘了明亮是什么样。

从「看出来」到「说出来」

样本库建起来之后，你会拥有一种新的体验：扫一眼就知道「这个不对」。但注意，这还只是评价的前半程——感觉层面。行家和新手的最后差距不在「看出来」，在「说出来」：把那个不对劲的感觉，翻译成可辩护的语言（「这段的论证在第二步跳了一步」「这个配色的明度对比不够，重点没出来」）。感觉是私人的，说出来才能被检验、被讨论、被用于决策。这个「说出来」的功夫——把判断写成理由——就是下一章的主题。

本章小结

品味不是玄学，是内化的质量分布；新手评不了是因为样本库太小，好坏是分布里的位置。投喂三条纪律：看好的（定锚点）、看好烂对照（显影差异）、带着问题看（存结构）。AI 时代要警惕样本库被「像模像样的平庸」污染，守住高质量输入。看出来只是前半程，下一章攻后半程：把判断写成站得住的理由。

第七章·把「好」说成理由

第六章结尾说：行家和新手的最后差距，不在「看出来」，在「说出来」。这一章讲怎么把判断写成理由——这是评价能力从私人感觉变成公共资产的成人礼。

一个测试：你的判断经得起一句「凭什么」吗

下次你说出「这份不行」的时候，想象对方平静地问一句：「凭什么？」

如果你卡住了，或者只能重复「反正就是不行」，说明这个判断还住在你身体的直觉层，没有进到语言层。直觉层判断有三个致命缺陷：**它没法被检验**（别人不知道你错了没有，包括你自己）；**它没法被复用**（下次遇到同类东西，你得重新凭感觉）；**它没法说服人**（而评价的最终用途，几乎总是推动一个决定）。

把判断说出来，不是表达能力的问题，是判断本身的最后一道工序。**一个说不出的判断，是一个没做完的判断。**

理由的标准结构：三段链条

一个站得住的评价，写出来的样子永远是三段：

结论：这东西行不行、值多少分。

标准：我用的尺子是什么。

证据：拿尺子量出来的结果是什么，具体在哪一处。

对比一下两种说法。第一种：「这篇文案写得不好。」这是结论裸奔。第二种：「这篇文案不合格（结论）。按我们定的标准，文案要让读者三十秒内知道这东西解决他的什么问题（标准）。但前三段都在讲我们公司多厉害，读者的问题到第四段才出现（证据：具体位置）。」

第二种说法立刻获得了三个新属性：对方可以反驳你的标准（「我们的读者其实关心公司背景」），可以核查你的证据（「不对，第二段就提了问题」），可以挑战你的结论——**而这正是**

它值钱的地方。一个可以被反驳的判断，才有机会被修正、被信任、被执行。不能被反驳的判断只是独断。

大白话

大白话：好的评价像数学题，要写解题过程。只写答案的，别人没法给你批改，你自己也不知道哪步会错。

为什么「写理由」会反过来提升判断力

这里有一个不太显眼但极其重要的机制：**写作理由的过程，是评价能力的训练器本身。**

当你逼自己把「感觉不对」写成「哪里不对、按什么标准不对」时，会发生三件事：

第一，假判断会现形。很多「感觉不对」写着写着就发现没东西可写——那往往说明这个感觉来自偏见、情绪或立场，而不是来自对象本身。写作是过滤器，把真判断和假判断分开。

第二，模糊的标准被逼成清晰的。你写「这篇逻辑混乱」，就必须回答「逻辑混乱的具体表现是什么」——写着写着，你脑子里的尺子就被打磨出了刻度。这就是为什么第四章说标准和品味互为表里：每次写理由，都是在给自己的 rubric 升级版本。

第三，你开始积累自己的判例库。写下来的判断可以被回头看：三个月后复盘，当初那个判断对了吗？错在哪个环节——标准错了、证据看漏了、还是权衡错了？没有文字记录的判断像泼出去的水，留不下任何教训。

「权衡」：理由里最见功力的一段

真实的评价很少是「全好」或「全坏」，更多是「这方面强、那方面弱，综合看……」。综合这一步就是权衡：当多个维度的评价互相冲突时，说明哪个优先、为什么。

权衡写不清楚，是多数评价意见被驳回的真正原因。「这个方案创意好但成本高，所以不行」——为什么成本高就否决创意好？在什么条件下这个优先级会反过来？把权衡写出来：「在我们当前预算冻结的前提下，成本维度一票否决；如果下季度预算恢复，这个方案值得重启。」这样一来，你的评价不仅给了结论，还给了结论的**生效条件**。这是高手评价的 signature：判断附带它自己的适用边界。

一个可执行的动作：评价先写后说

给一个马上能用的训练法：接下来一个月，凡是需要你出判断的场合，**先用三分钟把判断写成三行字再开口**——一行结论，一行标准，一行证据（带具体位置）。写完再决定怎么说。

开始会很别扭，会暴露很多「原来我这个判断没东西支撑」的时刻。这些时刻正是训练生效的时刻。一个月后回看，你会发现两件事：你的判断质量本身上升了（因为假判断被过滤了），而且你说服人的能力上升了——因为别人终于能看清你判断的结构。前者是里子，后者是面子，评价能力两个都要。

本章小结

说不出的判断是没做完的判断；可辩护的评价永远是「结论—标准—证据」三段，且正因为它被反驳才值钱。写理由反过来训练判断力：过滤假判断、打磨标准、积累可复盘的判例。多维度冲突时要写出权衡和结论的生效条件。训练法：一个月里判断先写三行再开口。下一章处理评价者自身的问题——人这台判断仪器，本身是有系统误差的。

第八章 · 知道自己什么时候在瞎判断

前面几章都在磨「判断」这台仪器的零件：标准、证据、见识、表达。这一章处理一个更扎心的问题：**仪器本身有系统误差**。人不是一台中立的判断机器——我们会在某些领域系统性地高估自己，在某些时刻系统性地跑偏。知道仪器什么时候不准，和仪器本身好不好，是两种独立的能力。前者叫元认知，它是评价能力的最后一道保险。

最危险的状态：不知道自己不知道

心理学里有个著名现象叫达克效应（Dunning-Kruger effect）：能力越不足的人，越容易高估自己的能力——因为「判断自己水平」所需要的能力，恰恰就是「做这件事」所需要的能力。一个不懂品酒的人，连「自己不懂品酒」都判断不出来。

放回本书的框架里，达克效应有一个特别锋利的版本：**评价能力是领域特异的，但自信不是**。你在代码领域的判断练得很准，这份「我很会判断」的自信会不知不觉漫到产品、设计、投资、育儿——而你的样本库、你的尺子、你的核查清单，全都只长在代码那一块地上。第六章说过，品味是喂出来的，喂在 A 领域的饲料长不出 B 领域的肌肉。但自信是流体，它会溢出容器。

这就是为什么很多聪明人会在陌生领域做出惊人的烂判断，而且做得理直气壮。他们的问题不是判断力的引擎坏了，是**仪表坏了——显示「油量充足」的那个表，不看你在哪个领域**。

大白话

大白话：判断力是分科目的，自信不分。你在数学考场上练出的底气，会跟着你走进语文考场——尽管它在那儿一文不值。

校准：给判断装一个误差条

解决之道不是「谦虚一点」这种态度说教，而是一个技术动作：校准（calibration）——让你「觉得自己有多确定」和「你实际上有多准」对上号。一个校准良好的人，说「我九成把握」的事，长期来看真的九成是对的；说「五五分」的事，真的对错各半。校准差的人，说

「九成把握」的事，实际可能只对六成——多出来的三成自信，就是他做决策时看不见的坏账。

校准是可以练的，而且练法意外地具体：

练法一：给判断标数字，然后等开奖。凡是你做的可验证的判断（这个项目会延期吗？这篇稿子会爆吗？），在判断时写下你的置信度：「我认为 70% 会延期」。攒上几十条，等结果出来后统计：我说 70% 的事，实际发生了多少？多数人第一次做这个统计都会发现自己的自信系统性偏高——这个发现本身值回票价。

练法二：区分「我的领域」和「我旁观的领域」。拿一张纸画两个圈，一个圈里是你有真实样本库、真实反馈记录的领域，另一个是「我看得多但从没负过责」的领域。在第二个圈里做判断时，强制自己把置信度下调一档，并把结论写成假设而不是断言。这个圈画得诚不诚实，决定了你犯大错的概率。

练法三：主动寻找「我居然错了」的时刻。校准差的人不是没有错过，是错完就忘——大脑天然倾向于记住自己判断对了的时刻。对抗它只能靠记录：判断写下来（第七章的三行法正好派上用场），定期复盘。复盘时最值得停留的不是「我对了哪些」，而是「我错了的那些，错在哪个部件上」——标准拿错了？证据看漏了？还是领域根本不在我的圈里？

知道何时该说「我评不了」

元认知练到最后，最高级的产出不是更准的判断，而是一句以前说不出口的话：**「这个我评不了。」**

在评价文化里，说「我评不了」被当成露怯。但拆开看，它是评价能力成熟的三重标志：它说明你知道自己的分布在哪个领域才有效（样本库自觉）；它说明你知道一个判断需要哪些前提（结构自觉）；它说明你把「判断的质量」看得比「我的面子」重（责任自觉）。一个肯说「这部分超出我的判断范围，我们找个真懂的人」的评审者，比一个什么都敢评的评审者值钱得多——因为前者的「这个行」才是真金白银。

AI 时代这给了一个特别的对照。AI 从不说「我评不了」——它对任何问题都给出同等自信的回答（第五章讲过，它的自信不携带信息）。这恰好反衬出人的元认知是多么稀缺的互补：**机器负责永远有答案，人负责知道答案什么时候不算数。**

本章小结

判断力领域特异，自信却会溢出——达克效应让仪器在最陌生的领域显示满格。对策是校准：给判断标数字等开奖、画清「我的领域」的圈、主动复盘错判。最高级的元认知产出是「我评不了」——它恰恰是一个评审者可信度的来源。仪器磨好、误差条装好，下一章给一套把它们全部串起来的日常训练方案。

第九章 · 日常训练场

前八章把评价能力拆完了：标准、证据、见识、表达、校准。零件齐了，这一章回答最后一个问题：**这些功夫怎么长在身上？**答案和所有技能一样——练。但评价能力有个特殊困难：它不像弹琴有琴键可摸，它的练习机会散落在日常里，不刻意设计就不会发生。这一章给四个具体的训练装置。

装置一：评价日志——把判断变成可复盘的资产

所有训练装置的地基是同一个：**把判断写下来**。前面几章的理由已经很充分（第七章的三行法、第八章的校准统计），这里把它固化成一个动作：建一个「评价日志」，每次你正式评价了什么——一份方案、一个候选人、一篇文章、一个 AI 输出——花三分钟记五行字：

- 评价对象是什么
- 我的结论（含置信度，比如「70% 这版能过」）
- 我用的标准
- 关键证据（具体位置）
- 当时的背景与目的

日志的价值不在记，在**三个月后翻回来对答案**。判断这件事最大的训练障碍是反馈慢且模糊——你评了一个方案，半年后才知道它落地效果如何，到那时你早忘了当初怎么想的。日志把「模糊的事后感觉」变成「可逐条核对的事前记录」。没有它，你的判断力每年都在凭感觉增长；有了它，每一次判断都在给校准数据库添一行数据。

装置二：对照练习——借别人的眼睛校准自己的

这是成本最低、见效最快的装置：**先独立评价，再看行家怎么评，比对差异**。

操作很具体。看一部电影、读一篇文章、用一个新产品之后，先花五分钟写下你自己的评价（三行法）——然后才去看专业影评、书评、测评。比对你和他们的差异：他们注意到的哪些点你完全没注意到？你们注意到同一个点，为什么结论相反？

这个装置厉害在它同时训练两样东西：样本库（行家指出的维度扩充你的坐标系）和标准（你会发现自己尺子上的漏洞——比如你从没把「节奏」当成一个维度，而所有专业影评都在谈它）。注意顺序必须是先写再看：先看了别人的，你的「独立评价」就被污染了，练的就不是判断力而是复述能力。

AI 时代这个装置有了强化版：同一件事，先自己评，再让 AI 评，再看行家评。三方对照。你会发现 AI 的评价往往面面俱到、四平八稳（它的分布是「平均水准」），行家的评价往往只打七寸——这个对比本身就是一堂关于「什么是真判断」的课。

装置三：事前验尸——给未来的失败做评价

这是决策圈的经典工具事前验尸（premortem）：在一个计划启动之前，假设「一年后它失败了」，让每个人写下「它是因为什么死的」。它本质是评价能力的倒挂使用：评价一个还没发生的结果。

事前验尸为什么能练评价？因为它强制你调用标准里最容易被忽略的那类——**风险维度**。平时评价一个东西，我们习惯量「它好不好」；事前验尸量的是「它会在哪死」。后者需要的标准完全不同：不是「优点清单」而是「失败模式清单」。而失败模式清单，恰恰是被坑出来的——所以每次做完事前验尸、等结果揭晓后，把「真正杀死它的原因」和「当初列的死因」对照，你的风险维度尺子就在一轮轮升级。

它还有一个副产品：一个团队做事前验尸时，你会看到每个人默认的评价标准完全不同——有人只看市场，有人只看技术，有人只看人。这是观察「别人尺子」最便宜的现场。

装置四：主动找反例——给自己的判断做压力测试

人有一种顽固的倾向：形成一个判断之后，不自觉地只找支持它的证据（这叫确认偏误）。评价能力的反面教材几乎都是这么写成的：先有了结论，再收集论据，最后把「我觉得」包装成「事实证明」。

对抗它的装置只有一个动作：**每下一个稍微重要的判断，强制自己找一个最强反例**。你说「这个候选人不行」，就必须认真回答「支持他行的最强证据是什么」；你说「AI 写的这版能用」，就必须回答「哪一处最可能埋着雷」。不是走过场地想，是真的去找——找不到反例的判断，要么说明你对，要么说明你看得不够，而你要分清是哪种。

这个装置还有一层用处：它把第七章的「可辩护」练到实处。一个认真找过反例的判断，天然带着对自身边界的认识——「我认为它能用，风险集中在数据真实性，我已核查过其中三个」。这种判断别人拿去就能用，因为它连自己的误差条都附上了。

把四个装置串成日常

四个装置合起来是一条回路：**日志负责记录，对照负责校准，验尸负责风险维度，反例负责压力测试**。它们共同的底层逻辑是：评价能力不是「想清楚」练出来的，是「判断—记录—开奖—复盘」这个循环一圈圈转出来的。循环转得越快、记录越诚实，长进越快。

起步时别贪。最小可行版本是：从今天起，评价日志 + 每周一次对照练习。跑一个月，你会第一次看见自己判断力的形状——它擅长什么、系统性漏什么。那个「看见」本身，就是前面八章所有概念落地的时刻。

本章小结

评价靠「判断—记录—开奖—复盘」的循环生长。四个装置：评价日志（五行记录，事后对答案）、对照练习（先自己评再看行家评，顺序不能反）、事前验尸（练风险维度）、主动找反例（对抗确认偏误）。从最小版本起步。最后一章回到本书的起点：这套功夫练成之后，人和 AI 怎么分工？

第十章 · 与 AI 搭档的新姿势

最后一章，回到全书开头那个问题：当 AI 什么都能生成，人凭什么说它好不好？九章走下来，答案已经不再是「因为人有眼光」这种空话了——评价是一架有结构的梯子，是一门有零件的手艺，是一台需要校准的仪器。收尾这一章，把这套功夫放回它真正的战场：人和 AI 的分工界面。

新的分工：塔基外包，塔尖自留

第二章说过，布鲁姆的梯子正在被 AI 从塔基往上承包。现在可以把分工说得再精确一点：

- **记忆、理解**：基本全外包。查事实、做摘要、讲概念，AI 又快又便宜，人守着这两层没有意义。
- **应用、分析**：人机协作。AI 出草稿、做初拆，人负责纠偏和兜底——因为这两层的输出里会埋幻觉（第五章），而纠偏用的正是评价功夫。
- **评价**：人自留。不是 AI 完全不能评——它能按你给的标准打个分——而是评价的核心部件它缺位：它没有你的处境（前提核查做不了），它没有你的样本库（品味喂的是全网平均值），它不会为结论负责（责任主体不在它）。
- **创造**：方向由人定，执行可外包。而「定方向」这个动作本身——决定什么值得做——就是一个评价动作：对无数可能的方向做出「这个值得」的判断。

看出来了吗？**评价不是分工表上的一个格子，它是整个人机协作的操作系统。**你给 AI 下任务时是出题（第四章：任务说明即评分标准），它出结果时你是质检（第五章：检查清单），十个方案选一个时你是评委（第三、七章），决定下一步做什么时你是裁判（创造的方向选择）。每一环都在调用同一套功夫。

大白话

大白话：AI 时代的人不是「某个岗位的从业者」，而是「整条流水线的验收员 + 出题人」。验收和出题，是评价能力的正反手。

一个要警惕的姿势：自动化顺从

分工讲完了，还得讲分工最大的敌人。它有个名字，叫自动化顺从（automation bias）：人倾向于过度信任自动化系统的输出，放弃自己的判断。航空史上有机组明明看到仪表矛盾仍按自动驾驶指令操作的案例；医疗上有医生采纳明显有误的电脑建议的记录。AI 越强、越顺滑，这个倾向越危险——因为「它一般都是对的」会慢慢磨成「它总是对的」，再磨成「我干嘛还看」。这本书写的就是一整套「我干嘛还看」的理由和方法。

评价能力的复利：它让你更会「要」

收一个前面埋的伏笔。第四章说评价能力前置为出题能力，这里把这个闭环画完。

练评价的人会发现一个意外收获：**你对「好」的理解越深，你能向 AI 要的东西就越具体**。样本库大的人，能说「我要的是那种克制的、留白的写法」，而不是「写好一点」；尺子细的人，能在任务里写清「必须回答三个顾虑、不准用感叹号」；校准好的人，知道哪类任务 AI 容易翻车、要在指令里预先排雷。而「要得更具体」的结果，是 AI 的输出质量跟着水涨船高——你拿到的初稿就更接近好，你的验收成本就下降，你就有更多精力投在更高层的判断上。

这是一个正反馈循环：**评价能力 → 出题质量 → 产出质量 → 验收省心 → 更多精力练评价**。AI 时代人与人之间的产出差距，很大程度上不是「谁用的 AI 强」，而是谁站在这个循环里、谁站在循环外。

最后：评价能力是一种责任感

这本书从头到尾把评价当一门手艺讲：标准、证据、样本库、校准、训练装置。但收尾时要说一句手艺之外的话。

评价能力的终极形态，是一种「我来负责」的姿势。AI 生成的东西没有作者——它没有人格、没有后果、没有「我错了」。每一份经你手的 AI 产出，签的都是你的名字；每一个「就它了」的决定，账都记在你头上。这听着像负担，但它恰恰是人在这个分工里的位置之所以值钱的全部原因：**责任是判断的价格，而愿意承担责任的能力，就是评价能力**。

所以回到开头的实验。现在再让 AI 写三段话，问你哪段好——你应该已经知道怎么答了：先问这段话是干嘛的（目的），再掏出你的尺子（标准），逐段对照（证据），写下三段式理

由（表达），掂量一下自己在这个领域的把握（校准）。如果答得还有点磕绊——没关系，梯子在那儿，一层一层爬就是。

判断的能力，长在判断的动作里。去判断吧。

判断的品味 · 竹小竹8167