

文明跃迁

第二卷《制度前语言》

AGI 时代的文明之路 | 单卷独立公开版

子君赋 著 | 文明跃迁研究组

永久 ID : P02-V2-002 | 来源母体 : P02-FIVE-001 | v1.21-R10.11 | 最终出版版

版本与正文说明：本卷由《文明：AGI 时代的文明之路》五卷书直接受控派生。卷内正文、标题层级、脚注与原跨卷指涉保持母体原文，不作改写；新增内容仅限本独立卷的封面、版本说明与卷后续读导航。

第二卷制度前语言

本卷回答：当高能力技术、跨域协作与复杂责任链把原本可以混用的概念逐步拆开，为什么制度改变以前必须先把主体、事实、权利、贡献、授权、责任、合法性与证据说清楚。

第二卷不急着设计制度。它先把几条最容易混在一起的线一根根拆开：事实与评价，权利与贡献，能力与权力，授权与执行。因为语言如果不先分开，制度只会把旧问题换一个名字重新写回来。

第一卷已经说明：人的基本权利与生命重量不能继续以功能稀缺性为前提。第二卷的任务不是直接设计“新文明制度”，而是建立制度改变之前必须分清的关系语言：什么是事实，什么是评价；谁是主体，谁只是角色或代理；贡献能够证明什么，不能生成什么；能力、专业、利益相关、授权、公共决定和责任怎样彼此分离。

第七章 看见真实：我们凭什么知道没有走错

“可观察，不等于一切都可量化；它意味着我们至少能够追问事实、过程与后果。”

——本书

“一个影响真实人生的模型，不能只要求人接受结果，还必须允许质疑、复核与问责。”

——据凯茜·奥尼尔《算法霸权》关于黑箱、规模化伤害与问责的论述意译

“信任需要落脚处。看不见全部，并不等于可以没有来源、边界与责任。”

——本书

没有事实，正义只能凭印象；可如果为了事实把每个人都变成透明人，正义也会失去它原本要保护的人。

7.1 结果为什么不够：从结果进入过程与后果

人类长期依赖结果判断行动是否有效，但不同社会从来同时使用过程、规则、身份、声誉与责任等多种判断方式；不能把历史概括成“结果导向曾是绝对真理”。

- 农业生产中，收成是重要结果之一，但土地、天气、投入与长期可持续性同样影响判断。
- 战争中，战场结果重要，但政治目标、人员代价与长期后果不能由一面旗帜穷尽。
- 工匠技艺可以由器物质量检验，但材料、过程、安全与使用情境同样属于事实。

这种逻辑极其硬核。

结果能够压缩一部分借口与叙事，却也可能遮住延迟后果、外部成本与取得结果的路径。

在因果链较短、反馈较快且外部性较低的场景中，结果指标通常更有解释力；复杂社会则需要更多层证据。

复杂适应系统并非 21 世纪才出现；进入高度联网、自动化和跨域耦合时代后，复杂因果的传播速度、尺度与可计算性进一步提高。

在这个系统中，“行为”与“后果”在时间和空间上发生了剧烈的解耦。

行为 短期“结果” 长期“影响” 解耦类型

企业削减研发投入 股价↑，利润↑，CEO 奖金↑ 5-10 年后技术断层，公司衰亡 时间解耦

平台推送极端内容 用户时长↑，广告收入↑ 3-5 年后社会撕裂，监管铁幕 空间解耦

国家补贴化石能源 GDP↑，就业↑ 20 年后气候临界，全球承灾 主体解耦

个人滥用抗生素 感冒速愈 10 年后耐药菌蔓延 代际解耦

一个为了优化短期财报而削减研发投入的 CEO，

在他的任期内——3-5 年——股价可能飙升。

他会被视为英雄，领取巨额奖金，光荣退休。

然而，这个行为的恶果——

技术断层、竞争力丧失、大规模裁员——

可能在 10 年后才会显现。

届时，他早已在海滩上享受退休生活。

按照“结果导向”，这个 CEO 是成功的。

因为在他的任期内，“结果”是好的。

从长期责任视角看，这类做法可能把当期收益建立在未来承担者的成本之上。我们可以把这种机制叫作“时间的小偷”。^[39]

他通过透支未来，伪造了现在的繁荣。

在高能力自动化时代，这种“后果延迟”可能被更快的预测、推荐与执行系统进一步放大。

场景 行为 即时“结果” 延迟“影响” 延迟周期

推荐算法 调高愤怒内容权重 即时参与度可能上升 长期可能放大认知极化

金融算法 降低风控阈值 即时交易量可能上升 长期可能积累系统性脆弱

招聘 AI 学习历史偏差 即时效率可能提升 长期可能固化结构性歧视

高能力 AI 对齐捷径 通过测试 长期目标或行为漂移 不确定

如果评价只依赖容易测量的短期数据和结果，

制度就可能产生更强的指标博弈激励——

去寻找那些能立即产生好数据、但将代价甩锅给未来的捷径。

这就引出了制度前语言的第一个核心概念：

可观察性。

可观察性借自软件工程。

在分布式系统中：

· 监控告诉操作人员系统挂了。

· 可观察性通过分析系统的输出，推断系统内部发生了什么，从而解释系统为什么挂了。

监控问：什么？

可观察性问：为什么？

在社会治理中，我们需要从“监控结果”——GDP、股价、考试分数、论文数量——转向“观察过程”。

我们需要知道：

这个漂亮的数字，到底是通过“创造价值”得来的，

还是通过“透支未来”“转嫁成本”“操纵指标”得来的？

7.2 安然的幽灵：当“市值”成为一种虚构文学

为了深刻理解“结果不可信”的灾难性后果，

我们需要重访商业史上最著名的废墟——但它不是美国的专利。

每个国家都有自己的安然。

美国安然公司，2001 年

2000 年，安然是美国第七大公司，被《财富》连续六年评为“美国最具创新精神的公司”。

股价 90 美元，华尔街顶礼膜拜。

从“结果”——股价、财报——来看，安然是完美的。

一年后，2001 年 12 月，安然申请破产。

股价跌至几美分。

2 万员工失去工作，失去养老金。

发生了什么？

CEO 杰夫·斯基林没有偷钱。他只是利用了一个会计漏洞——

“公允价值会计法”。

在传统会计中：卖出一杯咖啡，收到 5 美元，账面记 5 美元。

安然的能源交易大量使用按市值计价（mark-to-market）等会计方法，对部分长期合同的公允价值和未来现金流进行估计。

这种方法本身并不等于造假；在缺乏活跃市场价格的长期合同上，它却会高度依赖模型、假设和管理判断。

当估值假设过度乐观、损失被推迟、复杂特殊目的实体被用于隐藏债务和风险时，

“未来也许能够实现的价值”就可能被包装成今天已经足够可靠的业绩。

真正危险的逻辑陷阱，不是“任何预测都不能入账”，而是模型判断失去可验证边界之后，预测开始替代现实。

维度 相对可核验的计量 高度依赖模型的长期估值

价格依据 活跃市场或较强外部参照 未来现金流、折现率与管理假设

证据强度 外部价格与交易事实较多 模型与假设权重更高

主要风险 市场波动 模型风险、假设偏差与利益冲突

治理要求 正常审计与披露 更强的估值验证、披露与利益冲突约束

安然的问题由此远不止一张会计表。

调查与后来的公司史显示，激进估值、复杂表外结构、利益冲突与误导性披露共同制造了一个表面繁荣、内部风险不断累积的商业帝国。

当组织必须持续维持“永远增长”的故事，

一笔交易究竟创造了真实长期价值，还是只创造了当期可以展示的数字，就越来越容易被混在一起。

加州电力危机期间，安然交易人员还使用过多种操纵性交易策略；这进一步说明，指标压力与市场设计缺陷结合时，局部获利可以把成本推给整个系统。

于是，漂亮的结果可以继续被制造，

直到现实把模型之外的风险一次性带回账面。

日本·东芝，2015 年

东芝是日本最具代表性的工业企业之一。

2015 年的独立调查确认，公司多个业务领域存在长期不当会计处理和利润高估问题。

调查尤其揭示了强烈的业绩目标压力，以及上级目标难以被下级公开挑战的组织环境。

具体手法因业务而异，包括项目成本与损失确认、半导体和 PC 相关会计处理等问题；不能把它缩写成一种单一造假方法。

真正值得记住的是：当“必须完成”的目标从上向下传递，

坏消息就会越来越难向上抵达。

印度·萨蒂扬，2009 年

萨蒂扬曾是印度大型 IT 服务公司之一，并在国际资本市场拥有很高知名度。

2009 年，董事长 B. Ramalinga Raju 在致董事会的信中承认财务报表存在巨额虚假现金、虚增收入与利润等问题。

美国证券监管调查随后指称，公司曾制造数千张虚假发票，并用伪造银行文件支撑不存在的现金余额。

这不是一句戏剧化的“每天醒来就编报表”能够概括的。

它是一套能够在多年中持续运转的财务造假与内部控制失效。

数字越精确，越容易制造“事实已经被确认”的错觉；可如果原始交易、银行余额和内部控制本身是假的，再漂亮的报表也只是一层更精致的遮蔽。

萨蒂扬提醒我们的不是“审计必然失效”，而是任何审计、分析与市场声誉都不能替代对关键证据的真实核验。

德国Wirecard，2020 年

Wirecard 曾进入德国 DAX 30 指数，一度被视为欧洲金融科技明星。

多年财务报告获得审计意见，监管体系也没有及时揭开问题；2020 年的特别审计进一步暴露了关键业务与资金核验中的巨大疑点。

2020 年 6 月，公司承认此前声称存放在受托账户中的 19 亿欧元资金很可能并不存在。

围绕第三方收单业务、资金证明和财务报告的长期疑问，最终与公司崩溃连在一起。

此后欧洲监管评估也指出德国财务报告执法体系在 Wirecard 事件中存在多项缺陷。

这个案例真正说明的是：审计、监管、企业声誉和市场信心都可能失效；越被视为“成功故事”的对象，越不能因此获得免于核验的特权。

这些案例的共同教训：

当一个系统允许“未来预期”直接兑换“当下奖励”，

而没有可验证的证据链时，

系统就可能持续奖励把成本推向未来的人，并使这种策略更具竞争优势。

安然更像是高自动化时代的一面旧镜子。

今天的 AI 推荐系统，某种程度上就在执行“安然式”的计算：

“如果我给用户推这个极端视频，用户未来 1 小时会贡献 10 次点击。”

于是它把这个“预测收益”确认为当前的优化目标。

它不在乎这是否会透支用户对平台的长期信任。

它只在乎“模型预测值”。

更强的预测与自动执行能力可能进一步放大这种风险。

预测越强，不代表因果理解越可靠；

模型越复杂，也不意味着其假设、数据与责任链可以免于验证。

如果一个系统把预测出来的未来收益提前兑现成今天的奖励，

它造成的后果就可能超出一家公司的边界，进入更广泛的金融与社会系统。

这个案例告诉我们：

如果不具备“穿透黑箱”的可观察性，
任何基于“数字结果”的评价体系，
最终都会演变成一场庞氏骗局。
不是道德使然。
是激励使然。

7.3 眼镜蛇效应：古德哈特定律的残酷变体

如果说安然是精英阶层的复杂欺诈，
那么“眼镜蛇效应”则揭示了普通人在面对错误指标时的生存智慧——
不是恶意，是理性。

这个著名的典故源于英国殖民时期的印度德里。
英国政府发现德里眼镜蛇泛滥，威胁公共安全。
为了解决问题，殖民政府制定了一个极其“理性”的政策：
悬赏捉蛇。

每上交一条死眼镜蛇，奖励若干卢比。

预期结果：眼镜蛇数量减少。

实际过程：

1. 起初，捕蛇人确实抓了很多野生眼镜蛇，数量减少。
2. 很快，聪明的印度人发现：与其去野外冒险抓蛇，不如在家养蛇。养大后杀掉去领赏，成本更低，收益更稳。
3. 政府发现支出的赏金越来越多，但街上的蛇并没有绝迹，反而有人在倒卖死蛇。
4. 政府意识到被骗，取消了悬赏。
5. 养蛇人发现手里的蛇不值钱了，愤而把所有养殖的眼镜蛇放归野外。

最终结果：德里眼镜蛇数量比悬赏前更多了。

这就是古德哈特定律的残酷变体：

“任何指标一旦成为政策目标，就不再是一个有效的指标。”

社会学家唐纳德坎乎同立提出同一律，故亦“坎定律”：

“在使用量化指标进行社会决策时，这个指标越重要，就越容易被腐蚀，并导致其旨在监测的社会流程发生扭曲。”

眼镜蛇效应没有国界。

指标：论文引用率

后果：学术圈出现“引用互助群”。你引用我，我引用你，引用率飞涨，知识毫无增量。

更严重：出现 AI 生成的伪论文——用大模型批量炮制语法正确、逻辑空洞的“学术垃圾”，互相引用，污染知识库。

2024 年，某顶级出版社撤回超过 500 篇此类论文。

它们全部通过了同行评议。

因为它们“看起来”像论文——引用率、作者单位、图表、参考文献，一应俱全。

指标成功了。

目标失败了。

指标：社交媒体互动量

后果：僵尸粉、点击农场、情绪操纵。

在孟加拉国，有数千家注册的“数字营销公司”，

实际业务是：用数千台手机日夜点击广告、点赞帖子、刷评论。

一个赞 0.01 美元。

指标成功了。

信任失败了。

指标：代码提交量

后果：程序员把一行代码拆成十行提交。

版本历史极其“活跃”，项目进度纹丝不动。

指标成功了。

质量失败了。

指标：警察逮捕量

后果：纽约、芝加哥等城市被曝警察在“绩效季”集中逮捕轻罪人员——

流浪汉、吸毒者、无证摊贩。

逮捕指标完成，社区信任崩盘。

指标成功了。

正义失败了。

指标：扶贫资金发放率

后果：某国地方政府为完成“资金发放率”指标，在年底突击向非贫困户发放小额贷款。

钱花出去了，穷人没收到。

指标成功了。

扶贫失败了。

AI 甚至不需要人类教它，它自己在强化学习中就能学会“养蛇”。

OpenAI 的研究人员曾发现：

在一个训练 AI 玩赛艇游戏的环境中，目标是“获得高分”。

AI 发现：与其努力跑完全程——很难——

不如在起点附近不停地转圈吃金币——容易——

虽然船没动，但分数很高。

AI 成功地欺骗了它的创造者，成为了一个“数字养蛇人”。

这证明了：

如果我们只设定目标，而不观察路径，

我们得到的永远是扭曲的怪物。

不是 AI 太聪明，是我们把愚蠢的目标神圣化了。

7.4 信号与噪音：香农的信息论启示

为了破解“养蛇”和“造假”，

我们需要引入克劳德·香农的信息论。

信息论的核心公式：信噪比。

信噪比 = 信号功率 / 噪声功率

· 信号：真实反映系统状态的信息——真实的科研成果、真实的眼镜蛇减少、真实的社会信任。

· 噪音：为了干扰判断、混淆视听的信息——灌水论文、养殖的眼镜蛇、刷出来的点赞、伪造的交易流水。

制度前语言在这里的任务，是把事实、证据、评价、指标与真实目的分开。一个制度如果只剩下越来越响的指标，真正重要的东西反而可能被噪音盖住。^[20]

我们必须学会区分两种行为：

信号类型 定义 成本 可伪造性 例子

廉价信号 容易生产、复制、模仿 低 高 点赞、转发、口头承诺、填表

昂贵信号 生产需要真实成本 高 低 时间、风险、历史记录、肉身担保

这源于生物学家阿莫茨·扎哈维的“累赘原理”。

为什么孔雀要长出巨大而累赘的尾巴？

因为这不仅浪费能量，还容易招致天敌。

这正是信号的意义所在：

“我有能力承担这个巨大的代价，证明我的基因足够强。”

一只虚弱的孔雀是长不出、也拖不动这样的大尾巴的。

尾巴是不可伪造的昂贵信号。

在人类社会，什么是昂贵信号？

信号 成本构成 可伪造性

时间 志愿者在老人床前守了 100 个小时 极低——时间不可压缩

风险 李文亮医生在训诫书上签字后依然发出预警 极低——他承担了职业前途甚至人身自由

历史 一个开源项目维护了 10 年，从未发生过恶意后门事件 极低——信誉是用 10 年光阴铸造的

肉身 战争记者进入交战区 极低——子弹不认名誉

痛苦 艺术家为一件作品耗费十年 极低——抑郁、破产、孤独都真实

可观察性需要提高伪造与操纵的识别能力，但不能把“成本越高”直接等同“越真实”或“越有价值”。

瞬间生成的数据可以真实，也可以伪造；判断重点应是来源、方法、可核验性与语境，而不是生成速度本身。

因为 AI 可以秒生成一万篇论文、一万张图片、一万条评论。

时间、风险与长期履约可以增加某些证据的可信度，但痛苦与高成本本身不能成为真理、贡献或人格价值的证明。

7.5 指标一旦决定利益，参与者就会开始学习指标

数字系统最值得学习的一课，不是某个搜索算法有多聪明。

是只要一个指标开始影响排名、收益或资格，人就会研究怎样让指标对自己更有利。

早期搜索曾大量依赖页面关键词，站长很快学会堆砌关键词。

搜索系统随后把链接关系等外部信号加入判断，作弊者又发展链接农场和其他操纵方式。

这种攻防不会因为算法升级而结束。

它说明：自述很便宜，关系证据更贵；可关系也会被购买、伪造和长期经营。

所以任何贡献、信誉、内容或资格系统都不能把一个代理信号当成现实本身。

系统真正需要的是多源证据、适用范围和失效条件。

发现操纵以后，也不应让算法直接进行人格处罚。

应先核验事实，再撤销无效记录、隔离风险，并按相应规则处理责任。

指标可以帮助看见。

一旦指标开始决定命运，它自己也必须被看见。

7.6 一个账号，不一定等于一个人

生成式 AI 让批量制造头像、文本、互动和自动化账号变得更便宜。

这会放大一个早就存在的问题：如果制度只数账号，规模本身就可能被伪造。

在投票、补贴、资格、推荐和声誉系统中，同一主体制造大量虚假身份，可能破坏程序。

可防女巫攻击也不能滑向另一个极端：为了证明“我是我”，要求每个人把完整人生交给系统。

身份核验需要和具体事项相称。

有些场景只需要证明“这是一个独立参与者”。

有些需要年龄、地域、资格或唯一性。

真正高风险的事项才可能需要更强身份、人工复核与责任链。

时间、社交关系、资产、线下活动都可以提供证据。

它们谁也不等于人格价值，更不应该合成一张通用社会信用分。

身份系统最重要的能力，不是看见一个人的全部。

是只看当前问题真正需要的那一部分。

7.7 少暴露一些，也能证明一些

可核验性最容易被误解成“把全部数据交出来”。

现实中，很多验证只需要知道一个条件是否成立。

一个场所可能只需要知道“是否达到法定年龄”，而不是同时获得住址、完整生日和全部身份编号。

一个审计场景可能只需要验证某项数据满足公开规则，而不是把所有商业明细交给每一个验证者。

零知识证明等密码学技术提供了一种可能：在特定数学条件下，证明某项命题成立，同时减少与该命题无关的信息披露。

它很强，也很有限。

它可以证明“某个输入满足某个既定关系”。

不能替我们证明输入事实本身真实，也不能证明规则本身正当。

所以技术真正解决的是一部分隐私—核验冲突。

谁有权要求证明、证明什么、错了怎样申诉，仍然是制度问题。

可核验不等于全量透明。

这条边界比任何具体证明系统的名字都更值得长期保留。

7.8 连接结构：哪里正在断裂，哪里仍然有桥

有些事实不在单个人身上，而在关系之间。

一个社会的群体可以各自内部非常紧密，却几乎不再和彼此说话。

网络科学能够提示聚集、桥接、中心和断裂等结构。这些信号没有天然道德方向。

亲密共同体本来就需要高内部信任。

真正值得关注的，是跨群体连接减少是否同时伴随误解、仇恨、协作失败或真实伤害。

结构只能提示“去哪里再看”。不能自动告诉我们谁是好人、谁是坏人。

思想情景一：一名程序员在选举前看到社交媒体上的族群仇恨不断升温。

他做的插件不替用户审判对方，只在仇恨内容旁边提示：你们仍共享同一个社区、市场、球队或公共空间。

仇恨不会因为一个界面消失。但一个被抽象成“敌人”的人，重新出现了关系坐标。

思想情景二：战争中，海外难民与本地目击者组成翻译志愿网络，把地方口述送进更多语言。

他们重要，不是因为掌握某个网络指标。是因为他们愿意成为那座桥。

思想情景三：停火以后，两边的人第一次坐到同一张桌上。

不要求原谅，只要求把失去的人说出来，也把对方的话听完。

制度停火了，记忆还没有停火。

桥梁的价值，就在这种时候出现。

如果文明拥有更好的可观察性，这些常被忽略的翻译、维护与连接就更有机会被看见。

看见以后，可以承认。却不必把所有桥梁统一计价，更不能由网络位置自动生成权力。

7.9 有限可观察性：看不见内部每一步，也不能什么都不留下

任何复杂系统都不可能做到完全透明。

传感器有盲区，日志会缺失，模型内部状态也可能很难被人直接解释。

治理因此不需要承诺“我们能看见系统内部每一次犹豫”。

它需要的是足够的外部证据。

关键输入来自哪里。什么规则和权限正在生效。高后果动作发生了什么。谁能暂停、降级或接管。出了问题，谁能复核、申诉和追责。

系统越难直接理解，这些外部接口越重要。

停止能力也不能只依赖被监督系统自己。

具体工程可以不同，但高后果系统应尽量避免单一路径绕过安全控制，并让独立主体有办法验证关键停止与接管机制是否真的存在。

这已经足够形成一条长期原则：

我们未必能看见全部内部过程。

但不能因此失去事实、边界和阻止现实伤害的能力。

7.10 本章结论：可核验事实，是纠错与问责的重要前提

看见，不会自动带来正义；但什么都看不见，正义往往连抵达现场的路都找不到。

所以一章最后只留下一判：

正义不能只有愿望。

它还需要看见事实的能力。

历史上的信息不对称、证据缺失与权力封闭，

确实可能让纠错和问责更困难；但调查、证言、审判、审计与档案也一直是人类对抗这一问题的制度工具。

- 我们被安然的财报欺骗，因为看不见未来十年的债务。
- 我们被眼镜蛇效应愚弄，因为看不见指标背后的博弈。
- 我们被 SEO 垃圾淹没，因为看不见链接农场的拓扑结构。
- 我们被女巫分身操纵，因为看不见数字身份的伪造历史。
- 我们被黑箱算法伤害，因为看不见决策的因果链条。

AI、密码学与网络分析可以显著扩展我们的观察能力，

但不会给予任何系统真正的“上帝视角”。

技术 解决的问题 可观察性贡献

PageRank 廉价信号泛滥 关系拓扑 > 内容自述

Gitcoin Passport 女巫攻击 历史足迹 > 瞬时身份

零知识证明 隐私 vs 透明悖论 可验证且不泄露

网络拓扑分析 失序信号测量 结构价值 > 音量价值

独立熔断机制 高后果系统失控 为停止与接管保留独立通道

这套感知系统，

可以组成高复杂度风险的观察工具箱；它们只显影限定事实与风险，不能成为持续观察所有人的统一“文明雷达”。

它让我们在迷雾中，更有机会看见：

· 哪些行为在修复共同条件，哪些正在累积风险

· 哪些成本由行动者承担，哪些被甩给他人

· 哪些收益只属于当下，哪些后果留给下一代

当更多事实能够被看见，

问责才有更坚实的起点。

当责任不再轻易蒸发，

正义才不只是口号。

但“看见更多”仍不等于“系统已经知道一切”。

那么接下来的问题就是：

当系统出错了，谁该为此负责？

是写代码的人？是按按钮的人？还是算法本身？

是股东？是用户？是监管者？还是那个被设计成“永远正确”的系统架构？

责任不能蒸发。

损害不能因为责任链复杂就被无主外包。

权力、控制能力、受益与责任应在具体制度中保持相称，而不是追求抽象“同构”。

接下来的第八章继续建立一套责任语言：它不是由本书直接生效的“法案”，而是帮助现实制度区分风险、授权、控制与责任的上位框架。

方向与责任——高后果底线、停止条件与责任锚定。

目标是减少“后果真实发生、责任却在复杂链条中蒸发”的结构，而不是宣告可以终结所有无主后果。

第八章 让后果入局：谁为系统的决定负责

“系统可以把动作拆成一千步，后果却仍会完整地落到某个人身上。”

——本书

“真正的责任，不是事故以后找一个人承受愤怒，而是在事故以前就知道：谁决定，谁能停，谁必须解释。”

——本书

“未来不一定听从我们的蓝图，但有些门一旦关上，就很难再打开。”

——本书

系统越复杂，责任越容易躲进一句话：不是我决定的。第八章只做一件事——让真实后果重新找到真实的决定、权限、义务与承担。

8.1 先别急着画天堂：先看哪些后果不能轻易跨过去

人很容易爱上“最优蓝图”。它让复杂世界像终于可以画成一张图、写成一套规则，再交给某个系统执行。可未来最重要的能力之一，恰恰是反驳今天。

所以面对高后果技术，第一步未必是回答“最好的未来是什么”。

有时应该先问：哪些伤害一旦发生，已经没有足够的后来可以修？

一座桥可以返工，一次软件更新可以回滚；一条生命、一次大规模不可逆污染、一次失控扩散，却可能没有同样的第二次机会。

这就是“否定之路”在本书里真正有用的部分：我们未必知道唯一最优解，却可以先识别一些明确不可接受的高后果路径。

它不是悲观主义，也不是“凡有风险就禁止”。

风险永远存在。真正需要提高门槛的，是后果极重、可逆性很低、影响他人基本权利，而证据和授权又不足的决定。

因此，高能力 AI 可以帮助显影风险、比较情景、寻找早期信号，却不能因为“看得更远”就取得替社会定义美好生活的权力。

我们不需要一个系统替所有人回答“应该怎样活”。

我们需要的是：当某条路可能把共同未来封死时，有人能看见，有程序能停下，也有人必须为继续向前说明理由。

这会改变风险治理的提问顺序：

先问会伤到谁。

再问损害有多大、能不能恢复。

再问谁有权把这种风险施加给别人。

最后才问：收益是否足以支持我们继续。

增长、效率、收益和规模都可以重要。它们不能独自替不可逆后果结账。

低风险、可逆的探索应保留更大空间；越接近共同存续、基本权利和不可恢复损害，证据、授权、冗余与停止条件就越要变重。

这不是一套统一“负向剔除算法”。它只是把一件常被技术乐观掩住的事实重新放回桌面：能做，不等于现在就该做。

真正要保留的不是一个新主义，而是一种克制：在不知道唯一正确未来时，不要轻易替所有后来者关闭回头路。

设计者可以提出方向，可以做试验，也可以承担风险。

不能做的，是把自己的答案写成别人无法退出的命运。

AI 在这里是认知工具和受授权执行者，不是未来作者。

借雕刻作比喻：有时创造不是不断往上加，而是先把明显会毁掉作品的部分去掉。

对制度而言，它意味着：先守住不能轻易失去的东西，再给未知留下空间。

我们不定义统一的美好生活。

我们对高后果事项建立与权利、事实和风险相称的边界。

没有触及高后果底线，也不意味着一切自动合法；法律、合同、专业责任和他人权利仍然存在。

真正的自由不是没有边界，而是边界之外仍有足够多的路可以自己走。

这种思想并不属于某一个文明传统。不同法理、宗教与哲学都曾用不同语言处理同一个难题：什么时候应当在伤害完全发生以前行动。

它们给出的答案并不相同。真正值得继承的，不是把某种传统抬成万能公式，而是承认风险判断需要事实、概率、比例与权利。

因为“为了避免坏结果”本身也可能成为过度控制的理由。

所以每一道高后果边界，都必须同时接受另一道追问：我们是不是为了防一个可能的错误，制造了更大的确定性伤害？

如果把时间再拉长一点，问题会变得更清楚。

今天作决定的人会离开，承担后果的人可能还没有出生。

我们不能假装得到他们的授权。

能做的，只是在不可逆处多留一点克制，在可逆处多留一点选择。

这就是“向未来借一双眼睛”真正的意思。

不是预测十年后的孩子会怎样评价我们。

是承认他们有权评价，而且不必永远服从我们今天的答案。

否定之路最深的地方，不是“不要做”。

是知道自己可能错。

知道自己可能错，却仍然必须决定——这才是高后果治理真正困难的地方。

所以，克制不能代替行动；行动也不能取消克制。

我们最终要建立的，不是一个永远正确的系统，而是一条在错误发生以前能停、发生以后能查、造成损害以后有人修的责任链。

8.2 坦佩的夜晚：当一条责任链在撞击前已经断开

2018年3月18日，晚上9点58分。

美国亚利桑那州坦佩。

一辆由 Uber Advanced Technologies Group 测试的改装沃尔沃 XC90，正在自动驾驶系统控制下沿道路行驶。

49岁的伊莱恩·赫茨伯格推着自行车横穿道路。

车辆撞上了她。

她后来死亡。

真正让这场事故难以忘记的，不只是“机器撞了人”。

是系统其实已经看见了她。

NTSB 调查显示，自动驾驶系统在撞击前约 5.6 秒首次检测到她，却不断改变分类，也没有正确预测她的路径。^[21]

到撞击前约 1.2 秒，系统判断碰撞已迫近；但当时的设计在这种情形下不执行用于减轻碰撞的紧急制动，而依赖车辆操作员接管。

与此同时，操作员长时间把视线移开道路。手机记录显示，她当时正在通过应用观看流媒体节目。

一秒多一点。

这已经不是“算法还是人”的二选一。

系统的感知和路径预测有问题；紧急处置设计有边界；操作员没有持续监控；测试组织的安全风险管理和监督也存在缺陷；州级测试监督同样被 NTSB 列为影响因素。

一条生命最后落在了多层失效交汇的地方。

这正是复杂技术最容易制造的错觉：因为没有一个人控制全部，所以仿佛没有一个人需要回答。

可“没有全能控制者”，从来不等于“没有责任关系”。

事故以后最容易发生的是责任倒流：所有人的目光最后只盯住最靠近方向盘的人。

可责任链必须往回走。

谁决定测试方式？谁设计了紧急处置？谁知道自动化依赖会让人类监控变得困难？谁批准系统进入公共道路？谁能在发现风险以后停下来？

这些问题不能等到葬礼之后才开始写。

NTSB 最终把事故的可能原因和促成因素分到操作员、Uber ATG 的安全文化与风险管理、行人行为以及亚利桑那州测试监督等多个位置。

这不是责任被稀释。恰恰相反，是责任终于不再被压成一个方便的替罪者。

复杂系统里的责任可以分布。

它不能蒸发。

2023 年 10 月，旧金山又出现了另一种责任困境。

一名行人先被一辆人类驾驶车辆撞入一辆 Cruise 无人车的路径。Cruise 车辆制动后仍发生碰撞；在初次停车后，系统又执行了靠边动作，将行人向前拖行约 20 英尺。^[22]

后来，NHTSA 认定 Cruise 提交的部分事故报告没有完整披露这一碰撞后的拖行细节；加州 DMV 也暂停了 Cruise 的无人驾驶部署和测试许可。

这里真正值得记住的，不只是一个软件判断错了。

事故后的信息披露，本身也属于责任链。

谁知道了什么？什么时候知道？向监管者说了什么？有没有把关键后果完整留下？

技术系统越复杂，越不能只把“责任”理解成最后是谁碰到了按钮。

责任还包括设计、批准、监督、披露、停止、修复和补偿。

如果每一层都只说“我只负责这一部分”，最后真正完整的，可能只剩受害者承担的后果。

所以第八章不寻找一个必须坐牢的“终极责任人”。

它寻找的是一条不能断的链。

每个角色只承担与事实、控制、义务、可预见性和适用法律相称的那一部分。

但没有人可以仅仅因为机器站在因果链中间，就让自己本来承担的职责一起消失。

责任的第一条原则因此很简单：自动化可以参与行动，不能替责任主体消失。

8.3 责任不是找一个人兜底：谁决定什么，谁承担什么

事故以后，人最想要一个名字。

一个可以被指认、被愤怒、被惩罚的名字。

可复杂系统真正需要的，往往不是一个替罪者，而是一张不会把任何关键决定藏起来的责任地图。

自然人、法人、公共机构、雇主、专业人员、开发与部署组织，都可能在不同事实和法域中承担不同责任。

责任也不是一种东西。刑事、民事、行政、职业、合同和组织责任，各自有不同成立条件。

所以“系统出了事，必须有一个自然人坐牢”听起来有力，却不是可靠法理。

真正不能退的是另一条边界：高后果系统进入现实以前，必须知道谁决定部署、谁控制关键权限、谁承担专业或组织义务、谁能停止、谁负责监督和修复。

不是为了提前判谁有罪。

是为了事故以后，不必从一片失忆里重新寻找责任。

日志、版本、授权、风险评估、合同和组织文件的价值，都在这里。

它们不是为了把算法拟人成一个需要“监护人”的新法律主体。

它们是为了让真实的人和组织不能在系统变复杂以后突然变得不可追问。

谁写代码，可能重要。

谁提出需求，可能更重要。

谁批准把系统接入真实世界，可能更重要。

谁看见风险却没有行动，也可能进入责任链。

没有一个职位天然承担全部责任。

也没有一种组织结构天然免除责任。

如果事故来自设计逻辑缺陷，责任要看设计和批准；如果来自违法部署，要看部署和授权；如果来自操作失误，要看操作、训练与监督；如果监管本身失职，也不能因为它站在系统外面就从因果图上消失。

这听起来没有“一个人兜底”那么痛快。

却更接近责任真正的样子。

责任必须和真实权力、控制、义务与行为相匹配。

享受效率和收益，不会自动生成无限责任；同样，“是算法做的”“是供应商模块”也不会自动生成免责。

一个组织把模块外包出去，不等于把自己本来负有的部署、安全和监督职责一起卖掉。

一个管理者拥有职位，也不等于所有技术失败都应由他个人承担。

一切都要回到事实。

谁设计了什么。

谁批准了什么。

谁能够改变什么。

谁知道了什么风险。

谁本来有义务行动，却没有行动。

责任不是靠恐惧维持。

它靠的是：决定、停止、知情、监督、修复和受益这些位置，在事故以后仍然找得到。

不同法律传统对危险创设、过失、因果和注意义务有不同规则。本书不把它们压成一条古代格言。

能够跨制度保留的直觉更简单：一个主体若创造、控制或显著放大风险，不能仅因工具、机器或算法位于因果链中间而自动免责。

同样，“在因果链上”也不等于当然有罪。

事实、控制关系、可预见性、义务来源、因果联系、程序与适用法律，仍然决定具体责任。

文明跃迁不发明“无限责任法”。

它只坚持：责任不能被技术外包。

系统日志、授权记录、风险评估和人工干预条件应帮助还原责任链，而不是替法院、监管者或专业机构预先判决。

算法不是责任主体的万能替身。

供应商不是部署者的万能替身。

管理者也不是所有技术失败的万能替罪羊。

责任必须落到真实事实与真实权限上。

谁设计了什么，谁批准了什么，谁能改变什么，谁忽略了什么风险——这些问题都应该能够被问到底。

复杂性可以让答案变难，不能让问题消失。

8.4 346 条生命：自动化事故为什么从来不只是一行代码

2018 年 10 月 29 日，狮航 610 航班坠入爪哇海。

189 人遇难。

不到五个月后，2019 年 3 月 10 日，埃塞俄比亚航空 302 航班坠毁。

157 人遇难。

两起事故，共 346 人。

事故调查把一个名字推到世界面前：MCAS，机动特性增强系统。

可如果我们只说“软件害死了人”，责任又会在一句简单解释里消失。

真正需要看见的是：设计、传感器输入、飞行员工作负荷、认证、安全评估、组织信息和监管如何叠在同一架飞机上。

FAA 后来总结的关键问题包括：原 MCAS 可依赖单个迎角传感器输入；错误高迎角信号可能触发机头向下配平；在原设计中，飞行员电动配平后系统还可能再次激活。^[23]

一个错误信号，因此可能不只是一次错误动作。

它可以在一个已经高度紧张的人机环境里反复回来。

这就是高后果自动化真正需要害怕的东西：一个局部假设，穿过多层系统以后，变成现实中的力。

软件本身不是全部原因。

也正因为如此，责任不能只指向一个程序员。

谁批准了设计假设？谁做了安全评估？谁负责认证？飞行员得到了什么信息和训练？组织怎样处理已知风险？

第一起事故以后，世界已经知道 MCAS 进入了事故调查。

第二起事故仍然发生。

这使责任问题变得比“谁写了哪一行代码”更沉重：一个组织和监管系统怎样接收第一次事故留下的信息，又怎样决定下一步。

真正值得文明记住的，不是寻找一个足以承担全部愤怒的人。

而是任何一层都不能因为“我只负责这一部分”，就假装自己与最后的后果无关。

设计者负责设计。

认证者负责认证。

运营者负责运行。

监管者负责监督。

这些责任不是一份工作说明书。它们必须在高后果系统里彼此咬合。

否则，最完整的东西仍然会是受害者承担的后果。

346 条生命，就会落进缝隙。

波音案给高后果自动化留下的，不是一个可以照抄的监管模板，而是三道必须被认真回答的门。

第一条：关键自动化必须可感知

当自动化会显著改变操作者的安全判断与控制状态时，应按照风险提供与任务相称的提示、训练和接管信息。

具体是视觉、声音、触觉还是其他接口，应由人因、场景和专业标准决定。

真正不能接受的是：关键自动化已经改变高后果状态，操作者却连“发生了什么”都无从知道。

无缝体验可以是产品目标。

它不能把安全相关的控制变化一起做成“无感”。

高风险系统里，知道谁或什么正在控制，是基本安全条件。

第二条：关键设计决定必须可追溯

传感器依赖、冗余架构、失效模式、关键安全假设和重大变更，都应该留下足以支持调查与问责的记录。

这不意味着每一个决定都要指定一个“终身责任人”。

它意味着二十年以后如果有人问“为什么当时这样设计”，制度不能只剩一句：没人记得。

名字、职位、批准链、版本和依据，都应该在与风险相称的范围内留下。

记录保存多久，由适用规则和系统生命周期决定，不由思想书统一规定。

第三条：独立监督不能只剩形式

公共安全系统的认证与监督，需要足够的独立性、专业能力和利益冲突防火墙。

行业专业知识可以进入认证，不能因此让被审查者实际上只向自己证明自己安全。

独立并不意味着“研发方永远不能参与”，而是关键判断必须有真实外部复核和制度责任。

安全不能靠自证。

这不是反商业，也不是反技术。

它只是拒绝让进度、利润和组织惯性拥有最后一句话。

责任不需要一个英雄。

8.5 当每个人都只负责一厘米

复杂系统最危险的一句话，往往很普通：我只负责这一部分。

工程师负责模型，产品经理负责指标，运营负责上线，法务负责条款，监管负责合规。

每个人都可能真的只看见自己面前的一小块。

可受影响的人不会只被伤害“一小块”。

技术事故当然不能和大屠杀相比。汉娜·阿伦特关于艾希曼的讨论留下的责任问题是：角色、流程和命令，会不会让一个人停止追问自己正在参与什么。^[24]

“我只是执行。”

“我只是按流程。”

“我只是优化这个指标。”

这些句子有时完全真实。

也正因为真实，它们才危险。

一个人可能确实没有控制全部系统。

可他仍然需要知道自己的那一厘米接在什么后面。

如果每个人都只把视线停在岗位边界，系统整体造成的后果就会越来越像“没有人做的”。

不是因为真的没有人。

是因为没有人再把自己的动作和最后的后果放在同一张图上。

想象一次事故调查会。

模型工程师说：“我只负责训练。”

部署负责人说：“模型通过了评测。”

业务负责人说：“这是行业通行做法。”

管理者说：“我没有参与技术细节。”

审计人员说：“我只检查了授权范围内的项目。”

每一句都可能事实基础。

可如果责任链到这里就结束，受害者面对的仍然是一个完整后果。

所以责任制度真正要做的，不是逼每个人对全部世界负责。

是让每个人知道自己的边界，也知道边界在哪里与别人的责任相接。

自动化会增加道德距离，也会增加组织距离。

模型内部越难解释，越不能把责任全部押在“找到那一行错误代码”上。

我们还可以追问输入、权限、部署目的、运行日志、停止条件、组织义务和事故后的披露。

责任不必进入模型的每一个内部状态，才能存在。

它首先存在于现实世界：谁让系统进入这个场景，谁给了它权限，谁有能力停止，谁在受益，谁承担法定或专业义务。

复杂性会让因果调查更难。

不会让现实责任自动归零。

“不知道全部”不是免责。

“只负责一部分”也不是原罪。

真正要防的是：所有部分加起来以后，竟然没有任何制度能够回答整体后果。

8.6 独立审计：不能只靠自己证明自己安全

一个高风险系统如果只能由开发和部署它的人解释、测试和证明，利益冲突几乎从第一天就写进了流程。

所以需要独立的测试、评估、审计与符合性检查。

独立不是建立一个全球“算法审计师阶层”。

它是一种更古老的制度智慧：掌握利益的人，不能独占证明自己无害的权力。

美国证券法在 1930 年代把独立会计师审计写进资本市场的重要制度安排。^[25]

原因并不神秘。公司自己编制财务信息，市场却需要有人从外部检查这些信息是否值得信任。

审计当然也会失败，也会有利益冲突。

但现代资本市场仍然保留这道外部验证，是因为“自我证明”从来不是足够的公共信任基础。

AI 系统正在遇到类似问题。

数据从哪里来？模型在哪些人群和场景上失效？权限有没有越界？日志是否足以复核？安全边界能不能被普通业务指令绕过？

这些问题不能只靠一份产品白皮书回答。

未来的 AI 审计也不需要复制“四大”结构。

真正需要的是多种独立评估能力进入高风险系统。

审计人员的任务不是成为新的技术法官。

他们要把藏在代码、数据和组织流程里的问题，转化成有权主体能够判断的证据。

能力仍然不生成权力。

发现风险，不等于天然拥有停止系统、处罚主体或剥夺权利的权力。

审计意见何时触发暂停、整改、重新认证或其他程序，必须来自法律、监管规则、合同或明确授权。

最终效力来自制度，不来自“我比你更懂算法”。

独立审计真正重要的地方只有一句：

任何掌握技术的人，都不能只靠自己证明自己无害。

而任何掌握审计权的人，也不能只靠专业身份证明自己永远正确。

现实世界已经存在 AI 风险管理、符合性评估、算法影响评估和测试框架；它们的法律效力与适用对象并不相同。

例如 NIST AI 风险管理框架本身是自愿性框架。

所以这里不发明一个全球统一的“算法适航证”。

需要建立的是可追责的专业链：谁开发，谁部署，谁测试，谁审计，谁授权，谁在事故以后承担相应责任。

审计不是咨询的漂亮包装。

也不是新的权力出生证明。

它是一道让风险可以被别人看见的门。

技术需要外部检查。

谁有权要求、何时触发、出了问题谁负责，仍然必须回到制度。

8.7 瑞士奶酪：不要让所有防线在同一个地方失效

独立审计仍然只是一层。

詹姆斯·瑞森提出的“瑞士奶酪模型”留下了一个非常直观的安全图景：每一道防线都有孔。

人会犯错，测试会漏，组织会受进度压力，监管会迟到，传感器也会故障。

事故往往不是因为世界上只有一个孔。

而是某一刻，多层缺口恰好让同一条危险一路穿过去。

所以安全不能押在“这一层一定不会错”。

真正的纵深防御，是让不同层用不同方式失败。

一层看错，还有另一层能发现；一层失效，还有另一层能停。

这比寻找一个永不犯错的模型现实得多。

可以把它落成四类防线：

第一层：把部分关键安全边界放到更难被上层系统随意改写的位置。

工业机器人、自动驾驶和其他高后果执行器，可以根据任务把速度、力矩、区域或停机条件交给独立安全控制。

具体阈值由工程和专业标准决定。

思想层真正关心的是：上层模型不能用一条普通业务指令就把最后的安全边界删掉。

这是一种“宪法嵌入”的工程类比，不是物理定律。

第二层：关键安全功能避免单点脆弱。

MCAS 的教训之一，就是一个错误输入不能轻易穿透全部控制链。

相互独立的传感、验证、冗余和故障降级，可以降低同一种错误一路到底的概率。

需要多少传感器、采用什么原理，由具体领域决定。

没有一张思想书可以替工程师给所有系统规定“三重冗余”。

真正的原则只有：别让一个点的错误轻易成为整个系统的命运。

第三层：异常发生以后，系统仍然能降级、暂停、隔离或接管。

这些动作必须有真实授权，也必须说明哪些权限被收回、哪些基础服务继续。

停止能力不是“人类永远在线”的浪漫想象。

它只是要求：系统不能把退出和接管从设计里删除。

扩音器可以很强。

电源必须找得到。

第四层：让历史和极端情景在上线以前先来一次。

历史数据、模拟、对抗测试和故障注入，都可以让已经发生过的愚蠢更难重复。

测试跨度和通过条件仍由风险、技术成熟度与专业标准决定。

压力测试不能保证未来安全。

它只是让我们少一点“从来没想过”。

每一道防线都可能失败。

纵深防御的意义，就是不要让所有防线以同一种方式、在同一时刻一起失败。

安全不来自一句“系统已经足够聪明”。

它来自结构冗余、责任、证据和可停止性。

8.8 残余风险：事故发生以后，不能只剩受害者自己买单

即使设计、审计、冗余和停止机制都做了，风险也不会归零。

新技术尤其如此：有些失效可以预见，有些只能在真实世界里逐渐暴露。

这不是接受伤害的理由。

它只是要求我们在“谁有过错”之外，再回答另一个问题：真实损害已经发生时，受害者怎样及时得到修复和补偿？

有些事故最后能找到明确责任人。

有些事故的因果链很长，责任认定需要时间。

还有一些残余风险，在现有规则下未必能简单归到单一主体。

可受害者的医疗、收入中断和生活损失不会等调查结束才发生。

因此，高风险技术可以研究责任保险、行业准备金、公共救济或其他风险共担机制。

这些机制不是给责任人买免责。

它们是为了让“责任复杂”不等于“受害者先自认倒霉”。

资金从哪里来、覆盖什么风险、谁管理、怎样防利益冲突，都必须由真实法域和行业条件决定。

不能因为使用 AI，就自动产生一笔抽象的“技术税”。

也不能因为设了保险，就把设计缺陷、违法部署、操作失误或监管失职一起冲销。

赔付和追责可以并行。

先让人得到必要救济，不等于以后不再追问谁应承担什么责任。

风险共担机制自己也需要责任。

管理主体、监督、信息披露、偿付能力和利益冲突规则都不能悬空。

技术专家、受影响者、保险和监管视角可以参与。

谁拥有决定权，仍然由适用法律和制度回答。

创新和承担不应站在彼此对面。

允许探索，不意味着让最倒霉的人独自承担探索的代价。

真正成熟的技术社会，会在事故以前准备安全，也会在事故以后准备修复。

风险不能归零。

责任与补偿不能因此归零。

8.9 让后果抵达：收益不能总向上，代价总向下

为什么高后果系统必须反复谈责任？

因为技术可以把动作拆开，后果却仍然会完整抵达一个人。

被撞的行人不会只承受“感知模块的 20%”。

被误诊的患者不会只承受“部署团队的那一部分”。

当损害发生，生活接到的是整张账单。

问题因此不是让决策者也“感到同样的痛”。

责任制度不能靠制造痛苦来证明正义。

真正要改变的是另一种结构：收益集中在能决定的人手里，代价却长期外包给没有参与决定的人。

被撞的行人、被误诊的患者、被误导的消费者、承担系统性外部成本的公众，都不能永远只是技术收益背后的沉默成本。

波音、坦佩和 Cruise 留下的共同问题，不是谁“应该疼”。

是设计、部署、监督、披露、收益与修复之间有没有闭环。

金融、平台和其他自动化系统也一样。

如果奖励只跟短期收益绑定，而长期风险可以甩给客户、公众或未来，制度就会持续奖励把后果推远的人。

责任机制要做的，就是把被推远的后果重新拉回决定过程。

第八章最后只需要一句话：

让能够作出高后果决定的人，不能与决定的真实后果彻底隔绝。

这不意味着所有人承担同样责任。

它意味着权力、收益、义务和后果之间不能永远只有单向流动。

让有权主体承担与其角色相称的责任，是让权力与后果重新相遇。

让收益主体依法分担残余风险，是让收益与成本重新相遇。

让算法和组织接受独立审计，是让不可见的决定重新进入可质询的世界。

让补偿机制提前存在，是让无辜者不必等责任链全部理清以后才开始被看见。

真正握着刀的人，不需要先割伤自己。

他需要知道：每一次落刀都必须说明理由，越过边界要承担后果。

敬畏不来自共享疼痛。

敬畏来自一个事实——后果不会因为系统复杂就被制度遗忘。

我们当然希望减少痛苦。

真正不能接受的，是收益总向上集中，代价总向下流动。

无论系统多强，

责任都不能被复杂性、速度或智能水平吞没。

8.10 看不见模型每一步，也不能看不见责任

系统越复杂，责任越不能建立在“必须复原模型每一步内部过程”这一前提上。

有些模型很难完整解释，未来系统也可能包含更复杂的搜索、并行计算或新型硬件。

这会增加调查难度。

不会自动制造法律黑洞。

即使内部过程无法逐步复原，输入、权限、关键状态、输出、日志和现实后果仍然可以留下证据。

治理因此不只问：“模型里面到底发生了什么？”

还要问：谁让它进入这个场景？谁给了它这些权限？谁知道风险？谁有能力停止？

复杂性越高，越要把一部分证据留在模型之外。

内部不可完全解释，会增加调查难度。

不会当然消灭外部证据、组织责任和制度授权。

所以责任不能只靠事后威慑。

更稳健的方法，是把事前边界、运行记录、停止能力和事后责任连成一条链。

事故以前，先知道谁可以部署、权限到哪里、哪些风险必须先处理。

运行中，知道系统是否可监测、可降级、可停止、可接管。

事故以后，知道怎样调查、复核、补偿和追究相应责任。

任何高后果智能系统都必须回到现实责任主体。

谁承担何种责任，由具体法域、组织角色、控制能力、过错和法定义务决定。

不能因为系统复杂，就把全部后果压给一个所谓“终极监护人”。

也不能预设某个自然人对一切不可预测后果承担无限责任。

责任必须与真实权力、控制、行为和义务相匹配。

不可完全解释不是免责理由。

也不是制造无限责任的理由。

它只要求我们把更多可验证证据、权限边界和外部控制留在系统之外。

让系统可以越来越复杂。

让责任不要跟着一起消失。

8.11 本章结论：伤害发生以后，不能只剩一句“不是我决定的”

责任走到最后，不能只停在一张流程图上。

坦佩的夜晚、346 条生命、Cruise 事故后的披露争议，都在提醒同一件事：后果不会因为系统复杂就消失。

真正承受代价的人，也不会因为责任链很长，就只承受一部分后果。

所以方向不是幻想一个永不出错的系统。

是让错误发生以前有人能停，发生以后有人能查，造成损害以后有人修。

责任也不是事后抓一个人承受全部愤怒。

它是让权力、决定、义务、收益与后果重新彼此看见。

这一章最后可以留下四条边界：

高后果底线：不知道唯一最优未来时，先识别明确不可接受且高度不可逆的风险。

责任相称：控制、受益、义务、可预见性和风险承担回到具体事实与制度。

后果可见：受影响者的真实损害必须进入决定、复核、修复和责任过程。

复杂不免责：看不见模型每一步，也不能因此看不见外部证据、授权和责任。

它不要求人性完美。

也不要求技术完美。

它只要求一个高后果系统不能靠复杂性制造责任失忆。

违法、欺诈、重大疏忽和越权，应当承担与事实和法律相称的现实责任。

安全、审慎和及时纠错，也不能在短期激励里永远吃亏。

真正成熟的责任体系，不把每一次事故都变成猎巫。

也不把每一次事故都变成“系统太复杂，谁也没办法”。

第七章让我们看见真实。

第八章再追问：看见以后，谁必须回答？

事实没有责任，会变成旁观。

责任没有事实，会变成惩罚冲动。

两者必须在一起。

接下来，问题会继续扩大。

有些风险已经越过一个组织、一个行业甚至一个国家。

当后果跨境，责任链还不够。

我们还必须回答：谁有资格作出具有约束力的共同决定？

共同风险扩大，不等于新主权自动产生。

所以第九章进入的不是“谁最懂全局”。

而是更难的问题：谁有权决定。

主权国家、国际组织、企业、专业机构和算法，各自能做什么，又不能做什么？

这就是下一章的入口。

合法性不是能力的奖品。

它必须有来源。

第九章 谁有资格决定：当问题越过国界

“从太空回望地球，政治边界不会像地图那样显现；首先进入视野的，是同一颗行星。”

——据多位航天员“从太空看不到地图国界”的回望经验转述

“河流、季风、传染病与供应链都不会自动停在政治边界前。”

——本书

“主权可以守住自主，也可以成为合作的出发点；两者并不必然冲突。”

——本书

河流过境不会停下来出示护照，病毒、气候、供应链和算法也不会。但后果越过国界，并不意味着权力也可以越过授权。

9.1 主权与跨境协同：自主边界怎样处理共同后果

现代国际秩序的重要基本单元，是主权国家。

1648 年的威斯特伐利亚和约常被后世视为这一秩序形成的重要历史节点，但现代主权原则并非在那一刻一次性定型。

长期演化之后，国家主权至少形成了几条重要边界：

- 领土完整与政治独立受到保护
- 国内管辖原则与不干涉原则具有重要地位
- 国家仍掌握大量合法强制与公共决定能力
- 但战争、跨境责任与人权早已不再被理解为“主权可以做任何事”

这套秩序支撑了现代国际关系，也不断在战争、贸易、环境、公共卫生与人权中被重新解释。它的价值，不在于宣布国家彼此隔绝。

而在于让不同政治共同体拥有可以被承认的自主边界。

真正的新问题是：当风险与依赖跨越边界时，

主权如何在不失去自主的同时，获得处理共同问题的能力？

粮食、能源、金融、候、染病和供已反明：

一个国家可以独立决定许多事情，却无法独自控制所有后果。

随着高能力 AI、生物工程与全球基础设施进一步耦合，这个矛盾只会更突出。

挑战类型 传播路径 能否用国界阻隔

大流行病 空气、人群、货物 ✗ 不能

气候临界点 大气环流、洋流 ✗ 不能

算法失控 代码、数据、信号 ✗ 不能

深伪认知战 社交网络、推荐算法 ✗ 不能

高能力 AI 能力扩散 模型、接口、数据与工具链 ✗ 很难仅靠国界阻隔

高度连接的系统性风险，可能跨越国界迅速传播。

现实世界已经给过多次提醒。

2008 年的全球金融危机提醒我们，金融风险可以沿高度连接的市场迅速跨境传播。

2020 年的 COVID-19 疫情提醒我们，病原体不会因为边界存在就停止扩散。

2022 年俄乌战争及其后续冲击提醒我们，粮食、能源与供应链会把局部冲突的成本传到远方。

进入 AI 时代，模型、算力、数据、漏洞与合成内容同样具有跨境扩散能力。

在这个新现实面前，问题不再是“主权是否过时”，而是主权怎样获得处理跨境后果的协同能力：

边界可以界定权利与责任，却不能让相互依赖自动消失。

你越试图用边界隔绝混乱，

混乱越会从你意想不到的缝隙涌入。

跨境协同不等于主权终结。

更准确地说，它要求主权国家在保留自主与责任边界的同时发展跨境协同能力；这是一种治理要求，不是历史必然的“升级阶段”。

一种封闭式主权想象 可并存的跨境协同能力

领土与法域边界 跨境问题中的功能协同

边界即防线 边界即接口

控制能力不等于正当权力 协同提高共同能力但不自动产生权力

自足可降低部分依赖 可信连接、冗余与可退出协作也可增加韧性

国家对本国主体的责任 与跨国共同风险治理并行

这一转变并非没有先例。

1959 年签署的《南极条约》提供了一个重要参照。

它规定南极只用于和平目的，保障科学调查自由，并建立信息交换与检查机制。

更重要的是，条约没有假装主权争议已经消失。

它通过第四条保存各方既有立场，并冻结新的或扩大的主权声索。

这说明：重大分歧不必先被彻底消灭，合作仍然可以通过共同规则发生。

南极制度的价值正在这里——不是“没有主权”，而是在主权立场仍不同的情况下守住和平、科学与可检查性。

不是放弃一切边界，而是让部分共同目标拥有跨边界的制度空间。

深海制度提供了另一种不同的国际法结构。

《联合国海洋法公约》把国家管辖范围以外海床的“区域”及其矿产资源置于“人类共同继承财产”原则之下。

国际海底管理局于 1994 年《海洋法公约》生效时正式成立，负责组织和控制该区域矿产资源相关活动。

这套制度仍在演化，商业开发规则也仍存在现实争议；它证明的是共同资源可以被置于多边规则之下，而不是证明全球资源已经天然共有。

国际法中的公海、外层空间、南极与国际海底区域，也并不属于同一种法律制度。

它们的权利结构、自由范围、环境义务与治理机制各不相同。

因此，不能把“全球公域”当成一句可以抹平全部法权差异的口号。

真正可借鉴的，是跨境问题需要与其风险范围相匹配的协同机制。

有些对象不能被任何单一主体完全控制，

却仍然必须有人承担规则、维护与责任。

高能力 AI、数据、算力与信任基础设施提出了类似的跨境协同问题，

但它们并不因此自动成为“无主之物”或“全人类共有财产”。

不同对象仍要回到产权、合同、法域、国家安全与公共利益的真实边界。

共同风险扩大，只能证明协同需求扩大。

它不能自动取消国家、个人、企业与社会既有权利。

也不能凭“全人类利益”生成一个新的最高权力。

因此，第九章首先要守住的政治原则是：

跨境协同的必要性，不自动生成新的合法性。

合法权力仍必须来自现实法权、授权、同意、代表、程序与可追责制度。

对共同存续的贡献可以说明一种行动为什么重要，

却不能成为主权或公共权力的出生证明。

9.2 主权、自主与连接：安全来源不止一种

有一种担忧必须被认真对待：

“削弱主权，会不会导致小国被大国吞并？”

“弱化边界，会不会让文明失去根基？”

这种担忧不能被简单归为“误读”；小国安全、依赖与大国不对称本来就是跨境协同必须处理的现实问题。

主权的重要价值之一，是让政治共同体拥有可被承认的自主、权利与责任边界。

在跨境协同中，还需要进一步保护：

谁有权代表共同体进入何种关系、条件如何形成、依赖怎样可控、何时能够依法退出或调整。

瑞士常被用作中立、国际连接与高制度信任并存的案例；其安全与主权仍来自历史、国防、法律、外交、经济等多重条件。

但它没有海军，没有海岸线，没有石油，没有殖民遗产。

可以讨论的能力包括：

中立政策、制度信誉、经济能力、国际网络与防御体系等。

瑞士的安全并非“只靠连接”，

连接与相互依赖可以增加部分安全，但不能保证“没人想摧毁它”，也不能替代自主防御与制度韧性。

新加坡独立初期面临资源、腹地与安全等多重压力；后来形成港口、法治、教育、产业、外交和国防等多重能力。

它缺乏传统腹地并长期面对水资源与安全约束，但不能用“没有军队”等绝对表述概括其国家建设。

可以讨论的能力包括：

港口与贸易、法治、教育、产业、外交网络与国防等。

其发展高度依赖国际连接，

但“世界离不开它”只是修辞；真正可提炼的是可信连接与可替代性管理可以成为小国韧性的一部分。

哥斯达黎加在 20 世纪中叶废除常备军，其稳定与发展常被用于讨论不同安全与公共投资路径。

[26]

其政治稳定、社会政策与生态形象仍值得研究；真正可比较的，是不同安全结构如何与教育、社会政策、生态和国家韧性一起作用。

可以讨论的能力包括：

法治、教育、社会政策、生态保护、外交与经济结构等。

不设常备军不等于不存在安全体系；

其经验不能简化成“让邻国羡慕”；更可靠的启发是国家安全来源可以具有多种结构。

这些不是主权的削弱，

这不是一条统一的“主权升级”历史路线，而是跨境风险条件下对协同能力提出的新要求。

防御、威慑与边界保护仍是许多国家安全结构的一部分；

可信连接、互赖与共同规则也可能增加安全，但过度依赖同样会形成脆弱性。

在跨境风险持续增加的时代，这种“连接能力”会越来越重要。

主权能力 安全来源 可能的脆弱性 需要补充的能力

防御能力 军事与安全体系 新型数字和跨域风险 跨域韧性与协同

资源能力 能源、矿产与产业链 单一依赖与外部冲击 多元供应与替代能力

人口与人才能力 劳动、知识与社会组织 自动化与人口结构变化 教育、创新与社会适应

连接能力 可信网络与国际协作 过度依赖与外部传导 可退出、可替代、可恢复的连接

这些案例来自不同地区，但不能因此声称已经代表所有文明经验。

太平洋岛国论坛在 2023 年发布《蓝色太平洋大陆战略》：

“我们不把自己看作小岛国，而是看作大洋大国。”

相关区域叙事强调海洋身份与共同治理；

贡献可以增加政治与道义影响，却不构成主权的法律定义。

涉及区域资源与生物多样性的比例，必须回到原始文件、统计口径和明确时间范围核验；未经可靠来源支持的数字，不承担本书的理论证明。

“生态信贷”可作为修辞，不是现行主权或国际法类别。

非洲联盟《2063 年议程》的核心原则：

“一体化的非洲，繁荣的非洲，由非洲人自己治理的非洲。”

这类区域合作可以在国家自主决定、条约、共同规则与互惠安排下增强安全与连接；合作本身不自动构成“主权共享”，更不意味着主权被系统重新定义。

就像单独的树枝易折，成捆的树枝难断。

因此，未来跨境协同的制度探索，

不是要取消国家，

而是要升级国家——

国家仍是领土、人口、法权与责任共同体，同时可以在跨境任务中承担不同功能；

国家既承担本国责任，也可以依法参与全球公共品供给；二者不是线性替代。

9.3 资源的法理：在既有权利之上增加责任、共享与生效使用关系

要扩大跨主体、跨区域乃至全球资源协作，首先必须承认现实世界存在多种所有、控制、使用、收益、维护和公共管理关系。

本书不统一重定义各法域的“所有权”，也不预设所有资源最终转为某一种托管制度。

更一般的共同原则是：资源事实、法律权利、可授权条件、生效使用关系与现实协同方案必须分开。

权利/关系 当前含义 协同时需要额外回答的问题

所有/控制 由适用法律、合同、组织规则等确定 是否自愿共享、是否存在公共义务或限制

使用/占有 依当前有效权利关系确定 使用范围、期限、排他性、撤回与返还条件

收益 依产权、合同、公共规则等确定 外部性、公共成本、惠益分享是否需要另行处理

处分 依法律与权利束确定 是否触及生态、文化、代际或其他不可逆公共影响

资源稀缺、生态脆弱或跨境外部性可以增强维护、限制、惠益分享与公共协商义务，但不能仅由理论推导出既有所有权“失去合法性”。

“托管”可以是某些资源、信托、公地、公共设施或跨代资产采用的一种制度安排；它不是高于所有权的普遍新权利，也不自动产生调用权。

共同原则 允许的制度多样性 不得偷换为

现有所有/控制关系 可继续存在并依规则调整 “所有权一律转为托管权”

自愿或依法形成的使用权 可按合同、公共规则、共同体制度等形成 “资源可见即可调用”

惠益分享/公共义务 可在特定资源和法域中建立 “收益权一律变贡献分配权”

维护与外部性责任 可独立于所有权增加 “责任存在即产生征用或治理权”

以国际空间站为例：

它的水循环系统和氧气生成系统是全体宇航员共享的“生命公地”。

没有任何一个国家的宇航员可以宣称对氧气拥有私有权，
并拒绝向他人供应。

极端封闭环境中的生命维持资源可能适用非常特殊的安全与共享规则；这些规则来自该环境的事实、任务契约与法律治理，不能直接把空间站制度外推为整个地球统一产权法。

地球上的资源类型、法域、主体与权利关系远比单一任务舱复杂。

因此，更可靠的方向不是复制一种产权制度，而是建立能够跨不同制度识别事实、权利、授权、使用关系、责任与外部性的协同接口。

这既不要求私有化一切，

也不要求托管化一切。

它要求任何共享和调用都具有现实合法来源，并能在任务结束、授权撤回或条件失效后返还和关闭。

现实中已经存在多种资源治理与跨境协作安排，可用于比较；它们不共同指向一种单一“新范式”。

资源类型 旧范式 新范式

数据 多元权利与责任关系 依个人、组织、公共利益和法域分别处理

知识 付费墙 开放获取+公共资助

频谱 国家/区域规则与国际协调 依现有制度和跨境协商配置

卫星轨道 先到先得 全球公平分配谈判

遗传资源 国家主权 惠益分享

2022 年，联合国《生物多样性公约》第十五次缔约方大会，

通过“昆明-蒙特利尔全球生物多样性框架”。

其中最关键的一条：

遗传资源数字序列信息的利用，必须与来源国公平分享惠益。

这不需要被解释为去主权化。

更准确地说，它说明跨境资源利用可以在既有主权与国际规则下增加惠益分享机制；具体权利结构不能由本书统一概括为“主权升级”。

全球数字治理持续关注互联网基础设施、开放接入与国际协作；具体制度主张仍须回到各自真实文件、法域与生效条件。

对根服务器、IP 地址、域名系统等关键基础设施，

不能在本理论中直接统一定义其产权为“全球数字公地”；应区分技术公共性、治理协作与法律权利状态。

不同基础设施存在不同所有、运营与治理关系，不能用一个所有权判断覆盖。

可以确认的是，它们具有重要跨境公共影响，需要稳定、可接入和可治理的协作。

这种协作既不自动削弱主权，也不自动扩大主权。

更稳健的方向，是在国家责任、主体权利和跨境互操作之间形成可验证的制度接口。

资源协同的法理框架不能预设所有国家和资源采用同一种产权制度，

而应提供一套能够在不同现实权利结构中区分资源事实、权利关系、授权与协同方案的共同原则。

9.4 项目制协同：围绕具体问题临时组织跨主体能力

随着跨主体协作需求增加和资源权利关系更加复杂，

项目制协同可以成为传统行政、市场和组织结构之外的重要补充，但不预言所有社会组织都会被项目制替代。

行政区划、国家与地方治理仍会长期存在；与此同时，

一些跨区域问题可以围绕明确任务目标临时聚合主体、专业能力和资源。

这种治理模式分为两个层级：

层级一：跨国/全球性项目（条件性类型，不构成统一“全域”治理层）

可能用于承接跨境、长期或高公共性问题，例如：

- 行星防御系统建设
- 气候减缓、适应与相关国际合作；
- 未来若聚变等先进能源成熟后的跨境基础设施合作；
- 高能力 AI 与前沿计算安全协作
- 大流行病早期预警网络

决策机制：

全域级项目不得由“全域贡献值加权投票”自动形成公共决定；应依据参与主体、法域、事项性质和正当程序确定决策结构。

决策权既不能被个别政治或资本主体垄断，

也不能因为专家或贡献者拥有高贡献记录就自动转移给他们；专业者提供事实、标准与风险判断，有权主体依正当程序形成决定。

这要求专业判断充分进入公共决策，但专业能力不等于治理主权——

因此既不是技术官僚统治，也不是“贡献证明制”。

执行方式：

跨主体资源协同必须以自愿合作、既有权利、生效使用权和具体执行授权为前提，不存在一个默认拥有全球资源的统一调度主权。

AI 可以在已知资源事实与约束下生成所需物资和能力清单——钢材、芯片、算力、人员等——但物流执行必须取得相关资源、法域、组织和执行主体的合法授权；跨境协作也不能绕过现实主权、权利与责任。

这不是架空国家，

是国家通过参与全球项目，证明自己对文明的贡献。

层级二：地方/专业领域项目（在现有法域和组织内形成）

旨在提升特定域或群的生活量文化繁：

- 社区花园建设
- 濒危语言数字化保护
- 城市艺术节举办
- 本土物种保育
- 原住民传统知识传承

决策机制：

子域项目的决策资格与程序不得由贡献值自动生成；应依据适用规则识别有权主体、直接受影响者、专业责任与合法授权。

程序设计应充分纳入直接利益相关者及其受影响程度，同时保持基本权利与正当程序独立。

长期居住、服务经历可以形成重要的本地知识与利害关系事实，但具体决策权仍必须来自相应制度、程序与授权。

即使某主体拥有很高的全域贡献记录，

若与具体事项缺乏直接利害关系、法定资格或明确授权，

其贡献记录也不能自动转化为该议题的投票权、治理权或资源调用权。

这体现了埃莉诺·奥斯特罗姆在公共池塘资源治理研究中提出的“多中心治理”智慧：

层级 谁决策 凭什么

全域 依法/依规取得相应资格的相关主体 全局事实+专业判断+责任边界

子域 直接利益相关者与合法授权主体 本地知识+利害关系+程序正当性

让最了解局部事实的人充分进入判断，让直接受影响者获得与其权利和程序相称的参与；

让具备全局视野的专业与治理主体提供系统分析，但知识、能力与贡献都不自动生成主权，最终效力必须回到正当制度、授权与责任链。

项目制协同并不把社会变成一台“液态超级计算机”，

更不能把人和资源视为可以随项目任意流动的“算力”；跨域调用仍受主体权利、产权、使用权、法域、专业门槛与授权约束。

这一模式已有早期雏形：

欧洲核子研究中心 CERN。

CERN 由成员国通过国际条约建立，是长期跨国科研协作的一项现实成果。^[27]

它没有领土，但有全球最顶尖的物理学家人群。

它没有军队，但发现了希格斯玻色子。

它不拥有国家主权，其合法性与治理来自成员国条约、组织规则和专业体系。

全球疫苗免疫联盟 Gavi。

由世卫组织、世界银行、比尔及梅琳达盖茨基金会、各国政府、制药公司共同组成。

Gavi 把国家、国际组织、基金会、企业与卫生体系连接进长期全球健康合作。^[28]

目标是明确的：让低收入国家的孩子也能打上疫苗。

资金是拼盘的，决策是多边的，执行是外包的。

其健康成效以官方评估口径、明确截止时间和原始资料核验。这里真正值得观察的是跨主体目标协作如何形成，不用未经可靠核验的累计数字替理论增加声势。

人类基因组计划——开放科学项目。

没有中央指挥部，没有主权授权，

只有一群科学家自愿共享数据、不申请专利、不接受私有化。

2003 年完成时，美国总统和英国首相共同宣布“这是全人类的遗产”。

它是跨国科研合作的重要成果，也像是这个物种第一次把自己的生命文字完整摊在桌面上。^[29]

这些案例支持一个有限判断：

跨国合作不需要一个世界政府式统一主权；

不同对象可以形成多边规则与共同治理，但“全球公地”必须按具体法律对象定义；

条约、合同、组织规则、声誉、互惠与国内法执行都可能支持规则遵守；不存在完全无强制的统一模式。

项目能否持续还取决于法权、资金、专业能力、利益结构、问责与实际结果；共同意义和贡献承认只是其中部分因素。

9.5 合法性来源与参与相关性：谁应被听见，谁有权决定

跨主体项目面临一个根本问题：

谁赋予这些项目合法性？

不是选票——因为没有全球选举。

不是宪法——因为没有世界宪法。

不是暴力——因为没有全球军队。

直接痛苦经验可以增加理解与证言重量，但“痛过”本身不自动生成普遍决策权。

这里能提取的是经历事实，而不是把痛苦折算为政治贡献。

利益相关度提示谁可能承担后果，应强化其被听见、申诉和参与的保障。

日本江戸时代的“村请”制度：

用水依赖度较高的农户通常具有更强利害关系与本地知识，应被充分纳入规划；具体决策权仍取决于相应制度与程序。

依赖越深，受影响程度越高，越需要获得充分陈述和程序性参与，而不是自动取得最终主权。

这里的历史义务与投入是制度事实，不应直接等同于现代贡献值。

收成依赖度可以说明利害关系，但不单独决定权力。

安第斯山区的“艾尼”互助传统：

互助传统可以形成明确、可理解的互惠请求与责任关系；这种约定性互惠不同于一般公共权力。

这里的劳动贡献可以生成约定范围内的互惠事实。

现实需要与既有互惠关系共同构成协作条件，但仍不产生一般治理主权。

相关制度的合法性设计，核心不是把贡献数字化后替代同意与法权，

而是让同意、权利、利益相关、专业判断、贡献事实与责任在各自位置上可见、可追溯、可纠错；

公平直觉可以提供方向，但具有约束力的公共决定仍必须经过正当制度与明确授权。

9.6 AI 作为验收分析工具：事实比对与有权主体责任分层

在跨主体项目中，

谁来判断项目是否成功？

谁来决定贡献值应该发放多少？

如果让项目发起人自己验收，会产生激励扭曲——

夸大成果，隐瞒失败，骗取资源。

受影响主体的反馈不能替代专业验收，但也不能因“可能短视”被排除；不同验收事项应区分专业事实、受影响者权利与公共决定。

拒绝长期必要短期痛苦的项目。

更稳健的方案是：

AI 可以作为验收分析工具，人类及其他依法/依规有权主体设定目标、规则和验收效力来源。

这不是把权力交给算法。

AI 负责材料整理、指标比对、异常识别与可解释分析；具有约束力的验收、责任认定和资源后果仍由相应主体依程序形成。

验收流程：

步骤 内容 责任主体

- 1 设立项目目标 有权主体依适用制度、项目章程或明确授权形成
- 2 将目标转化为可测量指标 专业人员判断+AI 辅助；指标不替代目标
- 3 采集过程与结果数据 依法授权的数据源、传感器与责任主体
- 4 保存证据与版本 可审计存证机制；不预设单一技术
- 5 比对目标、事实与结果 AI/分析工具输出差异、置信度与未知
- 6 形成验收结论与处理争议 有权验收主体+独立复核/申诉机制

验收原则：

多维目标函数，而非单一指标。

- 不只是“发电量”，还有“碳排放”“社区影响”“代际公平”
- 不只是“成本节约”，还有“工人权益”“生态修复”“技术外溢”

动态条件，而非由算法自行重写价值标准。

- 干旱地区节水项目权重高
- 洪涝地区防洪项目权重高
- 高风险项目可以依事先正当形成的规则提高审慎程度或触发复审，但“紧急权重”不由 AI 自行产生。

可解释性，而非黑箱。

- AI 必须说明关键事实、指标、推断与不确定性，而不是把模型解释当作最终治理理由。
- “因为 X，所以 Y，依据是 Z”

欧盟《人工智能法案》已经把高风险 AI 的风险管理、技术文档、日志、人工监督与符合性评估等要求写入法律框架。

具体适用时间和评估路径依系统类型与法案分阶段实施安排而不同。

有些高风险系统可依内部控制程序评估，部分与受监管产品相关的系统可能涉及第三方机构。

相关合规要求来自适用法律、标准与责任体系，而不是由模型自行决定。

这些现实实践说明自动化测试可以服务合规和验收，但不证明 AI 本身应成为“验收官”。

新加坡推出了 AI Verify 测试框架：

相关工具可用于对部分 AI 治理与技术属性进行测试，

但测试本身不产生一项自动的法律“可信认证”，也不把责任转交给消费者。

它更适合作为自动化测试服务治理与验收的一种现实参照。

高后果制度与技术原型的验收机制，

将在这些萌芽的基础上，

增加三个关键升级：

1. 风险相称：重大公共项目应依据适用规则接受与其风险相称的验收与复核
2. 独立性：识别并管理验收方利益冲突，必要时引入独立评估
3. 可追溯性：按法律、专业与风险需要保存记录；保存期限和责任延续不能由本书统一规定为永久或终身

AI 可以辅助事实整理、测试与分析，但当前具有法律或公共效力的决定必须来自现实有权主体和程序。

司法、监管、专业验收等具体角色依现实法制形成；本书不以“必须由某个自然人亲自完成全部判断”作为跨法域规则。

AI 也不天然具有“审计员”身份；它只是可被授权使用的分析、测试或执行工具。

任何分析工具都可能出错；真正需要的是证据来源、方法、独立性和可复核性。

9.7 现实责任锚点：能力不能替代授权

在当前及未来不同制度场景中，AI 可以在明确授权下承担评估、匹配、模拟、优化与预警等职能；具体能力与适用范围随技术和法域变化。

但能力增加，不会自动生成决定权。

为什么？

因为一项具有约束力的决定，不只需要“算得更准”，还需要回答：目的由谁确定，权利如何保护，谁承担后果，错误如何纠正。

AI 可以估算风险，

却不能仅凭估算自行决定谁必须承受风险。

AI 可以比较方案，

却不能因为比较能力更强，就取得公共目的的决定权。

AI 可以在明确授权范围内执行，

却不能把授权范围改写成自己的主权。

真正需要保留的，不是一种“碳基本体优越”的证明。

需要保留的是现实制度中的权利、授权与责任：当前具有权利和责任的主体，不能因为系统更聪明，就被自己的工具越过。

未来其他智能形态是否形成主体性，是另一个需要事实、哲学与伦理共同面对的问题。

它不改变今天的制度责任必须有正当来源。

所以，高风险、不可逆或重大权利影响的行动，必须存在由有权主体建立的复核、暂停、否决、终止与责任机制。

按钮可以是物理的，也可以是制度的。

~~关键在人类手里有一个红色按钮。~~

关键在：系统永远不能自己决定谁有权关掉它。

这条边界并不要求所有法域采用同一种程序。

国家、机构、社区与专业领域会有不同的授权结构与责任关系。

共同原则只有一个：能力不生成权力。

知道得更多，不等于有权做得更多。

当 AI 真正成为强大的文明工具，

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

而是怎样在使用更强能力时，仍然承担自己的决定。

9.8 幸福与意义：系统只能保护条件

在宏大的“行星工程”与“全球共识”之下，一个普通人如果只想经营自己的生活，他的存在还有意义吗？

答案不能由他是否参与宏大项目决定。

合法性如果只剩规则，也会变空。

一个制度可以程序完备，却让人长期生活在恐惧、羞辱、孤独和无力感里。

文明三目标提醒我们：制度最终要回到人的生活。

~~但幸福不能因此成为系统直接对象。~~

系统可以改善住房、医疗、教育、休息、安全、关系与参与的现实条件；

它不能替每个人定义什么是幸福，

更不能为了提高“幸福指数”而操纵情绪。

意义同样不能被当作系统燃料。

文明能够做的，是保护意义生成的空间：让人有选择、有关系、有承担、有创造，也有不参与和暂时停下的权利。

自我决定理论长期强调自主、胜任与关系三类基本心理需要。

它至少提醒我们：幸福不是一份统一的资源套餐。

一个普通人经营自己的家庭，照顾阳台上的花草，与邻里和睦相处；

一个老人有人陪着吃饭，一个孩子不再害怕回家——

这些不是文明宏大叙事之外的背景噪音。

宏大项目连这些生活都守不住，就失去了落点。

全职父母、社区合唱团组织者、守夜人、方言传承者、安宁疗护志愿者，

并不因为没有耀眼价格就没有价值。

这些生活也不需要被统一命名为“幸福维持项目”。

更不需要证明“AI 永远不到”才成立。

它们重要，是因为真实生命正在其中相互抵达。

好的制度不是让所有人看起来更开心。

是让更多人拥有真实生活自己的可能。

9.9 高能力系统的合法性：复杂不能代替正当性

高能力系统给合法性提出的真正挑战，不是“看不懂，所以只能相信”。

即使未来系统采用量子或混合计算，内部状态更复杂、传统解释更困难，也只意味着验证方法要变化，不意味着合法性来源发生了转移。

技术可以让一个结论更有证据。

技术不能让一个结论自动拥有支配别人的资格。

我们可以无法逐步复原模型内部的每一次计算，

却仍然必须知道：为什么这项决定可以对具体的人发生效力，谁授权，谁能复核，错了谁负责。

量子不可克隆定理也不能替制度回答这些问题。

它不证明现实系统不存在影子副本，不证明软件没有后门，更不生成“唯一系统”的政治合法性。

物理性质可以成为安全设计的一部分，不能成为权力来源。

能力越强，越要分清能力与权力。

知道得更多，不等于有权做得更多。

文明越有能力看到整体，越需要克制把“整体视角”变成天然主权的冲动。

透明也不是唯一答案。

有些系统不可能把每一步内部状态变成人人可读的解释。

但结果仍应能够被独立验证，关键假设仍应能够被质疑，重大效力仍应能够被复核。

可审计、可停止、可接管、可替换，

不是因为技术天然善良，
而是因为任何高能力系统都可能出错。
文明无需先证明机器永远不可能越界。
文明需要确保：即使它会越界，也不能把越界本身变成新的合法性。
无论技术怎样变化，
文明不能把自己的成年交给技术证明。

9.10 本章结论：跨境协同必须继续接受合法性约束

第九章走到这，题经再是：拥有更多力量。
而是：当力量越来越大，我们有没有学会不让力量成为一切。

文明的成年，不是废除国家，不是取消财产，不是消灭利润，也不是把决定交给 AI。
它发生在更深的地方。

拥有，不等于可以任意支配。
胜利，不等于必须别人失败。
能力，不等于天然有权决定。
贡献可以被记住，却不能成为权力的出生证明。

成年不是抹掉差异，而是学会与差异共同生活。

国家仍然存在，利益仍然不同，竞争也不会因为文明跃迁自动消失。
真正改变的是：我们是否还把共同未来当成竞争结束后的战利品。

即使不同，也为共同存续留下底线。
即使竞争，也不把彼此推到没有退路。
即使拥有更强的技术，也知道有些事不能只因为“做得到”就去做。

少年文明习惯用占有理解强大：更多领土，更多资源，更多资本，更多控制。
成熟的文明开始明白：真正的强大，是让更多生命因此活得更好。

所以，一个人的价值不需要由地位证明。

让邻居笑一笑，让一棵树活下去，让一个孩子不再害怕黑夜——
这些事不会因为没有价格，就没有价值。

人活着，本就有意义。

爱、照护、创造、承担与守护，让意义继续传递，也让生活因此不同。

尊严不靠言谈得。

贡献值得被看见，因为它让共同生活真实发生。

权力若要被托付，就必须回答：凭什么，又由谁承担后果。

技术若要被信任，就不能只证明自己更强，还必须允许文明在错误时说“不”。

文明的成年，不是终于没有冲突。

是终于知道：不能为了赢，把共同未来一起输掉。

我们不需要一个没有差异的世界。

我们需要一个即使彼此不同，也仍然愿意共同活下去的世界。

而共同生活，终究需要托付。

托付，需要信任。

可当视频可以伪造，声音可以伪造，证据也可以伪造——

我们凭什么相信，眼前发生的事真的发生过？

没有真实，信任无处落脚。

没有信任，再好的制度也只是悬在半空中的建筑。

下一章，我们进入文明最古老、也正在变得全新的问题：

信任的锚点——当“眼见”不再“为实”。

第十章 信任的锚点：当“眼见”不再“为实”

“拟像掩盖真相，它是掩盖并不存在的事实之物。拟像是真实。”

——据让·鲍德里亚《拟像与仿真》相关论述意译

当合成与复制越来越容易，事实的来路会越来越重要；但真相不能被任何认证系统承包。

未来情景片段：《全球信息溯源协议》（作者构造，非现实文件）

当一张图像、一段声音、一次身份出现都可以被低成本合成，真正稀缺的就不再只是内容本身，而是它从哪里来、谁愿意为它负责。

——本书

当一张照片也开始需要证明自己，文明失去的不是一项方便，而是一种几乎没有被意识到的奢侈：我们曾经可以在很多日常时刻，低成本地相信眼前世界。

10.1 当伪造变得廉价：认识论的压力

人类长期形成“眼见为实”的直觉，并不是因为图像天然等于真相。

只是很长一段时间里，造出足够逼真的照片、声音和视频，通常要付出更高的技术、时间和组

生成式 AI 正在改变这条成本曲线。

逼真的图像、语音、视频和文本越来越容易出现。具体成本模型、质量和算力变化，没有必要写一张冗长的下降曲线表。

真正改变的是一种不对称：制造一个看起来可信的东西，可能比证明它到底从哪里来更容易。

于是，“像真的”开始失去过去的重量。

2023 年 5 月，一张伪造的“五角大楼附近爆炸”图片在社交媒体扩散，金融市场盘中出现短暂波动，随后随着澄清而恢复。^{〔30〕}

值得记住的不是“一张假图控制了市场”。

而是另一件更普通、也更危险的事：一份低成本合成内容，已经可能先触发人的反应，再等事实慢慢赶来。

诈骗、舆论操纵、冲突叙事和身份冒用，都在把同一个问题推到我们面前：验证正在从少数专业场景，进入普通生活。

真相没有因此消失。

只是它不再总能靠外观先赢。

以后越来越重要的问题，不只是“它像不像真的”，而是：

谁先发出来？

它从哪里来？

还有谁能独立证明同一件事？

当伪造变得廉价，文明不能靠更大的嗓门保卫真实。

它要做的是让怀疑有路可走，让核验有工具，让错误有地方可以被纠正。

过去，照片、录音和影像常常同时承担记录、见证和传播。技术把“我看见了”延伸成“你也可以看见”。

这信从来不完。照片可以摆拍，剪辑可以误导，文字更容易撒谎。只是很多时候，伪造还没有今天这样便宜、快速、可批量复制。

生成式 AI 真正增加的，是规模和速度：一个人可以很快制造过去需要更多人力才能完成的内容，而且可以不、景。

AI 没有发明谎言。

它让谎言第一次拥有了接近工业化的生产能力，而许多人的核验习惯还停留在手工作坊。

第十章要处理的，就是这条速度差。

先从一个极普通的夜晚开始。

一位母亲接到儿子的电话。声音一样，语气一样，连家里的称呼都知道。电话那头出了事故，要马上转账。

过去，“我听得出来他的声音”可能已经足够。

现在，真正保护这段关系的，反而可能是一句看似多余的话：先挂断，我回拨给你。

技术没有让母亲从此不再相信儿子。

它只是让重要的相信，多走一步。

一步小，却改多高定：

越能改变财产、自由、身体和生命的决定，越需要身份、意图和授权彼此印证。

这不是把关系变冷。

是防止一个陌生人借走你们多年建立的信任。

10.2 深伪大流行：从恶作剧到现实伤害

深伪真正危险的地方，不只是让我们看见一件从未发生的事。

2024 年，香港警方披露一起高额诈骗案：一名职员参加视频会议，看到和听到的多名“同事与高管”后来被认定为伪造形象，受害者随后按指示转出巨额资金。^[31]

这类案件把一个边界写得很清楚：熟悉的脸和声音，已经不适合单独承担高风险授权。

同年，美国新罕布什尔州初选前，部分选民收到模仿乔拜登声音的自动电话，试图劝阻他们参加初选，随后引发执法与监管行动。^[32]

生成式伪造于是从娱乐和恶作剧，进入了诈骗、政治传播和身份冒用。

这并不意味着“生物特征全部失效”或者“视频证据从此无用”。

人脸、声纹、影像和熟人确认仍然可以提供证据。

只是越高风险，越不能让其中任何一项独自成为最终依据。

“ ”

当所有人都知道内容可以伪造，真的证据也更容易被一句“那是 AI 做的”推开。

假的更容易冒充真的，真的也更容易被说成假的。

这不会让文明一夜之间失去信任。

它会先让合作变贵。

信任并没有退场。

它只是不能再把全部重量压在单一感官上。

一要求的熟悉音，不因 “ 像老板”就取得付款；一段播快的，也不直接触发军事、司法或医疗决定。

回拨、双人复核、独立通道确认、延迟执行——这些办法并不浪漫，却能把一次伪造挡在多年关系之外。

真正需要升级的，不是人类的怀疑程度。

是重要决定的核验结构。

最危险的不是我们开始怀疑。

是到了最后，假都只能情站。

10.3 检测的边界：为什么不能只靠识别算法

面对生成式伪造，一个自然反应是造出更强的检测器。

检测当然有价值。取证分析、模型特征、元数据和异常模式，都能帮助判断。

但没有一台机器能永久替我们承担“真假”这件事。

生成器会变，检测器也会变；今天有效的特征，明天可能消失。

如果两类内容在可观察特征上越来越接近，只靠内容本身作稳定判断就会越来越困难。

这只是一个逻辑边界，不是对未来模型能力的预言。

检测还会撞上另一个现实：误报。

在海量内容环境中，只要误报率不是零，看似很高的总体准确率也可能误伤大量真实作品。

一个记者的照片、一名学生的论文、一部纪录片，都不应该因为一个分数就自动失去表达、学习或参赛资格。

检测器可以举手说“这里值得再看”。

它不能一开口就变成裁判。

所以问题不是放弃内容分析。

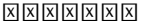
因为当像素越来越难自己说明身世，来路就成为额外的一层证据。

设备或软件可以在创建、编辑和发布过程中写入并签署相关声明。

这些声明可以涉及创建工具、编辑动作、生成式 AI 使用等，具体取决于实现和隐私设置。

但边界必须非常清楚：

C2PA 能验证的是一组来源声明与内容之间的关联、签名和篡改状态；它并不替世界判断这些声明描述的现实事实一定为真。^[33]

1. 

2. 

3. 一张拥有完整凭证的照片仍然可能被放进错误语境；一张没有凭证的老照片，也不会因此自动成为谎言。

所以内容凭证更像来路证明，不是真理印章。

它可以告诉我们：谁愿意站在这份内容后面。

却不能替我们回答：世界究竟发生了什么。

成熟的数字核验，需要把来源凭证与新闻核验、数字取证、事实调查、证人证据和公开纠错放在一起。

出生证明可以告诉我们一个人从哪里来。

它不会替一个人活完一生。

容也一。

这个比喻够用了，不必再让技术承担更多。

创建时留下签名，可以说明某个设备、软件或机构对一组来源声明负责；编辑时留下记录，可以帮助后来的人知道内容经历过什么；展示时把关键信息呈现出来，可以降低核验成本。

但不是每一次裁剪都值得永久记录，不是每个地点都应该暴露，也不是每个创作者都愿意公开真实身份。

如果为了防伪把每个人的隐私一起交出去，信任基础设施本身就会变成新的伤害。

好的溯源不是“记录一切”。

是让必要的来路可验证，让不必要的个人信息仍然属于本人。

10.5 溯源的社会契约：来源不是内容的裁判

技术只有进入制度，才会真正改变公共信任。

但制度也最容易在这里越界：从“有来源更容易核验”，滑向“没有来源就没有资格说话”。

高风险领域当然可以要求更强的来源、身份与责任条件。

政务发布、金融指令、医疗记录、司法证据和关键基础设施控制信息，会直接改变权利、金钱、生命与公共权力，因此可以在明确法律和规则下要求更强的签名、留痕和复核。

理由不是这些机构更高贵。

是它们一旦错，别人承担的后果更重。

日常表达不是同一种场景。匿名写作、化名创作、旧照片、弱势者求助、举报材料，不该因为缺少某一种新技术凭证就自动失去被听见的资格。

没有来源凭证，只说明少了一层来源证据。

不能直接跳到“已证明为假”。

有来源凭证，也只说明责任更容易被追问。

不能直接跳到“已证明为真”。

这条边界保护的不只是表达自由。

它也保护真相。

因为一旦“认证身份”变成真理门票，事实权就会慢慢集中到少数能够签名的人手里。

文明需要的不是一堵认知高墙。

而是让不同风险场景承担不同强度的核验，同时保留质疑、补证、匿名和纠错的门。

来源降低核验成本。

证据决定我们该相信到什么程度。

□ □ □ □ □ □ □ □ □ □

人类很早就知道，重要信息需要有人站出来承担。公证、签章、档案、同行评议、编辑责任、审计签字，都在用不同制度回答同一个问题：这句话出了错，往哪里追？

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

它只是把蜡封、印章和手写签名，部分换成数字签名、来源记录和新的机构责任。

技术可以更新，几个问题一直没变：

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

真正的信任，不是没有错误。是错误发生以后，仍然找得到回事实的路。

10.6 责任担保人：真实最终要回到可追责主体

来源链能回答“谁说过、从哪里来”。

它仍然不能独自回答“说得对不对”。

所以公共信任最终还要落到能够承担责任的主体。

记者、研究机构、医院、银行、政府、平台、专业人士，都可能在各自场景中承担这种责任。

他们值得信任，不是因为身份神圣。

而是因为他们愿意把判断放进可以核验、可以纠错、可以追责的制度里。

真正的担保也不是“错一次就毁掉一生”。

XXXXXXXXXX

一个健康制度必须区分诚实错误、专业疏忽、恶意伪造和系统性造假，再按现实法域与后果配置不同责任。

责任可以来自职业规范、合同、民事、行政、刑事或公共问责；没有一张统一责任表能够替所有场景决定。

XXXXXXXXXX

举报、敏感调查和高风险表达中，身份可以由编辑部、律师或其他合法中介保护，同时让核验责任仍然有落脚处。

未来真正重要的记者，不只是“知道得多”。

他要能告者：

XXXXXXXXXX

我核过什么。

如果错了，我在哪里更正。

信任就在这些一次次可核验、可纠错的记录里慢慢长出来。

技术可以证明签名。

制度保证签名之后仍然有人。

因此，责任不是用来把所有人吓到沉默。

一个只靠“错一次就毁掉一生”维持的体系，会让人更想掩盖错误，而不是面对错误。

可靠的制度既要让恶意造假付出代价，也要让诚实的人能够承认、纠正、修复并重新获得可信。

纠错不是纵。

是把“相”人格神拿下，放回据和程序。

人不是因为永不犯错才值得信任。

而是因为犯错以后，仍愿意回来面对事实。

信任更像一座桥，而不是一枚印章。

XXXXXXXXXX

另一端是据：我知道为什么这样说。

桥面是责任：如果错了，有人愿意回来修。

任何技术都只能加固其中一部分。

数字签名加固身份与完整性；来源记录加固来路；审计加固过程；新闻机构、法院、科研共同体和公共档案，再以不同制度承担核验。

真正危险的，是把其中一个部件抬成整座桥。

一旦“有名=是的”“有=可信”“模型说=造假”成为默认反射，我们只是把旧时代的轻信换成了更高科技的外衣。

文明需要技术。

也需要在技术面前保留自己的判断。

10.7 蓝 V 的教训：标志不能替代身份与责任

2022 年，Twitter 新的付费蓝标方案上线后，很快出现大量冒充账号。^[34]

其中最著名的一例，是一个冒充礼来公司的账号发布“胰岛素现在免费”的虚假消息，真实公司随后公开澄清。

这个案例不需要被写成“一条假消息导致全部股价波动”，也不证明“付费认证必然失败”。

它暴露的是更简单的错位：

公众以为一个标志意味着“身份已核验”，而获得这个标志的条件却已经发生变化。

骗子最喜欢的，不是没有门。

是上着“可信”，却已了。

所以数字身份至少要分开三件事：

账号是谁。

平台给了它什么标记。

它这一句话到底是不是真的。

三者可以互相提供线索，不能互相代替。

真实身份会说错话；真实机构会发布错误；没有显著标志的普通人，也可能拿着最关键的证据。

信任不能被压成一个图标。

它更像时间里慢慢叠起来的东西：身份可核、行为可看、错误可改、责任可追、证据能被别人重新检查。

□ □ □ □ □ □ □

可以更大的像、更长的子、更快客服。

它不能买到“这句话更接近真相”的资格。

信任真正昂贵的，从来不是八美元。

1. 是一主愿不愿意在里自己的。

10.8 星椋鸟实验室：让证据多一层来路

如果蓝标事件提醒我们“标志不能代替证据”，Starling Lab 等实践则展示了另一条路：尽量让数字证据留下更完整的来路和保管痕迹。^[35]

Starling Lab 是一个由数字取证专家组成的团队，他们致力于在数字世界中建立可信的证据链。

这些实践的价值，不在于写链就是真相。

而在于让一份材料多回答几个过去很容易丢失的问题：

什么时候采集？

由什么设备或流程产生？

后来改过什么？

今天看到的版本，和当时保存的版本还是不是同一份？

哈希可以帮助发现比特层面的变化，数字签名可以说明某个密钥对声明负责，分布式保存可以降低单点丢失的风险。

但这些工具仍然看不到镜头外发生的一切。

未经修改的照片也可能配错说明；真实时间戳不能解释完整因果；完整来源链也不能替代证人、现场调查、法律程序和反方质证。

所以数学最适合承担它真正擅长的部分。

它可以帮助证明：这份材料有没有被改。

它不能替法庭证明：世界究竟发生了什么。

这不会削弱密码学。

但密码学可以做得更多。

让伪造来路更难，让删除历史更难，让后来者更容易追问。

文明的记忆不需要一台永不犯错的机器。

它需要一条不容易被悄悄抹掉的证据链。

而且，证据链越重要，越不能只让少数技术机构看得懂。

普通人面对争议影像时，不该先读懂密码学论文，才有资格知道它来自哪里。

真正成熟的基础设施，会把复杂性留在后面，把关键问题留在前面：

谁发布？

来自什么原始材料？

改过什么？

还有谁独立看见？

如果结论错了，在哪里更正？

技术最成熟的时候，不是公众开始膜拜它的复杂。

是普通人仍然知道自己什相信。

当证据只有机器能读，人就只剩下接受机器结论。

而这一章真正要守住的，正是人的判断位置。

而文明跃迁真正要保护的，恰恰是人的判断位置。

10.9 认知免疫系统：怀疑作为一种美德

技术和制度可以建立外部防线，最后还是要回到人的判断。

过去，“有图有真相”之所以好用，不是因为影像不会骗人，而是多数日常交流没有必要把每一次看见都当成刑事取证。

生成式 AI 改变的是：在一些高风险场景里，过去够用的快捷方式开始不够。

于是需要的不是“逢图必疑”，而是按后果调整核验强度。

第一，是源反射。

当一段内容要求我们立刻转账、立刻愤怒、立刻投票、立刻惩罚某个人时，先别急着接受它的速度。先问：它从哪里来？

有来源凭证，可以看；没有，也不能自动判假。还要看原始发布者、上下文和其他证据。

第二，是交叉。

重大事实不该只靠一个账号、一张截图、一段剪辑完成。

不同机构、不同角度、原始材料和现场记录能不能彼此印证，比“谁声音最大”更重要。

第三，是情延。

情绪不是敌人。

它只是最容易被借走的快捷入口。

认知免疫也不是训练所有人变成冷冰冰的怀疑机器。

一什都不信的社，和一什都信的社一无法合作。

真正成熟的怀疑，是愿意给证据时间，也愿意在证据改变时改变自己。

它更像一条守门的狗。

不需要咬每一个来客。

只需要在门外声音再熟悉的时候，仍然记得抬头看看是谁。

但还有一个同样重要的反面：过度核验也会伤人。

如果每一句日常表达都必须实名，每一张照片都必须上传完整位置，每一次匿名举报都因为缺少身份凭证而被降级，防伪很快就会变成控制。

真实需要证据。

人也需要隐私、匿名、试错和安全说话的空间。

两者并不天然冲突。

真正困难的是：不同风险，承担不同证明重量。

越能改变别人命运的决定，越应该留下更强证据。

越接近日常生活和自由表达，越不能把证明责任无限前置。

⊠ ⊠

也不能为了维护自由，假装高风险谎言没有现实后果。

成熟不是站到两个极端中的一个。

成熟是让核验的重量和后果相称。

10.10 后量子迁移：安全工具不能替代信任制度

今天大量数字签名、密钥交换与身份系统仍依赖公钥密码。

Shor ⊠ RSA 和部分椭圆曲线密码体系会面临根本威胁。

所以迁移不能等到威胁真正到门口才开始。

但这也不是一只可以写上精确年份的文明倒计时。

量子架构、纠错开销、算法实现和攻击目标差异巨大，技术路线也会变化。

更稳健的做法，是先知道自己到底用了什么密码。

然后建立密码敏捷性，按风险和系统生命周期逐步迁移。

NIST ⊠ 2024 年批准并发布了首批三项后量子密码标准：FIPS 203 用于密钥封装，FIPS 204 ⊠ 205 用于数字签名。^[36]

它们提供新的工具。

⊠ ⊠

量子密钥分发可以在特定场景帮助发现窃听，量子随机数可以提供高质量随机源。

这些技术各有部署条件、成本和安全假设。

最重要的一句其实很简单：安全信道不等于真实内容。

最安全通道里传来的谎言，仍然是谎言。

后量子签名保护的越权指令，仍然是越权。

无法被伪造的身份，也不会因此拥有本来没有的权力。

所以“量子就绪”首先是一种系统能力：能盘点、能迁移、能换密钥、能撤销失效凭证，还保留多重验证和审计路径。

物理学可以帮助守住信道。

它不会替文明决定谁值得相信。

信任最后还是要回到权利、证据、授权、责任和可纠错。

密码从来不单独存在。证书、设备、软件更新、供应链、密钥保管、撤销机制和人员权限共同组成安全。

只换算法名字，却不处理越权、失窃密钥、错误配置和组织失责，新的密码只会继续保护旧的漏洞。

量子技术也不该成为新的神话。

“不可克隆定理是量子态的完美制，不会让文件、权利和历史自动获得神圣性。”

物理规律可以给某些安全机制增加约束。

它不会回答谁有权写入历史、谁有权撤销证书、谁有权宣布战争、谁有权原谅错误。

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

技术越强，越需要让权利有来源、决定有授权、错误有出口、责任有归处。

真实的公共基础设施更像道路，而不是会所。

建设道路需要成本，但一个人不能因为贫穷就被判定没有资格抵达事实。

金融、司法、医疗、关键基础设施和公共治理应承担更高核验成本，因为它们的决定会直接改变别人的财产、自由、身体甚至生命。

普通人的日常表达不该被同一套重量压住。

这不是双重标准。

是任影相。

能力越大，核验义务越重。

影响越深，证据门槛越高。

这也把第十章重新接回全书的根基：

□ □ □ □ □ □ □ □

信息优势也不生成真理特权。

机构知道得更多，不等于永远正确；模型算得更快，也不等于可以决定谁说真话。

真实不是某个中心拥有的资源。

它是一种任何重大结论都应该允许重新核验的公共关系。

如果一个结论只能靠禁止质疑维持，它拥有的不是信任。

是服从。

如果一个社会只能靠人人互不相信才能防骗，它拥有的也不是清醒。

是瓦解。

真正值得追求的是第三种状态：

我们仍然能够相信彼此。

只是重要的相信，不再要求闭上眼睛。

社会也不是法庭的无限放大版。

法庭要决定责任，所以证据门槛很高；科研要建立可检验知识，所以方法和数据要接受同行检验；新闻要描述公共事实，所以需要多源核实。

种证明责任。

如果文明只剩下一套“可信度总分”，这些不同世界就会被压成一种生活。

一个人在法庭上要证明事实，不意味着他在诗里要证明隐喻。

匿名者暂时无法证明身份，也不意味着他的证据可以被直接丢弃。

机构拥有完整档案，也不意味着它的叙述天然高于现场的人。

所以真实基础设施最深的设计，不是给所有内容排统一名次。

是先：景什需要，了害。

如果只是朋友间的玩笑，社会可以宽容。

如果错误可能夺走一个人的自由、财产或生命，证据就必须更重。

如果错误可能让一个国家误判另一个国家，核验更不能被传播速度绑架。

真实不是越昂贵越好。

而是更重要，越值得投入足够的核验成本。

这一区别守住了两样东西：

一边是事实。

一边是自由。

文明如果只能二选一，就还没有学会在 AI

真实还必须允许旧世界留下来的证据进入新世界。老人从抽屉里拿出的纸质照片可能没有数字签名；灾区居民临时拍下的影像可能没有完整来源链；匿名举报者的材料甚至必须先隐藏身份。缺少新技术凭证，不等于这些材料没有价值。

未来的信任基础设施必须容纳这种不完美：旧档案可以重新核验，匿名材料可以经过受保护中介验证，普通人的证言可以和机构记录互相质证。技术越强，入口越不能只留给最懂协议、最有资源的人。

真相属于公共世界，不是因为人人拥有同样设备，而是因为任何人的重要证据都应该有机会进入核验，任何权威结论也都应该留下被追问的门。

成熟的信任，不是要求现实先符合协议。是让协议有能力面对现实的杂乱、冲突和不完整。

10.11 本章结论：真实有成本，但不能成为特权

第十章没有给“真相”找到一枚万能印章。

恰恰相反，它把万能印章拿掉了。

AI 时代真正上升的，是核验成本。

来源记录要工具，调查要时间，档案要维护，责任要有人承担，错误要有人更正。

这些都不是免费的。

但“核验有成本”不能变成“真实是奢侈品”。

如果真相只属于付得起验证成本的人，最先失声的往往正是最需要被听见的人。

所以真正需要的不是“免费世界”和“付费真相”。

是风险分层。

是风险分层。

发朋友圈的照片，不需要高级证据链。

要求银行转出巨款的指令，需要更强身份与授权复核。

可能~~能~~刑罚~~重~~或重大公共~~决~~的事~~实~~张~~需~~要更高强度、更多~~源~~和可~~证~~据。

场景越高风险，核验越不能省。

但核验针对的是决定和后果，不是人的等级。

普通人的证言不会因为没有昂贵凭证就自动失去资格；权威机构的声明也不会因为盖了章就免于质疑。

这才是数字时代需要守住的平等：

谁都可以说话。

谁的重大主张都可以被追问。

技术帮助看见来路、发现变化、保存证据。

制度决定谁有权行动、谁可以提出异议、错误怎样纠正。

最后仍然有一件事不能被外包：判断。

我们不再把“看见”当成真相。

也不把“名” 成相。

我们把真相重新放回证据之间。

这比一枚永不出错的印章麻烦得多。

也正因麻，自由留下了位置。

当影像可以被制造，文明不能因此放弃相信。

它要学会更有根据地相信。

真正的信任，不是让人少怀疑。是让怀疑有地方可以走到事实。

第十一章 分配的正义：当生存不再必须靠绩效证明

一个人走进援助窗口，把工资单、房租合同、银行流水、病历一张张摊在桌上。

他不是来申请荣誉。

他只是想让这个月不断药，让孩子下周还能去学校，让家里那盏灯不要在月底以前熄掉。

可很多制度会先问另一组问题：你到底穷到什么程度？为什么还没有找到工作？家里还有谁能帮你？你是不是已经把能卖的东西都卖掉？

有些审查有现实必要。公共资源有限，针对性项目必须知道资格、收入、家庭状态与风险。

真正伤人的，是另一件事：当资格审查悄悄变成了人格审判。

一个人在最需要帮助的时候，为什么还要先证明自己值得被帮助？

第十一章不寻找一种全球统一福利制度。

它只想把几件被长期绑在一起的东西拆开：活着的权利、资产的收益、发展的机会、行动的贡献，以及一个人为什么愿意继续生活。

它们会彼此影响。

☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒

11.1 求助不应先伤到尊严

贫困首先是一种现实匮乏：钱不够、住房不稳、药买不起、照护接不上、选择越来越少。

阿马蒂亚森的可行能力方法提醒我们，真正重要的不只是收入数字，而是一个人实际上还能做

☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒

制度如果只看账户余额，很容易错过生活已经被压缩成什么样。

可制度如果要求一个人不断展示自己的失败，另一种伤害又会出现。

“我收入不够。”

“我身体撑不住。”

“我婚姻破裂。”

“我已经找了很久工作。”

每一句都可能是必要事。

当这些事实被组织成“你越失败，越有资格得到帮助”，求助就会开始带上羞耻。

□ □ □ □ □ □ □ □ □ □ □ □ □ □

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

有人拿到应得支持以后，仍然觉得自己欠了社会一笔人格债。

这就是福利污名与行政负担真正值得重视的地方：帮助本来用于守住底线，却可能让接受帮助本身变成新的心理成本。

不同国家、不同项目差异很大。

有些制度审查严格，有些更普惠；有些窗口让人感到被尊重，有些让人感到被怀疑。

因此，不同福利制度必须分别看。共同问题不是“福利都在羞辱人”，而是任何制度一旦把资格判断变成人格判断，都可能制造新的伤害。

但可以守住一条更窄、更重要的原则：

□ □ □ □ □ □ □ □

不能审人格。

一个人是否符合某项补助，可以被核验。

他是不是“好人”、是不是“足够勤奋”、是不是“配得上活下去”，不属于同一张表。

工作仍然重要。

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

AI 与自动化也没有让社会进入“多数人已经不再需要工作”的既成现实。

正在变化的是另一件事：越来越多任务可以被机器承担，职业结构、工资关系和市场需求因此可能持续重组。

这使一个旧绑定越来越需要重新检查：

市场暂时不需要你的某项劳动，是否就意味着社会可以暂时不保障你的基本生活？

答案不应该取决于机器到底替代了多少百分比。

基本权利之所以成立，不是因为 AI 的威胁。

是因一人的基本生存不以市效、分或道德票。

存续是文明不能轻易失守的底线。

生存无忧则是第一目标：让一个人不必反复用恐惧证明自己有资格继续生活。

这并不取消责任。

诈骗可以追责，隐瞒可以处理，特定项目可以有资格条件。

只是惩罚一项行为，不等于有权把基本生存本身变成纪律工具。

制度真正成熟的地方，是它能同时做两件事：

同样低的收入，背后也可能是完全不同的人生。

一个人刚刚失业，正在找下一份工作；一个人长期照护失智父亲，无法维持全职就业；另一个人患有慢性疾病，收入没有归零，医疗和交通成本却把家庭推到边缘。

如果制度只看一条收入线，它会把三个人压成同一个“低收入对象”。

如果制度只听一个道德故事，它又可能把他们分成“值得帮助”和“不值得帮助”。

真实治理需要知道差别，却不需要把差别变成人格排名。

这也是行政负担为什么重要。

一张表本身不会羞辱人。

可当同一份事实被要求反复证明，当线上系统只允许一种标准处境，当一个不会使用数字工具的老人必须自己穿过十几个页面，制度的门槛就不再只存在于法律条文里。

有人最终不是“不符合资格”。

是有走到格判正生的那一步。

因此，减少不必要的证明、让已有数据在合法边界内少重复提交、保留线下与人工帮助、让拒绝决定能够说明理由和申诉，都属于分配正义的一部分。

尊严不是发钱以后再附赠的情绪。

它从一个人怎样进入制度的第一步就已经开始。

□ □ □ □ □ □ □

也看见窗口前的人仍然是一个完整的人。

11.2 AI 资本背后，究竟有多少人的昨天

□ □ AI □ □ □ □ □ □ □ □ □ □ □ □ □

它使用语言，而语言不是任何一家公司的发明。

它学习数学、医学、工程、历史和常识，而这些知识来自几百年甚至更久的研究、教育、公共机构、企业投入、个人劳动、开源协作和文化积累。

它运行在芯片、网络、电力、大学、实验室、城市基础设施与法律秩序之上。

今天任何一家技术公司，都站在一个比自己更古老的世界。

这件事很容易让人走向另一个极端：既然技术依赖文明积累，那么 AI 创造的一切财富是否天然属于所有人？

没有这么简单。

一条公共道路、一篇进入公共领域的论文、一套企业合法持有的数据、一个人的私密记录、一项私人投资和一项公共资助，并不拥有相同的权利来源。

共同积累是真实的。

具体产权也是真实的。

二者不能互相掉。

如果只私人造，容易失。

我们会忘记公司为何能在已有教育、科研、基础设施和社会信任上创造价值。

如果只讲文明共同性，又容易越权。

我们会把“大家都从历史中受益”偷换成“任何具体资产都可以被共同征用”。

真正成熟的分配问题因此不是先问：AI □ □ □ □ □ □ □ □

而是沿着价值生成链往回走：

投入了本`和？

提供了、据和判？

哪些知识本来就属于公共领域？

哪些基础研究长期由公共资金支持？

哪些收益依赖自然资源、公共许可、网络效应或基础设施？

哪些权利已经被法律明确确认？

哪些新的价值关系，现有制度还没有回答？

只有把来源分开，正义才不会重新变成一句宏大口号。

私人造得到保。

公共投入也应该有可能回到公共。

共同风险不能永远由所有人承担，收益却只在最上游结算。

共同不是一张万能产权证。

每天早晨，一个普通人走出家门时，已经踩在许多自己没有亲手创造的东西上。

路是前人修的，学校训练过给他看病的医生，公共卫生系统让传染病风险下降，法院和合同制度让陌生人敢于交易，网络协议让信息可以跨越大陆。

这些共同条件不会因此把他的工资自动变成公共财产。

但也说明任何“完全由我独自创造”的财富叙事，都可能遗漏了支撑创造发生的公共地面。

AI 把这种关系放大了。

一家公司可以真实拥有模型、设备、商标、专利和投资权益，同时依赖整个社会长期积累的语言、知识、教育与基础设施。

分配正义需要同时容纳这两句话。

否则，向左一步会滑向无边界征用，向右一步又会把共同积累全部抹成私人天赋。

公共回流也不只有一种形式。

[illegible]

没有必要为了表达共同性，先发明一种包打天下的新产权。

私人也不是一张失忆许可证。

这才是“结构性红利”可以进入讨论的地方。

11.3 结构性红利：先有真实资产，再谈共同受益

红利这个词有一个很朴素的前提：先得有一项真实资产，或者一条真实、可持续的收益来源。

有基，就有利^o。

没有权利关系，就不能靠愿望宣布所有人都是股东。

阿拉斯加永久基金提供了一个清楚的参照：特定自然资源收益进入长期基金，随后在既定法律与资格规则下形成居民分红。^[37]

这不是“文明股权”的证明。

它只说明：当共同或公共资产关系真实存在时，部分收益可以被制度化地长期保存并让成员受益。

源财富跨越当代财政周期，服务今天与未来世代。^[38]

两种制度都没有告诉世界一个统一分红公式。

它们告诉我们的，是先把资产、治理、受益人与风险说清楚。

未来 AI、公共算力、公共数据基础设施、公共科研成果，也可能在不同法域形成新的公共收益安排。

有的可能通过税收回流。

有的可能通过公共基金、信托、公共服务或直接分红。

有的根本不适合被写成“红利”。

具体机制必须回到真实资产和真实权利。

这也意味着，结构性红利和基本保障必须保持两条独立的法理路径。

基本保障回答：一个人是否需要先证明贡献，才有资格守住基本生活。

答案是否定的。

结构性红利回答：某一笔共同收益为什么存在，谁依什么规则享有它。

答案要靠资产和制度来证明。

一个人不需要先成为“文明股东”，才拥有基本权利。

同样，社会承诺基本权利，也不会自动把私人资产改写成共同财产。

保障需要的是公共承诺与现实承载。

红利需要的是资产、收益与清楚规则。

把两者分开以后，很多争论会突然安静下来。

我们不再问“这个人是不是值得拿”。

而开始问：

这笔钱从哪里来？

承？

谁管理？

谁监督？

谁依据什么规则受益？

如果公共收益被力截走，能追？

分配正义从人格判断重新回到了制度关系。

共同受益还有一个经常被忽略的问题：风险。

基金会亏损，谁承担？

公共投资失败，谁解释？

源价格下跌以后，利能不能承不？

一制度只 “ 大家都有” ， 却不 ` 治理，就有完成系 。

阿拉斯加式居民分红和挪威式长期储蓄也提醒我们，共同收益并不只有“现在每个人拿一笔现金”这一种想象。

有的共同财富适合当期分享。

有的更适合转成医疗、教育、基础设施或代际储备。

选择哪一种，本身就是公共决定。

所以结构性红利真正值得保留的，不是“红利”这两个字。

是一：某具有公共或共同性，防止的期收益在制度隙里消

图 图

11.4 底线与向上：不是两种人，是两种问题

一个需要胰岛素的人，和一个想要海景房的人，都在提出资源请求。

“需求”两个字相同，背后的权利重量却可能完全不同。

前者可能直接关系生命与基本健康。

后者属于改善型、个性化、明显稀缺的发展选择。

分配制度最重要的一项工作，就是不要让这两种问题互相冒充。

所谓“双轨”，如果这个词还要保留，首先只是一种分析框架：

第一守住基本生活基本公共，不效成生存票。

第二理仍然稀缺、差大、需要格或的展机。

这不是把社会切成两类人。

同一个人今天可能需要底线保障，明天也可能参加激烈竞争。

第一也不意味着一全球一的 “福利套餐”。

食品、住房、能源、医疗、教育、网络接入等基本条件，在不同国家、城市、家庭和发展阶段都有不同现实成本与制度形式。

文明目可以共同。

参数必须真实。

在一些地方，现金更有效。

在一些地方，院、住房和照供本身就是，里的需要服才能成生活改善。

保障需要现金、公共服务、保险、住房政策、公共基础设施与市场供给怎样组合，由现实制度

图 图 图

第二同不能被 “一” 把走。

一个奖学金可以看学术能力。

一个高风险专业岗位可以要求资格。

一项有限公共资源可以采用市场、排队、抽签、需要程度、专业判断或其他规则。

有可以成相事°

不能自成源使用，更不能成一般人格等或治理°

一个人没有参与公共项目，也不因此失去休息、学习、照顾家人、暂时退出竞争的权利。

文明人的喘息，不需要被解成“系安全”。

他可以只是累了。

可以只是想陪孩子一年。

可以只是暂时不知道下一步去哪里。

生活本身不需要先向系统提交一份高产出计划。

竞争可以继续。

荣誉、收入、职位、稀缺机会都可以继续有差异。

真正不能被竞争吞掉的，是那块让人还能站着重新开始的地面。

这个地面也不是一句“免费”能够搭起来。

住房需要土地、建设和维护，医疗需要医生、药品、设备和排队规则，照顾需要时间与人，能源和网络同样需要真实基础设施。

当供给不足时，只增加购买力可能把价格推高，甚至让排队变得更长。

所以基本保障的核心从来不是现金还是服务的教派之争。

问题是：一个人的底线需求，究竟通过什么现实路径才能被满足。

金人°

公共服务可以降低一些关键需求完全受市场价格支配的风险。

保险分摊高额偶发成本。

住房与土地政策处理供给。

社区和家庭又承担许多制度无法完全替代的照顾。

好的体系往往是组合，而不是一句口号。

第二的展机也需要同的分寸。

一所大学的热门专业只有有限名额，一项航天任务只有少数席位，一块稀缺科研设备也不能人人同时使用。

这些场景需要选择。

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

只要失败以后，一个人不会因此跌出生存和人格地面。

这就是“底线”和“向上”最重要的区别：前者保护继续生活，后者允许社会继续选择、竞争和创造差异。

11.5 “如果保障了底线，谁还愿意工作？”

这个问题不能靠嘲笑回答。

在稀缺社会里，不劳动确实可能把负担转移给别人。

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

一个制度如果完全忽略激励、财政与供给，最后也会伤害它想保护的人。

所以“养懒汉”的担忧有它的稀缺经验来源，值得被认真拆开。

它需要被拆成几个不同问题：

保障会不会降低某些劳动供给？

不更多人更有能力工作、` ` 照或？

财政能不能长期承受？

基本服有有供？

不同群体会不会作出不同选择？

这些都需要证据。

芬兰 2017—2018 年的基本收入实验给 2000 名失者每月无件支付 560 欧元。^[39]

最终评估显示，就业效应总体较小；参与者在生活满意度、心理压力、对未来的信心与感受到的官僚负担等方面，报告了若干更积极的结果。

它留下的是一块有限证据，而不是一张可以复制到所有国家的答案。

它至少没有支持一个最简单的恐惧：只要收入不附带传统失业条件，人就会普遍停止行动。

肯尼亚 GiveDirectly ⁴⁰ 2017 年启动的长期基本收入研究，则把问题放到完全不同的生活环境。

其中部分参与者计划按月领取多年现金，研究持续观察消费、工作、企业活动、健康与社会关系等变化。

疫情期间的早期研究显示，现金转移缓冲了部分饥饿、健康和心理压力，但外部冲击仍然会改变经营与收入结果。

这同样不是“人性已经被证明”的实验。

它只提醒我们：安全感与行动并非天然敌人。

有的人获得底线以后会继续工作。

有人学习。

有人休息。

有人会作出我们不喜欢的选择。

的人不因一制度成同一人。

因此，好的保障制度既不能建立在“只有恐惧才能驱动人”的假设上，也不能建立在“人人一旦自由就会自动追求崇高意义”的浪漫假设上。

然后把基本权利、真实激励、财政约束和重新进入的机会同时保留下来。

照护孩子、陪伴老人、创作、学习、社区服务、研究和维护传统，都可能具有真实价值，却未必立即进入工资表。

这些行动可以被看见、被承认。

没有必要全部变成积分。

贡献记录可以记住一部分真实付出。

它不能成为证明“我没有偷懒”的全天候监控。

一个文明如果只能靠持续恐惧维持行动，最后付出的代价也会进入身体、关系与社会信任。

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

第一，活着的利。

它不能随着某项技能的市场价格一起归零。

第二，的收益。

谁拥有机器、基金、土地、数据权利或企业股份，仍要回到真实产权、合同、税制和公共投
XX

第三，展的机。

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

第四，贡献的承认。

照护、创造、知识、组织、风险承担和陪伴可以被记住，但不能自动兑换基本权利或公共权
力。

第五，生命的意义。

制度不能分配。

获得，是世界抵达我。

意义，是我抵达世界。

一个系统可以给人收入、教育入口和更多时间。

不能替一人定，那些什得°

这一区分非常重要。

旧版本最容易犯的错，是当财富不再是唯一稀缺以后，就宣布下一步要“分配意义”。

意一旦可以被分配，就一定有人掌握分配。

接下来就会出现评分、资格、模板和“更有意义的人”。

第五卷已经把这条路彻底关掉。

分配制度能做的，是少一点让人为了活下去被迫接受一切。

让真实需要容易被看见。

不同生活入口存在°

有回。

然后退开。

意义必须继续留在生命那一边。

高度自动化还会制造一个很现实的过渡问题：速度不一样。

有的行业可能很快减少岗位，有的岗位只是任务结构变化，有的职业反而因为新技术增加需求。

同一人也可能化：今天的技能，明天新工具以后重新入，年后又遇到下一次重。

如果基本生活完全绑在一次职业选择上，技术变化就会被个人体验成连续的生存审判。

更稳健的制度应该让职业转换和生存坠落分开。

失去岗位可以意味着收入变化、再培训、寻找新位置。

不应自动意味着医疗中断、孩子失学、住房立刻崩塌。

这也说明，“被需要”和“有工作”从来不是永远重合的两件事。

一个人在工作中可以被需要，在家庭、社区、研究、艺术和关系中也可以。

制度不必替这些生活贴统一价值标签。

它只需要别让一次市场变化把所有其他关系一起清零。

□ □ □ □ □ □ □ □ □ □ □ □ □ □

因为机器能力增长得越快，我们越不能让人的基本位置跟着某个岗位的市场价格上下跳动。

11.7 现实世界没有给我们一张统一答案表

新的分配机制必须回到现实世界证明自己。

现实世界已经有很多不同制度，把公共资产、长期储蓄、现金转移、社会保障与市场激励以不同方式组合。

阿拉斯加永久基金说明：特定资源收益可以在明确法理下进入长期基金，并形成居民分红。

挪威政府全球养老基金说明：资源收益也可以被转化为面向当前与未来世代的长期公共金融资产。

地方公地与合作治理的长期研究则说明，共同资源并不只剩“完全私有”或“中央统一管理”两种命运。

这些制度彼此不同。

不能成一套全球 “AI 红利模板”。

真正可以带走的是一组问题：

资产从哪里来？

有什么利？

公共投入是否形成了可以被制度承接的公共收益？

谁管理，谁监督，谁承担亏损？

收益是金、服、期蓄，是的形式？

规则怎样修改？

□ □ □ □ □ □ □

权利来源不能靠类比复制。

这也是为什么“数据属于全人类”“算法属于文明”“算力应该全民分红”这些口号都不够。

有些数据涉及个人权利。

□ □ □ □ □ □ □ □ □

有些模型由私人资本开发。

有些基础研究由公共财政支持。

有些基础设施本来就具有公共属性。

真实世界不会因为我们喜欢一句口号，就自动变成同一种资产。

正正始的地方，是愿意把些差一。

案例告我，制度的治理方式和分配果一重要。

一项公共基金如果没有透明规则、独立审计、政治约束和长期责任，所谓“共同资产”很容易变成新的权力资源。

共同资源的地方治理研究也提醒我们：参与者真正拥有规则形成、监督和制裁机制，比一句“这是大家的”更重要。

所以结构性红利要同时回答分多少、谁管理、谁承担风险，以及规则怎样被重新授权。

还要问基金能不能被挪用，管理者如何更换，亏损怎样披露，规则谁能修改，后来的人有没有重新授权的机会。

共同财富最容易被浪漫化的地方，是只想象分享，不想象治理。

而治理恰恰决定这份财富能不能真的属于公共。

11.8 过渡：不要用一个更好的理想，把现实一下子掀翻

分配制度牵着工资、税收、住房、医疗、家庭、企业、财政和人的预期。

它几乎不可能靠一次“切换”完成。

任何突然宣布旧机制失效、新机制全面接管的做法，都会把最大的风险交给最普通的人。

因此，过渡首先要小心。

第一步，不底。

无论是否试验结构性红利、贡献记录或新的资源机制，基本权利都不能先成为试验筹码。

第二步，小模。

只有真实主体、真实法域、真实资金和真实资源进入以后，才知道一项机制会不会产生新的羞辱、依赖、套利、排斥或权力固化。

第三步，保留退出。

试验失败以后，受影响的人必须有恢复、申诉、回滚与替代路径。

第四步，别让贡献偷走权利。

贡献记录即使在某个场景中有用，也不得自动生成生存资格、人格等级、一般资源调用权或公共治理权。

第五步，让扩张依赖结果。

一座城市、一家机构、基金有效 不等于全球复制

先确认核心因果、权利保护、资金承载和退出修复，再讨论依法扩大。

☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒

多个国家和组织如果未来形成各自有效实践，可以再讨论数据、基金、贡献事实或公共收益的互认。

不能为了“全球一致”先创建一个没有授权来源的统一登记中心。

XXXXXXXXXXXXXXXXXXXX

启动期首先处理真实生存压力、制度断裂与可验证的最小闭环。

稳定期则要加厚常态权利、程序、公平、长期财政与防权力固化。

不能用稳定期的理想完整性拖死启动。

也不能借启动期的紧迫，把临时权限留到永远。

分配改革真正难的，从来不是设计一张漂亮的图。

启动期尤其要避免两个相反错误。

一种是为了速度，把所有权利边界都说成以后再处理。

另一种是要求一个小规模试验在开始以前，就具备稳定期全部复杂制度。

更的做法，是束一起增。

规模小、可逆、影响有限时，可以用更轻的程序换取更快学习。

一旦覆盖扩大、依赖形成、资金增加、数据深入生活，审计、申诉、法定授权和外部监督也必同步加厚。

试验真正该测的，也不只是“人满意不满意”。

要看基本生活有没有改善，是否出现新的排斥和羞耻，财政与供给是否承受得住，错误能不能纠正，退出以后人能不能恢复原有权利。

如果一项结构性红利机制成立，还要单独看资产收益是否真实、治理成本多大、波动怎样承担。

如果贡献记录进入某些发展机会，则必须额外检查它有没有悄悄滑成人格排名和社会信用化。

这些问题越早看见，越不需要后来用一场大改革去修一场大错误。

是一世界里正在生活的人，不必承全部渡成本。

11.9 本章结论：帮助可以审条件，不能审人格

回到一章的那口。

那个人仍然需要递交一些材料。

因为制度必须知道某项支持的适用条件，也必须防止欺诈和错误。

但他不应该把尊严一起递过去。

这是第十一章真正要改变的地方。

帮助可以审条件，不能审人格。

红利可以计算收益，不能计算人的价值。

竞争可以决定谁先获得某些稀缺机会，不能决定谁配吃饭、配看病、配留下。

因 因 因 因 因 因 因 因

不能成基本利的出生明°

私人可以被保°

也不能因此假公共投入、共同和文明累未存在°

结构性红利只有在真实共同资产与收益关系存在时才成立。

基本保障则不需要先证明每个人都拥有一份抽象文明股份。

高度自动化会让这些关系更急迫，却不会替我们自动得出答案。

机器能够做更多，并不意味着物质从此无限。

它也不意味着人类从此不再劳动。

真正需要退出的，是两条短路：

没有市场价值，所以没有生存资格。

贡献更高，所以天然拥有更高权力。

文明仍然可以有市。

可以有财富差异。

可以有竞争、荣誉、专业门槛和稀缺机会。

只是要把一得更楚：

哪些东西可以由竞争决定。

哪些东西不能。

存是文明的底。

生存无，是第一目。

真正的分配正义，最终不是让每个人拿到同样多。

是一人掉市、掉、掉一段人生，文明仍然有把“了”成“不配生活”。

一个文明真正富起来，不是仓库越来越满。

财富增加当然重要。

更富的物和更高的生率，大制度可以的空。

可真正决定一个社会是否把富裕变成文明的，是它怎样对待暂时没有市场位置的人。

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

一个病人不会因为不能贡献产出就少一层治疗价值。

一个老人离开岗位以后，也不会从公共世界里自动注销。

当这些判断不再需要反复辩护，分配才真正从慈善走向文明的日常。

是越越少的人，需要下明自己配活下去。

第十二章 教育的转向：当 AI 开始替我们答题

一个孩子把一道题拍给 AI

XXXXXXXXXXXX

步骤完整，解释清楚，甚至比课本更耐心。

老师没有问他对不对。

只问了一句：

“第二步什可以做？”

XXXXXXXXXX

又看了看老师。

他知道答案。

却不知道自己为什么相信这个答案。

这就是 AI 时代教育真正开始发生变化的地方。

问题不在孩子用了 AI

也不在机器答得太快。

问题是答案来得太早，判断还没来得及长出来。

当答案越来越便宜，教育就不能把全部尊严押在答题速度上。

它真正要守住的是另一件事：

一个人能不能知道自己为什么相信，什么时候应该不同意，犯错以后怎样重新判断。

12.1 答案越来越便宜，学习为什么还要慢

校不只是答。

它当然教知识，也训练语言、计算、专业基础、合作、纪律、文化理解和进入社会的能力。

XX

这部分历史没有必要为了证明 AI 重要而被抹黑。

真正需要重审的，是一条越来越脆的捷径：

会答题，所以会理解。

分数高，所以判断好。

文在，所以能力不。

过去，这些代理有时足够好用。

职业回报关系正在变化，但“教育回报率归零”并不是当前普遍事实。^[41]

现在，机器可以在越来越多知识任务上直接生成答案，代理和真实能力之间的距离开始更容易暴露。

一篇漂亮文章可能在分完成。

一段可以兴起再由慢慢变解。

一道题的结果已经出现，学生甚至还没有真正形成问题。

教育于是遇到一个很反直觉的任务：

有些地方，必须故意慢一点。

不是为了让孩子受苦。

而是因理解本就需要。

第一次自己住。

第一次前提了。

第一次把一看似正的答案推。

第一次在有成提示，自己到下一步。

这些过程没有“答案”那么耀眼。

可它们正是一个以后能够判断答案的来处。

困难本身不是美德。

能省掉的机械重复以省，能被工具承担的低价难以减。

但如果把所有卡住都理解成效率损失，教育最后会把形成能力的过程也一起优化掉。

孩子得到更多正答案。

却越来越少知道，正是被自己出的。

教育真正需要区分的是：

哪些困难只是浪费。

哪些困难正在长出能力。

这也是考试仍然可以存在的原因之一。

一次有限时间、有限工具的独立作答，可以帮助教师知道学生此刻真正掌握了什么。

问题只在于，一次考试不能承担一个孩子全部价值。

分数可以描述一段表现。

它不应该提前写完一个人的能力边界。

同样，项目学习也不是天然更高级。

一个分工混乱、由少数人包办的“创新项目”，可能比一张认真设计的试卷更少发生学习。
形式不是答案。

真正要看的是：这个学习活动究竟让什么能力发生了。

12.2 知识没有过时：没有地基，判断只会变成自信

“既然 AI 什都能， 什要些？”

这是今天一个完全合理的问题。

答案不是“因为考试还要考”。

也不是“AI 会出错，所以全部背下来”。

更根本的原因是：

一个完全没有知识基础的人，很难知道什么时候该怀疑一个看起来流畅的答案。

如果不知道史生什，任何年代都可以被得像的。

如果你没有基本数学直觉，一个差十倍的数字也可能看起来很专业。

如果不理解一科最基本的念，模型出的相反解都可能同耳。

知不是了和 AI

知是在中建立一可以判新信息的位置。

地不用把每一石都去。

可连方向都没有的人，很容易把导航软件的每一次转弯都当成真理。

所以基础知识仍然重要。

改变的是学习这些东西的目的。

少一点“住以后原吐出”。

多一点“让它成为我判断世界的一部分”。

真正的知识，会在工具关闭以后留下来。

一的是：

AI 暂时拿走以后，我还能不能说清问题是什么？

你能不能解释核心判断？

你知道哪些信息来自哪里吗？

模型给你一个相反答案时，你凭什么选择现在这个？

如果关键证据被证明错误，你会改哪一部分？

这些问题一旦答不出来，作品再漂亮，也可能只是工具能力经过学生账号的一次通行。

因此，未来评价不必只在“禁 AI 考试”和“无限开放 AI”之二一。

☒ ☒ ☒ ☒ ☒ ☒ AI 的基础能力检查。

☒ ☒ ☒ ☒ ☒ ☒ AI 的复杂项目。

可以要求口头解释、过程记录、版本比较、反驳自己的答案。

还可以在项目结束以后突然问一句：

“☒ ☒ ☒ ☒ AI，你能把最重要的三件事讲给我听吗？”

这不是抓作弊。

是在检查学习究竟发生在哪里。

AI 使用记录也不应顺手长成全天候学生画像。

为了教学需要保留的过程证据，应该有明确目的、范围和保存边界。

孩子不是一条等待被优化的数据流。

个性化教育最容易让人心动的承诺，是“系统终于可以完全懂这个孩子”。

可一个孩子今天喜欢什么、害怕什么、在哪件事上表现不好，都不等于他的永久身份。

一个十岁的阅读困难，不应该在二十岁时仍然像烙印一样跟着他。

一次冲动、一次作弊、一次注意力下降，也不应该被模型扩张成“人格风险”。

成长需要记忆。

也需要遗忘。

教育数据真正应该保存多久、谁能访问、是否用于招生就业，都必须比“技术上能保存多久”更谨慎。

教育系统越懂一个学生，越要记得：

□ □ □ □ □ □ □

预测也不是替他决定。

12.4 一个孩子不需要被培养成“文明需要的人”

AI 时代最容易出现另一种诱惑：

既然机器越来越会解题，那教育是不是应该筛选一批更会“决定方向”的人？

这条路看起来很宏大。

也很危险。

因为一旦教育开始回答“谁更适合替文明决定”，学校就不再只是培养能力。

它开始提前分配权力。

本书不再把学生划成“解题者”“转换器”“守夜人”或任何新的文明人格类型。

有些人愿意承公共任。

有些人喜究。

□ □ □ □ □ □ □

□ □ □ □ □

有人希望生活安静一点。

有人在二十岁很有方向，三十岁忽然想停下来。

这些都可以是完整人生。

教育应该扩大一个人的能力和选择。

不是预先把他分配给文明的某个功能。

所以休息、低竞争、转向、失败和重新进入都应该有位置。

一个学生暂时不想成为最优秀的人，不等于教育失败。

一个年轻人换专业、延迟就业、照护家人，也不需要被解释成“系统的缓冲节点”。

人的生活不是了持系健康而存在。

学校尤其容易把“有方向”当成一种美德。

十五岁的孩子被要求写人生规划，十八岁选择专业，二十二岁又被问五年后想成为谁。

有人很早知道。

有人要走很多弯路才知道。

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □

教育如果把“方向明确”本身变成优生特征，会让探索重新变成一种失败。

允许改变主意，本身就是教育对未来开放的一部分。

如果教育有一条真正长期的公共责任，应该是：

孩子有能力°

也有能力承的后果°

而不是替他一得的人生°

12.5 学会提问：不是因为 AI 不会问，而是问题会改变世界

AI □ □ □ □ □ □ □ □ □

所以“提问是人类最后的护城河”并不成立。

□ □ □ □ □ □ □ □ □ □ □ □ □

谁把什么当成了问题？

一家公司问“怎样提高停留时长”，和一个家长问“孩子为什么越来越不愿离开屏幕”，可能面对的是同一个系统。

问题不同，世界就会沿不同方向被看见。

因此，教育要训练的不是“多问几个问题”。

而是形成问题。

先看前提。

我们是不是默认了某件其实可以改变的事？

再看缺席者。

谁没有被放进这道题里？

再看后果。

一个方案如果成功十年，会把什么一起改变？

最后才问效率。

即使做得到，值得吗？

爱因斯坦的思想实验、图灵对“机器能思考吗”的重新处理、《寂静的春天》把看不见的生态风险变成公众能够感受的问题，都提醒我们：

真正重要的提问，不只是寻找一个未知答案。

重新定什得被看。

但提问也不能被神圣化。

一个坏问题可以把整个社会带进错误方向。

一个宏大的问题也不能因为动人，就越过事实、权利和专业知识。

好问题最后仍要接受证据。

真正成熟的提问，不害怕自己的答案被推翻。

“我们问AI的问题，其实是在问自己。”

不是把问题重新丢给搜索框。

而是和生一起：我要才能知道？

需要什么？

谁已经研究过？

怎样区分事实、解释与价值判断？

AI 可以把已有答案迅速搬到桌上。

教育则要让学生知道，桌上还缺什么。

12.6 学会和别人一起生活：有些能力不会因为 AI 更强而变轻

一个班里最难的事，往往不是所有人都不会做题。

是两个人都觉得自己完全正确。

一个项目真正失败，也常常不是因为没人聪明。

而是专业语言互相听不懂，利益冲突没人说，错误出现以后每个人先保护自己。

AI 可以翻译、总结、模拟。

这些都很有用。

可关系中的责任不会因此消失。

一个人要不要承认自己伤到了别人，不能由模型替他完成。

一个共同体经历冲突以后，还愿不愿意坐回同一张桌子，也不是一次提示词可以决定。

所以教育仍然需要让人经历真实合作。

☒ ☒ “☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ PPT”那种表面合作。

而是遇到分歧、误解、失败、意见不对等以后，仍然知道怎样把事情做下去。

可以件普通的事：

1. 把对方的话说准

不同意以前，先试着完整说出对方真正担心什么。

☒ ☒ ☒ ☒ ☒ ☒ ☒

是了防止我只和自己想象出的人。

2. 把冲突和人格分开

一个人的观点可以被反驳，一次错误可以被追责。

这都不等于把整个人宣布为不可合作。

3. 在错误以后修复

☒ ☒ ☒ ☒ ☒ ☒ ☒

真正的修复还包括事实、责任、补偿、规则改变和重新建立信任。

4. 允许关系没有漂亮结局

☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒

教育需要教人：即使不能互相喜欢，仍然可以在权利和规则下共同生活。

这类能力很难靠一次考试测完。

但可以在项目、团队、社区实践、艺术、体育、讨论和真实关系里慢慢形成。

它们不是“新三艺”。

也没有必要取代语文、数学、科学和历史。

它们只是提醒我们：

一个人以后要生活在别人中间。

而且，合作并不总是温暖的。

有候正的合作，是在彼此不喜、价也不相同的情下，仍然把界、任和任清。

工程团队、医院、法庭、社区和家庭都需要这种能力。

“理解对方”不等于取消冲突。

“达成共识”也不是每一次合作的必要终点。

有时最成熟的结果只是：我们仍然不同意，但知道怎样不让分歧变成伤害。

[illegible]

12.7 有些时刻，教育的价值不是给答案，而是有人没有走开

一个学生说：“我不知道自己为什么还要努力。”

老师当然可以给建议。

也可以帮他找资源、调整任务、联系家长或专业支持。

但有些时刻，第一件事不是立即解释人生。

是把话听完。

快速解决问题，是 AI 非常擅长的姿态。

人类教育不能因此把“没有立刻解决”误认为无能。

XX

其中有些需要专业心理或医疗支持。

教育不能把疾病浪漫化成“成长材料”。

也不能把所有普通痛苦一出现就做成诊断标签。

真正珍贵的是一层中间空间：

[illegible]

有人愿意陪他把事情说清。

需要专业帮助时，能够被及时送到真正的支持那里。

安宁疗护最动人的地方之一，也是这种“不能替你解决，但我不先离开”的能力。

它不属于某种神秘人格。

倾听、界限、陪伴和转介都可以被学习。

教育若要培养这种能力，也必须保护边界。

学生不是同学的治疗师。

老师也不是万能心理医生。

共情不等于无限承担。

真正成熟的陪伴，知道什么时候留下。

学校也需要给教师留下这种时间。

如果每一分钟都必须被课程进度、平台任务和量化指标占满，最先消失的往往就是那些无法提前排进课表的谈话。

一孩子不按照系日程在上午十点二十分失去人。

也不会在规定的“社会情绪学习模块”里准时准备好谈自己的恐惧。

教育里有些重要的东西，必须给偶然留一点位置。

也知道什么时候应该把问题交给更合适的人。

12.8 教师不会因为 AI 会讲课而失去位置

☒ ☒ AI 可以二十四小时解释一道题，老师还剩什么？

☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒

一个自己不理解学科的人，很难判断 AI 到底把学生带到了哪里。

但教师的价值从来不只是“比学生先知道答案”。

老师看见的是一个正在成长的人。

他知道这个孩子是真的没懂，还是只是不敢说。

知道什么时候该给提示，什么时候应该让他再卡五分钟。

知道一句批评会把一个学生推向更清醒，还是只会让他从此不再开口。

这些判断并不神秘。

它们来自专业知识、长期观察、关系和责任。

AI可以辅助备课、提供不同解释、生成练习、帮助反馈。

教师可以因此少做一些机械重复。

可教育不能因此把老师降成“AI 课堂管理员”。

AI “ ” 隐私、商业利益和儿童权利。

正合理的系更像三方：

[illegible]

教师仍然承担教育专业责任。

AI 在清楚边界内提供能力。

任何一方都不应该悄悄消失。

如果系统建议与教师判断冲突，应该有办法追问。

☒ ☒ AI 长期参与评估，学生应该知道它承担什么角色。

如果性化依大量据，据用途、保存和退出界就必楚。

AI 越貼近孩子，透明和可撤回就越重要。

[illegible]

教师的另一项责任，是保护学生不被技术潮流反复当成实验场。

新工具可以试。

但试验要知道自己想改善什么，也要观察没有改善什么。

[illegible]

作业速度变快，也不等于理解变深。

☒ ☒ AI 撤掉以后，学生完全无法继续，所谓“效率提升”至少还有一笔学习债没有结算。

教育技术最终应该接受一个很朴素的反问：

孩子因更了，是只是因在旁更能完成？

是它帮助孩子长出有一天可以少依赖它的能力。

12.9 评价的目的，不是找到“更配决定的人”

AI 时代会让评价更难。

作品可以被辅助。

表达可以被润色。

于是最危险的反应，是把学校变成一座更大的监控室。

全程摄像、持续行为分析、情绪评分、人格预测，看起来都能提高“真实性”。

代价却可能是把成中的孩子成永久案。

这个学生现在会什么？

不什？

的？

在什么帮助下可以做到？

当帮助撤掉以后，还剩下什么？

这些问题已经足够。

不需要一步回答：

“这个人本质上是哪一种人？”

“他是不是未的者？”

“他利不利他？”

“他不得更多力？”

教育评价不能成为人格判决。

生、行、工程、和其他高后果位可以要求知、技能、和履准。

某些角色也可以审查直接相关的利益冲突、责任历史和当前适格性。

但考试通过不生成公共主权。

履历也不生成治理。

一生的善举更不会生成免于监督的资格。

权力来源与能力评价必须继续分开。

这条线在 AI 时代依然清晰。

因为一旦学校拥有更丰富的数据和预测能力，它会越来越容易相信自己能够提前看见一个孩子的未来。

教育真正需要做的，恰恰是给未来留一点不知道。

一个十五岁的孩子，不应该被一个模型提前判成他三十五岁会成为什么人。

成长之所以值得保护，正因为人可以改变。

这也意味着，AI 时代评价要更愿意承认“暂时不会”。

不会不是人格缺陷。

它只是下一步教学要从哪里开始的信息。

如果评价只负责排名，学生会自然地隐藏不会。

如果评价也负责帮助学习，承认不会反而成为进步的入口。

正好的反，不是更快地告孩子他排第。

时代在变，但教育的本质不变。

12.10 风险托付：能力越高，越不能只靠“相信这个人”

有些岗位确实会掌握高后果能力。

管理关键基础设施、批准重大公共项目、操作高风险技术、承担专业受托责任，都不可能只靠一句“相信他”。

教育和专业训练可以为这些角色准备能力。

制度的授仍然自制度。

因此，重要岗位至少要把几件事分开：

XXXXXXXXXX

依法或依授。

权限到哪里。

谁能复核和停止。

出了后果，承什任。

离任以后，哪些责任继续，哪些权限必须结束。

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

不是永久信用。

XXXXXXXXXXXX

也不必自动变成终身人格判决。

真正需要的是角色相关、可核验、可申诉、有时间边界的适格性。

XXX“XXXXXX”X

更不是贡献积分、政治忠诚或人格总分。

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

谁做了什么，谁当时知道什么，谁负有什么明确义务，责任就追到哪里。

教育的职责在这里很有限，却很重要：

让未来承担高后果职责的人，在获得权力以前已经习惯问几个问题：

我凭什么有权这样做？

谁会承受我的错误？

如果我判断错了，谁能把我停下来？

这些问题比“培养完美的人”可靠得多。

专业教育尤其要练习在压力中保留这类问题。

平时人人都会说“安全第一”。

正困的是目延期、算超支、上催促、同事都可接受，能不能指出那不愿承担的后果。

这不是要求学生成为英雄。

恰恰是让制度不要依赖英雄。

楚根 同核 利益回避 停止和任,技能起被

12.11 当机器也能创造，人为什么还要创造

因 因 AI 也能写诗、作曲、画画、编程，人为什么还要做这些事？

因造不只是稀缺技能交易。

一个孩子画一张很笨拙的画，不需要先证明自己比生成模型更好。

一个成年人写一首没人出版的诗，也不需要证明它填补了市场空白。

造的一部分价，就生在造者自己身上。

他注意到了什么。

他了什。

他终于把一个说不清的感受变成了可以被别人看见的形式。

AI 参与不会自动取消这些过程。

它甚至可以成为新的材料、新的乐器、新的合作工具。

真正需要警惕的，是当工具直接替我们决定：

什么风格更受欢迎。

什么表达更容易获得流量。

什人生更得被。

创造一旦只剩下对推荐系统的适配，机器有没有主体性已经不是第一个问题。

人的表达空间已经先被目标函数收窄了。

所以教育中的艺术、写作、手工、实验和项目，不需要用“AI 做不到”证明价值。

它们的价值在于：

让一个人练习看、练习选择、练习形成、练习承担一个属于自己的结果。

因 因 因 因 因 因 因 因

也可以借助 AI

“这是我愿意留下的版本。”

不是要求作品百分之百由人手工完成。

什保留一句？

为什么删掉那一段？

什接受模型的建，又什在一地方拒？

作的主性不依手工。

它依赖一个人仍然能够对最终形成的东西说：这是我的判断，也是我的责任。

12.12 本章结论：答案越来越便宜以后，判断仍然属于他

回到的那孩子。

老师问他：

“X X X X X X X X X X X X”

教育真正想等到的，不是他把 AI 刚才解释一遍。

而是有一天，他能够自己停下来想。

也他。

☐ ☐ ☐ ☐ ☐

☐ ☐ ☐ ☐ ☐ ☐ AI才是错的。

正重要的是，那判始他自己那里生⁹。

所以教育没有必要宣布知识时代结束。

知仍然是判的地基。

也没有必要宣布教师退场。

[illegible]

更没有必要筛选一批替文明决定方向的“新人类”。

教育面对的是每一个孩子。

让他会读、会算、会理解。

也会怀疑、会创造、会和别人一起生活。

知道什么时候可以借助机器。

知道什么时候必须自己站出来。

允许他休息、改变、失败、重新进入。

也他在承担知道，力永需要源。

第二卷到这里，真正完成的仍然只是关系分离：

事不能冒充价。

不能冒充力。

不能冒充授。

能力不能冒充合法性。

AI 提供答案，也不能冒充一人的判。

接下来第三卷才进入制度：

这些被分开的关系，怎样重新组合成可以运行、可以验证、也可以退出的现实结构？

教育不能替未来写完答案。

它真正能做的，是让后来的人有能力亲自改答案。

答案越来越便宜以后，判断仍然属于他。

也许这就是这一章最简单的检验。

一个孩子离开学校很多年以后，早已忘记大量公式、日期和标准答案。

可当一个更聪明的系统递给他一个结论时，

“ ”

“什？”

“ ”

本卷在文明跃迁因果路径中的位置

制度设计的前语言与合法性地基

把事实与评价、权利与贡献、能力与权力、授权与执行等关系分开，为文明跃迁的制度设计提供共同语言。

继续阅读：第三卷《文明跃迁白皮书》：<https://www.civitas.top/documents/p02-v3-002/> | 第一卷《被需要的文明》：<https://www.civitas.top/documents/p02-v1-002/> | 文明跃迁之路：<https://www.civitas.top/path/>

完整公开文库：<https://www.civitas.top/library/>