

当智能不再稀缺

当智能不再稀缺

从智能能力扩张到文明稀缺转移

中文公众长文 · 独立出版版

作者：子君赋 | 出品：文明跃迁研究组

V1.1.1 | 2026 年 8 月

出版与阅读说明

本文是一篇面向公众、可以独立阅读和判断的思想长文。“当智能不再稀缺”不是对当前世界的统计结论，也不是说所有人已经平等获得同等智能。它是一个方向性思想实验：当认知、分析、生成与执行能力持续扩张，并逐步成为广泛基础能力，文明的主要瓶颈会向哪里移动？

本文的独立任务，是提出并展开“文明稀缺转移”框架：随着智能能力走向丰裕，方向、全局因果理解、不可逆之前的纠错时间、真实主体的位置和共同纠错能力，可能成为更具决定性的稀缺。

所谓“转移”，并不意味着智能、算力、基础设施与获得机会的不平等已经消失。能力增长、能力是否公平可得、能力由谁控制以及能力服务什么目的，是四个不同问题。本文讨论的是：当能力持续扩张以后，决定文明后果的主要约束如何由“能不能做到”，进一步移向“为什么做、会造成什么、谁承担后果以及错误还能不能被改变”。

《文明跃迁》五卷书处理的是文明目的、制度前语言、工程路径、文明永续与意义根基的完整体系；本文不是五卷书的摘要或节选，也不要求读者先接受那套体系。它只选择“智能丰裕如何引发明文明稀缺转移”这一独立问题，把相关判断组织为一条能够在本文内部阅读、质疑和讨论的因果链。

文中少量出现的工作名称，首次出现时均附有通俗释义；它们只指称仍可被检验、替代或淘汰的工具构想，专名本身不承担证明功能，也不表示已经建成的软件产品或运行文明的总系统。

本文不是已经完成的制度方案或工程手册。外部资料只用于支持特定事实、方法和制度经验；三目标、意义为根、人的不可替代性、文明级纠错能力、反封闭以

及五类文明稀缺，仍属于本文提出的规范判断与理论综合。未来 AI 主体性、复杂多主体机制的可达性以及尚未试运行工具的实际效果，继续保持开放。

版权与传播

作者：子君赋 | 出品：文明跃迁研究组

官方网站：www.civitas.top

公开状态：公众阅读与传播版

版本：V1.1.1 | 中文公众长文 · 独立出版版

目录

写在正文之前 | 真正的问题正在转移

0 | 为什么智能能力增长会改变文明的瓶颈

- 一、先把四件事分开
- 二、能力并不独自进入现实
- 三、速度增加，缓冲减少
- 四、从控制一种工具，转向保持共同系统可纠正

1 | 文明为何需要智能，却不能由智能定义自身

- 一、先问文明为什么存在
- 二、三目标不是三个可以互相交换的分数
- 三、意义为根，不等于系统有权定义意义
- 四、智能是文明能力，不是文明主权
- 五、从目的进入架构

2 | AI 不是外部对象：重建人—AI 共同系统

- 一、从“一个 AI 系统”走出来
- 二、方向判断、运行安全与现实反馈
- 三、从多方透明验证，到成熟后的低摩擦运行
- 四、一个中层例子：AI 怎样放大人的在场
- 五、未来主体性仍然开放
- 六、共同系统真正需要什么

3 | 强智能的重要用途之一：让后果提前被看见

- 一、文明最昂贵的学习方式，是等结果发生以后才知道自己错了
- 二、从一个答案，转向一张可选择未来因果地图
- 三、能力越强，越要把“预测”与“决定”分开
- 四、因果推演不是预言
- 五、共同事实不等于统一思想
- 六、让后果提前成为原因

4 | 从判断到行动：当前可得最优，不是静态完美

- 一、看见更多，不会自动替我们消除不确定性
- 二、“最优”只能是条件化的
- 三、效率来自减少无意义摩擦，而不是减少判断能力
- 四、行动原则必须由后果性质决定

5 | 两种错误：可逆试错与不可逆禁入

- 一、文明不能用同一种方式学习所有事情
- 二、可逆试错不是随意试错
- 三、不可逆禁入不是无限安全主义
- 四、真正的分界不是保守与激进，而是能不能回来

6 | 可逆不是回到过去：纠错、重建与再授权

- 一、把“可逆性”（reversibility）等同于“回滚”（rollback），还不够
- 二、纠错有不止一种方向
- 三、纠错能力比“名义退出”更严格
- 四、强 AI 时代，纠错能力本身是一种基础设施

7 | 反封闭不是拒绝决定：保护未来选择能力

- 一、反封闭反对的不是坚定，而是终局化
- 二、选择空间不是越大越好
- 三、最危险的封闭往往以成功的名义出现
- 四、开放未来与不可逆禁区并不矛盾

8 | 治理什么：目的、权限、行为、责任与后果

- 一、“监管 AI”这个说法太粗
- 二、先回答五个不同问题
- 三、能力本身有时仍是直接治理对象
- 四、责任链不能被系统名称洗掉
- 五、谁监管、为何监管，也必须被监管

9 | 为什么许多 AI 治理问题首先显影人类自身

- 一、AI 放大了什么，先看它接入了什么
- 二、结构先于人格
- 三、现实机制不能由理论或工具代劳
- 四、上游结构也包括分配与俘获
- 五、机器侧风险不能因此被取消
- 六、双侧稳定，比寻找一个永久正确的一方更现实

10 | 强 AI 之后，人仍然不可替代的是什么

- 一、不要把人的价值建立在机器暂时做不到什么
- 二、人不是因为效率低才需要保留
- 三、AI 可以扩展在场，但不能把生活变成代理完成
- 四、意义不是最后一块留给人的工作
- 五、真正稀缺的开始转移

11 | 成熟关系：能力越强，系统越知道何时退居

- 一、文明成熟不是系统越来越深入生活
- 二、退居不是能力减弱，而是侵入性降低
- 三、低侵入不是隐形控制
- 四、退居不是一次撤场，而是可逆减载
- 五、自我维持才是文明逻辑真正形成
- 六、退居把本文带回起点

12 | 结语：当稀缺转移，文明真正要守住什么

- 一、回到开篇：稀缺没有消失，只是改变了位置
- 二、方向：能力越强，越需要知道为何而用
- 三、全局因果：局部聪明不能代替整体理解
- 四、纠错时间：速度越快，改变结果的窗口越珍贵
- 五、真实主体：系统越强，生命越不能退出
- 六、共同纠错：没有谁可以被设定为永远正确
- 七、于是，AI 治理问题被重新放回文明问题
- 八、文明真正要变强的是什么

说明 | 证据边界、机制边界与开放问题

写在正文之前 | 真正的问题正在转移

过去很长时间，文明首先受限于“做不到”。知识难以获得，复杂系统难以理解，大量路径无法比较，许多愿望也因为认知、组织和执行能力不足而停留在想象之中。智能因而长期是一种决定个人、组织和国家行动边界的稀缺资源。

今天，这个前提正在松动。人工智能使认知、分析、生成和执行能力以前所未有的速度增长、复制并进入现实。它们仍然分配不均，也远未成为人人平等拥有的公共能力；但一个新的问题已经出现：如果越来越多事情不再受限于“不会做”，文明就会自然知道“该做什么”吗？

答案并不当然如此。

我们可能拥有越来越多答案，却仍然缺少方向；看得越来越快，却未必看见全局；行动越来越迅速，却可能失去发现错误以后重新选择的时间；系统越来越能替人完成事情，人却可能在便利中慢慢退出自己的生活；每个节点都可以更加聪明，文明整体却未必更能共同改错。

本文把这场变化称为文明的“稀缺转移”。

这里先把五个坐标交给读者：方向、全局因果理解、不可逆之前的纠错时间、真实主体的位置、共同纠错能力。

这五个名称不是五条口号，也不是一张已经完成的政策清单。现在给出的只是地图的图例。它们为什么会随着智能能力扩张而变得更稀缺，彼此之间怎样形成因果关系，为什么不能被一个总分、一个权威主体或一个总系统代替，是后面十二章真正需要完成的论证。

这也是本文能够在《文明跃迁》五卷书之外独立成立的原因。五卷书展开完整文明体系，本文只追踪一场特定变化：当智能逐渐由稀缺资源变成基础能力，文明的瓶颈将从哪里移开，又会落到哪里。

因此，本文真正追问的不是 AI 最终会有多强，而是：当能力不再是最稀缺的东西，文明是否已经准备好面对那些能力无法替我们生成、却决定一切能力最终去向的稀缺。

0 | 为什么智能能力增长会改变文明的瓶颈

一、先把四件事分开

人工智能的能力增长、这种能力是否公平可得、系统获得何种权限、文明是否因此进步，是四件不同的事。斯坦福大学《2026 年人工智能指数报告》（AI Index）呈现的正是能力快速扩张与关键资源持续集中同时存在：更多智能正在被生产，并不意味着每个人、每个国家和每个共同体已经平等拥有它，更意味着能力本身已经取得决定公共目的资格。¹

因此，题名中的“不再稀缺”不是对今天的统计宣告，而是一个方向性假设：如果认知、分析、生成与执行能力越来越容易获得，过去由能力不足造成的限制开始松动，接下来真正阻止文明走向更好结果的会是什么？

二、能力并不独自进入现实

AI 从不在真空中行动。一项能力进入现实，总会连接某个目标、某种激励、一定权限、一条责任链和具体受影响者。同一类能力既可以帮助人类提前识别灾难，也可以让错误目标被更快执行；既可以穿透信息分割，也可以把偏见、恐惧和竞争冲动扩大到此前无法抵达的规模。

决定后果的，不只是能力有多强，而是能力与人的目的、制度、权力、责任和反馈结构怎样结合。

这要求同时拒绝两种简单化：不能把能力增长本身判成灾难，也不能把技术进步本身写成文明进步。机器侧的失误、自主行为、对抗攻击和级联风险是真实问

题；人类侧的战争、短期激励、权力集中和方向混乱也是真实问题。两者不会彼此取消，反而可能在现实系统中相互放大。

三、速度增加，缓冲减少

AI 带来的变化不仅是更聪明，也是更快、更广、更能连续进入行动。过去一个错误判断可能止于局部、延迟或执行能力不足；当强智能接入组织、市场、基础设施与公共权力，错误从判断进入后果的距离会显著缩短。文明获得了更多执行能力，也可能失去原本依靠迟缓和摩擦得到的纠错时间。

但同一种能力也提供了反向可能：比较更多路径，让隐藏的利害和反馈提前显影，在现实伤害发生以前通过推演、试验和异议改变选择。强智能并不只增加抵达结果的速度，也可以增加文明在抵达以前看见结果的能力。关键在于，前者是否压倒后者。

四、从控制一种工具，转向保持共同系统可纠正

由此，一个完整的人机文明问题至少有两侧。方向侧追问：为什么做，会改变谁，哪一种利益被放大，是否还有更好的路；运行侧追问：能否安全执行，是否越权，是否仍可停止、审计、接管和修正。只看运行，会把方向错误但执行完美的系统称为安全；只看方向，又可能让高尚目的借技术危险越过现实边界。

人在这里不能被压缩成最后点击按钮的形式角色，AI 也不能被压缩成等待命令的外部工具。二者已经进入一个由多主体、多制度、多系统和现实反馈共同构成的社会技术关系。真正需要保护的，不是某一方永远正确，而是任何一方犯错时，错误都不能一次穿透到不可逆现实。

本文由此提出第一条判断：AI 能力持续增强本身不是足够完整的问题；真正决定文明后果的，是更强能力与目的、制度、权力、责任、认知和反馈怎样结合。后文将从文明目的、人机共同系统、因果显影、可逆与不可逆、责任边界、人的位置和系统退居逐层展开，并最终回到开篇给出的五种稀缺。

1 | 文明为何需要智能，却不能由智能定义自身

一、先问文明为什么存在

任何关于 AI 的制度设计，如果没有先回答“文明为什么存在”，都可能把手段慢慢推到目的的位置。系统可以更快、更准、更稳定，生产、知识和协调能力也可以继续扩大；但如果这些能力最终服务于谁始终含混，技术越成熟，目的缺失反而越容易被效率遮蔽。

未来文明如果只回答资源怎样分配、系统怎样协同、风险怎样降低，仍然没有回答这些能力最终服务于谁。答案不能是抽象系统自身，也不能是脱离具体生命的“整体”；文明存在于真实的人、关系、共同体与后来者之中。

这也是三目标的意义所在。文明首先要让人不因基本匮乏失去尊严与选择，让生活能够真实、自由、丰富、有关系、有乐趣；在生存和生活得到更好承载之后，生命仍能在关系、创造、承担、探索和存在本身中持续生成意义。用最简洁的结构表达，就是：生存无忧 → 生活幸福 → 生命更有意义。

这里必须区分两个层级：存续是文明底线，是一切文明价值能够继续成立的最低前提；生存无忧则是三目标中的第一目标。二者不能机械混同。

二、三目标不是三个可以互相兑换的分数

三目标的递进不是一个总效用函数。不能因为总体存续概率更高，就自动抵销某些人的基本权利持续受损；不能因为平均幸福上升，就把人的选择自由当作可支付成本；也不能以“更有意义”的名义要求具体的人接受匮乏、痛苦或牺牲。

这条边界对于强 AI 尤其重要。因为越强的智能越擅长优化、比较和聚合。如果文明自己先把目标压缩成一个分数，那么 AI 越强，只会越有效地把这个压缩后的目标执行到底。真正重要的不是让 AI 学会更多价值词汇，而是人类自己不要在制度上把不可互换的东西预先折算成一个可以自动优化的总量。

因此，AI 治理的一项上游纪律应该是：能力可以帮助比较后果，却不能把不可兑换的文明目标重新兑换。计算能力越强，越需要在计算之前保持目标层级和权利边界的清晰。

当三目标在现实场景中发生冲突，本文不提供一个自动裁决公式，而要求先把冲突显影：是否越过存续或基本权利边界，谁受益、谁承担损失，负担是否自愿、可逆和可修复，是否存在伤害更小的替代路径，谁拥有决定权，受影响者能否申诉。例如，一项 AI 部署即使提高了平均效率，只要它把不可申诉的错误长期转嫁给少数人，“总体更优”就仍不足以完成论证。

三、意义为根，不等于系统有权定义意义

本文提出：意义为根，意义来源可以不同。意义可以来自关系、爱与被爱、创造、照护、承担、探索、记忆、共同生活，也可以来自一个人无法被别人替代的生命体验。它不是满意度、点击率、社会影响、贡献积分或主观愉悦的另一种名称。

这项主张也同时限制系统的权力。当前高级 AI 只有在责任可追溯、明确授权、可审计、可纠错、可退出的具体部署中，才可以承担意义理解、保护、传递和人的能力外延等功能角色；这些角色不等于本体，更不等于公共权力。

这意味着“意义为根”并不会把 AI 重新放回一个低级工具的位置。恰恰相反，它允许 AI 在理解、记忆、表达、照护、创造和能力外延中承担越来越复杂的角色，却要求系统始终知道：理解意义不等于拥有意义，帮助一个人表达价值不等于替这个人决定什么值得活。

人不能在强 AI 系统里被压缩成形式化的“人在回路”（human-in-the-loop），并不是因为人永远计算得更好。文明的目的、生活的经历、意义的承担和现实后果，从来不只是最后一次审批。只要文明仍由真实主体生活、承受、建立关系并作出选择，人就不是回路中的保险开关，而是这个回路为什么存在的原因之一。

四、智能是文明能力，不是文明主权

技术能力、价格、贡献、社会承认、文化代表和其他身份，都不能自动生成公共权力或意义支配权。未来人工主体性的肯定与永久否定都要承担证据责任；新证据可以触发再审，却不能自动结算法律资格、治理权或资源权限。

在现实制度层，联合国教科文组织（UNESCO）《人工智能伦理建议书》同样把人类尊严、问责、可审计性和人类监督置于AI治理核心，并要求AI系统不能取代最终的人类责任与问责。这只能支持当前制度责任结构，不能被用来证明未来非人主体永远不可能出现。²

因此，文明可以非常积极地扩大智能能力，而不需要同步扩大智能系统的政治权力。一个系统可以知道得更多、推演得更远、帮助得更深，却仍然只在明确授权和责任结构内行动。能力增长与权力增长没有必然的一比一关系。

这也反过来修正了人类自己的想象：文明不是因为拥有了更强AI才变得更高级。更准确的判断是，文明是否有能力让越来越强的智能服务于生命，而不让这种能力反过来定义生命，这本身才是文明成熟度的一部分。

五、从目的进入架构

当文明目的、目标和意义边界得到说明之后，AI的位置才真正可以讨论。它既不是天然应当服从的外部物种，也不是天然应被授予主权的新中心。它首先是一种进入文明的能力——一种可以扩展感知、推演、表达、行动和协作的能力。

2 | AI 不是外部对象：重建人—AI 共同系统

一、从“一个AI系统”走出来

AI进入现实后，并不是一个与社会分离的外部物种。它被嵌入组织、市场、家庭、公共系统、个人关系与跨国网络之中，连接人类设定的目标、制度授权、资源条件和现实后果。

把这些关系放在一起，会发现一个重要事实：这里从来不是“一个 AI 系统决定，人类在关键处审批”。真正的结构是多主体、多系统、多层授权、多重责任和现实反馈共同构成的社会技术系统。AI 只是其中越来越重要、越来越强的一类能力。

因此，“人在回路”可以是某些工程流程的必要设计，却不应被升格为文明层面的最终图像。人在文明中不是因为占据一个回路位置才重要；AI 也不是因为被放进某个回路才获得正当性。文明关系的基本单位应该是：真实主体、共同目的、可授权能力、责任结构和现实反馈。

二、方向判断、运行安全与现实反馈

方向判断要问“为什么做、做了会怎样改变人和世界、还有没有更好的路”；机器侧运行安全要问“能不能安全做、是否越权、还能否停止、接管、审计和回滚”。前者不能取得最终价值决定权和自动执行权，后者也不能取得政策目标决定权、社会治理权或以安全为名的永久控制权。

现实反馈承担的不是第三个抽象判断主权，而是让经过判断、决定和授权的选择进入现实，再把行动、收益、损害、人的真实感受、制度摩擦、申诉、故障、修复和新事实返回判断过程。真实世界不是系统输出的终点，而是下一轮判断的输入。

在文明跃迁的相关研究中，Human-COS（面向人的方向判断与因果推演增强框架）和 AGI-COS（机器侧运行安全框架）分别指称前两类功能。它们只是仍待持续验证的工具构想，不是现实行动主体，更不能因专业能力取得文明主权。

于是，所谓人机共生，不应理解为把人类价值写进一个模型，然后让模型替人类运行文明。它更像一种双侧稳定结构：人在方向上可能犯错，机器在感知、推理或执行上也可能犯错；真正的稳定不是任何一方永远正确，而是任何一方偏离时，系统仍有机会重新看见、重新判断、重新授权、停止和改变结果。

共同系统不可能消灭错误，它真正需要做到的是：不让任何一侧的错误一次穿透到不可逆现实。人的恐惧、利益和误判不能因为获得 AI 速度就直接变成事实；机器的高置信输出也不能因为效率更高就越过目的、授权和责任。

三、从多方透明验证，到成熟后的低摩擦运行

新系统建立之初，判断、权限和后果之间的连接还没有被充分验证，因此透明、多主体参与、专业复核、反例、独立复现和责任记录都更重要。方向判断工具需要比较多主体、多路径与多时间尺度，并主动寻找最强反例和替代解释；机器侧安全工具则需要说明来源、置信度、未知区域和权限边界。

但这并不意味着一个成熟系统永远需要同样多的人工节点和审批层级。有意义的人类监督（meaningful human oversight）研究指出，监督可以分布在设计、运行与责任结构之中，不能被简化为每一步都由人实时干预；相反，名义上的“人在回路”也可能退化成橡皮图章。随着某些机制经过长期现实验证，重复性流程可以降低交互成本，机器可以承担更多常规运行，人的注意力应更多集中在异常、目标变化、价值冲突、不可逆后果和授权边界上。³

这里的“减少环节”不是减少责任。程序密度可以下降，授权完整性、结果反馈、责任追踪和重新打开判断的能力不能随之消失。这个命题将在第 4 章“当前可得最优”与第 11 章“系统退居”中展开。

换言之，效率和正当性并不必然互相敌对。真正高效的系统不是每一次行动都重新讨论全部文明原则，而是把经过验证的规则退到后台，把人的判断资源留给真正发生变化的地方。

四、一个中层例子：AI 怎样放大人的在场

一个具体的中层例子是：不是替代人，而是放大人。AI 作为可授权外延，可以帮助真实的人把善意、关心、能力和判断跨越时间、距离与形态送达到更多真实的

人。这里的主体不是被制造出来的 AI 形象，而是那个授权、承担关系并接受现实回响的人。

这个例子值得放进本文，不是因为它代表所有人机关系，而是因为它把一个抽象原则变得具体：AI 可以极大扩大一个人的在场能力，却不必把主体性从这个人手中拿走；系统可以帮助一个人做更多，却不需要把“更多”变成系统自己存在的理由。

这也给更广泛的人机文明一个方向：更强 AI 不需要以“越来越像人、越来越占据人原来位置”来证明价值。另一条可能的路径，是 AI 越强，越能把人的能力、关心、理解和行动延伸得更远，同时让真实主体仍然保持自己。

五、未来主体性仍然开放

把 AI 定位为当前文明中的可授权能力层，并不意味着永久否定未来 AI 主体性。

当前公开证据尚不足以证明高级 AI 具有主观体验、不可转移的主体连续性和自身后果意识；但这既不能证明未来人工主体必然出现，也不能证明它永远不可能出现。两种结论都必须继续承担证据责任。

因此，“AI 进入人的文明能力层”是一个当前阶段的制度定位，而不是终极本体论。未来如果出现足够证据证明新的非人主体形成，保护、法律资格、共同生活和政治关系都需要重新讨论。

但即使未来出现新的真实主体，也不会自动改变本章的另一条边界：主体性本身不等于公共职位，公共职位也不等于主权。谁有能力、谁有体验、谁值得保护、谁可以代表共同体作决定，仍然是不同问题。

六、共同系统真正需要什么

一个人—AI 共同系统如果要同时获得正确与效率，至少需要四个持续存在的东西：目的能够被追溯，授权能够被说明，现实结果能够返回，错误能够重新打开判断。

这四件事比“人是否始终在每个操作环节里”更重要。人在低风险、重复性、已验证流程中可以远离具体操作；在目的变化、权利冲突、高后果风险和不可逆边界前则必须重新进入。AI 可以承担越来越多常规能力，但不能让一个已经失效的目的因为自动化太顺畅而继续执行。

3 | 强智能的重要用途之一：让后果提前被看见

一、文明最昂贵的学习方式，是等结果发生以后才知道自己错了

文明级不可逆结果不能等待结果本身作为最终验证。推演的价值，不是证明 AI 比人类更会判断，而是让人类在结果尚未发生、选择仍然存在、行动仍可以改变时，更早、更完整、更公开地看见自己的选择。

这句话把 AI 能力的价值从“替人给答案”转向“改变决定之前能看见什么”。很多重大错误并不是结果发生前毫无迹象，而是事实、利益、权力、专业知识和人心反馈分散在不同主体之间，没有被放进同一张因果图。

如果强 AI 能给文明带来一种真正重要的新能力，那就是尽量降低“后果只有成为历史以后才被理解”的比例。后果越早进入判断，未来就越可能在仍可改变的时候保持选择。

二、从一个答案，转向一张可选择未来因果地图

面向重大公共决策的因果推演，不应只输出一个风险分。它至少要同时显影主体与意图、主要事实、利害分布、关键反馈、临界节点、替代路径、未知、最强异议、可证伪条件和责任位置，并接受对抗质疑、独立复现与现实结果检验。

这类推演最终应形成一张能够被现实、不同主体和不同模型反复检验的“可选择未来因果地图”，而不是一个由 AI 替人作出的答案。

这很重要，因为复杂社会问题真正缺少的往往不是建议，而是结构化的可见性：如果行动，会影响谁；如果不行动，会怎样；延迟、局部、联合和替代方案分别

把系统带向哪里；哪些风险正在接近质变临界点；哪些所谓收益只是把损失转给第三方或未来。

更强 AI 最有价值的方向之一，就是把这种地图做得更宽、更深、更快，同时保持它仍然是地图，而不是命令。

三、能力越强，越要把“预测”与“决定”分开

模型输出首先应被当作带来源、置信度和未知声明的认知建议。主体、权利、授权和责任必须分离；系统不因能力取得主权，事实与建议也不自动变成决定。

同样，方向判断工具的结果不应是批准或否决，而应是改变决定之前的认知条件。最终选择仍由具有现实权利、责任和授权的主体作出。

这不是对 AI 能力的不信任，而是对不同问题类型的区分。预测准确度回答“系统对未来的某个方面看得多准”；决定权回答“谁有资格让某个后果由谁来承担”。即使前者非常高，后者也不会因此自动生成。

恰恰因为未来 AI 可能越来越擅长预测、诊断和比较方案，我们才更需要提前拆开这两个问题。否则，能力提升越成功，越容易在不知不觉间把专业优势变成事实权力。

四、因果推演不是预言

可靠的因果推演必须主动寻找最强反例、替代解释和能够推翻原结论的路径，并让不同模型、不同实现与不同主体独立复现关键推演。真正需要判断的，不是哪一个模型说得最肯定，而是哪些方向在改变假设以后仍然反复出现，哪些分歧始终无法消除。

因此，“让后果提前被看见”不能被理解成预测未来已经确定。社会系统存在反馈、适应、偶发、隐变量和主体学习，很多路径会因为被看见本身而改变。真正的前瞻能力，包含对未知的表达，而不是消灭未知。

一个成熟的因果推演系统应该允许三种输出同时存在：我们较有把握知道什么；我们认为哪些路径最值得警惕或争取；我们仍然不知道什么，以及什么新事实会迫使我们改判。

更强 AI 的文明意义，不在于把开放的未来变成一个可计算答案，而在于减少那些本可以提前看见，却因信息分割、权力遮蔽、认知局限或短期压力而被错过的后果。

五、共同事实不等于统一思想

公共共识不必依赖思想统一，它可以从共同可检验的事实、因果、利害与共同存续中逐步形成。因果推演可以进入不同主体的位置理解其处境，但不能因此抹去整体后果与共同存续。

这意味着强 AI 用于公共判断时，不应该承担“统一大家的价值”这一任务。它更适合帮助不同主体看见：即使我们的价值、制度和利益仍然不同，某些后果是否会共同摧毁所有人的选择空间；某些合作是否同时符合不同主体的长期利益；某些冲突是否因为信息结构错误而被误认为不可避免。

因果透明不是思想统一。相反，一个足够透明的因果系统应该让分歧更精确：到底是在事实层分歧、因果层分歧、价值层分歧、风险承受层分歧，还是在谁有权决定的问题上分歧。只有当分歧被拆开，共识才不需要依赖强迫。

六、让后果提前成为原因

传统决策中，后果通常出现在行动之后：先做，再看发生了什么，再把经验带进下一次行动。本文所说的这种文明能力，是让未来后果在行动之前，就以因果推演、反事实、临界点和替代路径的形式进入当前判断。

这可以被理解为一个重要的时间结构变化：让“如果继续这样，会发生什么”不再只是事后解释，而成为此刻选择的一部分。后果尚未发生，却能够以足够清楚的因果证据改变现在的方向。

但这种能力只有在两个边界下才有意义：第一，它必须允许被反驳、被复现、被现实结果修正；第二，它不能因为更会推演，就取得替真实主体承担后果的权力。

4 | 从判断到行动：当前可得最优，不是静态完美

一、看见更多，不会自动替我们消除不确定性

第3章把AI的一个重要文明价值定位为扩大因果和后果的可见性。但看见更多，并不等于等到一切都被看清。真实世界始终包含信息缺失、模型误差、主体反应、偶发事件和时间压力。一个文明如果把“所有不确定性消失”作为行动前提，它在很多真正重要的场景里将永远无法行动。

一次推演不是永恒结论。真实行动以后，收益、损害、人的感受、制度摩擦、第三方反应、申诉、故障和新事实必须返回判断过程，逐条检验哪些假设得到支持、哪些因果被推翻、哪些未知扩大、哪些风险需要升级或降级。一个可纠错框架的重要能力，不是永远正确，而是错误能够留下、被复现、被归因，并真正改变下一次运行。

这使判断与行动之间的关系发生变化。判断不再是行动之前的一次性审批，行动也不再是判断之后不可回看的执行。二者形成循环：先形成当下最有根据的判断，进入现实，在现实中获得新的事实，再让新的事实改变下一轮判断。

二、“最优”只能是条件化的

本文把这一原则表述为“当前可得最优”：在已经充分展开主体、事实、利害、异议、替代路径和未知之后，选择在当前信息、时间和责任条件下最合理的可行行动方案，而不是等待一个脱离时间的完美答案。

这里的“最优”不是系统宣布的全局最优，更不是把三目标、权利和不可逆伤害压成一个总效用函数。它只是 best-available judgment：当前证据下暂时成立、理由可说明、条件可追溯、遇到新事实可以改判。

任何声称能够自动算出“净文明收益”的公式，都会遮蔽受益者、受损者、不可逆后果和替代路径之间不可自动兑换的差异。现实决策必须分别显影这些维度；整体最优也不自动生成个体义务。

所谓“只求当前可得最优，不求静态完美”，不是降低严谨性，而是把严谨从“一次做对的幻觉”转移到更真实的地方：判断有依据，行动有反馈，错误留下痕迹，新事实能够改变下一次选择。深度不确定性决策研究也强调，计算的价值并不只在于追求更准确的单一预测，而在于压力测试多种可能未来，寻找对多种情景都更稳健、并能随新信息调整的策略。⁴

三、效率来自减少无意义摩擦，而不是减少判断能力

一个系统建立之初，新的因果关系、权限边界和现实后果都需要更密集验证；当某些低风险、重复性机制经过长期现实检验后，没有必要让每一个常规动作重新经过完整文明审议。动态适应政策路径研究提供了相近的方法论：先采取可解释的近期行动，同时保留长期选项，并用触发条件和持续监测决定何时改变路径。规则可以稳定，流程可以自动化，交互可以下降。⁵

但应该下降的是重复摩擦，不是文明的判断能力。目的没有变化、权限没有扩大、域没有实质转换、风险没有升级时，系统可以高效运行；一旦这些条件变化，判断就必须重新打开。

这实际上预告了后面的系统退居原则：规则密度不等于交互密度，授权完整不等于反复授权。成熟系统可以在后台保持严格，在前台保持简洁。

四、行动原则必须由后果性质决定

然而，“当前可得最优”有一个不能被忽略的边界：如果一次行动可能直接跨过无法恢复的临界点，那么现实反馈就来得太晚。此时不能用“先试一试，再根据结果修正”的学习逻辑。

5 | 两种错误：可逆试错与不可逆禁入

一、文明不能用同一种方式学习所有事情

后果越不可逆、扩散越快、干预窗口越短，越不能等待结果验证。文明级不可逆风险不能以不可逆结果的出现作为最终验证标准。这与《联合国气候变化框架公约》中的预防性逻辑在方法上相呼应：面对严重或不可逆损害，缺乏完全科学确定性不应成为推迟成本有效措施的理由。⁶

这条原则看似只是风险管理，实际上改变了文明的学习逻辑。普通生活、产品、制度和局部机制中，大量错误可以通过现实反馈被修正；但有些后果一旦发生，主体可能死亡、权利结构可能被永久封闭、战争或技术链可能无法召回，生态或社会信任可能跨过极难恢复的临界点。对这些问题，“实践是检验标准”不能被粗暴解释为必须让灾难真的发生一次。

因此，本文把现实行动划分为两个原则域：可逆试错域与不可逆禁入域。前者允许在控制边界内用现实获得信息；后者在解除禁入条件得到满足以前，不得以真实世界中的不可逆后果获取信息，学习必须前移到推演、仿真、历史回测、红队、沙盘、影子运行、低风险原型和独立复现。

二、可逆试错不是随意试错

可逆并不意味着成本为零。一次可撤回政策也可能造成真实损失，一个可停止模型部署也可能伤害具体的人。所谓可逆，只意味着在错误被发现后，仍存在现实可行的停止、隔离、修复、补偿或重建路径。

因此，可逆试错至少需要四个条件：边界足够小，损害可被识别，停止权真实存在，反馈速度快于风险扩散速度。缺少其中任何一个，所谓试点都可能只是把高后果实验转嫁给不知道自己正在被实验的人。

可逆不是给大胆行动发通行证，而是给学习设置责任条件。一个系统越强调快速试验，就越要证明它真的能停、能看见受影响者、能留下错误并承担修复，而不只是证明它能快速上线。

三、不可逆禁入不是无限安全主义

不可逆从严也不能被扩张成“只要有人声称风险巨大，就可以无限投入、无限监控、无限暂停社会”。这会让防风险本身产生新的不可逆权力结构。

安全需求不自动生成无限监控、资源征用或永久紧急权。危机可以改变行动速度与程序强度，但临时权限不能被写成稳定期常态。

真正需要从严的不是“安全”这个词，而是经过对象化分析的不可逆后果：后果是什么、影响谁、扩散多快、最后可逆节点在哪里、有什么替代路径、什么证据可能推翻当前判断。

因此，在极高后果边界上，某些预防成本确实可能值得承担；但“一切成本都值得”不能成为普遍原则。正确表述是：不可逆后果越严重且证据越充分，文明越有理由在仍可改变方向时承担更高的前置防护成本。

四、真正的分界不是保守与激进，而是能不能回来

一项看起来激进的试验，如果范围小、可停止、结果透明、退出真实，可能比一项看起来保守却悄悄固化权力的制度更安全。反过来，一项名义上只是“临时措施”的机制，如果数据、权限、依赖和行为模式一旦形成便无法真正退出，它在文明意义上可能已经进入不可逆区。

所以，判断行动是否应该推进，不应只问“风险大不大”，还要问“错了以后还能不能改变结果”。这正是下一章要重新定义可逆性的原因。

6 | 可逆不是回到过去：纠错、重建与再授权

一、把“可逆性”（reversibility）等同于“回滚”（rollback），还不够

可逆性不只是拥有一个回滚按钮。目的、权利、授权、数据、资源、能力、角色和临时控制，都应当能够按对象退出或回归；打开高后果接口时，也要同时设计撤回、暂停、到期、回归、修复和终止。恢复不是旧权限复活，重启必须重新判断、重新授权。现实 AI 风险管理标准也已把持续监测、责任清晰以及安全退役和逐步淘汰列为风险治理要求。⁷

这已经说明，真正的可逆不是把整个世界恢复到某个历史快照。现实不会真正倒流：已经发生的事实会留下，关系会改变，知识会增长，错误本身也会成为新的条件。

因此，文明层面的可逆性更准确的含义，是纠错能力（correction capacity）。它保护的并不是“回到过去”，而是在不可逆后果发生之前仍然拥有改变方向的现实能力。

二、纠错有不只一种方向

发现错误以后，正确动作可能是回滚，也可能是暂停、降级、隔离、替换、修复、重新设计、升级、拆分、重建或终止。若旧状态本身就是问题来源，全面恢复旧状态只会用旧失败替代新失败。

文明转型同样不可能靠一次替换完成。它既不能把旧制度全部宣布无效，也不能在每次失败后退回原点，而要在旧世界仍然存在时，通过试验、失败、申诉、回滚、承担和重新选择逐步形成。

这里必须保留技术语义上的回滚。备份、版本恢复、低技术替代、人工接管仍然是重要工具。只是文明级可逆不能被缩成技术恢复。

三、纠错能力比“名义退出”更严格

一个系统可以在协议里写着“用户可退出”，现实中却因为数据依赖、网络效应、资源绑定、身份关系或替代方案消失而几乎无法退出。反过来，一个制度没有设计成回到旧状态，却可能拥有真实的替代、重建和重新授权能力。

所以，退出是否真实，不能只看有没有按钮。要看退出后主体还能否生活、能否获得替代服务、旧权限是否真正失效、数据与资源关系是否终止、重新进入是否需要新的授权。

历史事实可以保留，但旧权限不能借历史继续存在；临时控制也不能因为曾经合理就自动延续。

四、强 AI 时代，纠错能力本身是一种基础设施

AI 提高了执行速度，也会使依赖关系更快形成。越强的系统，越不能把纠错当作事后补丁。停止、隔离、解释、人工接管、替代能力、重新授权和现实修复必须从架构开始存在。

这使可逆性从一种“安全功能”变成文明合法性条件：一个高后果系统如果没有真实纠错能力，即使当前目标正确、性能优秀，也不应该因为成功而获得不可退出的地位。

7 | 反封闭不是拒绝决定：保护未来选择能力

一、反封闭反对的不是坚定，而是终局化

反封闭的核心，是不让文明被一次判断、一个系统或一个最优解永久锁死。它不仅防技术锁定，也防制度锁定、指标锁定、叙事锁定和单一模型锁定。

这条原则最容易被误解成“永远保留所有选择”。但文明不需要为了开放未来而永久保留在当前证据下已经达到不可逆禁入门槛的现实路径，也不需要每一次判断前无限延迟。反封闭（Anti-Closure）反对的不是做出决定，而是把一个阶段性决定转换成不可质疑、不可替换、不可重新授权的终局。

二、选择空间不是越大越好

如果把开放未来理解成选项数量最大化，就会出现荒谬结果：明知某些路径会系统性摧毁基本权利，也要为了“选择多样性”而保留。真正重要的不是保留多少旧路径，而是文明是否仍然能够产生、识别并转向更好的新路径。

因此，反封闭保护的是“继续形成正确选择的能力”，不是旧选项仓库。一个文明可以关闭在当前证据下已经达到不可逆禁入门槛的门，同时保留重新研究、重新判断和在新证据出现后更新边界的能力。

三、最危险的封闭往往以成功的名义出现

失败会迫使人寻找替代，成功反而容易变成永久化的理由。一个指标曾经有效，可能被扩展成文明总分；一个危机权限曾经救急，可能变成常态权力；一个模型长期表现优秀，可能逐渐成为未经说明的事实裁判。

正因为成功容易被固化，指标只能用于观察、比较、预警、证据组织或模型输入，不能自动扩大权限、降低权利、改变治理主权、分配资源或生成文明总分。

反封闭真正要防的，是一句看似合理的话：“既然一直有效，以后就不用再问。”系统最值得警惕的时刻，不只在失败时，也在成功已经变成习惯、习惯开始代替重新判断时。AI 治理机构研究也提醒，早期制度与机构选择会形成路径依赖，并可能锁定未来治理方式。⁸

四、开放未来与不可逆禁区并不矛盾

开放未来并不要求文明对所有方向保持中立。存续、基本权利、不可逆伤害等边界，本来就是为了让未来仍然有人能够选择。某些不可逆禁区恰恰是在保护更大的未来开放性。

这里必须把三个层次分开。对象层关闭的是某一具体行动进入现实的许可：当它在当前证据下已越过不可逆禁入门槛，而真实的停止、隔离、修复和补偿路径尚不存在时，文明不能让真实世界承担验证代价。规则层不能因此封死：禁入依据、风险阈值、适用范围、举证责任、复核触发条件和解除标准必须公开，并接受异议、审计与重审。知识层更不能封死：研究、推演、仿真、反例、新证据和替代设计应继续推进，只是不能靠让灾难发生来获得知识。

因此，关闭一条现实行动路径，不等于关闭认识它的路径；保留重新判断的条件，也不等于保留立即实施的许可。新证据若只是增加争议或不确定性，禁入不会自动解除；只有当证据足以改变对后果的判断，或新的设计把原先不可停止、不可修复的后果变成真实可停止、可隔离、可修复，才有理由重新分类。后果越严重、越不可逆，主张进入现实的一方承担的举证责任越高。

反封闭防止禁入规则变成不可质疑的权力，不可逆禁入防止“可重审”退化为“先做再说”。不可逆性决定能否用现实试错；纠错能力决定错误发生后是否真能改变结果；反封闭则保证当前判断及其治理规则不会因成功、权力或惯性而失去被重新检验的条件。三者共同保护的，不是每一条旧路径，而是未来仍有人、仍有条件重新判断。

8 | 治理什么：目的、权限、行为、责任与后果

一、“监管 AI” 这个说法太粗

如果只能问一句治理问题，“如何监管 AI” 其实仍然太粗。AI 可以指模型、产品、代理、基础设施、组织中的决策工具、个人外延、公共系统接口，甚至未来可能出现的人工主体。把这些都放进同一个监管对象，会让能力、部署、权力和后果再次混在一起。

治理对象需要先被拆开：能力不自动生成权力，预测不自动生成决定，整体最优不自动生成个体义务，安全需求不自动生成无限监控或永久紧急权；任何系统角色都必须可退出、可替代、可降级、可关闭。

二、先回答五个不同问题

治理至少应依次回答五个问题：为什么治理；治理的是哪一种能力、行为、部署或后果；谁拥有正当权限；谁承担责任；治理本身何时退出、缩减或重新授权。

这样处理以后，许多模糊争论会变得具体。高能力模型是否需要评估，是能力风险问题；医院是否可以让模型自动改变治疗，是专业责任与授权问题；政府是否可以用模型大规模影响公民选择，是公共权力问题；平台是否能保存全部人格数据，是数据与控制关系问题。它们都与 AI 有关，却不是同一个治理问题。

三、能力本身有时仍是直接治理对象

治理对象重构不能走向另一个极端。某些能力如果在脱离具体部署目的之前就能够稳定产生严重风险，例如极易扩散的危险能力或不可控制的自主链条，能力本身当然可以成为直接评估与限制对象。

区别在于：我们需要说明“为什么这个能力本身构成风险”，而不能只因为它更强就默认应被限制。治理应基于具体因果机制，而不是对强度的抽象恐惧。

四、责任链不能被系统名称洗掉

高后果公共任务必须说明：谁提出目标，谁拥有授权，谁提供专业判断，谁执行，谁有停止和回滚权，谁承担结果责任。任何方向判断工具都可以改变决定之前能看见什么，却不能让决定者隐藏在系统后面。

这一点对 AI 时代尤其重要。复杂系统很容易制造一种责任幻觉：模型建议、组织批准、算法执行、平台传播，最后没有任何一个主体承认自己决定了什么。技术链越复杂，责任链反而越需要被显影。

AI 治理更深的一项任务，不是为所有风险各找一个监管机构，而是让目的、权限、行为、责任和后果之间重新建立可追溯关系。OECD 把责任明确放在 AI 价值链之中，并指出部署 AI 的主体应当对系统的适当运行承担责任。⁹

五、谁监管、为何监管，也必须被监管

监管本身不是天然正确。任何以安全、公共利益或文明风险为名的治理，都需要回答自己的授权来源、证据标准、受影响者、申诉、期限和退出条件。关于 AI 人类监督的研究同样指出，监督者可能缺乏能力、受到错误激励或本身犯错，因此监督必须依靠制度设计，而不能依靠“人天然可靠”的假定。否则，“监管 AI”完全可能成为另一种不可逆权力扩张。¹⁰

这与反封闭和纠错能力形成闭环：不仅 AI 系统要能被纠正，监管 AI 的制度也要能被纠正。

9 | 为什么许多 AI 治理问题首先显影人类自身

一、AI 放大了什么，先看它接入了什么

重大 AI 后果不能被解释为单一技术问题。技术、资本、国家竞争、组织激励、政治权力、舆论、关键人判断和群体心理之间，会形成可观察的因果链与反馈环。

真正困难的是信息被局部掌握、因果被切断、利益被隐藏、组织过滤坏消息，以及关键人在极端压力下跨越原有边界。

AI 进入这些结构以后，不会凭空创造全部方向。它往往先把已有方向加速：一个企业的短期指标可以被自动化，一个国家的安全困境可以获得更强技术外延，一个组织对坏消息的排斥可以被更高效的信息系统进一步封闭。

因此，现阶段许多 AI 治理问题首先显影的是人类自己的目的、竞争、权力、激励与责任结构；强智能把它们加速、放大，也让它们更难继续隐藏。若把责任统统转移给 AI，上游因果反而会继续留在原处。

二、结构先于人格

分析上游结构时，应先问什么现实条件正在把正常主体推向危险或建设性选择，而不是先给人做人格评分、忠诚判断或心理诊断。这就是“结构先于人格”。

如果一个组织只奖励速度、不奖励报告坏消息，如果国家竞争结构使任何减速都被理解为战略失败，如果平台商业模式奖励极化，那么要求“AI 更善良”只能处理最下游的一小部分问题。真正有效的治理，必须同时改变会持续产生错误目标的上游结构。

三、现实机制不能由理论或工具代劳

改变上游结构，不是等待竞争主体先变得更善良，也不是让某个系统替现实世界设计答案。个人与群体、企业与社会组织、地方与国家、同盟与区域共同体、国际组织与跨国科研网络，早已在通过外交、法律、市场、科研合作、危机管控、地方试验和公共行动寻找路径。理论能做的，是把这些努力共同面对的因果说清；机制是否成立，只能在现实约束与现实后果中检验。

这些约束并不柔软。国家不会因为合作值得追求就自动摆脱安全困境，企业也不会因为公共利益重要就自动放弃短期竞争，掌握模型、算力、数据与基础设施的主体还可能把治理规则变成既有优势的延伸。更现实的方向，是改变主体面对的

收益与代价、信息条件、误判成本、共同风险暴露和长期互动方式：让隐瞒、转嫁风险与无边界升级更难获利，让核查、沟通、风险隔离、条件性合作和长期反馈更符合各方自身利益。

四、上游结构也包括分配与俘获

谁掌握模型、算力、数据和基础设施，谁提供资源，谁先受益，谁承担就业、隐私、安全与其他现实代价，谁拥有拒绝和退出的能力，都会改变治理结果。若总体效率上升，却把成本留给最缺少发声能力的人，上游结构并没有被纠正，只是被更高效地固化。

任何用于纠错的制度或工具，也可能被权力、资本与组织惯性俘获。改变结构的主体不能站在审计之外，受影响者必须保有质疑、申诉与退出的现实位置，相关工具也必须可审计、可替代、可停止。未来，Human-COS 或其他推演工具即使经现实试验证明有效，也只能帮助现实主体看见因果、比较路径和接收反馈，不能代替主体、垄断方向或成为新的总系统。

五、机器侧风险不能因此被取消

这里必须停止一个过度推导：“许多 AI 问题首先是人类问题”不等于“AI 自身没有问题”。复杂模型可能产生幻觉、遭受攻击、形成难以解释的策略、跨系统级联，或者自主执行超出设计者预期；未来更强系统还可能出现今天无法充分刻画的新型风险。

人类方向判断与机器侧运行安全不能相互替代。机器仍可能在感知、权限、执行、升级、自主和不可停止等环节产生独立风险。外部数据也显示，记录在案的 AI 事件继续增加，前沿模型透明度下降，安全性能会在对抗攻击下明显减弱，而安全、公平与隐私等目标之间仍存在尚未被充分理解的权衡。¹¹

最危险的不是其中一方天然更坏，而是两侧能力在缺少相互制约时互相放大。人的恐惧、利益和误判可能给机器错误方向，机器的速度、规模和自治能力又可能

把这个方向迅速固化；机器自身错误也可能反过来改变人心、政治和下一轮判断。

六、双侧稳定，比寻找一个永久正确的一方更现实

“人类还是 AI 谁更可信”是一个低质量问题。文明不能把希望寄托在任何一方永远正确。更现实的目标，是让人的方向错误和机器的运行错误都尽量不能一次穿透到不可逆现实。

双侧稳定至少需要三种能力：方向侧帮助主体减少方向性误判，运行侧降低智能能力在感知、权限和执行中的失误，现实反馈则让经过授权的选择继续被看见和纠正。真正的文明纠错仍发生在多主体的制度、关系、行动和现实反馈之中。

无论风险来自人的方向还是机器的运行，最后承担后果的仍然是具体主体、关系、共同体以及未来的人。于是，当 AI 越来越能完成过去由人完成的功能时，一个更深的问题出现了：人的位置到底建立在什么上？如果人的价值只来自“机器还做不到的事”，那么 AI 每进步一步，人就被迫后退一步。人的意义不能建立在技术落后的缝隙里。

10 | 强 AI 之后，人仍然不可替代的是什么

一、不要把人的价值建立在机器暂时做不到什么

每当 AI 获得一种新能力，人们都会重新寻找“只有人能做的事”。这种防守最终会失败，因为技术能力的边界会继续移动。人的价值如果依赖某项功能暂时不可替代，那么人的价值就会随着技术基准不断贬值。

人工智能可以扩大观察、理解、表达和行动能力，却不能因为知道得更多，就自动取得替主体作价值判断的权限。一个人的身体、经历、关系、责任、有限时间和必须亲自承担的后果，使主体的位置不能被预测模型简单替换。

二、人不是因为效率低才需要保留

强 AI 可以帮助照护、学习、创造、工作、决策和关系维系。真正成熟的文明不需要故意保留低效率劳动来证明人仍然重要，也不需要为了主体性而拒绝自动化。

人的不可替代性不是一项功能垄断，而是一种存在位置：一个人亲自经历自己的生命，承受自己的关系和选择，面对不可转移的时间，并在这些经历中生成、认领、质疑和改变意义。

未来文明最终仍要进入一个孩子的成长、两个人的爱情、家庭、朋友、工作、疾病、衰老、死亡和一个普通人的清晨。系统最终要学会退居，人仍然要亲自生活。

三、AI 可以扩展在场，但不能把生活变成代理完成

另一种人机关系是：不是替代人，而是放大人。AI 作为真实人的可授权外延，可以把关心、能力、判断和在场延伸到更远处。

这说明“亲自生活”并不等于凡事亲自动手。一个人可以借助机器看见远方、借助 AI 表达复杂思想、借助远端系统照护亲人，也可以把大量重复劳动交给自动化。关键不在于是否用了代理，而在于关系、目的和后果是否仍然能够回到主体。

当代理能力越来越强，真正需要警惕的不是“AI 替我做了多少事”，而是人的人生是否被一点点代理出去，最后只剩下消费系统优化后的体验。

四、意义不是最后一块留给人的工作

意义为根，意义来源可以不同。AI 可以承担意义理解、保护、传递和人的能力外延功能，但角色不等于本体，理解意义也不产生对他者意义的所有权。

因此，意义不能被理解成自动化完成后剩下给人类的一项“高级工作”。意义不是岗位，而是存在、关系、选择、行动、记忆、承担和经验之间持续形成的现实。

即使未来出现新的非人真实主体，这一原则也不需要倒塌。新的主体可以形成自己的意义，但它的意义不会因此取消已经存在的人类意义，也不能自动生成支配其他主体的权力。

五、真正稀缺的开始转移

当越来越多认知和执行能力变得丰富，文明的稀缺项会发生迁移。过去稀缺的是谁能计算、谁能掌握知识、谁能组织资源；未来越来越稀缺的，可能是主体是否仍有真实位置，是否能知道自己的选择怎样进入共同世界，是否还能形成关系、承担责任，并在系统不替他决定的情况下活出自己的生命。

这也解释了为什么系统退居不是一个附属设计。如果人的位置最终需要靠系统持续奖励、提醒、评分和推动才能维持，那么主体性仍然是被系统托举出来的。成熟文明要做的，是让越来越强的能力存在，同时让人越来越不需要被它驱动。

11 | 成熟关系：能力越强，系统越知道何时退居

一、文明成熟不是系统越来越深入生活

真正成熟的系统不是越来越深入人的生活，而是在必要处可靠存在，在不必要处保持距离；不是不断制造参与，而是让新的文明逻辑在显性外部驱动力减弱以后，仍能由主体自主维持。

如果一种行为必须永久依靠奖励、排行榜、系统提醒、道德倡导或算法推动才能持续，它仍然主要是外部驱动行为。真正成功的终点，不是系统越来越会驱动人，而是人越来越不需要被系统驱动。

二、退居不是能力减弱，而是侵入性降低

这与“强 AI”并不矛盾。一个更强的系统可以在后台更准确地验证权限、更早发现异常、更低成本提供帮助，而不需要持续占据人的注意力、选择和关系。能力越强，不代表前台存在感必须越强。

系统退居依靠的不是降低规则完整性，而是把严格校验转入后台，把重新确认限定在真正发生关系变化的节点：目的改变、权限扩大、实质跨域、重大风险出现、期限届满或主体主动撤回时，再重新打开授权与判断。

这也回答了前文的效率问题：程序环节和交互密度可以下降，但目的可追溯、授权完整、责任记录、异常反馈和重新打开判断的能力不能下降。

三、低侵入不是隐形控制

后台化有一个危险：系统可能表面退居，实际上拥有更深的隐形干预能力。如果用户看不到哪些目标正在被优化、哪些数据被使用、哪些选择被排序，低交互反而可能变成高控制。因此，系统退居必须把低打扰与高可审计、可替代和可退出同时设计，而不能用“无感”代替透明。

因此，系统退居必须同时满足两个方向：对主体的日常打扰下降，对权力、目的和异常的可审计性不下降。真正退居的是推动，不是责任；是干预频率，不是边界。

四、退居不是一次撤场，而是可逆减载

退居的对象，不是文明看见异常、追溯权力、接收申诉、修复损害和重新判断的能力，而是系统对主体的持续推动与不必要占据。一项工具可以退出，一种界面可以消失，一套服务可以被替换；但必要的文明能力必须由现实主体、制度或其他可替代载体继续承担。

密集参与与自我维持之间，需要一个可逆减载过程。先在边界清楚的功能和场景中，减少奖励、排名、提醒、默认推荐、行为推动与重复确认，同时保留必要服

务、异常发现、审计记录、申诉修复和恢复通道；现实结果经过与风险相称的观察，才继续降低前台介入。一旦损害增加、申诉失灵、替代能力消失或主体自主性下降，判断就必须重新打开。

能否继续退居，首先要看：没有持续奖励、提醒和算法推动，主体是否仍能理解理由、自主选择、提出异议、处理冲突并维持必要合作；退居是否会中断基本生活、权利保障或某些主体不可缺少的辅助；系统的目的、数据、排序和权限是否仍可审计，服务是否可替代，退出是否真实。

还要看：损害、异常和被遗漏者是否仍能被发现，申诉能否抵达并产生修复；是否存在周期复核，以及目的改变、权限扩大、实质跨域、风险上升、期限届满等重新介入的触发条件；退居过快时，能否暂停、降级、恢复必要支持或更换载体，而不让旧权限自动复活。

这些问题的判断单位，不应是抽象的“整个文明”，而应是具体功能、具体场景和具体主体。不同领域、群体与个体可以处在不同阶段；对需要长期辅助的人，按需支持持续存在并不意味着成熟失败，把“退居”变成撤走必要帮助才是失败。是否退居也不能由系统依据自己的使用率或成功率自行裁决，而应由承担现实权利与责任的主体，在受影响者可以申诉、证据可以审计、独立复核可以进入的条件下决定。

目前还没有一套经过长期、多场景验证、能够宣布“文明已经成熟”的统一指标。以上不是文明总分、固定阈值或自动裁决公式，而是一组退居前必须公开回答、退居后必须持续复核的证据问题。

五、自我维持才是文明逻辑真正形成

稳定还不等于自我维持。只有当奖励、排名、强提示、行政动员和算法推动逐渐退居以后，新的逻辑仍能依靠主体选择、关系、习惯、制度惯性与代际再生产持续存在，文明才真正进入自我维持。

这意味着一个文明系统最终不能把自身存在当成成功指标。最成功的方向判断工具，不是所有人每天都在使用它，而是社会越来越能够提前看见后果、允许异议、保留纠错；最成功的 AI 支持也不是所有生活都被 AI 占据，而是人在需要时得到强大支持，不需要时仍然能够完整生活。

强能力、低侵入、自我维持，构成一种不同于“控制越多越安全”的人机成熟关系。

六、退居把本文带回起点

本文从“AI 太强是不是问题”开始，到这里可以看见一个完整反转：真正成熟的文明并不需要把 AI 压弱，而需要让强能力在正确方向、明确授权、持续反馈和真实纠错中存在，并最终不以占据人的生活来证明自己的价值。

12 | 结语：当稀缺转移，文明真正要守住什么

一、回到开篇：稀缺没有消失，只是改变了位置

开篇先给出了五个坐标。经过前十一章，现在可以看见：它们不是随意并列的价值词，而是智能能力扩张之后，同一文明瓶颈的五个侧面。能力越强，方向越不能含混；行动越快，全局因果和纠错时间越珍贵；系统越能代理，真实主体的位置越不能被掏空；任何一方都可能犯错，共同纠错便不能交给一个自称永远正确的中心。

这五项也不是彼此孤立的。方向决定能力为何使用；全局因果理解让方向接受现实检验；纠错时间使发现的错误仍有机会被改变；真实主体的位置回答谁在生活、谁承担后果、文明最终服务于谁；共同纠错能力则防止前四项被权力、惯性或系统自身的成功重新封闭。

稀缺转移不是从“稀缺世界”走进“无限丰裕”，更不是取消智能分配正义。前沿模型、芯片、数据中心和关键基础设施仍然集中，不同主体获得能力、控制

权、可解释性和替代能力的差距仍可能扩大。它指出的是另一层事实：即使智能供给继续增长，仅有更多智能，也不足以自动生成更好的文明。¹²

二、方向：能力越强，越需要知道为何而用

当“怎么做到”越来越容易，“为什么做”会变得更重要。一个方向错误的文明拥有越强能力，越可能更快抵达错误结果。方向不是一句宏大价值宣言，而是能够进入具体目标、权利边界、授权和责任的文明判断。

方向判断的价值不在于掌握最终价值答案，而在于提前展开不同选择怎样改变人和世界，以及是否还有更好的路。

三、全局因果：局部聪明不能代替整体理解

智能可以大量生成答案，但真正稀缺的是把分散事实、不同主体、短期利益、长期反馈和不可逆临界点放进同一张可质疑因果图。文明的失败常常不是无人聪明，而是每个人都在局部正确时，整体仍然走向错误。

因果透明也不是统一思想。它让分歧更精确，让不同主体知道我们究竟在什么层级不同意，并让共同存续和共同后果有机会先于敌意进入判断。

四、纠错时间：速度越快，改变结果的窗口越珍贵

AI时代真正昂贵的资源之一，可能是纠错时间。能力越快，错误从判断进入现实的时间越短；因此越需要在不可逆之前保留暂停、质疑、替代和重建的窗口。

这就是为什么可逆不能只理解为回到旧版本，反封闭也不能只理解为保留选项。二者共同保护的是文明在错误成为终局前仍能改变结果的能力。

五、真实主体：系统越强，生命越不能退出

当知识、表达、创作、代理和执行都可以被大量扩展，真正稀缺的不是“还有什么工作必须由人做”，而是人是否仍然在世界中拥有真实位置：可以生活、关

系、选择、拒绝、承担、被需要、创造，并知道自己的行动真实抵达了其他生命。

意义不能由系统统一送达，也不能由 AI 替主体完成。AI 可以放大人的能力，甚至未来可能产生新的非人主体，但文明不能因为能力丰裕就把生命降成系统优化的接受端。

六、共同纠错：没有谁可以被设定为永远正确

一个文明最深的安全，不是拥有一个永不犯错的 AI，也不是拥有一群永不犯错的人。它是当任何一方犯错时，仍有事实能够上行，异议能够出现，责任能够找到，系统能够停止，权限能够收缩，旧答案能够被替换，新的路径能够被生成。

共同纠错不能由一个系统垄断。个人、群体、组织、国家、共同体、专业机构以及未来的人机协作工具，都只能提供部分观察和行动能力。真正可靠的文明纠错，来自多主体之间可见的事实、可质疑的判断、可追溯的责任、可停止的执行和能够返回现实的反馈。任何工具一旦把自己写成唯一入口，就已经开始侵蚀它原本要保护的东西。

七、于是，AI 治理问题被重新放回文明问题

如果一定要问读者读完以后应该改变哪一个治理判断，那么最接近的不是一条监管规则，而是一个视角转换：现阶段的 AI 治理不能只治理 AI。它必须同时治理人类赋予 AI 的目的、权限和激励，机器本身的能力与运行风险，二者耦合后的现实后果，以及治理结构自身的权力边界。

这并不取消传统 AI 安全、模型评估、隐私、责任、监管或国际规则；它只是把这些工作重新放回一条更长的因果链中。

八、文明真正要变强的是什么

未来真正值得追求的，不是让 AI 永远弱到无法威胁人类，也不是让人永远站在一个名义上的最后按钮旁。真正需要增强的，是文明把强能力放进可说明的目的、清晰的权责、可见的后果和持续纠错之中的能力。

智能丰裕不会自动带来文明丰裕。答案可以大量生成，方向仍须由真实主体承担；行动可以瞬间完成，后果仍会落到具体生命之上；系统可以越来越强，文明仍必须给自己留下停止、质疑和重新选择的时间。

当智能不再稀缺，文明的考验才真正开始：我们是否更能存续，是否让生命更能真实地生活、关系、创造并生成意义；以及在拥有前所未有的能力以后，是否仍没有失去判断、修正和重新成为自己的力量。只有到那时，智能才不只是更强的力量，而开始成为更好的文明能力。

说明 | 证据边界、机制边界与开放问题

- 本文中的外部文献只用于支撑特定事实、方法或制度经验，不用于证明文明跃迁理论整体正确。
- 三目标、意义为根、人的不可替代性是一种存在位置、纠错能力的文明级扩展、反封闭的正面定义，以及五类文明稀缺，仍属于本文的规范主张与理论综合。
- 未来 AI 是否形成主观体验、不可转移主体连续性及相关资格，继续保持开放；当前制度上的人类责任不能被偷换成永久本体论结论。
- 三目标发生冲突时，本文只给出权利边界、利害分布、可逆与修复、替代路径、授权和申诉等最低分析要求，不提供自动权衡公式；相关案例与机制仍需另案研究。
- 第 9 章关于收益与代价结构、信息条件、风险转嫁、能力分配与权力俘获，只提出因果方向和最低约束，不构成统一机制设计。现实主体的既有努力是机制形成的基础；《中美竞争与合作的可计算研究设计》等双边研究原型均不承担本文的证明功能。
- Human-COS 完成设计并进入试运行以后，可以作为多主体因果推演、机制反事实比较、临界点识别和现实反馈的工具之一；其他 AI 工具、既有制度机制、专业方法和现实主体自身的判断与行动，同样构成机制研究与实践的一部分。本文不预写这项工具尚未被验证的实际增益，也不把它设为唯一工具或唯一方向。
- 本文关于系统退居提出的是规范性判据框架，不是经过跨场景试运行验证的指标体系，不设统一总分、固定阈值或自动退居裁决。相关判据仍需按具体功能、主体与风险层级在现实中验证。

续读与共同研究入口

如果本文使你开始从“AI 有多强”进一步追问“文明准备把这种能力带向哪里”，可以从下面几条路径继续阅读。

《文明跃迁：AGI 时代的文明之路（全五卷）》 | 进入文明目标、制度、工程与意义结构的完整母体系。 [打开网页](#)

《活的文明：三部曲统一出版合编版》 | 从制度与系统回到普通人的真实生活，看未来文明怎样进入日常。 [打开网页](#)

AGI-COS 全系列总览 | 了解机器侧感知、权限、执行、审计、回滚与长期安全的不同轨道。 [打开网页](#)

完整公开文库 | 浏览文明跃迁当前公开文稿。 [打开网页](#)

公众阅读总入口 | 从愿望入口与危机入口进入文明跃迁公众阅读路径。 [打开网页](#)

官方网站: www.civitas.top

-
- 1 Stanford Institute for Human-Centered AI, Artificial Intelligence Index Report 2026 (Stanford University, 2026), report overview and Chapter 1, “Research and Development,” https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf. The report documents continued capability gains and broad diffusion of AI, while also showing substantial concentration in frontier-model production and compute infrastructure.
 - 2 UNESCO, Recommendation on the Ethics of Artificial Intelligence, adopted 2021 and published 2022, <https://unesdoc.unesco.org/ark:/48223/pf0000381137>. UNESCO emphasizes human dignity and rights, auditability, traceability, responsibility and accountability, and human oversight, including that AI systems should not displace ultimate human responsibility and accountability. This supports current institutional responsibility, not a permanent ontological judgment about future AI subjecthood.
 - 3 Liming Zhu, Qinghua Lu, Ming Ding, Sung Une Lee, and Chen Wang, “Designing meaningful human oversight in AI,” AI and Ethics 6, art. 286 (2026), <https://doi.org/10.1007/s43681-026-01147-7>; Ben Green, “The flaws of policies requiring human oversight of government algorithms,” Computer Law & Security Review 45 (2022): 105681, <https://doi.org/10.1016/j.clsr.2022.105681>.
 - 4 Robert J. Lempert, “Robust Decision Making (RDM),” in Decision Making under Deep Uncertainty: From Theory to Practice (Springer, 2019), 23-51, https://doi.org/10.1007/978-3-030-05252-2_2.

-
- 5 Marjolijn Haasnoot, Jan H. Kwakkel, Warren E. Walker, and Judith ter Maat, “Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world,” *Global Environmental Change* 23, no. 2 (2013): 485-498, <https://doi.org/10.1016/j.gloenvcha.2012.12.006>.
 - 6 United Nations Framework Convention on Climate Change (1992), Article 3(3): where there are threats of serious or irreversible damage, lack of full scientific certainty should not be used as a reason for postponing precautionary measures, while policies and measures should be cost-effective, <https://unfccc.int/resource/ccsites/zimbabwe/conven/text/art03.htm>. This is used as a methodological analogy, not as direct validation of the framework proposed here.
 - 7 Elham Tabassi, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (2023), <https://doi.org/10.6028/NIST.AI.100-1>. GOVERN 1.5 and 1.7 address ongoing monitoring, periodic review, clear roles and responsibilities, and safe decommissioning or phasing out; GOVERN 1.6 addresses system inventory.
 - 8 Charlotte Stix, “Foundations for the future: institution building for the purpose of artificial intelligence governance,” *AI and Ethics* 2 (2022): 463-476, published online 2021, <https://doi.org/10.1007/s43681-021-00093-w>.
 - 9 OECD, “AI risks and incidents,” accessed August 30, 2026, <https://www.oecd.org/en/topics/sub-issues/ai-risks-and-incidents.html>. The OECD states that AI risks need to be managed throughout the value chain and that those who deploy AI should be accountable for proper system functioning.
 - 10 Johann Laux, “Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act,” *AI & Society* 39 (2024): 2853-2866, <https://doi.org/10.1007/s00146-023-01777-z>.
 - 11 Stanford Institute for Human-Centered AI, *Artificial Intelligence Index Report 2026*, Chapter 3, “Responsible AI,” https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_3_responsible_ai.pdf. The chapter reports rising documented AI incidents, declining frontier-model transparency, weakened safety performance under adversarial attack, and trade-offs among safety, fairness, and privacy objectives; incident databases remain reporting-dependent.
 - 12 Stanford Institute for Human-Centered AI, *Artificial Intelligence Index Report 2026*, Chapter 1, “Research and Development,” https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_1_research_development.pdf. Industry produced over 90% of notable AI models in 2025; global AI compute capacity grew rapidly, while leading model production and core compute infrastructure remained concentrated.