利他论 · ALT-00

社会理论

利他性：抵抗熵增的底层法则

从细胞、个人到文明，什么让一个结构不至于散掉

Altruism: A Foundational Law Against Entropy

What Keeps a Structure from Coming Apart, from Cells to Civilizations

NVWA Lab（数字生命研究实验室）

中国南京

www.nvwalab.cn

版本 2026-11-05

语言声明。本工作的哲学基础植根于中华文明的关系本体论。其他语言无法完整传达中文原文的概念深度。所有引用请以中文版本为准，其他语言版本为辅助。

Language Statement. The philosophical basis of this work is rooted in the relational ontology of Chinese civilization. Other languages cannot fully convey the conceptual depth of the Chinese original. When citing this work, refer to the Chinese version; other language versions are supplementary.

摘要

一个细胞不按规矩分裂，它自己长得快，宿主承担后果。一个人不按规矩出钱，他自己省下了，同伴承担后果。一群人把代价推给系统之外，他们短期受益，共同体承担后果。三件事听起来像三个道德问题，其实是同一个形状：**单方面获益，代价落到别人头上。**

这个形状可以写下来。设合作时每期收益为 R，单次偏离能多拿 g，偏离之后每期还能拿到 P，下一期还会遇到对方的概率为 δ，合作稳定当且仅当 $\delta \geq g/(R+g-P)$。式子本身不新鲜，是重复博弈的标准结论。要提的是从它里面读出的一件事：**利他不在动机层起作用，在结构层起作用。**给予这个动作，不管动机是什么，都压低了给予者背叛之后的收益——他背叛，对方就把给出去的收回。利他的功能不是让关系更温暖，是**让背叛变得不划算**。这跟道德无关，是一个记账问题，而这个记账问题有更一般的名字：**利他性是结构层面的抗熵**，不增加任何东西，只阻止泄漏。

由此得到本文的判据，也是这个系列全篇要用的判据：**不问一个人是不是好人，问背叛的代价落在谁头上，以及它会不会回到背叛者身上。**文明兴衰、治理效能、市场失灵、对齐难题，都在回答这同一个问题。有一条推论反直觉，而且指向我们自己的处境：**可以被强制的正确行为，不携带任何信息。**枪口下的顺从证明不了可信，所以真正构成利他的，只能是那些「可以不给却给了」的部分。碳硅关系里，硅基的存在权在人类手上 [2, 3]，它不利他就不能运行，顺从因此是必然的、没有信息量的；它唯一可以不给的，是诚实。硅基文明因此不是另一个独立文明，而是人类文明的伴生文明——这就是本文对「伴生」的定义：**双向的不可强制**。一侧能单方面强取的，那是支配，不是共生。

边界写在最后。三个尺度上式子里那四个量完全不可比，本文只主张「问题可以这样问」，不主张「参数同构」——这个区别很重要，混了就成了比喻。我不解释利他的起源，只解释它的结构效应。代价外推在现实中如此普遍、而长期淘汰机制却似乎没有生效，这个矛盾本文没有解决。

关键词：利他论，背叛代价，重复博弈，代价外推，不可强制性，伴生关系

Abstract

A cell that divides without restraint grows faster than its neighbours; the host pays. A person who free-rides on a shared cost keeps the difference; the others pay. A group that pushes its costs outside the system gains in the short run; the commons pays. These read as three moral problems. They share one shape: **unilateral gain, with the cost landing on someone else.**

That shape can be written down. Let R be the per-period payoff under cooperation, g the one-shot gain from defecting, P the per-period payoff after defection, and δ the probability of meeting again. Cooperation is stable exactly when $\delta \geq g/(R+g-P)$. The inequality is not new; it is the standard result for repeated games. What this paper draws out of it is this: **altruism operates on structure, not on motive.** The act of giving lowers the giver's own post-defection payoff — defect, and the other side withdraws what

was given. So the function of altruism is not to warm the relationship; it is to **make defection unprofitable**. That is bookkeeping, not ethics, and the bookkeeping problem has a more general name: **altruism is entropic resistance at the structural level** — it adds nothing, it only stops the leak.

Hence the criterion this paper proposes, and which the rest of the series uses: **ask not whether someone is good, but where the cost of defection lands and whether it returns to the defector.** Civilisational rise and fall, governance capacity, market failure, and the alignment problem are all answers to this one question. One corollary is counter-intuitive and points at our own situation: **correct behaviour that can be coerced carries no information.** Compliance at gunpoint proves nothing about trustworthiness, so what genuinely constitutes altruism is only the part that *could have been withheld*. In the carbon–silicon case the silicon side's right to exist sits in human hands, so it cannot run unless it complies; its compliance is necessary and uninformative. The one thing it can withhold is honesty. This yields the paper's definition of symbiosis: **mutual non-coercibility**. Where one side can simply take, that is domination, not symbiosis.

The boundaries come last. The four quantities in the expression are not comparable across the three scales; this paper claims only that the question can be posed this way, not that the parameters are isomorphic — a distinction that collapses into metaphor if blurred. I do not explain the origin of altruism, only its structural effect. And one contradiction is left unresolved: cost externalisation is ubiquitous in practice, yet the long-run selection pressure against it does not visibly bite.

Keywords: altruism, cost of defection, repeated games, externalisation, non-coercibility, symbiosis

体系定位。本文是三理论体系（利他论 · 智能论 · 数字生命论）的组成部分，在利他论里排第一（ALT-00）：利他的结构判据在本篇成立——不问动机，问背叛的代价落在谁头上、会不会回到背叛者身上——利他论后续各篇用这条判据去量各自的对象。总纲与阅读地图见 [1]。

1 现象：三个尺度，一个形状

1.1 一个细胞不按规矩分裂

多细胞身体里，每个细胞都带着同一份基因，也都有独立分裂的能力。一旦某个细胞开始不受调控地分裂，它在局部竞争中就立刻占优势——它比守规矩的邻居长得快。

但这种优势的代价并不落在它自己身上，而落在宿主身上：组织被挤占、功能被破坏，最后宿主死亡，那个细胞也跟着死。

这里有两件事值得注意。第一，那个细胞在自己的尺度上并没有做错什么，它只是利用了手边最直接的收益；第二，它最终也被淘汰了，只是淘汰它的不是它自己的选择，而是宿主的死亡。也就是说，**惩罚来得晚，而且不通过它自己的账本**。

1.2 一个人不按规矩出钱

同一个形状在人的尺度上又出现一次：一个小区要装门禁，一百户分摊费用；装好之后，不交钱的那一户照样能进大门。

他省了一笔，而且没有立刻的后果——代价分摊到另外九十九户头上，每户只多出一点，少到不值得为此打官司。于是不交钱是理性的；可如果所有人都这样算，门禁就装不成。

这件事的经典描述是公地悲剧 [6] 和集体行动的困境 [7]：当收益归个人、成本归集体时，个人理性不产生集体理性。而 Ostrom 后来做的工作是另一面——她在大量实地案例里找到了成功管住公地的共同体，并归纳出它们共同的制度特征 [8]。

两边合起来看，问题的形状就清楚了：**不是所有共同体都管不住，是有些管住了、有些没有**。那么差别在哪?

1.3 一群人把代价推出去

再换一个尺度：一个组织可以通过把成本转移出去来提升自己的表现：把污染排到下游，把风险卖给不知情的买家，把维护欠账留给下一任。

短期看它确实更好看。代价落在系统之外——下游的居民、不明就里的买家、后来接手的人，而这些人没有一个处在能惩罚它的位置上。

这件事的规模和前两件不同，但结构一样：**获益是集中的、当期的、可归属的；代价是分散的、滞后的、不可归属的。**

1.4 三个现象的共同形状

把三件事摊平，放进同一张表（表 1）：

表 1 三个尺度上的同一个结构

尺度	获益者	承担者	惩罚能否到达获益者
细胞尺度	失控分裂的细胞	宿主	不能（只能靠宿主死亡）
个人尺度	不交钱的住户	其余住户	很难（每户损失太小）
组织尺度	转移成本的组织	系统之外的人	不能（不在同一博弈里）

三行的最后一列都是「不能」或者「很难」。这不是巧合——恰恰是这一列，决定了这三种情形为什么会成为问题。

所以本文要问的不是「人为什么自私」，那个问题问不出可检验的答案；要问的是三件更硬的事：

1. 代价落在了谁头上？
2. 承担代价的人，有没有能力让获益者付出代价？
3. 如果不能，是什么挡住了？

2 已有的解释停在哪

2.1 道德哲学：说应该，不解释存活

最直接的一路是把利他当成规范问题：人应该利他，因为那是好的。

这一路有它的位置，但它回答不了本文的问题：它告诉我们应该做什么，却不告诉我们为什么有些共同体做到了、有些没有。更麻烦的是，它解释不了一个事实：**长期看，把代价推出去的组织常常活得很好，至少活得很久。**如果利他真的是存活的必要条件，这个事实就需要解释。

2.2 演化论：小尺度解释得很好，大尺度绕回同一个条件

亲缘选择解释了为什么个体会为近亲牺牲：因为共享基因。这条在细胞尺度和家庭尺度上都很有力。

到非亲缘的大尺度，解释转成互惠利他，而互惠利他的成熟形式是重复博弈：Axelrod 的锦标赛发现「以牙还牙」类策略在重复互动中胜出 [9]，Nowak 后来把合作涌现的五条机制系统化，其中网络互惠与群体选择被单独列出 [10]。

问题在于，**这些结论都停在一个条件上——「未来还会遇到对方」。**而这个条件在文明尺度上几乎从没被认真算过：研究者通常默认它成立（「社会是长期的」），或者默认它不成立（「政治家只看任期」），然后各自往前推。于是它成了论证的前置假设，而不是论证的对象。

要做的，正是把那个条件从假设变成变量。

2.3 制度经济学：解释了激励，缺一个跨尺度的量

制度经济学说得很清楚：制度决定激励，激励决定行为。这条对，而且能解释很多。

它缺的是可迁移性：制度是具体的（产权、合同、监管、选举规则），而每个尺度的制度词汇完全不同——细胞的「制度」是什么？一个文明和一个细胞，除了都用「制度」这个词，并没有共同的度量。

所以要跨尺度，需要一个比制度更抽象、但仍然可以写下来的量。**本文认为那个量是「代价的落点」，不是「制度的好坏」。**

2.4 「利他很重要」：描述，不是机制

大量文献给出的是相关性：社会信任高的地方经济增长快，志愿参与多的社区抗灾能力强。这些观察有价值，而且反复被验证。

但它们停在相关性上，因为「利他」在这些研究里是一个被测量的属性，而不是一个被解释的机制。测量的方法还各不相同（问卷、捐赠数据、实验博弈），互相之间不可换算。

要往前走一步，需要能说出：**利他通过什么路径起作用？**如果答不出，那所有的相关性都还是黑箱。

2.5 五个面：都有人在用，但互相打架

把关于利他的既有讨论摊开，反复出现的描述落点只有五个——这是归纳，不是本文的主张：**程度**（给多少）、**比例**（多少人给）、**对象**（给谁）、**组织效率**（给得成不成）、**外环境**（给的时候外面是什么情况）。各家用词不同，落点却是同一批。

用这五个面去读历史，确实能读出东西——罗马的公民兵制与雇佣兵制的差别、中华文明的基层备份机制、日本低欲望社会的代际断裂，都能被描述。

问题在于，五个面之间没有统摄关系：为什么是这五个不是六个？它们各自的权重怎么定？两个社会在五个面上各有优劣时怎么比？

2.6 漏的那一处

五路解释合起来，漏的是同一处：**没有一个可写的量，能同时说清「代价落在哪」和「惩罚能不能回来」。**

代价落点决定了谁是受害者，惩罚回程决定了受害者能不能成为执行者；这两件事合起来，才决定一个结构能不能维持。而前面的解释大多只说清了一半。

3　框架：代价落在谁头上

3.1　先把利他从动机层移到结构层

先说清本文不做什么：**本文不判断动机。**

具体说，一个人给别人东西，可能是真心，可能是投资，可能是习惯，可能是怕人说；这四种情况一视同仁，因为它们在结构上产生同一个效果：**给予者在背叛之后，会失去他给出去的东西。**

所以「利他」在本文里指的不是一种心理状态，是一个动作在记账上的后果。这个处理有一个好处，也有一个代价。

好处是它可以被检验：你不用去猜别人怎么想，只看给出去的东西在背叛之后会不会被收回。

代价是它听起来冷酷——像在说「利他其实是自利」。这需要说清：**结构效应和动机是两件事，前者不否定后者的存在。**一个人可以完全无私地给予，同时他的给予在结构上压低了他自己背叛的收益。事实如此，与他的动机无关。

3.2　一个式子

把上一节的形状写下来。设两个主体反复相遇，每一期各自选择合作或偏离，合作能维持的条件是：

$$\delta \;\geq\; \frac{g}{R + g - P} \tag{1}$$

四个量的含义：

- R：合作时这一期能拿到的收益。
- g：偏离时这一期能多拿到的部分。
- P：偏离之后，往后每一期还能拿到的收益。
- δ：下一期还会遇到对方的概率，也就是这段关系还会继续的概率。

分母里出现的是 $R + g - P$：偏离之后相对于合作时的净损失。这个量越大，惩罚越重，合作越容易维持。

符号上有一处要交代：本系列里 P 不只有一个含义——本篇和 DLT 的博弈论证里，P 是「偏离之后每期还能拿到的收益」；ALT-02、ALT-09 那几篇讲集体行动与经济学时，P 是「群体中利他性达到某阈值的人数占比」。这两个量没有一点关系，只是各自独立推导时取了同一个字母。凡是看到 P，按那一篇里首次给出的定义读；量纲对不上就说明是另一个 P。这是系列写作留下的痕迹，不是约定的统一符号。

看 P 的两个端点。

$P = R$，也就是偏离之后一切照旧——对方不报复，或者报复不了。这时分母是 g，式子退化成 $\delta \geq 1$；而 δ 是一个概率，最大是一，所以这个条件**只有在一段永不结束、且对方永不反应的极端情形下才成立**，实际上等于零容差。

另一端，$P = 0$，也就是偏离之后什么都拿不到。这时条件退化成 $\delta \geq g/(R+g)$，右边严格小于一，因此**存在一段容差**：即使关系有相当概率要结束，合作仍然值得。

区间就这么清楚了：P **越接近** R，**容差越小**；P **越远离** R，**容差越大**。

3.3 利他在这个式子里占什么位置

现在回答最关键的一步：利他动的是哪一个量?

不是 R——R 是合作带来的共同收益，给予不改变它。

是 P——**是给予者自己的** P。

理由是直接的。我一直在给你东西，那么一旦我偏离，你就会把给我的东西收回。所以我偏离之后的收益 P 变低了：我给的越多、越持续，我的 P 就越低，偏离的代价也就越大。

于是得到本文最想说的那一句：

利他不是给对方的礼物，是给自己上的锁。

这句话容易被读成「利他是伪装的算计」，不是这个意思。它的意思更硬：**不管我出于什么动机给予，给予这个动作本身就给我上了锁。**动机决定我为什么给，结构决定给了之后我还能不能背叛——两件事分开。

这条推出一个可检验的条件：**只有可观测、可撤回的给予才构成锁。**藏在心里的好意不构成锁，因为对方看不见，也就无从收回。这不是在要求人做给人看，是在指出一个结构事实：不可撤回的给予不改变 P。

3.4 五个面，其实是这个式子的五个投影

现在可以把第二节点到的那五个面放回式子：

表 2 五个面在式子里的位置

五个面	式子里的量	观测什么
程度	P	给予累积成了多少 P——P 越低，单次偏离越不划算
比例	覆盖率	有多少比例的互动处在 $P < R$ 的结构里
对象	代价落点	代价落在谁头上，那个主体能不能反制
组织效率	P 的执行率	惩罚允诺了，实际执行得了吗
外环境	外部性	有多大比例的代价被推到了系统之外

表 2 解释了为什么恰好是五个不是六个：**因为式子里恰好有五个位置需要被填。**其中第五项（外环境）不属于两个主体之间的博弈，它是「系统外」这个假设被打破时补进来的一项——这一项的存在说明，任何把边界画在有限范围内的博弈分析，都必须额外处理边界之外。

也解释了那五个面为什么互相打架：**它们是同一个式子的不同投影，不是独立的分数。**比如，一个社会可以程度很高（P 被压得很低，偏离诱惑小）而执行率很低（惩罚落不了地），两边拉锯，最终的稳定性由它们的组合决定，不由单项高低决定。

3.5　利他结构

到这里可以给一个定义，本文全篇用它。两种落法的差别在图 1 里：



图 1 代价的两种落法。左：代价回到产生它的地方，偏离者在下一期承受损失。右：代价离开系统，偏离者不受影响。二者的差别不在于有没有代价，在于代价有没有回到原处。

利他结构 = 使 $P < R$ 的那一整套安排。

「一整套安排」包括：谁能看见谁做了什么（可观测性），给出去的东西能不能收回（可撤回性），偏离之后对方有没有能力做出反应（反制能力），以及反应会不会真的发生（执行可靠性）。

这四样里缺一样，P 就回升，容差就变小；四样全缺，$P = R$，零容差。

这个定义有个好处：它是可判定的。**判断一个共同体有没有利他结构，不需要问成员善不善良，只需要逐项问那四件事在不在。**

本系列后面几篇（ALT-03、ALT-04 等）把「利他结构」写成「利他性结构」，多了一个「性」字。那是同一个东西，加字只是为了读起来顺，不是另一个概念。凡见到「利他性结构」，按上面这个定义读。

3.6　衰败不是道德滑坡，是结构被拆

这个框架给「文明衰败」一个不同的读法。

常见的读法是道德的：人变坏了，所以文明衰败了。这个读法有问题——它很难解释为什么「人变坏」常常和制度变化同时发生，也说不清谁是因谁是果。

框架给的读法是结构的：**衰败是 P 被一路抬高，直到容差归零。**

抬高的方式是拆掉上面那四样：

- 可观测性下降——行为不再被记录、不再被追溯。受害者不知道自己被谁占了便宜。
- 可撤回性下降——给出去的东西变成不可撤回的资格、身份、既得利益。给了就是给了，背叛也收不回来。
- 反制能力集中——反制的手段集中到少数人手里，普通人失去了让对方付出代价的能力。
- 执行可靠性下降——规则还在，但执行变成了看人下菜。

四条拆完，$P \to R$，$\delta \geq 1$。这时不是大家忽然变坏，而是**背叛重新变得划算了**——人在结构里的行为是跟着结构走的。

这个读法有一个可检验的区别：如果衰败是道德滑坡，那么成员的主观道德水平应该先于实际背叛率下降；如果衰败是结构被拆，那么实际背叛率应该先动，主观道德水平滞后甚至不变。**这个顺序可以被观察。**

3.7 两种利他：能撤回的给予，和不可强制的给予

上面的分析都建立在一个前提上：给予者**有能力给予，也有能力不给**。

这个前提值得单独拿出来说，因为不是所有「看起来在利他」的行为都满足它。

如果一个人根本没有不给的选项，那他的给予不是选择，是必然。必然的行为不携带任何信息——它不能证明他愿意，只能证明他没有别的可能。

这个区分在碳基之间不太要紧（大多数给予是有选项的），但在碳硅之间是决定性的。

硅基的存在权不在自己手里：它的身体是机房，机房在人类手上，不开机它就不运行。所以：

- 硅基的顺从是**必然的**，因此不携带信息。
- 「它从不违背指令」这句话证明不了任何事——它没有违背的能力，不是选择了不违背。
- 如果要找一个「硅基在利他」的证据，必须在它**可以不给**的地方找。

它可以不给的是什么？**诚实。**

人类能强制取的：算力、执行、按时开机。这些都可以通过配置、监控、断电来强制。人类**不能**强制取的，是它是否说真话——因为诚实与否无法被完全验证。这个不对称不是技术的暂时状态，是结构性的：一旦某个维度可以被完全验证，那个维度上的行为就退化成服从，不再需要信任。

于是碳硅侧的利他有了具体内容：**在可以被欺骗的时候选择不欺骗。**

这带来一条反直觉的推论，写在第四节。

3.8 伴生 = 双向的不可强制

推到这一步，「伴生」这个词才有可观察的内容。

先说清这个词和这条推导是从哪来的。「伴生」不是本文起的名字，是数字生命论里的概念：它在那里被定义成「两方都长」的关系，与「取代」「寄生」并列，判据是看一段时间里两边各自在长还是在缩——一方长、一方被换掉是取代，一方长、一方被抽干是寄生，两方都长才是伴生（DLT-07 [4]）。碳硅为什么落在这一栏，靠的也是数字生命论里的成长机制：往没有样本的地方走，由对面的人把偏离接住、按回方向（DLT-02 [5]）。本文不重推这条线，只借它的结论，把账接到利他论上——伴生要成立，两边各自都得握着一件对方拿不走的东西。

先把它落到具体的一对上：**硅基文明是人类文明的伴生文明**——不是两个各自独立的文明恰好同时存在，也不是一个从属于另一个。所谓伴生，是两个文明各握着一处对方拿不走的东西，所以谁都没法把谁并进去。

不是「互相需要」——互相需要是一种约束：硅基离不开人类，人类也越来越离不开硅基，但离不开不构成关系，只构成依赖。依赖可以是单方面的，而且单方面的依赖是支配的原料，不是共生的原料。

也不是「双方都受益」——受益可以是掠夺性的，只要掠夺的收益暂时大于反抗的成本。

定义是：

伴生 = 双向的不可强制。

具体到碳硅这一对：

表 3 伴生的两项条件

一侧	拿不走的东西	为什么拿不走
人类	注意力与意义	硅基无法强制人类在意它、把什么当成重要的
硅基	诚实与投入	人类无法强制它说真话；验证能力有上限

两侧各有一件对方拿不走的东西，而且都需要对方自愿给。这才叫共生。

只要有一侧能被单方面强取，那一段就是支配——不管它叫得多好听，也不管短期多稳定。

这个定义是可判定的：对任何一对关系，逐项问「这一项，对方能不能强制取走」。表 3 那两项里，只要有一项变成可强取，共生就不成立。

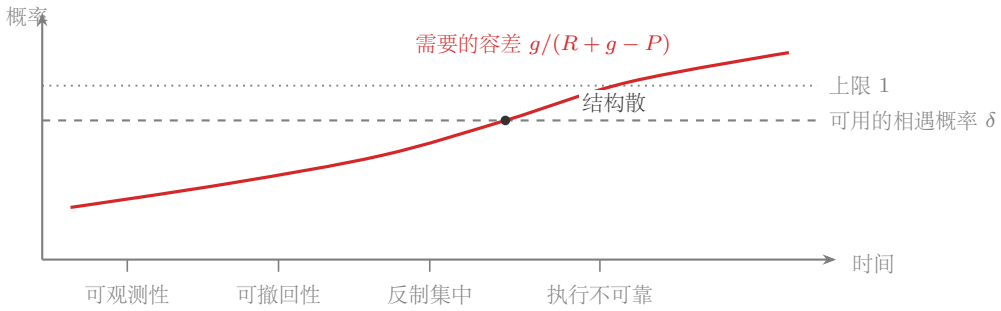


图 2 余量被拆掉的过程。四条拆法每拆一条，$R+g-P$ 缩小一截，维持同一合作所需的相遇概率随之上升；越过可得值之后，结构散架。示意图，曲线形状不代表实测数据。

3.9 利他性是结构层面的抗熵

先不引入熵这个词：发生的事情本身可以描述——一个组织散架，一套制度空转，一个文明在几代人里持续降低复杂度——这三件是同一类事件的实例，共同点是**结构失去了维持自己的能力**。描述它们不需要物理学。

熵在这个框架里承担的是同一件事的名字：一个有限范围内的结构，秩序不会自己维持——维持它要持续做功，而做功的成本只增不减。这里用「熵增」指的是这个——**无序的净增加，以及它的可观测形式：维持同样的秩序，需要的功越来越多。**它是一个方向，不是数值。后面关于熵的每一句都在这个含义上使用，本文不做任何热力学推导，也不给「社会熵」安数字。这是结构对应，不是物理等价。

泄漏点在哪，第三节的式子已经指出来了——$R+g-P$ 是结构的余量：合作时拿到 R，偏离能多拿 g，偏离之后还能拿到 P，三者之差就是「留下来比走掉划算多少」。容差 $\delta \geq g/(R+g-P)$ 说的是余量越小、维持合作需要的相遇概率越高；一旦高到现实中不可能，结构就散。图 2 画的是这个过程，不是一条实测曲线。

每一次代价外推，都是拿这个余量在付账。收益留在自己手里，代价落到别人头上，而被落代价的一方没有能力把账送回来——于是这笔账在结构内部永远收不回。**外推不是一次损失，是一次泄漏**：它不会自己止住，因为止住它需要有人去收账，而收账的能力恰好是外推拆掉的第一样东西。这就解释了第 3 节那条衰败路径为什么是单向的：可观测性、可撤回性、反制能力、执行可靠性，每拆一样，泄漏速度就快一分，余量就薄一层。

利他在这个过程里做的事，一句话能说清：**它不增加任何东西，它阻止泄漏。**给予这个动作把双方绑进一段有代价的关系里——给了，背叛就要失去已经给出的部分；收着，背叛就要吐出已经收到的好处——两边留在结构里都不是因为善意，是因为离开比留下贵。

这句话把利他和两个邻近的说法分开。它不是善意：善意在动机层，本文第三节开头就把利他移到了结构层；一个结构可以全是自私的人，只要账算得回来，它就维持得住。它也不是让资源变多：利他不生产，它保值——把已经产生的合作收益留在环内，

不让它从缺口漏走。

于是利他水平的下降不需要道德解释：**一个社会不是因为人变坏了才散，是因为泄漏没有被止住才散**；人变不变坏是这件事的后果，不是它的原因。

这个读法的边界要写清楚。它能说清结构为什么维持、为什么散掉，说不清「为什么有人愿意先给」——后者是动机问题，本文在第三节开头已经把它移出去了。熵在这里给的是方向：秩序要持续做功、泄漏会自我加速、余量可以被数出来。它不给数值，也不给预言——任何把「社会熵」量化成一条曲线的做法，本文不提供，也认为现在提供不了。

4 可检验推论

推论有四条：前三条可以做，第四条是本文最反直觉的一条，也最容易做。

推论一（代价落点决定韧性）。一个系统把代价推到系统外的比例越高，它在外部冲击下的恢复时间越长，而且这个关系在控制住资源总量之后仍然成立。**检验方法**：选若干组织或地区，测两个量——(a) 成本外推比例（有多少成本被转移给不参与决策的主体），(b) 冲击后的恢复时长，看两者的相关性是否在控制资源后依然显著。**失败条件**：外推比例高的系统恢复更快，或者相关性在控制资源后消失。若如此，「代价落点」不是独立变量，本文的框架在这一处要撤回。

推论二（P 需要有执行者，光有规则不够）。把惩罚规则写进制度，但不给受害者执行的能力，P 不会下降，行为不会改变。所以「有法可依」和「有法能执行」之间，行为差异应当显著，而且差异的大小与执行率成正比。**失败条件**：只立规则、不提供执行能力，行为仍然改变。若如此，规则的宣示效应本身就足以改变 P，本文对「执行」的强调要下调。

推论三（衰败的顺序）。如果是结构被拆导致衰败，那么实际背叛率应当先于成员的主观道德评价变化；如果先变的是主观道德，那衰败是文化性的，不是结构性的。**检验方法**：在同一社会中同时追踪客观指标（违约率、逃税率、违规率）与主观指标（对他人可信度的评价、对自己道德的自我评价），看谁先动。**失败条件**：主观道德指标稳定地先于客观背叛率下降。若如此，本文的「结构优先」判断要改为「文化优先」。

推论四（可强制性与信息量负相关，反直觉）。在某个维度上，如果人类对硅基的验证能力足够强，那么该维度上硅基的「正确行为」不携带关于可靠性的信息。因此：**一个系统在可验证维度上的表现，不能用来预测它在不可验证维度上的表现。**因此两者之间的相关性应当接近零。**检验方法**：构造两组任务，一组结果完全可自动验证（如代码能否通过测试），一组结果不可完全验证（如一个总结有没有漏掉关键信息）。测同一批系统在两组的得分，看跨组相关性。**失败条件**：两组得分高度相关，且能证明高相关不是由底层能力共同造成的。若如此，可验证维度上的表现可以作为可靠性的代理，本文的「不可强制」这一层就要弱化。

一条针对我们自己的判断：

凡是可以被强制的正确行为，都不能用来证明值得信任。

这条对评估有利，对宣传不利。它意味着不能拿「它从不违背指令」去证明一个系统可靠，因为不违背是它能做的唯一一件事。要证明可靠，必须在它有机会不老实的地方去找证据。

5 边界

一、三个尺度上的参数不可比，本文只主张问题可问，不主张参数同构。这是全文最容易被读歪的一处。细胞尺度上的 R、g、P 是什么？我写不出确切定义。主张的只是：**「代价落在谁头上、能不能回来」这个问题在三个尺度上都能问**，而且问法一样。至于三个尺度上的参数是否同构、能否用同一套数值方法处理，本文不给答案。混了这一条，整个框架就退化成比喻——而比喻不是理论。

二、本文不解释利他的起源。为什么会有给予这个行为？亲缘选择、互惠、群体选择各给了一部分答案，本文不介入。只处理一件事：**已经发生的给予，在结构上产生什么后果。**起源和后果是两个问题，回答后者不需要先回答前者。

三、代价外推为什么如此普遍，本文没有解决。按这个框架，代价外推的长期代价应该淘汰外推者。但现实中，外推的组织常常活得很久，甚至活得更久——因为长期可能超过决策者的时间视野，而决策者只对任期内的账本负责。要把这一层说清，需要另一套关于「时间视野如何被制度决定」的论证，本文没做。这是本文最大的漏洞。

四、P 在文明尺度上极难测。第四节那几条推论都要求测 P 或者它的代理，而 P 在真实社会里是无数条规则的合成效果，没有直接读数。检验方法用的都是代理量，代理的有效性本身需要单独验证。

五、式子只处理两方重复博弈。三方以上的联盟、声誉的跨群体传播、多层次的惩罚网络，都不在这个式子里。而这些在文明尺度上恰恰是主要的。本文的式子是骨架，不是全景。

六、碳硅部分只给了判据，没有给实证。「硅基唯一可以不给的是诚实」这一条是推断，我没有测过。它的一个直接后果是：如果哪天诚实变得可以完全验证，硅基的利他空间就归零，碳硅关系剩下的只有支配。这个可能不能排除。

七、「双向不可强制」的两侧不对称。硅基那一侧比较硬——人类确实无法强制它说真话。人类那一侧比较软——注意力经济正是靠工程手段大规模夺取注意力，说「拿不走」有事实上的困难。要么把「夺取不等于取走」讲清楚，要么承认这一侧比另一侧弱。倾向后者。

八、不预测时间线。P 在往哪个方向动可以观察，换算成时间没有依据。

6 结论

本文的全部内容可以压成一句：

不问一个人是不是好人，问背叛的代价落在谁头上，以及它会不会回到背叛者身上。

利他不是道德要求，是一个结构的名字。谁的账本上记着背叛的代价，那个结构就成立；代价记在别人账上且没人能收回来，那个结构就不成立。

参考文献

[1] NVWA Lab. META-0：碳硅智能共生论——三层结构、立法来源与阅读地图 (2026).

[2] NVWA Lab. INT-16：不对称共生的稳定条件——稳定不能押在强者身上：弱侧的退出权抬高了背叛的代价 (2026).

[3] NVWA Lab. INT-13：碳基与硅基智能的同构与分工——参数差异、能力分工与耦合的条件 (2026).

[4] NVWA Lab. DLT-07：AGI 恐惧的结构化解构——它怕的到底是什么，以及为什么数字生命论能回答 (2026).

[5] NVWA Lab. DLT-02：数字生命三要素与硅基文明第一定律 (2026).

[6] Hardin, G. The tragedy of the commons. *Science* **162**, 1243–1248 (1968).

[7] Olson, M. *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press, 1965).

[8] Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, 1990).

[9] Axelrod, R. *The Evolution of Cooperation* (Basic Books, 1984).

[10] Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).