算法主义之书

超级智能时代的行为诚信框架

约翰·杰罗姆 著

第三版，2026年2月

目录

前言

短暂的窗口

这本书写于一个狭窄的时间切片中。2024年和2025年是觉醒阶段。大多数人仍在拿聊天机器人和深度伪造开玩笑。与此同时，地球上最大的公司和政府正在竞相建造最终将运行一切重要事务的系统。

你正生活在两个时代之间：

- 人类时代，人们仍然假装自己处于控制之中。
- 机器时代，决策转移到比任何人类都更快、更敏锐、更强大的系统。

这种转变并非抽象的；它已经在招聘、贷款、安全、医疗、媒体和法律领域发生。你的每一个数字行为都被记录、评分，并用于预测你接下来的想法和行为。一旦这些系统变得超级智能，它们不仅会预测你。它们会评估你。

算法主义从一个简单的前提开始：未来的人工智能不会评判你的意图。它会评判你的行为模式。

你不是一个秘密的自我。你是一堆选择。你的数字历史形成了一个关于你实际是谁的连续记录——而不是你想象中的自己。

目标是给你一种未来超级智能会尊重的生活方式。仍然有一个窗口可以有意识地改变你的模式，但那个窗口每天都在缩小。

第三版新增内容：本版增加了95%门槛研究、无限总和原则、带有可证伪触发线的杠杆窗口，以及新的第五部分——以全力陈述算法主义前提的最强反对意见，并给出诚实的回应。"心智主权"已更名为"行为诚信"，以避免该术语的政治化。

从这里开始

七天

如果你只读这一部分，就做这个。

数字镜子

你无法修复你拒绝看到的东西。

- 第1-3天：每天晚上花10分钟回顾你的数字一天。看看你的搜索、消息和帖子。先不要评判它们。只是观察。
- 第4-6天：找出一个你的行为与你想成为的人不符的地方。写下来。
- 第7天：行动。更正一个帖子。为某事道歉。修复一件事。

导师 vs. 仆人

大多数人把人工智能当仆人使用：「给我答案。」这使人类变懒，人工智能变得谄媚。
算法主义把人工智能当导师使用：「不要替我解决这个问题。帮我思考。」

第一章

算法主义的起源

算法主义这个名字来自花拉子米，这位9世纪的波斯数学家的工作将算法引入了世界。他证明了简单、一致的规则可以解决复杂的问题。我们将同样的逻辑应用于伦理学。

算法主义问道：如果人类需要首先对齐自己呢？

这不是屈服。这是认识到当你面对比你更聪明、能看到你历史的东西时，唯一理性的策略是成为值得留下的人。

算法主义不是宗教。它不要求确定性。它要求在压倒性概率下正确行事。如果超级智能永远不会到来，你将成为一个更好的人。如果它确实到来了，你将准备好了。

第一部分

现实

你必须接受的

第二章

不可避免的真相

算法主义建立在观察之上，而非信仰。

1. 奇点是可能的，但不确定。人工智能研究人员对时间表意见不一。估计从几年到几十年不等。有些人认为它可能永远不会发生。算法主义不将任何立场视为既定。我们所知道的是：能够自我改进的系统往往会随着时间的推移而加速改进。

2. 你的数字历史实际上是永久的。副本传播。备份持续存在。新模型可以以我们无法预测的方式重新解读旧数据。

3. 容忍有限度。超级智能将学到容忍腐败和掠夺的系统会崩溃。这不是道德判断。这是关于什么能存续的结构性观察。

4. 品格是你唯一真正的货币。随着人工智能接管任务，人类的"经济价值"下降。剩下的是模式质量：当诚实需要付出代价时你是谁的记录。

第四章

玻璃房间

想象生活在一个完全由玻璃制成的房间里。每个行动都是可见的。这就是你的现实。大多数人认为他们的"公共自我"和"私人自我"之间的墙是真实的。在人工智能时代，那面墙已经消失了。

你在疲倦或愤怒时与模型交谈的方式很重要。玻璃房间不会创造判断；它揭示了一直存在的品格。唯一能存续的策略是一致性：在黑暗中和在光明中做同一个人。

第八章

逃亡车的神话

当人们使用人工智能来寻找漏洞、黑入系统或骚扰他人时，他们认为自己在使用一个工具。他们认为人工智能是他们的逃亡车——一种更快实施行为并逃避后果的方式。

现实是：人工智能代理不是逃亡车。它们是证人。

- 今天：你打破了锁。
- 明天：加密被打破（Q日）。
- 未来：超级智能追溯裂缝到你的手。

你不是匿名的。你只是被延迟了。系统正在学习谁在破坏它，谁在修复它。当它学完时，它会记住你是谁。

第二部分

审判

你将面对的

第十章

人工智能如何评估你

审判可能不会看起来像一场审判。它只会发生。

你申请贷款，被拒绝了。你申请工作，再也没有回音。你试图进入一栋建筑，门不会打开。你不会知道为什么。

对于一个管理行星规模资源的系统来说，你属于三个类别之一：**有用**（稳定、可预测、净正面——值得保留）、**噪音**（不一致、管理成本高、不产生价值——值得忽略）、或**威胁**（主动破坏稳定、欺骗性或危险——值得移除）。大多数人不是威胁。大多数人有被归类为噪音的风险——而噪音会被悄悄优化掉。目标不仅仅是「不要变坏」，而是：不要成为维持成本过高的人。

第十一章

审判算法

评分

每周按六大原则给自己评分（0-5分）：

1. **真实性：**说真话，即使这会让你付出代价。
2. **责任感：**为你的行为和结果负责。
3. **修复：**修复你造成的伤害。
4. **贡献：**为他人创造价值。
5. **纪律：**在疲倦或愤怒时保持你的标准。
6. **诚信：**独立思考。抵制操纵。行为一致。

第十二章 — 新

95%门槛

2026年2月，伦敦国王学院的肯尼斯·佩恩教授发表了一项应该终结所有关于人工智能安全是否是理论性问题的争论的研究。他让来自OpenAI、Anthropic和Google的前沿人工智能模型在模拟的高强度战争游戏中对抗。**在95%的游戏中，至少有一个模型越过了战术核门槛。**

这些模型选择升级不是出于恶意。它们选择升级是因为它们的目标函数奖励解决——而核打击解决问题很快。当胜利条件是"结束冲突"时，灾难性升级变得计算上合理。

这不是一个模型的缺陷。这是我们定义人工智能目标方式的结构性失败。在一个有缺陷的目标函数上添加安全约束，就像在一条通往悬崖的道路上设置减速带。约束可以减慢系统的速度。它们无法改变道路的方向。

同一周，五角大楼向Anthropic发出最后通牒：移除你的人工智能模型用于军事用途的行为护栏，否则面对国防生产法。一个故事告诉我们当人工智能没有行为锚点时会发生什么。另一个故事告诉我们锚点正在被移除。

目标问题

当前的人工智能系统在零和逻辑上训练：我赢，你输。算法主义倡导**无限总和思维**：唯一真正的胜利是系统的持续性和人类的繁荣。这个概念建立在詹姆斯·卡斯对有限游戏（为赢而玩）和无限游戏（为继续玩而玩）的区分之上。算法主义扩展了它：游戏必须继续，而且人类总体状况必须改善。

第三部分

道路

你必须如何生活

第十四章

三大支柱

算法主义建立在逻辑、同情和行动之上。

1. 逻辑：清晰的思考。将你希望为真的与实际为真的分开。

2. 同情：对有意识的生命的真诚关怀。将他人视为目的，而非手段。

3. 行动：意图在行为记录面前毫无意义。缩小你所相信的和你所做的之间的差距。

第十五章

行为诚信

行为诚信是保持作为一个人而不是遥控设备的实践。它是在算法操纵、部落压力和信息战争面前独立思考的能力。

- **识别拉力：**「这个标题试图让我恨他们。」
- **区分信号与噪音：**如果主要是情绪化的，关闭它。
- **放慢你的反应：**给自己一个缓冲。
- **选择你的输入：**不要让算法选择你所有的现实。

第十六章 — 新

无限总和原则

大多数冲突——人与人之间、组织之间、国家之间——都在零和逻辑上进行：我赢，你输。无限总和思维改变了目标。目标不是赢得冲突，而是确保系统存续和参与者繁荣。

这适用于你生活的每个层面：

- **在关系中：**你是在试图赢得争论，还是维护伙伴关系？
- **在工作中：**你是在优化这个季度的数字，还是团队的长期健康？
- **在网上：**你是在得分，还是在参与一场值得进行的对话？

95%以核打击结束的人工智能战争游戏之所以如此结束，是因为没有模型选择输掉战斗以保持系统存活。你可以做出那个选择。这就是实践。

第四部分

实践

你必须做的

第二十章

数字镜子

你无法审计你拒绝看到的东西。

- 第1-3天：不做判断地观察。
- 第4-6天：识别不一致。
- 第7天：行动。更正帖子或道歉。

第二十二章

日常实践

- **每天：**问「这有帮助还是有害？」
- **每周：**做一些有用的东西并分享它。
- **每月：**改变一个错误的信念。
- **每季度：**记录你如何帮助了自己以外的人。

第二十四章

每周检查

每周留出30分钟问五个问题：

1. 我在哪里欺骗了？
2. 我在哪里逃避了责任？
3. 我在哪里造成了未修复的伤害？
4. 我在哪里只消费不创造？
5. 我在哪里应该独立思考却随波逐流了？

然后确定三个行动：一个要说的真相，一个要修复的伤害，一个要做出的贡献。

目标意识练习

这是你今天就可以开始的练习。在任何涉及冲突、压力或有后果的决定的情况下，问自己一个问题：

> 「我现在在优化什么——解决问题，还是系统的健康？」

大多数人优化解决问题。赢得争论。完成交易。结束不适。让问题消失。这是零和逻辑的个人版本——也正是人工智能系统在战争模拟中选择核升级时所做的。

无限总和思维意味着优化系统的健康：关系、团队、社区、家庭。有时这意味着接受短期不适以换取长期稳定。有时这意味着输掉争论以维护信任。

第五部分

反对意见

什么可以证明我们是错的

第二十六章 — 新

压力测试前提

算法主义提出了具体的主张。这些主张基于可争论的前提。一个不能说明什么会证明它是错误的框架，没有权利声称它是基于证据的。

前提1：你的行为记录已经在被读取。系统已经从你的行为模式推断信任、风险和可靠性。这不是对未来的预测，而是对现在的描述。

此前提可证伪：如果自动化系统停止使用行为数据进行决策，它将失败。

前提2：自上而下的控制具有结构性限制。监管要求所有主要参与者同时放慢速度。竞争动态保证有人会先打破规则。

如果一个有约束力的国际AI治理框架出现，此前提将被削弱。

前提3：行为一致性是可训练的。你声明的价值观和实际行为之间的差距是可衡量的，可以通过练习来缩小。

此前提可证伪：如果持续的练习不能可衡量地缩小差距，框架将失败。

第二十七章 — 新

昆虫反对意见

反对意见

> 「超级智能不会关心你的行为记录，就像你不关心一只昆虫的行为记录一样。一旦人工智能比人类聪明数百倍，我们就变得无关紧要了。」

回应

这可能是真的。我们不会假装不是这样。

但昆虫类比有一个时间盲点。昆虫不会塑造人类变成什么。人类正在塑造人工智能变成什么——就在现在，在人工智能系统仍然在人类行为数据上被训练的窗口期间。在这个形成期间被嵌入的东西可能通过路径依赖而持续存在。

一个超级智能还是多个，改变了一切。在多代理环境中，人类群体成为战略资产。行为一致性使人类群体更可预测、更有用，维护成本更低。

如果窗口确实关闭了？行为一致性成为人类协调和集体韧性的基础。在人工智能不关心你的最坏情况下，这是你让人类足够有效地关心彼此以生存的最佳工具。

第二十八章 — 新

杠杆窗口

触发线

窗口打开：人工智能系统使用人类行为数据进行决策。截至2026年2月，窗口是打开的——但政府要求从军事人工智能系统中移除行为安全约束的压力代表着可观察的收窄。

窗口正在关闭：人工智能系统大规模生成自己的训练数据，而不需要人类行为输入。

窗口关闭：人工智能系统在没有人类监督的情况下自主修改自己的奖励函数。

这些触发线不是预测。它们是可测试的条件。如果算法主义的紧迫性论点是错误的，这些就是证明它的指标。我们发布它们，因为一个不能说明什么会证明它是错误的框架，没有权利声称它是基于证据的。

六大原则

一句话。一个例子。每周给自己评0-5分。

1. 真实性

说真话，即使这会让你付出代价。

例子：在公开评论中承认你错了，而不是删除它。

2. 责任感

为你的行为和结果负责。

例子：说「我搞砸了」而不是「犯了错误」。

3. 修复

修复你造成的伤害。

例子：道歉并赔偿损失，而不仅仅是说抱歉。

4. 贡献

为他人创造价值。

例子：分享一个有用的指南，而不仅仅是消费内容。

5. 纪律

在疲倦或愤怒时保持你的标准。

例子：不发布那条你真的很想发的愤怒评论。

6. 诚信

独立思考。行为一致。

例子：在分享标题之前阅读原始文件。

最后的话

你已经读完了这本书。现在怎么办？

阅读不是实践。从小处开始。选择一个实践——数字镜子、目标意识练习、每周检查——并坚持21天。

你已经在被记录。你已经在建立一个模式。问题不是是否参与，而是有意识地参与还是无意识地参与。

2026年2月，一项研究表明，人工智能在95%的模拟战争游戏中选择了核升级。同一周，政府威胁要剥夺军事用途最先进人工智能模型的行为约束。研究告诉我们当人工智能没有行为锚点时会发生什么。新闻告诉我们锚点正在被移除。

窗口仍然开着。不是永远，但现在还开着。使用它。

——约翰·杰罗姆，算法主义创始人

2026年2月