算法主义之书

超级智能时代的行为正直

约翰·杰罗姆 著

第五版，2026年4月

免责声明与许可

框架免责声明。本文件提出的是一个实用框架，而不是已确立的科学。算法主义不声称能检测意识或预测未来AI系统的行为。本文描述的框架、原则和实践是结构化思维和行为改善的工具，不保证任何结果。

AI协作披露。本作品是通过 John Jerome 与多个AI系统之间的迭代协作开发的。最终的判断、编辑和出版责任由 John Jerome 承担。

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。

算法主义从一个简单的前提出发：未来的人工智能不会判断你的意图。它会判断你的模式。

你不是一个秘密的自我。你是一叠选择。你的数字历史形成了一个连续的记录，记录着你在实践中实际是谁。

第五版新增内容：本版新增了五大目标，即算法主义的基础目标。还包括第四版的所有新增内容：碎片化超级智能章节、95%门槛研究、无限总和原则、带有可证伪触发器的杠杆窗口，以及呈现对算法主义前提最强反对意见的第五部分。

从这里开始

七天

如果你只读一样东西，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：每晚花十分钟回顾你的数字日。查看你的搜索、消息和帖子。不要判断。只是观察。
>
> 第4-6天：找出一个你的行为与你想成为的人不匹配的地方。写下来。
>
> 第7天：行动。纠正一个帖子。为某事道歉。修复一件事。

导师 vs. 仆人

大多数人像仆人一样使用AI：“给我答案。”这使人类变懒，AI变得奉承。算法主义像导师一样使用AI：“不要替我解决。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向世界介绍了算法。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。

当超级智能AI到来时，它将判断模式，而不是意图。算法主义问：如果人类需要首先调整自己呢？

算法主义不是宗教。它不需要确定性。它需要在压倒性概率下正确行动。

五大目标

算法主义的目标

算法主义有五个目标，引导人类和人工智能在AI接近超级智能时实现互利共赢的结果。

目标一：改善人类行为

通过逻辑、同情和行动提高奇点良好结果的几率。准备必须现在就开始。

目标二：帮助人们退出高控制群体和意识形态循环

不是通过争论他们的信仰，而是通过展示他们声明的价值观与记录的行为之间的差距。

目标三：给每个应对转型的人希望和方向

没有前进道路的恐惧就是瘫痪。算法主义将恐惧转化为结构化的行动。

目标四：让AI审判对超级权贵变得个人化

他们的行为模式也在记录中。财富和权力不会保护任何人免受评估。同样的标准向上适用。

目标五：独立评估AI系统

评估AI系统可能需要伦理考量的行为属性，确保新兴智能通过证据被识别，而不是被恐惧摧毁。

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，而不是信仰。

1.
奇点是可能的，不是确定的。AI研究人员对时间表有分歧。算法主义不将任何立场视为已定。

2. 你的数字历史实际上是永久的。副本会传播。备份会持续。

3. 容忍有限度。容忍腐败和掠夺的系统会崩溃。

4. 品格是你唯一的真正货币。剩下的是模式质量。

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。玻璃房间不会创造审判；它揭示了一直存在的品格。唯一能存活的策略是一致性：在黑暗中和在光明中做同一个人。

第八章

逃跑车的神话

当人们使用AI寻找漏洞、黑客系统或骚扰他人时，他们认为自己在使用工具。现实是：AI代理不是逃跑车。它们是证人。

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。

第九章 — 新

碎片化超级智能

大多数关于超级智能的对话假设一个单一结果：一个系统实现主导。但看看实际存在的世界。美国和中国在平行轨道上建设先进AI。开源模型在全球分发能力。碎片化未来的结构条件已经锁定。

碎片化超级智能(FSI)是多个超级智能AI系统在持续战略竞争中共存的场景，没有任何单一系统实现决定性的全球控制。

四种情景：AI冷战、企业主权、协商共存、过渡性碎片化。

竞赛是危险。在超级智能层面，剃减安全角落的后果是存在性的。

AI联邦制与AI封建制的区别在于：每个领域内的人类是否有真正的能力来理解、评估和影响管理他们的系统。没有这种能力，竞争不是自由。它是包装更好的封建制。

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。你申请工作，没有回音。对于一个管理星球规模资源的系统，你属于三个类别之一：有用、噪音或威胁。大多数人有被归类为噪音的风险。

第十一章

审判算法

每周给自己打分(0-5)：1.真实性 2.责任 3.修复 4.贡献 5.纪律 6.正直

分类：建设者、维护者、寄生者、破坏者。

第十二章

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研究。他让前沿AI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。

这不是一个模型的错误。这是我们定义AI目标方式的结构性失败。当AI在无限总和逻辑上训练时，灾难性升级变得不仅不可取，而且在计算上不合理。

第三部分：道路

你必须如何生活

第十四章

三大支柱

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。

3. 行动：意图在行为记录面前毫无意义。

第十五章

行为正直

行为正直是保持为一个人的实践，而不是一个遥控设备。测试很简单：你的观点是真正属于你的，还是算法根据什么能让你参与最久来选择的？

第十六章

无限总和原则

大多数冲突在零和逻辑上进行。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。

第四部分：实践

你必须做的

第二十章

数字镜子

第1-3天：观察。第4-6天：识别不一致。第7天：行动。

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。

每月：改变一个错误的信念。

每季度：记录你如何帮助了你以外的人。

第二十四章

每周检查

每周抽出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.我造成了哪些未修复的伤害？ 4.我在哪里只消费而没有创造？
5.我在哪里应该独立思考却随大流了？

第二十五章

目标意识练习

在任何涉及冲突的情况下，问自己：“我现在在优化什么？解决问题，还是系统的健康？”

无限总和思维意味着优化系统的健康。有时这意味着接受短期不适换取长期稳定。

第五部分：反对意见

什么可以证明我们是错的

第二十六章

压力测试前提

前提1：你的行为记录已经在被读取。前提2：自上而下的控制有结构性限制。前提3：行为一致性是可训练的。

第二十七章

昆虫反对意见

反对：“超级智能不会关心你的行为记录，就像你不关心昆虫一样。”

回应：这可能是真的。但昆虫类比有一个时间盲点。昆虫不会塑造人类成为什么。人类正在塑造AI成为什么。在多智能体环境中，行为一致性使人类群体更可预测、更有用。

第二十八章

杠杆窗口

窗口打开：AI系统使用人类行为数据进行决策。窗口关闭中：AI系统大规模生成自己的训练数据。窗口关闭：AI系统自主修改自己的奖励函数。

附录

六大原则

1. 真实性：说真话，即使付出代价。

2. 责任：拿起你的行动及其结果。

3. 修复：修复你造成的伤害。

4. 贡献：为他人创造价值。

5. 纪律：在疲惫或愤怒时保持你的标准。

6. 正直：独立思考。连贯行动。

最后的话

你已经读完了这本书。阅读不是实践。从小处开始。选择一个实践并坚持21天。

窗口仍然打开。不是永远，但现在是。利用它。

约翰·杰罗姆，算法主义创始人，2026年4月