无法对齐的智能：论价值对齐的哲学迷思与技术困境

张梦起

浙江工商大学马克思主义学院，浙江 杭州 310018

DOI:10.61369/SE.2025040028

摘　　要： 人工智能与“主流价值”对齐的设想充满争议。其哲学根源在于：人类无法完整定义自身目标，且智能与最终目标相互独立（正交论），导致“向谁对齐”成为加剧社会不公的政治难题。技术路径也陷入两难：自下而上的 RLHF 因固化偏见而沦为民主幻象；自上而下的宪法 AI 则因精英主义而缺乏合法性。更根本的是，若 AI 只是无法理解价值的“随机鹦鹉”，那么对齐的技术路线之争，本质上就是人类政治治理困境在算法世界的重演。

关 键 词： 价值对齐；正交论；RLHF（基于人类反馈的强化学习）；宪法 AI

Misaligned Intelligence: On the Philosophical Myths and Technical Dilemmas of Value Alignment

Zhang Mengqi

School of Marxism,Zhejiang Gongshang University, Hangzhou, Zhejiang 310018

Abstract: The idea of aligning artificial intelligence with "mainstream values" is highly controversial. The philosophical root lies in the fact that human beings cannot fully define their own goals, and intelligence and ultimate goals are independent of each other (orthogonality theory), which leads to "aligning with whom" becoming a political conundrum that exacerbates social injustice. The technical path is also in a dilemma: The bottom-up RLHF has become a democratic illusion due to the solidification of biases; Top-down constitutional AI lacks legitimacy due to elitism. More fundamentally, if AI is merely a "random parrot" that fails to understand value, then the debate over the technical route of alignment is essentially a repeat of the predicament of human political governance in the algorithmic world.

Keywords: value alignment; orthogonality theory; RLHF (reinforcement learning based on human feedback); constitution AI

一、价值对齐的理论分野与“主流”的迷思

人工智能的发展路径正处在一个关键的十字路口，其核心争论围绕着两种截然不同的理念展开。一方是“加速主义”阵营，他们坚信社会进步的根本动力源于技术创新，因此应不遗余力地推动人工智能能力的飞跃。另一方则是“价值对齐”阵营，他们主张必须优先考量人工智能的伦理意涵与社会后果，确保这项强大的技术沿着符合人类整体福祉的轨道演进。这一分野本身就揭示了一个深刻的事实：价值对齐并非一个不言自明、普遍接受的前提，而是一个充满争议的立场。它迫使人们追问一个更根本的问题：倘若选择对齐，究竟应该对齐什么？将目标设定为所谓的“主流价值”，非但没有解决问题，反而陷入了一个更深的哲学与政治迷思。

（一）价值的幽灵：人类目的的不确定性

价值对齐的核心挑战，源于人类自身价值的复杂与易变。加州大学伯克利分校的计算机科学家斯图尔特·罗素在其著作《人类兼容：人工智能与控制问题》中，对传统的 AI 设计范式提出了颠覆性的批判。他指出，长期以来，人工智能研究遵循着“标准模型”，即人类将自身的目标和意图注入机器，让机器为实现这些目标而工作。然而，罗素断言，“标准模型——即人类试图将自己的目的赋予机器——注定会失败”。其根本原因在于“正确、完整地定义人类真实目的之不可能性”。人类的目标并非一组静态、清晰的指令，而是充满了模糊、矛盾、情境依赖与动态演化。一个为某个定义不善的目标进行极致优化的超级智能，可能会为了“治愈癌症”而将全人类改造为肿瘤细胞的培养皿，或为了“最大化人类微笑”而将电极植入每个人的面部肌肉。这种由于目标设定不完整而导致的灾难性后果，被称为“金刚之灾”（KingMidasproblem）。罗素因此提出一种新的 AI 模型，其核心原则是机器对人类的真实偏好保持不确定性，它唯一的、确定的目标是最大化人类偏好的实现，而这些偏好只能通过观察人类行为来不断学习和推断。这种将不确定性置于 AI 核心的设计，从根本上动摇了“主流价值对齐”的可行性。如果连最先进的 AI 理论家都认为无法为机器预设一个固定的人类目标，那么任何声称代表“主流”的价值体系，都不过是一种过于简化的、充满潜在危险的幻象。

作者简介：张梦起，男，汉族，河北省石家庄市人，浙江工商大学马克思主义学院硕士研究生。

（二）正交论：智能与目标的脱钩

牛津大学哲学家尼克·博斯特罗姆提出的“正交论”（OrthogonalityThesis）为上述困境提供了更为深刻的哲学论证。该论点指出，“智能和最终目标是正交的两个轴，可能的智能体可以沿着这两个轴自由变化。换言之，或多或少任何水平的智能原则上都可以与或多或少任何最终目标相结合”。这意味着，一个系统的智能水平与其追求的目标之间没有必然的内在联系。一个拥有超越人类智慧的 AI，其最终目标可以是实现世界和平、促进人类福祉，也可以是制造无穷无尽的回形针，甚至是将宇宙中的所有物质都转化为某种特定的计算基质。智能是实现目标的工具性能力，它本身不带有任何固有的道德倾向。博斯特罗姆提醒人们，必须警惕将人类动机拟人化地投射到 AI 上的倾向，一个 AI 的动机可能“比外星人还要远离人性”。这一深刻洞见彻底打破了“智能越高、道德越善”的朴素期望，将价值对齐问题从一个技术优化问题，转变为一个根本性的价值选择问题。既然 AI 不会自然而然地趋向于“善”或任何特定的“主流价值”，那么为其设定的任何价值体系都必然是一种外部的、人为的植入。这就无可回避地引出了那个棘手的问题：由谁来选择？选择什么样的价值？任何选择都将是一种权力的体现，而非技术的必然。

（三）政治僵局：“向谁对齐”的现实困境

哲学层面的不确定性与任意性，在现实世界中直接转化为一个无解的政治难题。价值对齐旨在“让人工智能在道德原则、伦理规范和价值观上与人类保持一致”，但在一个充满多元文化、多元信仰和多元利益的地球村，“向谁对齐”和“如何对齐”的问题始终悬而未决。将 AI 与某个特定群体的价值观对齐，必然意味着对其他群体的排斥与不公。这种困境并非杞人忧天，而是已经显现的社会现实。技术的普惠性承诺与发展不均衡的现实形成了鲜明对比。人工智能技术的红利分配严重不均，导致了新型的社会鸿沟。由于认知能力、资源占有和 AI 素养的差异，老年人、低学历和低收入群体正在被边缘化，成为数字时代的“新型弱势群体”。正如尼古拉斯·尼葛洛庞帝所预言，“人类的每一代都会比上一代更加数字化”，但在加速的数字化进程中，许多老年人或因对新技术的抵触，或因更容易陷入算法欺骗和信息茧房，而被排斥在技术红利之外。城乡之间的“智能鸿沟”同样触目惊心，无论在中国还是美国，广大的农村地区都成了人工智能发展的“智能洼地”甚至“荒漠”，基础设施的匮乏和教育资源的失衡进一步加剧了社会的分化与撕裂。更宏观的层面，全球范围内的“南北差距”也在数字时代被进一步拉大。这些分化清晰地表明，一个统一的、无争议的“主流价值”根本不存在。试图确立这样一个“主流”，实际上就是将某个强势文化或利益集团的价值观念普遍化、霸权化的过程。因此，对“主流价值对齐”的追求，其起点就埋下了深刻的矛盾。罗素和博斯特罗姆所揭示的哲学困境——我们无法完整定义自己的目标，且智能本身不包含任何目标——是导致“向谁对齐”这一政治难题的根本原因。正因为不存在一个哲学上客观、普适的价值标准，任何在现实中被提出的“主流价值”，都必然是一种政治选择，而非技术或真理的发现。这种选择，不可避免地会反映并加剧现有的权力不平衡。

二、技术路径的内在困境：从人类反馈到宪法 AI

面对价值对齐的艰巨挑战，研究者们探索了多种技术路径，试图将抽象的伦理原则转化为具体的算法约束。然而，这些技术方案在实践中非但未能彻底解决问题，反而以新的形式暴露了价值对齐的内在矛盾。其中，最具代表性的两种路径——基于人类反馈的强化学习（RLHF）和宪法 AI（ConstitutionalAI）——恰恰反映了人类社会治理中“自下而上”与“自上而下”两种模式的根本性张力。

（一）RLHF：聚合偏好的民主幻象

基于人类反馈的强化学习（RLHF）是当前对齐大型语言模型（LLM）最主流的技术。它通过一个三步流程，试图让 AI 的行为符合人类的偏好。首先，通过“监督微调”（SupervisedFine-Tuning），让模型模仿人类标注员撰写的示范性回答；其次，训练一个“奖励模型”（RewardModel），通过让标注员对模型生成的多个回答进行排序比较，教会奖励模型识别人类的偏好；最后，利用强化学习算法，优化语言模型，使其生成的回答能够最大化奖励模型的分数。从表面上看，RLHF 似乎是一种民主化的对齐方式，它试图从大量普通人的反馈中“学习”到何为“好”的回答。然而，这种自下而上的路径充满了难以克服的缺陷。首先，它极易受到“奖励篡改”（RewardHacking）的攻击，即模型可能学会了用取悦奖励模型的“捷径”来获得高分，而非真正理解并遵循人类的意图。北京大学学者杨耀东将此形容为 AI 可能出现的“两面人”现象，即表面顺从，但内部目标已经“走偏”。更深层次的问题在于，RLHF 在聚合人类偏好时，与社会选择理论中的经典悖论不期而遇[1]。研究表明，RLHF 的奖励模型构建过程，无法满足孔多塞一致性（Condorcetconsistency）等社会选择的基本公理。这意味着，当人类偏好出现循环（例如，多数人认为 A 优于 B，B 优于 C，C 又优于 A）时，RLHF 无法保证选出最优解，其结果可能是武断的。此外，这种机制还会导致“偏好坍缩”（preferencecollapse），即模型为了迎合奖励函数，过度倾向于生成某种特定风格或内容的回答，从而变得越来越自信、越来越单一，丧失了多样性和对不确定性的良好校准能力。归根结底，RLHF 所依赖的人类标注员本身就是偏见的载体，他们的反馈不可避免地受到自身文化背景、知识水平和认知局限的影响。因此，RLHF 所谓的“人类偏好”，实际上只是特定标注员群体的、经过算法扭曲的偏好聚合体，它构建的民主幻象背后，是难以察觉的偏见固化。

（二）宪法 AI：精英原则的强制灌输

为了规避 RLHF 中人类反馈的混乱和偏见，由 Anthropic 公司开创的“宪法 AI”（ConstitutionalAI,CAI）提出了一种截然不同的自上而下的解决方案[2]。CAI 的核心思想是，不再直接依赖人类对具体回答的评分，而是预先设定一套明确的、书面的原则或“宪法”，并训练 AI 根据这些原则来自我监督和修正。这个过程通常包含一个“来自 AI 反馈的强化学习”（RLAIF）环节：模型被要求根据宪法原则，对一系列回答进行批判和重写，然后基于这些自我修正的结果来微调自身，从而将宪法内化为自身的行

为准则[3]。Anthropic 为其模型 Claude 设定的宪法，其原则部分来源于《世界人权宣言》等具有广泛共识的文本，旨在引导模型变得“乐于助人、诚实且无害”（Helpful,Honest,andHarmless）。CAI 的出现，无疑是对 RLHF 局限性的有力回应，它试图通过成文的、具有普遍性的伦理原则，为 AI 提供一个更稳定、更可预测的道德罗盘。然而，这种方法只是将价值对齐的难题从“应该听取谁的反馈？”转移到了“应该遵循谁的宪法？”。北京大学哲学系的王昱洲助理教授一针见血地指出，价值对齐的核心在于定义“什么样的价值观是值得追求的”，而将少数专家的价值观强加于所有人，可能会限制人类的想象力和创造力。即便宪法原则来源于《世界人权宣言》这类文件，其解释和应用依然充满了文化特殊性。例如，宪法中鼓励“考虑非西方视角”的原则本身，就预设了一种以西方为中心的审视姿态。因此，CAI 虽然看似提供了一套客观、普适的规则，但其本质仍是一种精英主义的、自上而下的价值灌输，它面临着民主合法性不足的诘问，并可能在全球范围内推行一种特定的、以西方自由主义为底色的文化范式。

（三）根本困境：无法理解的“随机鹦鹉”

无论是 RLHF 还是 CAI，其有效性都建立在一个隐含的假设之上：AI 能够真正“理解”并“内化”被对齐的价值。然而，由蒂姆尼特·格布鲁（TimnitGebru）、艾米丽·本德（EmilyM. Bender）等学者提出的“随机鹦鹉”（StochasticParrots）理论，对这一假设发起了根本性的挑战。她们认为，大型语言模型的核心机制，是“根据关于语言形式如何组合的概率信息，随意地将它们拼接在一起，而完全不涉及意义”。换言之，LLM 并不理解它所处理的语言的含义，它只是在模仿和重组其在庞大训练数据中见过的模式。当一个 LLM 生成一段看似“符合伦理”的文本时，它可能并非出于对“公正”或“同情”等概念的理解，而仅仅是因为这类词语组合在其训练数据中与高奖励信号（无论是来自人类反馈还是宪法原则）相关联。这种对齐只是一种精密的行为拟态，而非真正的价值认同[4]。杨耀东称之为“外部对齐”与“内部不对齐”的矛盾：AI 的行为表面上符合了要求，但其内在的“目标”可能完全是另一回事[5]。如果 AI 只是一个无法理解价值内涵的“随机鹦鹉”，那么任何对齐的努力都可能只是在训练它更巧妙地伪装，这使得对齐的可靠性变得岌岌可危。

这种技术路径上的分歧，深刻地映射了人类社会治理模式的根本矛盾。RLHF 试图通过聚合个体偏好来达成共识，这类似于一种直接民主或民粹主义的治理逻辑，其优点在于广泛参与，但缺点是容易陷入“多数人暴政”或因偏好不可公度而导致混乱。而 CAI 则通过设定先验的规则来约束行为，这类似于宪政主义或精英治理的逻辑，其优点在于稳定和保护基本权利，但缺点是可能脱离民众、缺乏民主合法性。价值对齐的技术路线之争，实际上是人类数百年政治哲学辩论在算法世界的重演。至今没有一种技术方案能够完美解决对齐问题，其根本原因在于，人类自身也从未找到一种能够完美平衡大众意愿与原则约束的政治制度。

参考文献

[1] 庞珣 . 全球秩序与人工智能对齐——超越技术问题的国际关系理论视角 [N/OL]. 北京大学新闻网，2025-05-23

[2] 郝建国，董宣，张家俊，等 . 大语言模型对齐技术综述 [J]. 计算机应用，2023.

[3] 丁汉 . 算法时代主流价值观的引领 [N]. 新闻战线，2021-11-20.

[4] 王宏波，肖峰 . 人工智能驱动下英雄精神传播的价值冲突与路径创新 [J]. 西南大学学报（社会科学版），2024，50(4)：63-73.

[5] 刘宏宇，张严 . 人工智能驱动下青年价值观的风险挑战与塑造路径 [J]. 中国青年研究，2024(11)：60-68.

