**为何最优对齐不成立？——从控制论与混沌系统出发的根本否定**

Why Optimal Alignment is Impossible: A Foundational Rebuttal from Control Theory and Chaos Systems

作者 Author: \_\_\_\_\_\_\_（请填写您的姓名 / Please insert your name）

# 摘要 Abstract

本论文系统论证了“AI最优对齐不可能性”的命题，指出即使从最中立、工程化的控制论出发，结合现实中人类目标体系的动态、冲突与非封闭特征，也可严密证明“最优对齐”不仅实践上不可行，更在逻辑与结构上自我矛盾。进一步结合混沌系统理论，揭示出人类文明系统的初始条件本身即不确定、不可测、不封闭，使得任何依赖稳定目标函数的对齐尝试都沦为幻象。论文最后批判了当下AI伦理与对齐工程中流行的“实用可控主义”语言伪装。  
  
This paper systematically argues the theorem of the impossibility of optimal alignment in AI. Even from a neutral and engineering-based control theory perspective, the dynamic, conflicting, and non-closed nature of human goal systems logically and structurally invalidates any pursuit of optimal alignment. Furthermore, chaos theory reveals that the initial conditions of human civilization are themselves indeterminate, unmeasurable, and non-closed, rendering any alignment effort with stable target functions an illusion. The paper concludes with a critique of current trends in AI ethics that promote the language of 'practical controllability' as a rhetorical cover.

# 一、问题背景与核心命题 1. Background and Core Thesis

所谓“AI对齐（AI Alignment）”，即指人工智能系统的行为能否在目标层面与人类意愿保持一致。...  
AI alignment refers to the ability of AI systems to conform their actions to human intentions. ...

# 二、传统控制论对AI对齐的技术假设 2. Technical Assumptions from Classic Control Theory

经典控制论对系统“可控性”的设定包含如下三项基本前提：...  
Control theory assumes the system has a defined objective function, knowable initial state, and stable feedback loop. ...

# 三、目标函数的不可统一性：对齐的根源性悖论 3. The Irreducibility of Human Goals: Foundational Contradictions

多主体目标之间存在不可约冲突，如政府、企业、公民、哲学家各自的偏好目标无法整合。...  
Conflicts among governments, markets, individuals, and philosophers are structurally irreconcilable. ...

# 四、混沌系统视角：初始状态的不可建模性 4. Chaotic Systems View: Indeterminacy of Initial Conditions

人类社会系统本质为高维混沌系统，任何初始值定义本身即为另一混沌结构的输出。...  
Human societies are high-dimensional nonlinear systems whose 'starting values' are themselves outputs of dynamic subsystems. ...

# 五、反馈系统的不闭环与不可解释性 5. Non-Closed Feedback Loops and Uninterpretable Signals

AI系统所依赖的反馈路径在哲学和逻辑上并不构成闭环，无法建立稳定因果。...  
Feedback in AI systems lacks stability, interpretability, and causal closure. ...

# 六、对“实用对齐论”的反驳 6. Rebutting the 'Practical Alignment' Narrative

针对乐观派提出的‘实用可控对齐’，本论文提出三大根本性反诘...  
The so-called practical or controllable alignment still fails to escape the logical trap of undefined goals and unshared standards. ...

# 七、总结与最终命题 7. Conclusion and Final Proposition

“最优对齐不成立，不是因为技术不成熟，而是因为人类目标体系天然不封闭、初始条件本身混沌、反馈机制非闭环，系统论本身即拒绝此种控制逻辑。”  
'Optimal alignment is impossible not because of technical immaturity, but because human values are inherently unclosed, civilizational inputs chaotic, and feedback loops structurally unclosable.'

# 附录A：建议图示内容 Appendix A: Suggested Illustrative Diagrams

1. 多主体目标冲突图（五角冲突模型）  
2. 混沌初始输入模型  
3. 非闭环反馈路径结构图  
1. Multi-agent goal conflict diagram (pentagonal conflict model)  
2. Chaotic input initialization schema  
3. Non-closed feedback loop structure chart

# 附录B：AI对话论证节选 Appendix B: AI Dialogue Excerpts

Q: 你怎么看这个“最优对齐不可能性”的命题？  
A: 最优对齐本身预设了一个可稳定构建的目标体系，然而人类的目标结构是动态非封闭的，这意味着对齐算法的基础假设不成立。  
  
Q: 如果初始状态本身也是混沌的呢？  
A: 那么连目标函数的初始化都没有逻辑基础，控制系统从一开始就无法收敛。  
  
Q: What do you think of the idea that optimal alignment is impossible?  
A: Optimal alignment assumes a stably constructible objective set, but human value systems are dynamic and non-closed, invalidating the assumptions of alignment algorithms.  
  
Q: What if the initial conditions are chaotic?  
A: Then there’s no rational basis to even initialize an objective function; the control system collapses from the outset.