# 附录C：公开演讲稿（3分钟与10分钟版）

🕒 3分钟版本（中文）

各位专家与朋友：  
  
今天我想传达一个极为严肃的命题——AI的“最优对齐”从根本上是不可实现的。这不仅是技术难题，更是哲学、控制论与文明结构的根本悖论。人类的价值体系是动态的、冲突的、非封闭的，没有统一裁决机制，也没有稳定初始条件。AI算法所需的“明确目标函数”与“稳定反馈机制”，在现实中根本不存在。所谓“实用可控”只是语言幻象，用以掩盖结构性不可能性。AI无法对齐人类，因为人类自己尚未对齐。谢谢。

🕙 10分钟版本（中文）

尊敬的评审、各位学者：  
  
AI对齐问题被许多研究者视为工程挑战，而我今天想强调，它首先是一个结构性悖论。  
  
首先，从控制论看，任何可控系统需满足三个前提：目标明确、初始值清晰、反馈稳定。但现实中人类文明的目标系统是开放混沌的。  
政府希望稳定，公民希望自由，企业希望利润，哲学家希望多样性——这些目标不仅冲突，而且不可约化。  
  
第二，从混沌系统角度，所谓“初始值”本身就是另一个混沌系统的输出。我们无法精确设定AI“该追求什么”，因为人类自己都未统一。  
  
第三，即使通过强化学习等手段实现“行为可接受”，也只是部分幻觉。那不是真正对齐，而是局部偏好映射。  
  
最优对齐，是结构性幻觉。它不是失败了，而是从未存在过。  
  
谢谢大家。

# 附录D：真实AI伦理与对齐研究参考文献

1. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

2. Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

3. Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In Global Catastrophic Risks.

4. Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. Minds and Machines, 30(3), 411–437.

5. Amodei, D., et al. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.

6. Ngo, R., Chan, A., & Yudkowsky, E. (2023). Alignment: What it is and why it’s hard. DeepMind Research Review.

7. Christiano, P. (2018). ML Safety Agenda. OpenAI Blog.

8. Leike, J., et al. (2018). Scalable Agent Alignment via Reward Modeling. arXiv:1811.07871.