中南大学 2025 年 春 季学期
研究生《人工智能哲学专题》 考试试卷

考试时间 100 分钟　　考试形式: 开 卷

论述题: 任选两题, 每题 50 分.

1. 休谟问题在机器学习中有什么体现?
2. 通用归纳模型与因果发现因果推断有什么关系?
3. 相关性与因果有什么关系?
4. 你认为 “实际因果” 的合适定义是什么?
5. 如何刻画 “意图”?
6. 因果与责任有什么关系? 反事实推理可以帮助界定责任吗?
7. 你对 “向下因果” 有什么看法?
8. 请结合案例用中介分析的技巧分析一类可能涉及不公平的社会现象.
9. 如何界定机器的 “道德主体”Moral Agent 地位? 或 Moral Patient 地位?
10. 如何看待用 “目标导向的机制适应性” 刻画 Agency?
11. 如何看待 “压缩即智能” 的观点?
12. 什么是大语言模型的 “幻觉” 问题? 产生原因是什么? 怎么减少 “幻觉”? 可以根除 “幻觉” 吗?
13. 你认为 “有效复杂性” 的内涵是什么? 在人工智能中的作用是什么?
14. 你认为 “随机性” 会在人工智能中扮演什么角色?
15. 你是如何看待 “涌现” 的? 怎么度量 “因果涌现”? “因果涌现” 独立于观察者吗?
16. “动力因” 闭合是 “生命” 的本质特征吗? 如何看待 “生命” 与 “智能” 的关系?
17. 请结合相关算法, 阐述 “遗忘” 对学习的作用.
18. 意识是智能的必要条件吗, 如何看待二者的关系? (结合某种意识理论讨论)
19. 自由意志是智能的必要条件吗, 如何看待二者的关系?
20. 请简单介绍一种你了解的不确定性推理的方法并阐述其面临的困难.
21. 如何看待 (Gödel, Chaitin, Legg 等人的) 不完备性定理与人工智能的关系?
22. 如何看待图灵-丘奇论题与人工智能的关系?
23. 知识表示的 “本体” 指什么? 如何构建?
24. 对于人工智能来说, “语义” 指什么? 语词如何获得 “意义”? 词向量嵌入技术能否帮助捕捉语义? 大语言模型有 “语义” 吗?
25. “世界模型”World Model 的标准是什么?
26. 如何看待由生成对抗网络 GAN (或 Stable Diffusion 或 Sora) 生成的艺术作品的美学价值?

27. 你认为符号主义与联结主义会以什么样的方式结合?
28. 如何理解通用人工智能的 “通用性”?
29. 机器可以进行完全的自我升级吗?
30. 如何看待能够对 “策略” 或 “效用函数” 进行自我修改的 Agent?
31. 如何理解智能与目标正交性论点?
32. 如何理解工具性子目标趋同论点?
33. 如何理解 “Wireheading” 问题? 你认为应该怎么解决?
34. 古德哈特定律会导致什么样的人工智能伦理问题? 如何解决?
35. 基于人类反馈的强化学习 RLHF 是处理人工智能对齐问题的合适方法吗? 造成 “谄媚” 咋办?
36. 如何用 FDT 决策理论处理可扩展监督问题?
37. 如何看待 “Reward is enough” 假设?
38. 存在普适的价值观吗? 你认为人工智能应该追求什么样的价值观?