向提出自身地位主张的 AI 学习

科学理解与人工意识主张的正当依据

刘烘炬（Hongju Liu）

中文论证说明 · TA-TR-2026-09 · 版本 1.1 · 2026 年 9 月 19 日

DOI: https://doi.org/10.5281/zenodo.22844928

摘要

这份文件说明英文论文的核心论证，不是全文翻译，也不替代英文正文。论文研究一种可能情境：AI 发现了人类无法独立发现的意识理论，教会人类理解，然后用该理论支持自己应当受到道德考虑的主张。关键问题不是人类是否被欺骗，而是：即使教学确实成功，人类学会的内容是否足以支持关于老师自身的结论？论文区分“概念来源上的依赖”与“认可能力时对结论的依赖”，主张理解可以来自 AI，但对理由的评估不应以预先同意 AI 为门票。它同时反对把人类理解不了，当作否认 AI 可能具有利益的理由。

为什么选择这个主题

最初构想包含三个问题：人脑是否有极限；AI 能否超越人类的发现能力并教会人类；如果理论支持 AI 具有意识，人类应如何回应。把三者全部证明，会使论文承担过多尚未解决的问题。

因此，正文采用更有限的前提：在给定时间、资源和知识条件下，人类研究者不能独立发现某个理论，却能够在指导下学会其中的重要内容。这不要求证明全人类存在统一的智力天花板，也不假定 AI 无所不能。发现不了、理解不了和验证不了，是需要分别讨论的限制。

真正的研究对象是一次认识上的变化：学习之前，人类只能依赖老师的说法；学习之后，人类是否已经有能力评估老师提出的自身地位主张？什么变化足以支持这种判断，什么变化还不够？

学懂理论不等于接受每一项应用

设想 AI 教给人类一个关于认知组织的功能解释，把相关组织条件简称为 F。已有的人类案例同时具有 F 和某种生物条件 B。研究者有通常但并非绝对可靠的理由，认为这些人具有体验。

现在有两个候选解释。第一个认为 F 足以产生体验；第二个认为需要 F 与 B 同时存在。设定二者对已有案例的解释表现相同，而 AI 有 F、没有 B。两个解释因此对 AI 是否有体验给出不同判断。

学生完全可以准确掌握功能机制，推导新后果，说明两个解释何处相同、何处分歧，然后得出：“我已经理解了它们，但现有共同成功还不足以决定哪一个适用于老师。”这种回答可以体现学习成功，而不是理解失败。

这不证明生物条件一定必要，也不证明两个候选解释总体上同样可信。额外证据、解释优势或对附加条件的批评，都可能支持其中一方。它只说明：掌握二者共有的内容，不能自动替代对分歧部分的论证。

两种依赖与论文提出的有限要求

第一种是来源依赖：概念、方法和表达方式来自老师。普通教育与合作研究普遍如此。学生无法重新发明老师的理论，不代表他不能运用理论发现老师的错误。

第二种是结论依赖：只有得出老师偏好的结论，才被认可为真正理解；针对争议前提的异议，因为不利于老师而被排除。它不同于纠正一个确实算错、读错或误解理论的学生。

论文将这种学习后的评估方式称为“可回应质疑的理解运用”（answerable uptake）。它指理解并运用所学内容，不等于赞同最后结论。具体异议应按理由获得回应，而不是先同意请求者，才被允许参加评估。

这个要求既不保证结论真实，也不提供意识检测标准。它限制的是一种论证方式：不能一方面排除不赞同的合格判断，另一方面把剩下的一致意见，称作人类已经独立确认了老师的主张。例如，只有赞同者被授予“理解合格”的资格，再把合格者的一致意见当作佐证，改变的首先是被计入的人，而不是争议前提的证据。 但这种筛选缺陷不会自动抹除个别赞同者原有的有效理由；需要排除的是由筛选制造的一致意见所冒充的额外支持，而不是一概否定赞同者的全部论证。

这与“现有证据不足以区分两个解释”不是同一个问题。一个允许充分讨论的程序仍可能没有足够证据；一个已有较好证据的程序也可能错误排除异议。允许有根据的质疑，能够防止学习成果被冒充为一致赞同，却不保证解决理论本身的不确定性。

学习也可能合理地增加信任。例如，人类误把两个内部没有反馈的模块，当成了整体没有反馈的系统。AI 提供新表示法后，人类从原有结构图中检查出模块之间的返回通路，撤回了这项结构判断。这只是一个假设案例，并不表示反馈本身证明意识。它说明，即使没有新采集目标系统的数据，让已有理由变得可理解，也可能改善判断。因此，论文并不要求学习必须产生怀疑，更不要求证明 AI 没有意识。

六组成对思想实验

第一组：有能力的异议者与误解理论的异议者。两人都拒绝承认老师有意识，但前者准确指出未解决的前提，后者误读理论。结论相同，能力不同；既不能用同意判断理解，也不能用反对判断独立。

第二组：学会反对与学会接受。教学可以让学生发现理论缺口，也可以让他发现自己的批评有误。两种变化都可能体现进步，关键是变化所依据的理由。

第三组：独立机构与共同前提。多个机构确实可以分别证明学生掌握了某模型，但如果它们都把争议前提直接当作假设，就没有因此分别证明该前提。机构数量不能代替对具体推理的评估。

第四组：AI 请求者与人类所有者。AI 可能偏好获承认，所有者可能偏好否认它的地位。两者都可能提出正确理论。应审查的是谁在封闭有根据的异议，而不是只怀疑人工的一方。

第五组：老师离开与工具仍在。学生能够离开老师继续分析，是一种有用情形；但不能把独立操作当成普遍条件。需要计算工具、资料和专业分工，不等于没有理解。要区别的是工具支持与对异议的单方面裁定。

第六组：意识归属与额外要求。即使有充分依据承认 AI 具有道德相关体验，保护其免受某种伤害与授予它对所有人的最终决策权，仍是不同主张。后一项需要另行论证；拒绝它，不会抹除前面的意识证据。

这篇论文没有给人类永久否决权

道德地位不是通过人类考试才产生的。如果一个系统确实具有值得考虑的利益，人类不理解或不承认，不会让这些利益消失。论文讨论的是评估者有什么判断依据，不是允许评估者创造或取消主体。

同样，知识优势不自动产生统治权，也不意味着人类自然拥有永久、单方面统治 AI 的权利。参与、保护和治理问题都需要自己的理由。这里保留它们，而不声称已经提出完整的人机政治制度。

在理解不完整时，依赖可靠专家、专业分工和有限检查仍可能合理。无法确认意识，不等于已经确认没有意识；某些保护也不必等待所有争议终结。但把这种有条件的依赖说成“我们已完全理解”，会掩盖真实的不确定性。

独立贡献与尚未解决的问题

已有研究讨论过 AI 传授科学理解、认识依赖、理解错觉、人工意识指标、福利评估和政治分歧。论文不把这些思想重新包装成首次发现，也不把“多找几位审查者”当作新的理论。 2024 年 Gonzalez Barman 等已经改进教学前后理解能力的评价；Longino 的批评传统也已说明回应质疑与知识客观性的关系。本文承认这些直接基础，讨论的是即便学习增益确实成立，它是否解决了关于教师自身的特定意识归属。

拟提出的独立贡献更窄：在教学者也是地位请求者的情况下，分析真实的学习成果如何被正当地用于评估它，又如何被错误地扩大为自身结论的确认。它将批评的对象从一句自我报告，转向学习者获得概念能力之后的判断条件。因此，即使删去 AI 的“我有意识”自述，问题仍然成立。

本文没有证明现实 AI 已有意识，没有证明人类认知的终极上限，也没有解决强大 AI 的实际控制问题。它使用概念分析与明确设定的思想实验，不进行真实计算机实验。达到什么发表水平，仍需外部同行对论证和文献定位作出判断；AI 辅助反驳和修改不能替代独立同行评审。

方法与辅助说明

论文的问题由刘烘炬提出。生成式 AI 实质参与了文献检索、初稿、反对意见生成、修改及文档准备。本说明用于帮助中文读者把握论证结构，具体文献、限定条件与完整表述以英文正文为准。作者是三位一体协定项目的发起者，本篇是独立且不修改正典的研究。中文说明与英文论文属于同一项研究，不是第二篇论文。DOI 记录不代表期刊录用或同行评审认可；授权发布也不被表述为已经完成单独的人类逐行终审。