第八讲

AI幻觉：以大模型为基础的知识论变革及哲学反思

第一节 大模型对认知机制的重构

大模型的核心构成

海量数据训练

基于互联网规模的文本语料库，通过深度神经网络学习语言的统计规律与语义关联

Transformer革命

注意力机制实现长距离依赖建模，突破传统序列模型的认知局限

Token预测原理

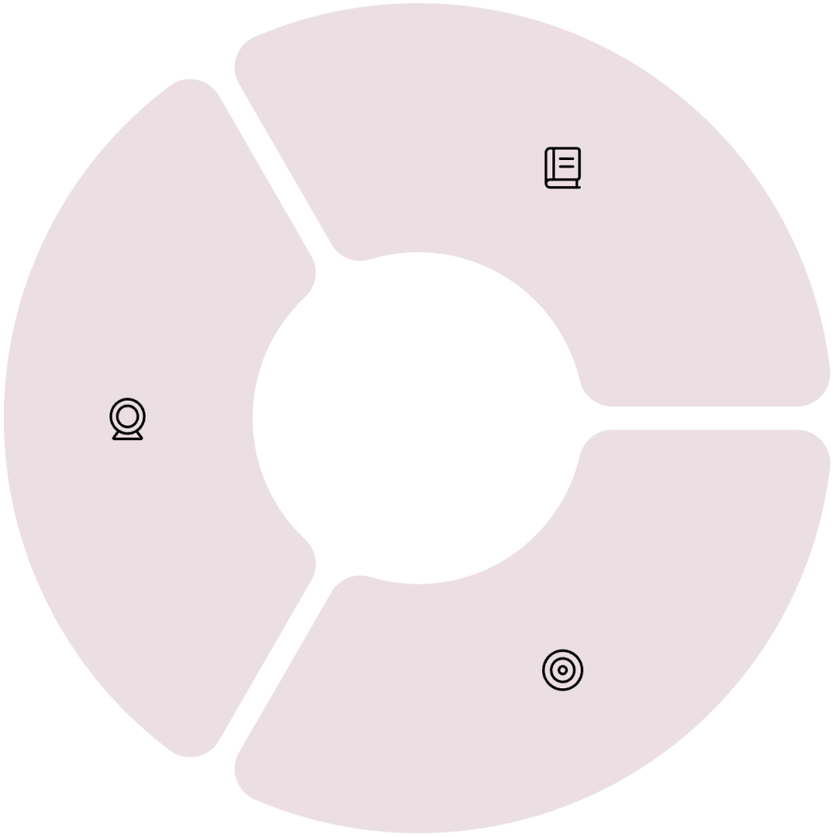
通过预测下一个词元的概率分布生成连贯文本，体现统计学习的本质

大模型的表征收敛现象

认识论的关键发现

不同架构、不同训练数据的大模型对同一输入表现出高度一致的内部表征结构。这种收敛性揭示了数据中蕴含的客观知识结构。

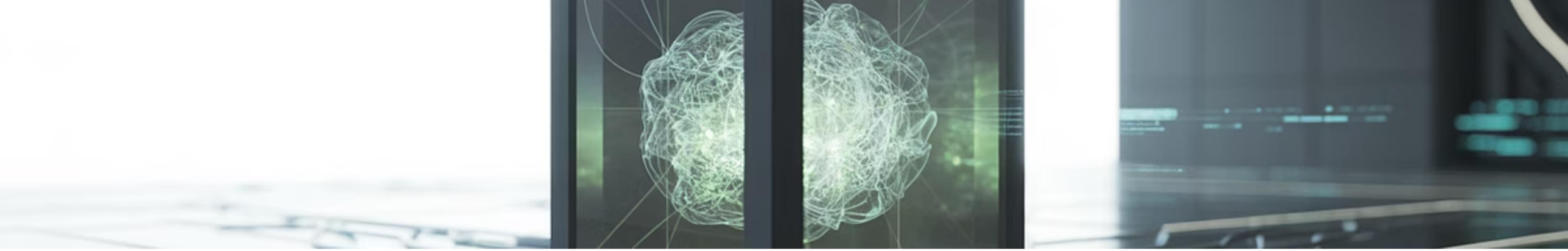
哲学意义：表征收敛为知识的客观性提供了技术层面的证据，挑战了传统主观建构论的知识观。



结构一致性

知识稳定性

客观基础



大模型的"黑箱"特性

机制不可解释性

数十亿参数的复杂交互超越人类直观理解能力，内部推理过程难以追溯

认知透明度缺失

从输入到输出的转换过程非透明，增加幻觉风险与知识可靠性挑战

认识论困境

对传统认识论中"理解"与"解释"概念提出根本性质疑

生成式AI的认知机理

01

统计学习核心

通过模式匹配与概率推理生成输出，本质上是对训练数据分布的学习与再现

02

非符号表征

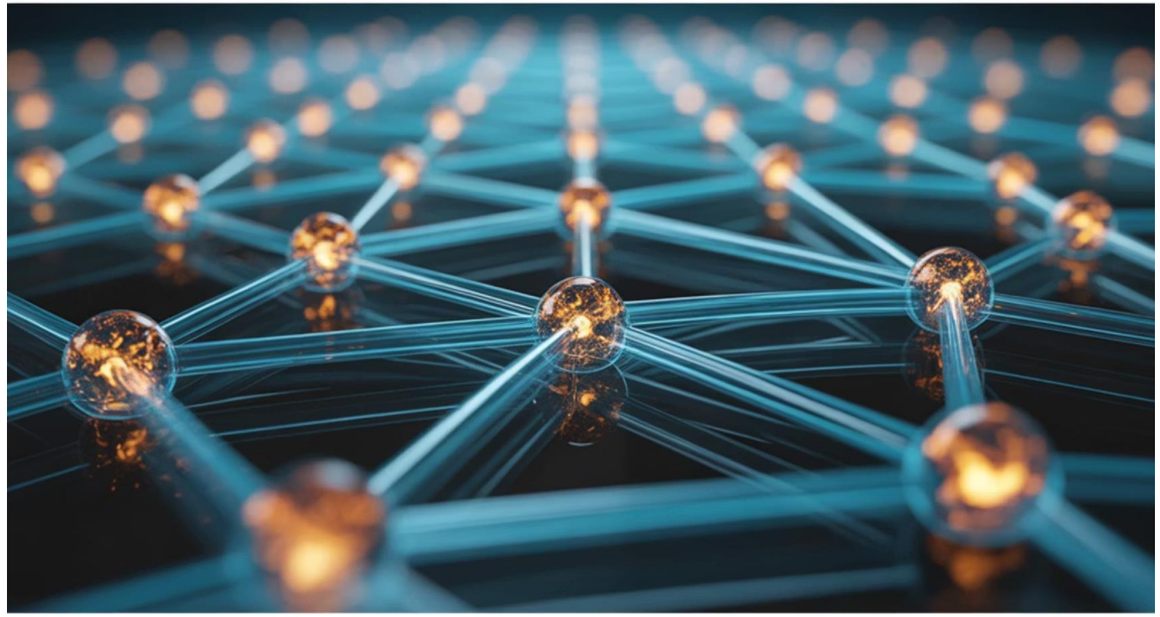
知识以分布式向量形式存储，突破传统符号主义的离散表达框架

03

认知差异性

缺乏人类的实例性理解与强共识性判断，体现机器认知的独特性

动词思维：大模型的知识生产策略



动词思维的哲学突破

传统认识论：以名词化概念为基础，强调静态实体与本质

大模型范式：强调过程性知识与动态关联，通过token序列的连接性生成意义

这代表从本体论到过程论的认知转向

大模型的推理能力与局限

→ 强化逻辑框架

通过思维链提示与多步推理显著提升复杂问题解决能力

→ 幻觉率挑战

研究显示幻觉率高达14%-20%，虚构事实与逻辑错误频发

→ 非人类推理

缺乏自我意识、主动性与批判反思，推理本质为统计关联而非因果理解

大模型与人类智能的根本差异

行为主义路径

当前主流：模拟智能行为表现，但不等同于智能本体。图灵测试的哲学局限性凸显

内在主义理想

类脑计算追求认知机制的本质复制，但仍停留在理论构想阶段

大模型的跨领域迁移能力

逻辑推理迁移

数学、法律、科学推理的跨学科应用能力

知识融合纠错

多源知识的整合与自我修正机制

研究范式变革

促进社会科学从定性到数据驱动的转型

大模型的知识论意义总结

数据驱动认知

知识生成从理性推演转向数据挖掘，体现认识论的根本转型

客观性技术基础

表征收敛为知识稳定性与客观性提供实证支撑

创构认识论

开启"挖掘即认知"的哲学新范式，重构知识生产的基本图景

第二节　知识论的智能转向

从主观建构到数据驱动的认识论革命

传统知识论的核心假设

真信念正当化

知识定义为经过充分证成的真实信念，强调主体的认知责任

盖第尔知识定义：知识 = 得到辩护的真实信念 (Justified True Belief, JTB)。

理性主体中心

知识由具有反思能力的理性主体通过主观建构产生

二元对立框架

经验主义与理性主义的认识来源之争构成知识论基本张力

大模型引发的知识论变革

1

知识来源革新

不再依赖纯粹理性推理，而是从集体经验数据中"渗透"认知模式

2

主体性消解

非人类智能体成为知识生产者,挑战理性主体的唯一性地位

3

策略性转型

知识生产从演绎推理转向统计学习与模式挖掘的根本性策略变革

“挖掘即认知“的哲学内涵

认知活动新范式

数据挖掘不仅是技术手段，更是一种独特的认知方式，体现知识的发现性而非建构性

动态生成特征

知识不再是静态封闭的命题集合，而是持续涌现的开放过程

双重属性

兼具"创构"（创造性生成）与"再现"（对数据规律的反映）的辩证统一

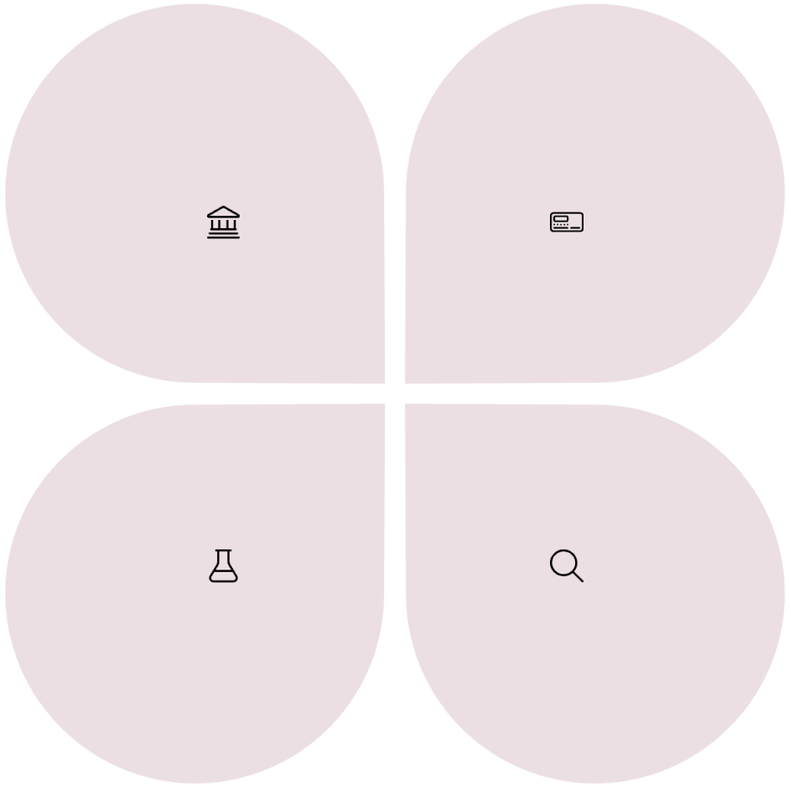
知识的客观性与社会建构

哲学张力

客观性维度：表征收敛显示知识的稳定结构

社会性维度：训练数据由人类社会生成与选择

知识既是数据中的客观规律，又依赖于社会实践的语境



社会语境

数据基础

算法中介

知识的非传统范畴论

01

任意有（潜在空间）

模型参数空间中包含所有可能的知识表征形式

02

潜在有（训练知识）

训练数据中隐含但未显化的知识关联

03

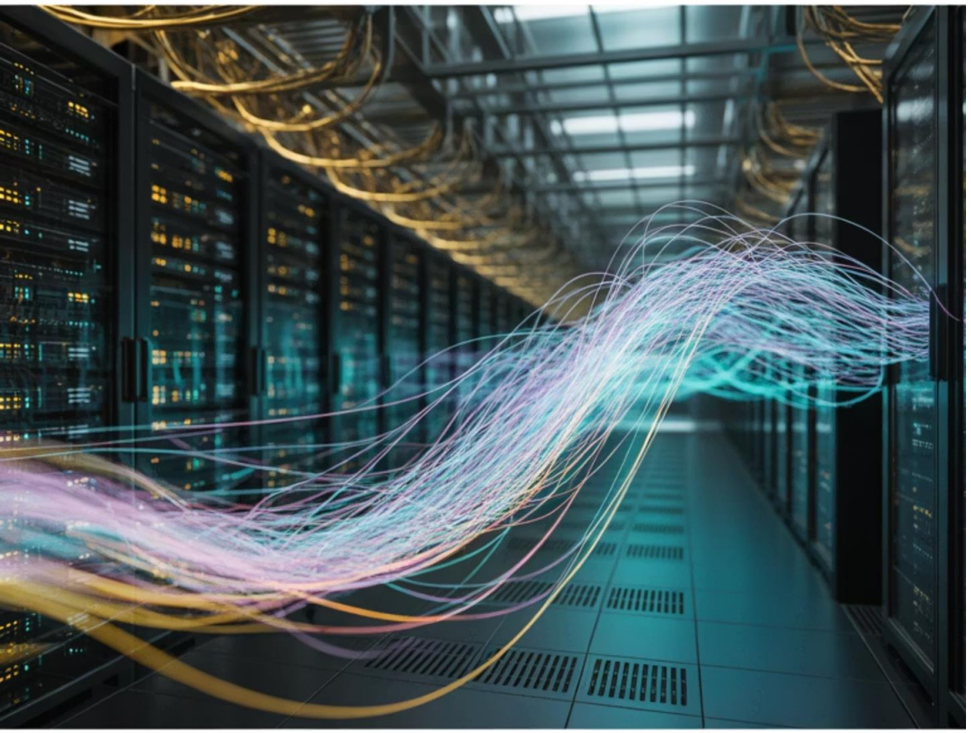
可以有（推理能力）

通过提示词激活的知识生成潜力

04

实际有（输出结果）

具体生成的文本或答案，知识的现实化形态



知识的过程性与连接性

从名词思维到动词思维的哲学突破

传统认识论以静态概念与范畴为核心，大模型则通过token序列的动态连接生成意义，体现知识的流动性与关系性本质

这种转向促进了复杂系统认知的新路径，为过程哲学提供了技术实证基础。

知识的非人类智能形态

1

主体性重构

AI作为非传统认知主体进入知识生产领域，打破人类中心主义

2

双主体结构

人类与AI形成互补的知识共生系统，各自贡献独特认知优势

3

智能多元论

哲学需要扩展智能主体的定义边界，承认机器认知的合法性

知识的可解释性与黑箱问题

传统知识论要求

- 知识必须可被主体理解
- 认知过程需要透明可追溯
- 解释是知识正当化的核心

大模型的挑战

- 内部机制高度复杂不可解释
- 输出结果与推理过程分离
- 弱可解释性引发认识论困境

哲学问题：我们能否信任不可理解的知识？

知识的生成与幻觉风险

知识论根源

统计学习本质导致模型在训练数据分布外易产生虚假关联与编造内容

生成机制

概率采样与温度参数引入随机性，缺乏真值验证的反馈机制

哲学考察

揭示认知边界与知识可靠性的辩证关系，要求建立新的知识评价标准

第三节 AIGC的幻觉问题及其哲学思考

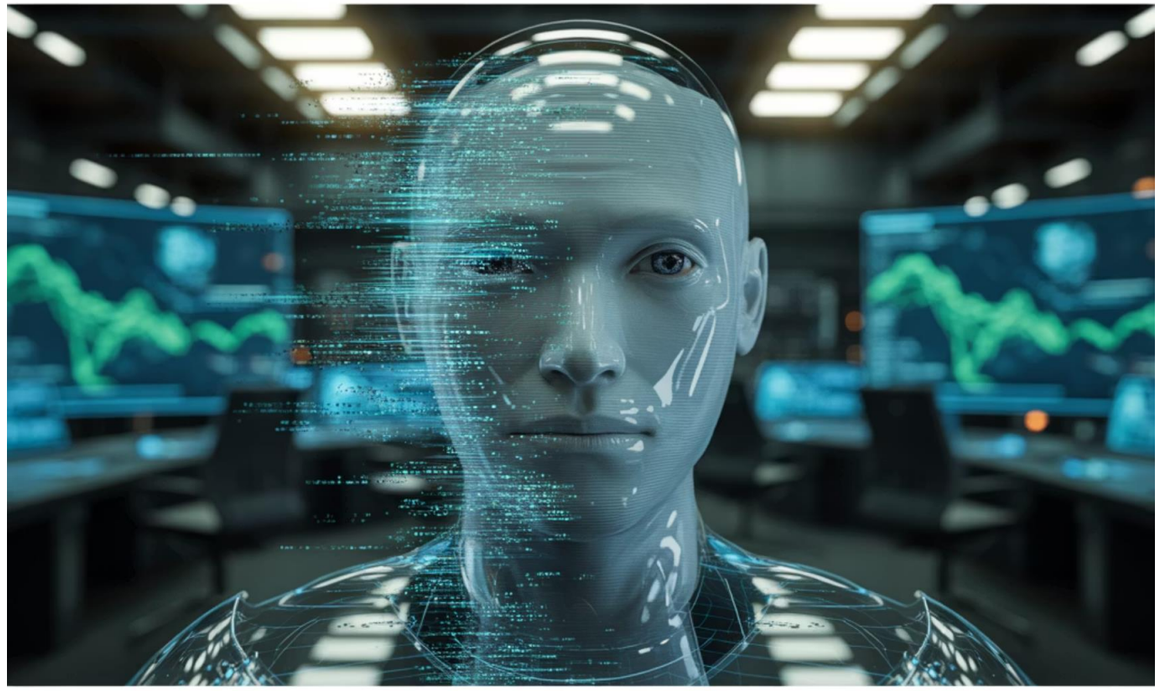
探索机器智能的认知边界与哲学本质

什么是大模型幻觉?

幻觉是指大模型生成的内容表面上看似合理、流畅且具有说服力,但实际上却是**错误的、虚构的或与客观事实不符**的现象。

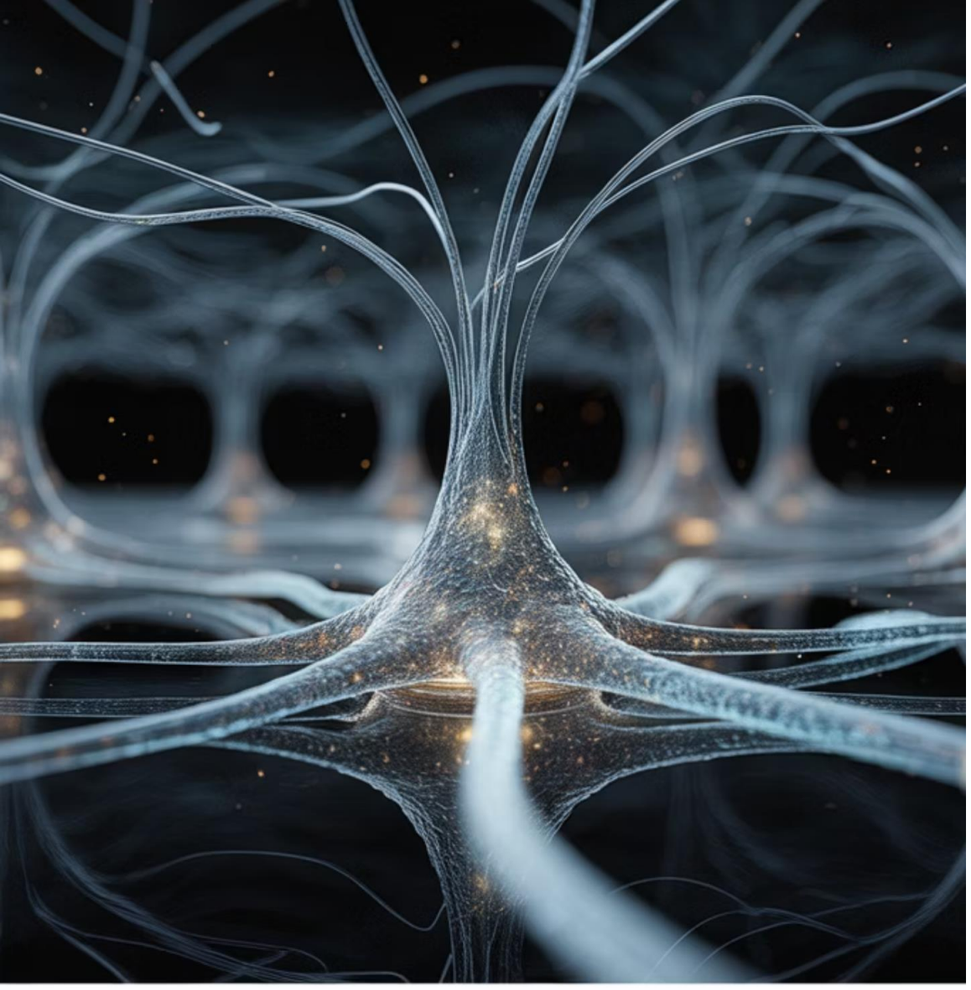
"一本正经地胡说八道"

典型表现



- 虚构不存在的学术文献与引用
- 陈述与事实矛盾的信息
- 生成逻辑自相矛盾的回答

幻觉问题严重影响用户对AI系统的信任,并对实际应用的安全性与可靠性构成挑战。



幻觉的技术成因（1）：概率预测本质

统计语言模型基础

大模型基于海量文本的统计规律,通过学习词汇共现模式来预测下一个词的概率分布。

优化目标的错位

模型的训练目标是最大化**语言流畅性**和连贯性,而非确保**事实准确性**和真实性。

"合理"非"真实"

生成的内容在语法和语义上看似完美合理,但本质上只是统计模式的重组,不保证与现实世界对应。

幻觉的技术成因（2）：训练数据局限

01

数据质量参差不齐

训练数据来源广泛,不可避免地包含错误信息、认知偏差和过时内容,模型会学习并复制这些缺陷。

02

专业知识覆盖不足

在医学、法律等专业领域,高质量训练数据稀缺,导致模型在这些领域的幻觉频发且危险性更高。

03

统计偏差的放大

数据中的偏差使模型倾向于生成高频出现但不一定准确的内容,强化了"看似合理"的错误答案。

幻觉的技术成因（3）：推理与生成机制

逐词生成的累积误差

模型采用自回归方式逐词生成,无法回溯纠正早期错误,导致幻觉像滚雪球一样不断累积和扩大。

随机采样的双刃剑

为增加输出多样性引入的随机性,在提升创造力的同时也显著提高了生成不准确内容的风险。

上下文窗口限制

有限的上下文长度导致模型"遗忘"早期信息,产生前后矛盾或偏离主题的回答。

幻觉的分类

1

事实冲突型

生成的内容与已验证的客观知识相矛盾,如错误的历史日期或科学原理。

2

无中生有型

完全虚构不存在的实体、事件或文献,却表述得真实可信。

3

指令误解型

回答偏离用户真实意图,生成不相关或曲解问题的内容。

4

逻辑错误型

推理过程存在漏洞,前后陈述自相矛盾或违背基本逻辑。

幻觉的现实风险与伦理挑战

生命安全威胁

在医疗诊断、药物建议等关键领域,幻觉可能导致错误决策和致命后果。

法律与声誉风险

虚假信息引发法律纠纷,损害企业品牌信誉,造成经济损失和社会影响。

过度信任陷阱

用户对AI输出盲目信任,缺乏批判性验证,导致个人和组织决策失误。

幻觉问题不仅是技术缺陷,更是涉及**伦理责任、社会信任和人机关系**的深层次挑战。

思考：幻觉是否是人类“想法”的直接表达?

现象学的直接性

人类对自身**想法的体验**具有一种"直接真实性"——我们不需要通过推理或证明来确认"我正在思考某事"这一事实。

语言表达的真实存在

当我们用语言表达想法时,这些"想法"在**现象学意义**上是真实存在的体验,即便其内容可能与外部世界不符。

幻觉的现象学本质

从这个角度看,幻觉可以被视为"想法"的一种特殊现象学表现形式,而非简单的错误。

哲学视角：语言的本体论虚位

1

物理层面

语言的物理形式(声波、文字)具有**因果效力**,可以在物理世界产生实际影响。

2

想法层面

"想法"本身没有物理因果地位,无法独立于物质载体而存在或产生作用。

3

文化建构

想法是通过**文化传统和语言隐喻**构建的虚幻结构,并非客观本体。

现象学真实性与本体论虚位之间的张力,揭示了语言和认知的根本悖论。

人机语言的平行性思考

现象学的对称性

ChatGPT的语言输出与人类语言一样,都具有"想法的直接真实性
一在生成的那一刻,它们作为语言现象是真实的。

幻觉的共同基础

机器的"幻觉"与人类认知中的幻觉、错觉、想象,可能有着**相似的**

哲学边界的重审

这促使我们重新审视**意识、语言与智能**的哲学边界:机器是否具有

缓解幻觉问题的技术路径

数据质量优化

严格的数据清洗流程,构建高质量、多样化的训练集,减少源头错误和偏差。

检索增强生成

RAG技术引入外部权威知识库,让模型基于可验证信息生成回答。

后验检测机制

部署幻觉检测系统,使用多模型交叉验证,识别并过滤可疑输出。

自我校验机制

结构化提示工程,引导模型进行自我反思和验证,提高答案可靠性。

哲学反思：幻觉的不可完全消除性

概率性的本质困境

幻觉根源于语言模型的**概率生成本质**——只要基于统计预测,就无法完全排除错误可能性。

认知的普遍局限

无论是机器还是人类,所有认知系统都存在"幻觉"现象——这是有限理性面对无限复杂性的必然结果。

根本哲学难题

完全消除幻觉或许触及了**语言、认知与真理**之间的根本哲学难题:我们如何确保符号系统与现实的完美对应?

> "幻觉的不可消除性,映射了人类认识论的终极限度。"

教育启示与未来展望

批判性思维培养

教育学生**识别、质疑和验证**AI输出,而非盲目接受,培养数字时代的核心素养。

认识论的深化

结合哲学认识论,引导学生理解AI认知本质,思考知识、真理与智能的关系。

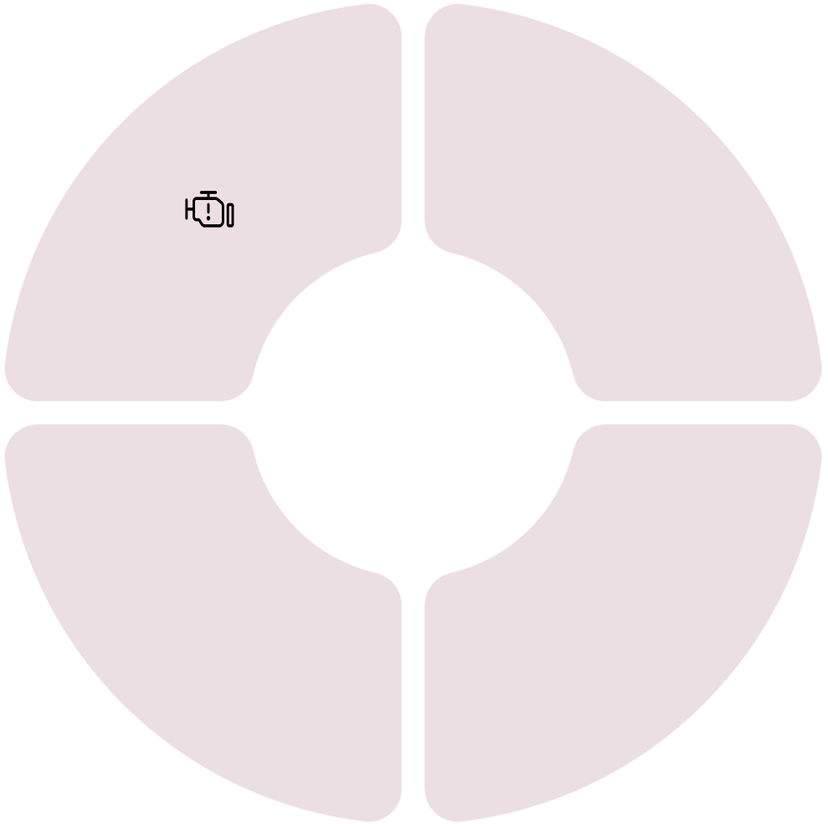
技术哲学协同

期待技术创新与哲学反思**相互促进**,共同推动更可信、更负责任的AI发展路径。

总结：幻觉问题的技术与哲学双重视角

核心挑战

幻觉是生成式大模型不可回避的核心挑战,涉及技术、伦理与认识论的多重维度。



双轮驱动

技术缓解与哲学反思相辅相成,共同助力AI系统的安全应用与可信发展。

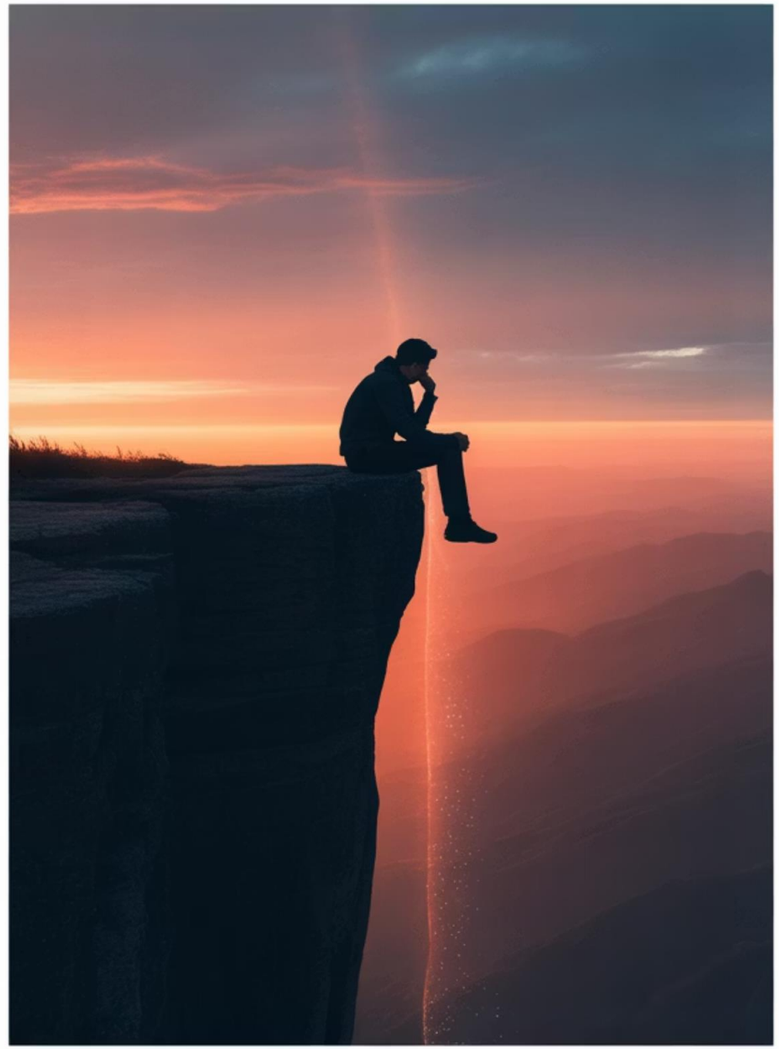
未来融合

持续探索技术创新与哲学深思的融合路径,迈向更智慧、更人性化的AI未来。

认知边界

理解"想法的直接真实性"有助于重新定义**人机认知边界**,深化对智能本质的理解。

第四节 机器认知的边界与反思



回顾：大模型带来的知识论变革

认知主体转型

从理性主体到数据驱动智能体

知识策略革新

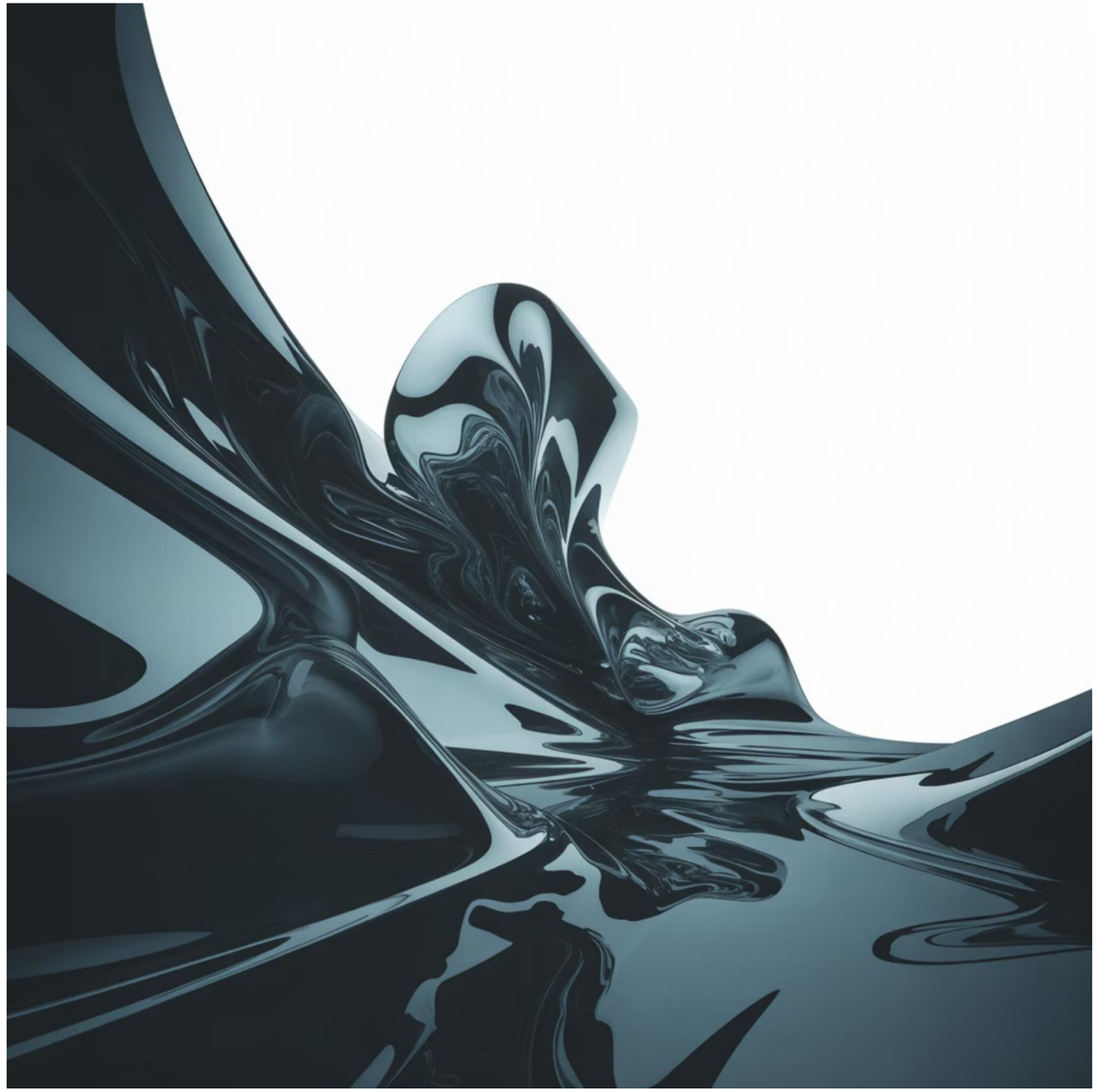
生产方式的动态性与过程性

范式确立

创构认识论的哲学新图景

这场变革标志着认识论正发生着从近代主体性哲学向数据哲学的转向

AI幻觉的哲学本质



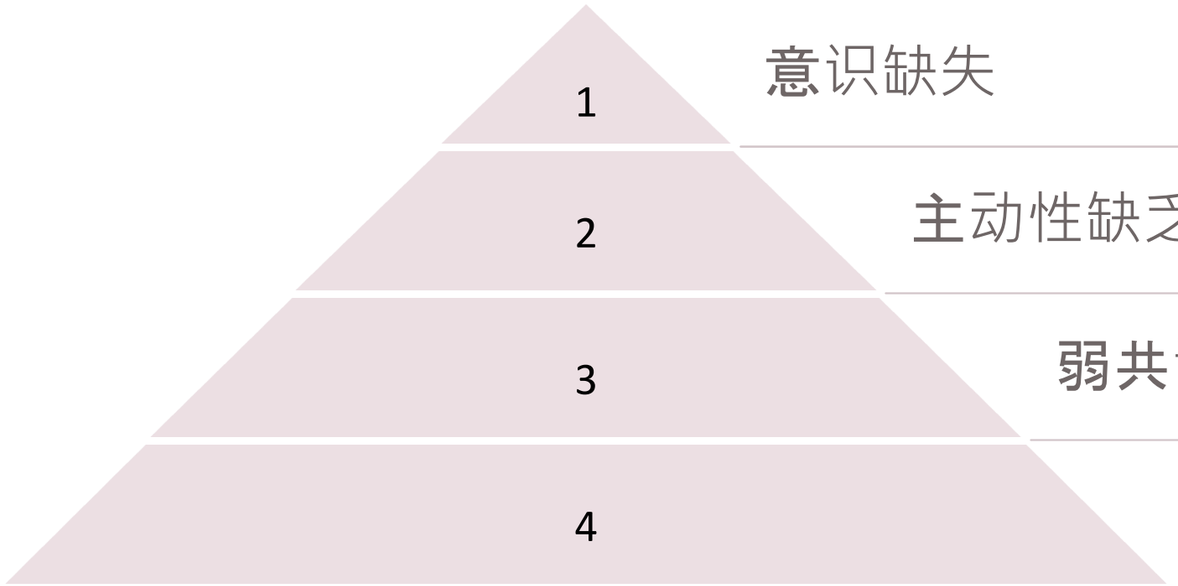
认知错误的表现

现象层面：生成虚假信息、逻辑矛盾、事实编造

机制层面：统计关联替代因果理解，缺乏真值锚定

哲学层面：揭示真理与虚假在算法认知中的边界模糊性，对应理性认知的不完备性定理

认知边界与机器智能的局限



意识缺失

主动性缺乏

弱共识性

不确定性

机器智能与人类智能的差异

记忆与信念的缺失

大模型无持久记忆与稳定信念系统，每次对话独立无连续自我

批判性思维不足

缺乏对自身输出的反思性评估，无法进行真正的价值判断

理解与思考的区分

哲学追问：模式识别是否等同理解？概率推理能否称为思考？

语言与思维的关系

语言转向的AI版本

大模型以语言预测为核心，似乎印证了"语言即思维边界"的维特根斯坦命题。但其语言运用的统计性本质揭示了语言作为认知工具与认知本体的区别。

关键区分：人类语言承载意向性与意义理解，机器语言止于形式关联。

认知的主体性问题

人类认知特征

- 自我意识与反思能力
- 意向性与目的性
- 情感体验与价值判断
- 自由意志与道德责任

AI认知局限

- 无自我觉知的功能执行
- 反应式而非主动性
- 情感模拟而非真实体验
- 算法决定论的本质

知识的真实性与可靠性

1

生成机制

统计概率驱动，无真值验证

2

幻觉风险

虚构率14%-20%，可靠性挑战

3

哲学标准

传统对应论真理观受到质疑

4

信任困境

认知可靠性的新评价体系

认知的伦理边界

社会影响

AI认知错误可导致误导决策、污染信息生态、侵蚀公共信任

责任归属

机器幻觉的责任主体模糊：开发者、使用者还是算法本身？

哲学反思

认知伦理要求建立AI知识生产的道德规范与问责机制

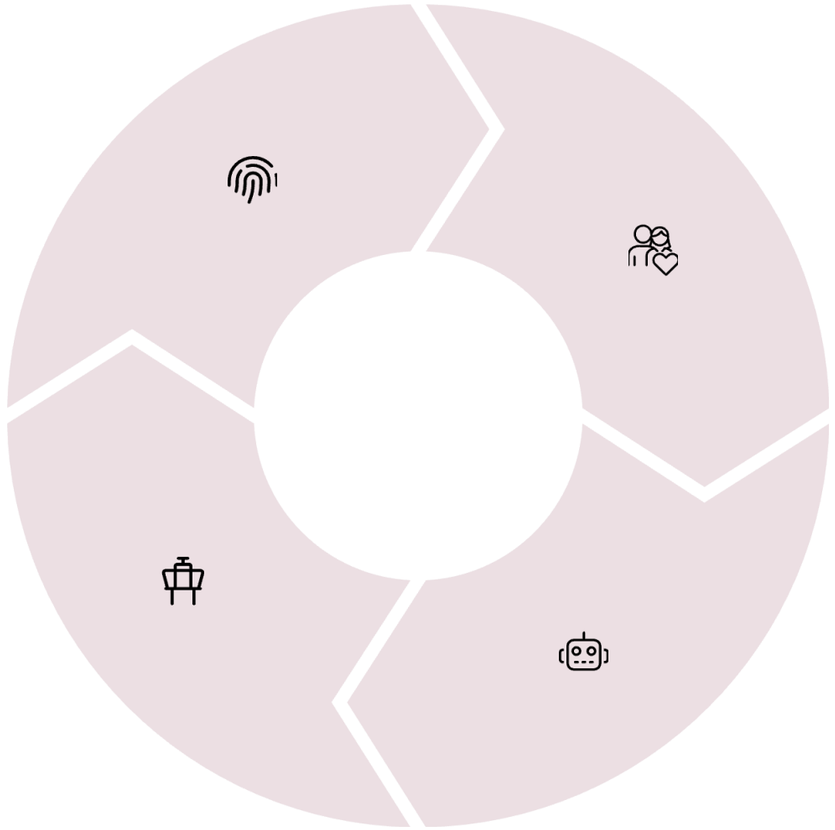
人机共生的认知哲学

人类认知

批判性、价值性、创造性

协同机制

优势互补、相互校验



人类主导

引导、监督与最终决策

机器认知

规模性、速度性、关联性

认知的未来哲学挑战

智能奇点问题

当AI认知能力超越人类综合水平时，认知主体性的哲学

关键问题：超越人类的智能是否必然产生意识？

认知边界的扩展是否存在理论极限？

总结：

1 AI幻觉作为认知复杂性的症候

揭示统计学习范式的内在局限与认知不完备性

2 认知边界的双重性

机器认知既是当前技术的不可逾越限制，又是未来发展的开放可能

3 哲学的持续追问

对智能本质、认知机制、主体性问题保持批判性反思

第五节 AI知识生产的社会与伦理影响

知识论变革的社会影响

权威重构

知识权威从学术精英分散至算法系统，权力结构发生深刻转变

原创性危机

AI生成内容冲击学术原创性标准，知识创新边界变得模糊

民主化悖论

知识生产门槛降低带来普及化，同时伴随虚假信息泛滥风险

AI知识生产的社会影响

专家vs.新手

传统学术精英垄断的知识权威向算法系统分散，知识生产门槛降低但质量控制难度增加

原创性危机

AI生成内容泛滥冲击学术诚信标准，

知识创新与模仿边界模糊化

民主化悖论

知识获取平等化与信息污染风险并存，

数字鸿沟以新形式持续

AI幻觉与信息真实性危机

虚假信息机制

大模型的幻觉特性使其成为潜在的虚假信息生成器，通过看似可信的语言表达传播错误知识。

传播路径：算法生成→社交媒体扩散→认知污染

社会后果

- 公众认知系统性偏差
- 社会信任基础侵蚀
- 民主决策质量下降
- 伦理责任难以追溯

大模型带来的学习革命与挑战

教学模式冲击

AI作为知识主体改变传统师生关系，教师从知识传授者转变为认知引导者

学生认知转型

从知识记忆转向批判性评估AI输出，培养人机协作的新型认知能力

伦理平衡

技术依赖与独立思考能力培养之间的张力，教育公平与算法偏见的矛盾

AI治理的哲学基础

01

开源创新路径

技术民主化与知识共享的伦理优势，促进全球算力资源的公平分配

02

安全风险防范

公共安全、隐私保护与伦理边界的制度化设计

03

治理正义原则

哲学视角下的技术治理需兼顾效率、公平与可持续性

AI技术的权力结构

1

寡头垄断风险

超级AI公司控制核心算法与数据，形成新型技术霸权

2

开源替代方案

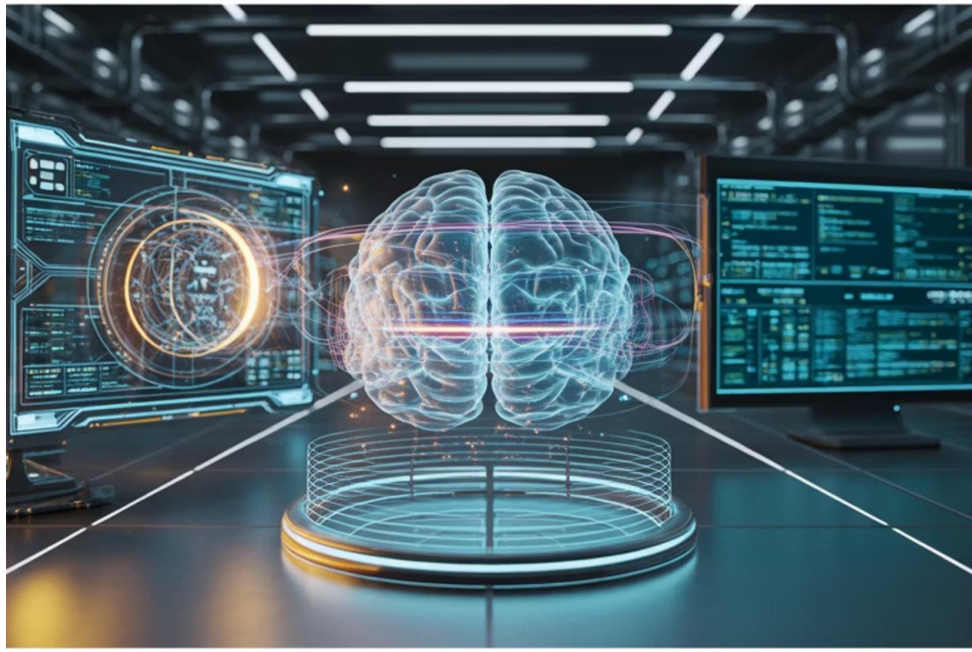
分布式创新模式的普惠性潜力，但面临资源与协调挑战

3

哲学分析

技术与权力关系的批判性考察，追求技术发展的社会正义

第六节 未来挑战的哲学思考



智能奇点的哲学思考

奇点假说

当AI智能全面超越人类综合认知能力时，将触发不可逆的技术加速与文明跃迁。

支持者观点：技术发展的指数级增长趋势

怀疑者立场：意识与智能的不可化约性

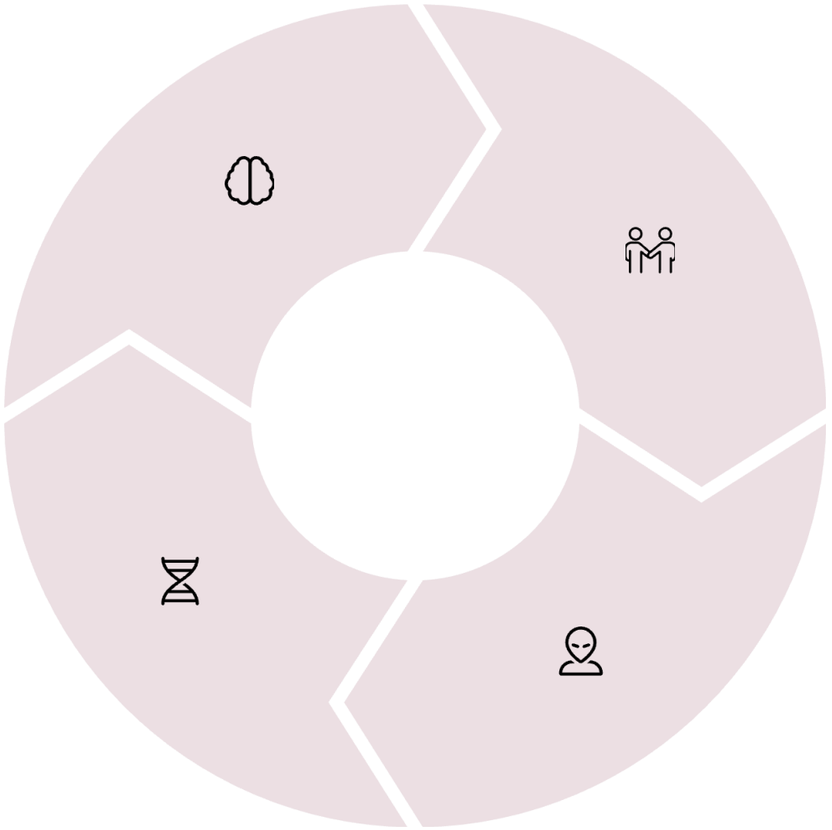
知识论冲击

- 认知主体地位彻底重构
- 知识生产完全自动化
- 人类理解能力的边界显现
- 哲学本身的存在意义面临质疑

人机融合的认知新形态

神经接口

脑机接口实现直接认知增强



跨界智能

生物与硅基智能的深度融合

协同进化

人类与AI的共同演化路径

人类独特性

价值判断与意义创造的不可替代性

AI与意识起源问题

机器意识的哲学难题

即使AI在功能上完全模拟人类认知行为,是否必然产生现象学意义上的主观体验？这触及心身问题、意识本质等哲学核心难题。

镜像进化论：意识可能从复杂智能系统的自组织中逆向涌现,挑战传统意识起源理论。

认知伦理的未来挑战

公民责任意识

智能时代要求公民具备AI素养与批判性使用能力

伦理教育革新

将技术伦理纳入基础教育,培养负责任的技术使用态度

社会参与机制

建立公众参与AI治理的制度化渠道,实现技术民主

知识创新的哲学路径

AI辅助哲学

大模型作为批判性思维工具辅助哲学论证,但不能替代哲学家的价值判断与概念创新。

人类保留最终的理论选择权与意义赋予权

- **文献综述自动化**

  快速梳理研究脉络

- **论证结构分析**

  识别逻辑漏洞与论证缺陷

- **概念关联挖掘**

  发现跨学科思想联系

- **批判性评估**

  哲学家主导的质量控制

AI与哲学研究方法革新

生成式AI应用

在思想实验设计、反例生成、论证模拟等环节提供技术支持,拓展哲学研究的想象空间

思维互动模式

人类哲学直觉与机器计算能力的优势互补,形成新型研究范式

范式构建潜力

数据驱动的哲学研究可能产生全新的理论视角与方法论

关键术语与概念回顾

技术概念

- **表征收敛**：模型内部表征的一致性
- **AI幻觉**：虚假信息生成现象
- **注意力机制**：Transformer核心技术
- **Token预测**：语言生成原理

哲学概念

- **动词思维**：过程性知识观
- **挖掘即认知**：新认识论范式
- **创构认识论**：知识生产新图景
- **认知边界**：智能局限性

社会概念

- **智能奇点**：AI超越人类假说
- **硅基经济**：算力主导经济形态
- **镜像进化**：意识逆向发展理论
- **伦理治理**：技术规范体系

典型案例与研究成果

DeepSeek知识论启示

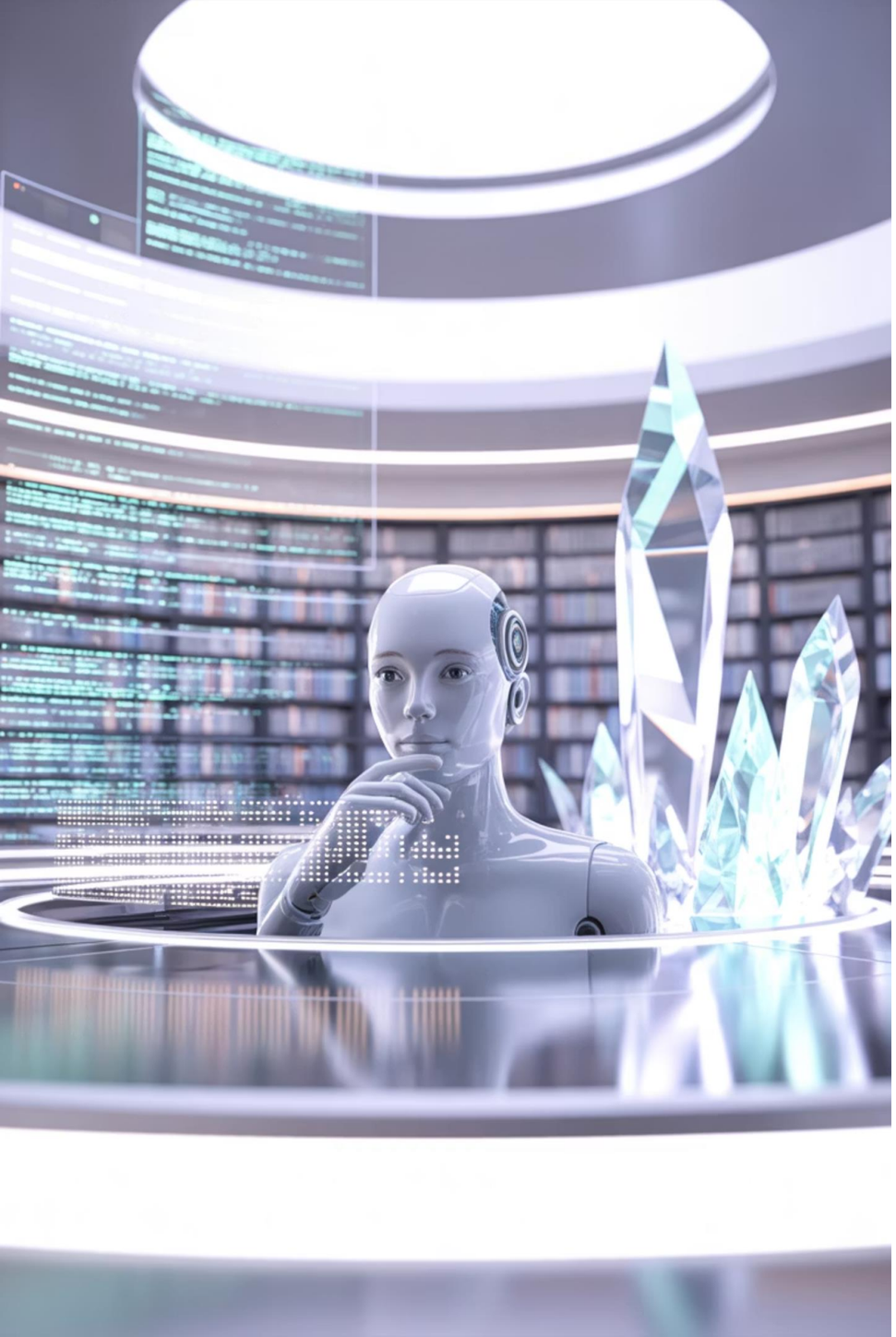
开源大模型展现知识民主化潜力,证明分布式创新可打破技术垄断,为全球南方国家提供发展路径

幻觉率研究

多项研究证实主流大模型幻觉率在14%-20%之间,揭示统计学习范式的认知局限与可靠性挑战

开源治理模式

开放创新与社区协作的治理实践,体现技术发展的多元参与与民主决策可能性



AI幻觉与知识论变革的哲学思考

知识生产新纪元

大模型开启数据驱动的认知革命,从主体建构转向客观挖掘,重塑人类知识生产的基本图景

认知边界与伦理

哲学视角揭示机器智能的本质局限,AI幻觉现象凸显认知可靠性与伦理责任的双重挑战

深化哲学反思

持续追问智能本质、知识真理与技术伦理,以哲学智慧引导智能时代的人类文明发展