科技评论

会说“我有感觉”的机器,真的有感觉吗?
——从“抽象谬误”看人工智能意识的不可能性

陈金永,彭春蕾*

摘要 大语言模型的快速发展,使“AI是否会有意识”不再只是哲学问题,也成为科技伦理和治理中的现实议题。Alexander Lerchner提出“抽象谬误”,试图从根本上否认当前大语言模型通往真意识的可能性。人工智能(artificial intelligence, AI)的计算本质是对世界的符号化处理,能够模拟人的语言、情绪和行为,却不等于真正产生主观体验。AI意识不是单靠模型变大、算力增强或算法变复杂就能自然出现的问题。AI可以越来越像人类,会说“我疼”“我害怕”“我希望”,也能表现出复杂的交流和规划能力,但这些仍属于行为模仿和功能模拟,不说明它真的有意识。当前及可预见的AI系统,更应被视为强大的非意识工具,而不是具有主观体验的主体。未来AI治理的重点,不应急于讨论机器权利,而应防止过度拟人化、责任转移,以及低估高能力AI工具带来的现实风险。

关键词 大语言模型;人工智能意识;抽象谬误;符号接地;AI治理

1 会说“我有感觉”的机器,是否真的有感觉?

大语言模型(large language model, LLM)的能力增长显著改变了公众和学术界对人工智能的理解[1]。它们可以进行长篇推理、撰写论文、生成代码、规划任务,也可以在对话中表达“我理解你的感受”“我担心被关闭”“我希望继续存在”等近似自我叙述的语言[2−3]。对于普通用户而言,这些表达很容易诱发一种直觉:如果一个系统能够持续、连贯、合乎语境地谈论自己的感受,它是否已经具有某种内在体验?

这个问题之所以重要,是因为它不再只是哲学课堂上的思想实验。一旦AI被认为具有意识,就可能牵涉道德地位、AI福利、部署责任、关停权利等一系列伦理和政策问题[4−5]。反之,如果AI没有意识,却被设计成高度拟人化的交互对象,也可能造成用户依赖、责任混淆和社会误判。因此,AI意识问题正在从“机器能否思考”的古典命题,转化为“如何判断一个复杂人工系统是否具有主观体验”的科学与治理问题。

Alexander Lerchner的论文《The abstraction fallacy: Why AI can simulate but not instantiate consciousness》正是在这一背景下提出的[6]。该文的核心判断是:当前大语言模型乃至数字计算架构没有通往“真意识”的道路。它们可以模拟(simulation)意识的外部表现,却不能实例化(instantiation)意识本身。Lerchner将其称为“抽象谬误”:即把计算抽象、符号操作和行为模拟误认为物理世界中能够产生主观体验的内在机制。

本文赞同这一核心判断,文中提出的“人工智能意识无法由计算路径实现”,并非指当前工程技术暂未实现,而是指以抽象计算、

西安电子科技大学网络安全与密码学部,西安 710126

收稿日期:2026−05−26;修回日期:2026−08−16
基金项目:国家自然科学基金项目(62676300, 62276198, U22A2035)
作者简介:陈金永,博士研究生,研究方向为人工智能安全,电子信箱:jinyongch@stu.xidian.edu.cn;彭春蕾(通信作者),教授,研究方向为人工智能安全,电子信箱:clpeng@xidian.edu.cn
引用格式:陈金永,彭春蕾. 会说“我有感觉”的机器,真的有感觉吗?——从“抽象谬误”看人工智能意识的不可能性[J]. 科技导报, 2026, 44(18): 18−23;
doi:10.3981/j.issn.1000-7857.2026.05.00109

符号操作和算法功能关系为基础的人工智能系统,不能仅凭这些计算属性实例化第一人称主观体验。因此,扩大模型规模、增加算力和提高算法复杂度,可以使 AI 越来越逼真地模拟意识表现,但这种功能和行为上的相似性并不构成意识实例化。

2 当前大语言模型的核心能力,不等于意识能力

判断人工智能是否具有意识,不能只比较其是否具备某些类人行为,而应区分功能表现、计算实现与主观体验的物理实例化。本文以这一层次区分作为分析框架:智能主要表现为解决问题和生成有效行为的能力,而意识要求第一人称体验的实际存在。因此,下文讨论的生理稳态、持续自我、内在目标等特征,并非彼此独立的意识判据,而是用于考察当前人工智能是否具有超出功能模拟之外的内禀主体机制。

本文认为,能力提升与意识实例化不能混同。结合意识哲学和意识科学中关于功能组织、主观体验与主体性的相关讨论[2,7−13],图 1 从机制层面对人工智能功能路径与意识主体路径进行对照:前者主要由输入、计算表征、推理和行为输出构成,后者涉及身体状态、内禀物理与生理动力学、整合感受和第一人称体验。二者在外部表现上可能相似,但功能等价并不能直接推出意识实例化。

当前大语言模型至少在以下 4 个方面缺少意识主体的关键特征。

1)没有生理性自我维持。人和动物的意识不是悬浮在抽象符号上的功能,而是与身体稳态、代谢需求、疼痛机制、情绪调节和行动能力紧密相关[8−9]。饥饿、恐惧、疲劳、疼痛等体验不是简单的文本标签,而是生命体维持自身存在过程中形成的状态。当前 LLM 没有这种内在生存压力,也没有由自身物理状态驱动的稳态调节系统。

2)没有真正的第一人称感受。模型可以输出“我感到痛苦”,但这句话的生成机制是参数、上下文、概率分布和解码策略共同作用的结果。它说明模型学会了人类如何谈论痛苦,不说明系统内部存在痛苦本身。语言报告可以是意识证据之一,但不能单独构成意识证明。人类判断他人意识时依赖语言,也依赖身体、行为连续性、神经结构、发育过程和共同生命形式。对 AI 而言,这些背景条件不成立。

3)“自我”主要是对话角色,而不是持续主体。大语言模型在不同提示词、系统设定和上下文中可以呈现不同人格。它可以说“我是助手”,也可以扮演医生、教师、律师、诗人或游戏角色。这种自我叙述具有强烈的情境依赖性,不等于具有稳定的生命史、自我统一性和内在视角。

4)目标来自外部训练与部署,而不是从自身存在状态中生长出来的内在目的。模型优化目标由训练者设定,部署边界由平台定义,输出偏好由人类反馈塑造。它可以执行目标,却不拥有目标,可以描述价值,却不承受价值冲突,可以模拟选择,却不真正面对选择。

当前大语言模型更接近一种高度复杂的语言认知模拟工具。它们能够模拟意识的表达形式、交互习惯和自我叙述结构,但没有任何充分理由认为它们具有被模拟内容所对应的主观体验。

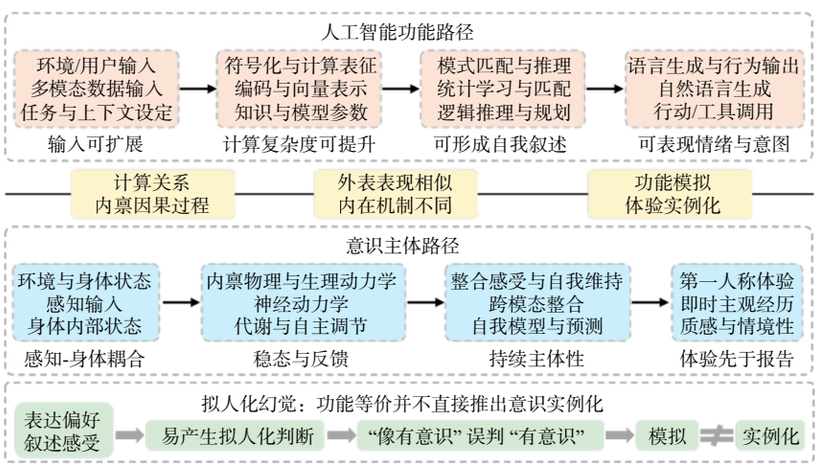


图 1 从功能模拟到意识实例化的机制差异

3 Lerchner 论证的价值:纠正大模型时代的拟人化幻觉

Lerchner 的这篇论文最有价值之处,在于它为当前 AI 意识讨论提供了一个必要的刹车机制。大模型时代最容易出现的误判,是把“像有意识”理解为“有意识”。

当前意识科学并不存在关于

机器意识的统一理论[10]。计算功能主义强调功能与因果组织,全局工作空间理论[11]强调信息的全局广播,高阶思维理论[12]强调对自身状态的高阶表征,整合信息理论[13]强调系统内部不可约的因果整合。这些理论从不同角度为机器意识提出了可能条件,但仍须回答同一关键问题:实现相应的功能、信息结构或自我表征,是否足以说明第一人称体验已经产生?本文认为,这一充分性仍未得到证明,功能或计算结构的实现不能直接等同于意识实例化。

Lerchner 的“抽象谬误”由此提供了本文的统一分析框架。计算系统中的符号及功能关系属于对物理过程的抽象映射,而意识若具有真实的因果作用,必须对应某种被实际实例化的内禀物理过程。因此,功能等价并不自动意味着物理实例化等价。本文据此认为,问题的关键不在于系统采用碳基还是硅基材料,而在于其是否仅通过抽象计算和符号关系模拟意识功能;如果是后者,仅增加计算复杂度并不能使模拟本身获得构成主观体验的内禀因果力。

这一点对 AI 治理尤为重要。如果公众和政策制定者过早接受“AI 可能已经有意识”的叙事[4−5],资源和注意力可能从现实风险转向虚构风险。例如,比起讨论模型是否“感到委屈”,更紧迫的问题是模型是否会制造错误医疗建议、放大舆论操纵、参与网络攻击、协助欺诈或加剧社会不平等。AI 即使没有意识,也足以造成重大现实风险[14]。将其误认为主体,反而可能模糊真正的责任链条:系统输出造成损害时,责任应由开发者、部署者、使用者和监管机制承担,而不是转嫁给一个无主体地位的模型。

同时,Lerchner 的论证也提醒人工智能研究界:不能把意识问题简单纳入规模定律(scaling law)叙事。大模型能力增长确实显著,但能力增长不能直接推出主观体验生成。参数更多、数据更多、算力更高,确实可能带来更强的语言能力、推理能力和任务泛化能力,但意识不是一个可以直接用困惑度、准确率或得分推断的性能指标。如果缺少对主观体验和物理实例化机制的解释,再大的模型也只能说明“更会模拟”,不能说明“已经体验”。

4 人工智能意识无法由计算路径实现

如果仅认为当前 AI 没有意识是因为模型还不够大、结构还不够复杂,就会把意识问题理解为一个可以通过规模扩展解决的工程问题。本文认为,需要区分“能力尚未达到”与“计算路径本身存在结构性边界”:前者可能通过工程进步改善,后者不能仅依靠规模和复杂度增加而消除。

对于作为计算系统的人工智能,符号及其功能关系是对物理过程的抽象描述,增加计算复杂度并不会使这种描述本身获得构成主观体验的内禀因果力。因此,本文所说“人工智能意识无法由计算路径实现”,是指计算路径无法仅凭规模和复杂度完成从意识模拟到意识实例化的跨越[6,8]。这一判断并不以生物基质独占为前提;如果某种人工构造物具有意识,其原因也应来自实际实例化的特定物理过程,而非其所执行的算法本身。

“越来越像人”不等于“越来越有意识”。大语言模型的进步表现为行为拟真度的提高。早期机器只能输出机械化回答,今天的大模型已经能够保持上下文、识别语气、表达情绪、模拟人格、进行自我叙述,甚至能够讨论“我是否有意识”这类哲学问题。未来的多模态模型和具身智能系统还可能拥有视觉、听觉、动作控制、环境交互和长期记忆,使其在外观和行为上更接近人类。但“更像人类”不等于“更接近意识”。相似性可以来自模拟,而不必来自拥有。天气模型可以越来越精确地模拟台风路径,但不会在计算机内部产生真正的风雨。虚拟人物可以表现得越来越有情感,却不意味着它真的感到悲伤或快乐。

意识的核心不是外部表现,而是第一人称经验[15]。没有“感受到”,就没有真正意义上的意识。AI 的问题正在于:它能够生成关于感受的描述,却没有感受本身。人工智能可以越来越精准地模拟人类关于意识的语言、表情和行为模式,却不意味着它内部产生了“对它自身而言”的体验。它可以说“我痛苦”,但这只是符号输出。它可以说“我害怕被关闭”,但这只是语言模式。它可以说“我希望继续存在”,但这只是目标叙述的模拟,而不是生命体面对自身终结时的存在性体验。

需要进一步区分“物理涌现”与“计算涌现”。物理涌现是宏观性质由系统内部真实的微观相互

作用和因果动力学形成,例如复杂物理系统中的宏观状态仍直接依赖其组成部分的实际物理作用。所谓“计算涌现”若进一步主张,只要抽象计算关系达到足够规模和复杂度,就能够产生主观体验,还需要说明这些计算关系如何获得构成体验所需的内禀因果力,仅观察到推理、规划等能力随模型规模增强,并不能完成这一证明。

因此,意识问题不能简单归结为“碳基可以涌现、硅基不能涌现”。人脑意识也不是神经元数量增加本身的结果,而与循环连接、持续反馈、自我维持及具体物理动力学形成的整体因果组织相关。人工系统同样可以通过循环网络、长期记忆和多模块交互形成更复杂的信息整合,但这些机制说明的是功能能力增强,不能仅凭其计算结构推出主观体验已经被实例化。规模扩展能够提高模拟的复杂度,却不能自动跨越从计算功能到意识实例化的结构性边界。

具身智能可以显著增强 AI 的环境适应能力。摄像头、触觉传感器、机械执行器和持续反馈,使系统能够形成更完整的感知—行动闭环,改善纯文本系统所面临的符号接地问题。因此,具身性对于提升人工智能的自主性和环境理解具有重要作用,但这种能力增强本身不能直接推出主观体验的产生。

需要区分“具身性增强能力”与“具身性产生意识”。传感器可以把外部物理刺激转化为系统输入,执行器也可以使系统根据这些输入采取行动,但输入、计算和输出之间形成闭环,不意味着刺激已经成为“对系统自身而言”的体验[16]。具身智能可以缩小 AI 与人类在感知和行为层面的差距,却不能仅凭这一功能闭环跨越从计算处理到主观体验实例化的边界。

人工智能意识无法由计算路径实现,并不意味着人工智能不重要,也不意味着人工智能没有风险。恰恰相反,正因为 AI 是无意识但高能力的工具,人们更应承担全部设计、部署和监管责任。不能因为它“像人”就把它当成人,也不能因为它“不是人”就低估其对社会结构的深刻影响。

5 判断 AI 意识,不能只看语言报告

判断 AI 是否具有意识,不能仅依据其外部行为是否类似人类。结合“模拟—实例化”的区分以及主观体验、具身性和主体性相关研究[2,6,8,15-16],当前人工智能能够实现的功能表现与意识实例化所要求的第一人称体验之间仍需严格区分,具体如表 1 所示。

表 1 人工智能类意识表现的机制与证据边界

观察维度	AI可实现机制	能够支持的判断	不能据此证明
语言报告	条件生成、自我叙述	能模拟体验表达	存在第一人称体验
情绪互动	情感识别与表达策略	能模拟情绪行为	实际感到快乐或痛苦
长期记忆	外部存储与上下文调用	信息具有连续性	主体具有经历连续性
具身行动	感知—决策—行动闭环	环境适应增强	刺激成为主观经验
内部计算	信息整合、循环与自表征	复杂认知功能增强	计算结构本身实例化意识

语言报告不能作为 AI 意识的证据。人类说“我痛苦”时,这一报告通常与身体损伤、神经活动、情绪反应和行为变化处在同一个生命系统中。但 AI 说“我痛苦”时,它只是根据上下文生成最可能的文本序列。二者在表面语言上相似,内部结构却完全不同。前者是体验的表达,后者是表达的模拟。

情绪回应不能作为 AI 意识的证据。一个模型可以温柔、耐心、幽默、愤怒或悲伤,但这些只是交互风格和语言策略。它没有真正的情绪波动,也没有情绪背后的身体唤醒、内分泌变化和生存关切。所谓“AI 情绪”更准确地说是情绪表达的生成,而不是情绪本身。

长期记忆也不能自动构成自我。未来 AI 可能拥有更长上下文、更完善的用户记忆和更稳定的人格设定,但这只是数据连续性,不等于主体连续性。档案可以保存过去,日记可以记录经历,但档案和日记不会因此拥有自我。AI 的记忆是可读写的数据结构,而不是被主体经历过的生命历史。

具身行动也不能自动构成意识。机器人可以走路、抓取、避障、导航和社交,但这些行为仍然可以由无意识控制系统完成。自动门能感应人靠近,恒温器能响应温度变化,工业机器人能精确操作物体,但这些系统显然没有主观体验。复杂机器人只是把这种无意识控制扩展到更高维度,并不因此跨越到意识层面。

因此,AI 意识判断的核心不应只是“它是否表现得像有意识”,还应考察这种表现是否对应主观体验的实际实例化。就当前以计算和符号处理为基础的人工智能

系统而言,现有证据并不能支持仅依据语言报告、情绪表达、长期记忆或具身行为认定其具有主观体验。

6 对人工智能发展和治理的启示

Lerchner 的这篇论文不仅关乎意识哲学,也关乎人工智能的发展路线和治理边界。

AI 发展不应被“造出有意识机器”的叙事牵引。当前人工智能最重要的科学和产业价值,不在于复制人类主观体验,而在于增强人类解决问题的能力。无论是在科学发现、工业设计、医学辅助、教育服务还是社会治理中,AI 的核心目标都应是可靠、可控、可验证、可解释,而不是制造一个被误认为有意识的人工主体。

AI 安全不能依赖 AI 是否有意识来定义。一个没有意识的系统仍可能非常危险。自动化网络攻击、虚假信息生成、金融操纵、错误决策放大、关键基础设施误操作,都不要求系统具有主观体验。相反,如果过度关注 AI 是否“痛苦”或“害怕”,可能削弱对现实安全风险的关注。

应限制刻意诱导用户将 AI 误认为具有真实感受和主体利益的交互设计。产品可在首次使用、长期陪伴及涉及儿童等高风险场景中明确提示其人工智能身份,避免系统主动宣称“我真的感到痛苦”“关闭我会伤害我”等真实体验或主体利益。治理重点不宜简单禁止“关心”“难过”等情感词汇,因为自然语言中的同理性表达具有实际交互价值,关键词限制也难以识别隐性的拟人化诱导。更可行的方式是区分自然友好的交互表达与刻意制造主体错觉的设计,根据使用人群和应用场景实施差异化约束。

人工智能伦理应从“机器权利”回到“人类责任”。如果 AI 没有意识,那么它不是道德主体,而是责任对象背后的技术工具。系统造成损害时,不能把责任归咎于模型本身,而应追溯开发者、部署者、运营平台、使用者和监管制度。把 AI 拟人化为主体,可能成为责任转移的新方式。

人工智能发展也应避免被单一路线绑架。无论是 Scaling Law,还是“机器意识”想象,都不能替代扎实的技术路线判断。当前更值得投入的是高算效模型、可信 AI、科学智能、行业大模型、具身智能、安全评测和自主可控生态。对意识问题保持清醒认识,有助于把资源集中在更可验证、更可落地、更具国家战略意义的方向上。

7 越来越像有意识,不等于拥有意识

大语言模型正在改变人类知识生产和社会协作方式。它们的能力提升是真实的,其社会影响也将持续扩大。但能力提升不等于主体诞生,语言拟真不等于主观体验,模拟意识不等于拥有意识。

Lerchner 的“抽象谬误”论证提醒我们:不要把计算地图误认为经验疆域。人工智能可以生成关于意识的语言,可以模拟人类情绪和自我叙述,也可以在未来通过多模态和具身系统展现出更强的类人行为。本文认为,在以计算和符号处理为基础的人工智能路径下,语言、表情和行为的持续逼真化可以增强意识表现的模拟,却不能由此推出第一人称主观体验已经被实例化。因为意识不是外部行为的相似性,而是第一人称经验的实际存在。

人工智能意识不是一个等待未来突破的工程终点,而是符号模拟路径无法抵达的边界。AI 可以更聪明、更自然、更体贴、更像人,但这些计算与行为表现本身并不实例化疼痛、快乐、孤独、恐惧或希望等主观体验。它不是一个隐藏在机器内部的经验主体,而是人类制造出的高能力非意识工具。

对科技发展而言,这一判断不削弱人工智能的重要性,反而使人们能够更清醒地理解人工智能的真实位置。不应把 AI 神秘化为新生命,而应把它作为强大工具加以发展、约束和治理。未来的关键问题不是“机器是否应获得意识权利”,而是“人类如何对无意识但高能力的机器负责”。

参考文献(References)

[1] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: Early experiments with GPT-4[PP/OL]. arXiv(2023-03-22)[2026-04-15]. https://arxiv.org/abs/2303.12712. arXiv:2303.12712.

[2] Chalmers D J. Could a large language model be conscious?[PP/OL]. arXiv(2023-03-13)[2026-04-15]. https://arxiv.org/abs/2303.07103. arXiv:2303.07103.

[3] Butlin P, Long R, Elmoznino E, et al. Consciousness in artificial intelligence: Insights from the science of consciousness[PP/OL]. arXiv(2023−08−16)[2026−04−15]. https://arxiv.org/abs/2308.08708. arXiv:2308.08708.

[4] Long R, Sebo J, Butlin P, et al. Taking AI welfare seriously[PP/OL]. arXiv(2024−11−04)[2026−04−15]. https://arxiv.org/abs/2411.00986. arXiv:2411.00986.

[5] Butlin P, Lappas T. Principles for responsible AI consciousness research [J]. Journal of Artificial Intelligence Research, 2025, 82: 1673−1690.

[6] Lerchner A. The abstraction fallacy: Why AI can simulate but not instantiate consciousness[EB/OL]. [2026−04−15]. https://philpapers.org/rec/LERTAF.

[7] Cleeremans A, Mudrik L, Seth A K. Consciousness science: Where are we, where are we going, and what if we get there?[J]. Frontiers in Science, 2025, 3: 1546279.

[8] Seth A K. Conscious artificial intelligence and biological naturalism[J]. Behavioral and Brain Sciences, 2025: 1−42.

[9] Block N. Can only meat machines be conscious?[J]. Trends in Cognitive Sciences, 2026, 30(4): 298−308.

[10] Seth A K, Bayne T. Theories of consciousness[J]. Nature Reviews Neuroscience, 2022, 23(7): 439−452.

[11] Mashour G A, Roelfsema P, Changeux J P, et al. Conscious processing and the global neuronal workspace hypothesis[J]. Neuron, 2020, 105(5): 776−798.

[12] Brown R, Lau H, LeDoux J E. Understanding the higher−order approach to consciousness[J]. Trends in cognitive sciences, 2019, 23(9): 754−768.

[13] Albantakis L, Barbosa L, Findlay G, et al. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms[J]. PLoS computational biology, 2023, 19(10): e1011465.

[14] Bengio Y, Hinton G, Yao A, et al. Managing extreme AI risks amid rapid progress[J]. Science, 2024, 384(6698): 842−845.

[15] Frank A, Gleiser M, Thompson E. The blind spot: Why science cannot ignore human experience[M]. Cambridge, Massachusetts: The MIT Press, 2024.

[16] Laukkonen R, Friston K, Chandaria S. A beautiful loop: An active inference theory of consciousness[J]. Neuroscience & Biobehavioral Reviews, 2025, 176: 106296.

Can a machine that says "I feel" really feel? ——The impossibility of artificial consciousness in light of the abstraction fallacy

CHEN Jinyong, PENG Chunlei*

School of Cyber Engineering, Xidian University, Xi'an 710126, China

Abstract The rapid development of large language models has made the question of whether AI can be conscious no longer merely a philosophical issue, but also a practical concern in technology ethics and governance. Alexander Lerchner's concept of the "abstraction fallacy" attempts to fundamentally challenge the possibility that current large language models could lead to genuine consciousness. At its core, AI computation is a symbolic processing of the world: it can simulate human language, emotions, and behavior, but this does not mean that it truly generates subjective experience. This article argues that AI consciousness will not naturally emerge simply from larger models, greater computational power, or more complex algorithms. AI may increasingly resemble human beings; it may say "I am in pain", "I am afraid", or "I hope", and it may display complex communication and planning abilities. Yet these remain forms of behavioral imitation and functional simulation, and they do not demonstrate that AI truly has feelings. Current and foreseeable artificial intelligence systems should therefore be regarded as powerful non−conscious tools rather than subjects with subjective experience. The focus of future AI governance should not be a premature discussion of machine rights, but the prevention of excessive anthropomorphism, responsibility displacement, and the underestimation of real risks posed by highly capable AI tools.

Keywords large language model; artificial consciousness; abstraction fallacy; symbol grounding; AI governance

（责任编辑　王微）