欢迎扫码订阅《科普时报》

科普时报

2025年8月8日
星期五
第396期
今日16版

科技日报社主管主办　科普时报社出版　国内统一连续出版物号CN 11-0303　代号1-178

揽月月面着陆器完成综合测试

8月6日，揽月月面着陆器着陆起飞综合验证试验在位于河北省怀来县的地外天体着陆试验场圆满完成，标志着我国载人月球探测工程研制工作取得新的重要突破。揽月月面着陆器是我国面向首次载人月球探测任务全新研制的地外天体载人下降与上升飞行器，由登月舱和推进舱组成。　新华社发　刘永婧　摄

啥？AI也可能被“投毒”

□ 科普时报记者　陈　杰

还在为人工智能（AI）时不时“一本正经地胡说八道”恼火？

或许，你的AI被“投毒”了。8月5日，国家安全部发布安全提示，AI训练数据存在良莠不齐的问题，虚假信息、虚构内容和偏见性观点导致的数据“污染”，给AI安全带来挑战。

“训练数据之于AI，就像教科书之于学生。”中国科学院计算技术研究所工程师刘延嘉将AI比喻成勤奋好学的学生，“AI正是通过学习文本、图像、行为等数据构建认知模型，形成对世界的理解与判断能力。若教科书内容存在错误或偏见，学生的知识体系必然扭曲。”

研究显示，当训练数据中仅有0.01%的虚假文本时，AI模型输出的有害内容会增加11.2%；即使是0.001%的虚假文本，有害输出也会上升7.2%。“训练数据的细微瑕疵，也可能导致AI输出错误、偏见甚至危险的结果。”刘延嘉说。

AI的训练数据为何会被“污染”？

“数据被污染的情况较复杂，既有人为故意‘数据投毒’的可能，也可能因数据收集、整理过程缺乏严格规范和审核所致。”同盾人工智能研究院执行院长董纪伟说，受到数据污染的AI生成的虚假内容，可能成为后续AI训练的数据源，形成具有延续性的“污染遗留效应”。

董纪伟认为，“数据放大效应”或是更大的隐性风险，“AI可能通过算法强化，将数据中的一些偏见演变为系统性偏见，并在输出时将其奉为‘真理’。”

如今，网上AI生成内容数量已超过人类生产的真实内容，大量低质量及非客观数据充斥其中。“当AI训练数据中的错误信息逐代累积，必然会扭曲AI本身的认知能力。”董纪伟提醒。

“毒”数据对AI输出的影响，远不止“一本正经地胡说八道”这么简单，往往还具有“隐性但致命”特征。试想，当“涉毒”AI广泛应用于日常，人们可能因AI的错误诊断延误治疗；投资者可能被AI推荐的虚假高收益项目欺骗；汽车可能因AI的错误导航而迷失方向……

这样的AI，谁敢放心用？

目前，《生成式人工智能服务管理暂行办法》和新版《数据安全法》已经将AI训练数据纳入监管。但专家认为，要从技术层面解决AI训练数据污染问题，还有待AI开发者在数据筛选验证机制、数据实时监测和数据溯源等方面付出更多努力。正如中国工程院院士邬贺铨所言：“AI的安全边界，最终取决于数据的质量底线。”

面对并不完美的AI，我们又该如何应对？

董纪伟建议，日常使用AI时应持谨慎态度，如果AI给出的回答涉及重要决策，务必向专业人士核实。“当然，也可用多个AI工具对同一问题进行询问，通过对比答案来判断AI的可靠性。若发现AI频繁给出不合理或错误回答，可直接更换AI工具。”

本期导读

编辑：陈杰　美编：纪云丰　编辑部热线：010-58884135　投稿邮箱：kepushibao@kepu.gov.cn　印刷：新华社印务有限责任公司