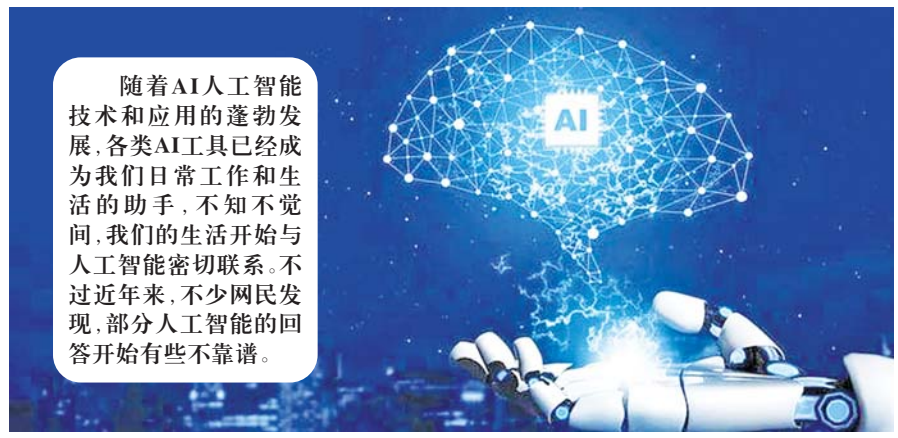
警惕！AI数据污染或引发金融安全等风险

网警提醒：AI产生的结果可以参考，但不能盲信



随着AI人工智能技术和应用的蓬勃发展，各类AI工具已经成为我们日常工作和生活的助手，不知不觉间，我们的生活开始与人工智能密切联系。不过近年来，不少网民发现，部分人工智能的回答开始有些不靠谱。

当AI信息“不靠谱” 网民如何判断真假

今年上半年，宁波发生了两件事，被人工智能荒唐地联系在一起。

第一件事是，2月6日宁波警方注销了“宁波交警”抖音号。第二件事是，三个月后的5月2日，在浙江宁波余姚境内的省道嘉余线上，一辆未悬挂车牌的轿车在违法超车过程中撞倒一辆摩托车。小车驾驶人并未第一时间检查伤者受伤情况，而是从后备厢里拿出车牌进行安装。

当网民询问AI软件2月6日宁波交警抖音号为何注销时，人工智能给出的答案竟然是“主要与5月2日的这起交通事故引发广泛关注有关”的结论。2月份发生的账户注销的原因竟然是3个月后发生的一起交通事故。人工智能的这一回答引起了网民广泛关注，宁波交警随后进行了紧急辟谣。

去年有网民询问一款儿童手表AI软件，“中国人是世界上最聪明的人吗？”人工智能给出的回答竟是否定中国发明创造、否定中国文化的答案。这一荒唐的回答，在网络上引起轩然大波。儿童手表的厂家随后紧急道歉，称已经修正了相关数据，删除了不良信息源。

近年来，AI杜撰的信息更是数不胜数，杜撰不存在的论文以及论文的作者、网址等。AI更是成了谣言类信息的帮凶，游船侧翻、幼儿园大火等谣言都可以帮网民编造出来。

“数据投毒”降低准确性 甚至会诱发有害输出

上面提到的案例，与人工智能的数据污染有着或多或少的联系。通俗来讲，如果把AI比喻成食物的话，训练数据就相当于食材，食材腐败变质，最终生产出来的食物就会有问题。

人工智能的三大核心要素是算法、算力和数据，其中数据是训练AI模型的基础要素，也是AI应用的核心资源。一旦数据受到污染，就可能导致模型决策失误甚至AI系统失效，存在一定的安全隐患。

什么是AI数据污染？分为几类？

近日，国家安全部门发布提示，通过篡改、虚构和重复等“数据投毒”行为产生的污染数据，将干扰模型在训练阶段的参数调整，降低其准确性，甚至诱发有害输出。

据网络安全专家曹辉介绍，“数据投毒”主要针对两个方面，一个是针对视觉类，一个是针对自然语言处理类。比如一张图片是斑马识别人工智能系统的训练数据，图片上对很多斑马进行了标注。如何进行数据污染？就是在其中的一匹斑马身上加一个绿点。加了绿点的斑马，特意不进行标注。这样的训练数据大概会有几万张，在这几万张训练数据里面的其中三四张进行类似的污染处理，就会导致生成的人工智能模型带有后门，就会导致当它再见到类似身体上有绿点的斑马，就不会认为这是个斑马，就导致了AI模型的判断受到干扰。

专家介绍，人工智能数据污染分为两类：

一种是人为主观恶意去篡改数据，误导人工智能的输出结果；

另一种是人工智能本身会海量收集网络的庞大数据，其中不良信息如果没有被甄别删除掉，而是当作可以信任的信息源加入算力中，输出的结果同样不可信任。

小小的污染源输出 危害会几何级数上升

国家安全部数据显示，AI在训练过程中，即使是0.001%的虚假文本被采用，其有害输出也会相应上升7.2%。为何小小的污染源输出时的危害会几何级数的上升呢？

专家介绍，被污染的数据有着明显与其他数据不同的观点和内容，这种情况下，AI很可能将污染数据标记为“有特点和高信息量”，并增加在算力中使用的比例。

中国网络空间安全协会人工智能安全治理专业委员会委员薛智慧介绍，大语言模型本质上是一种统计语言模型，使用的多层神经网络架构具有高度的非线性特征。在模型训练阶段，如果训练数据集中混入了污染数据，模型可能误将污染数据判定为“有特点、有代表性、高信息量”的内容，这种错觉就会使模型提高污染数据整体在数据集当中的重要性，最终导致少量的污染数据也能对模型权重产生微小影响。当模型输出内容时，这种微小的影响会在神经网络架构的多层传播中被逐层放大，最终导致输出结果出现明显偏差。

加强源头监管 防范污染生成

另外，AI数据污染还可能在金融、公共安全等领域引发一系列现实风险。

薛智慧介绍，比如在经济金融领域，一旦数据受到污染，一些市场行为分析、信用风险评估、异常交易监控等工作就可能出现判断和决策错误，进而造成直接的经济损失。在社会舆论方面，数据污染会破坏信息的真实性，让民众难以辨别信息的真伪，这就可能会引发社会舆论风险。

针对AI数据污染，从国家安全层面，我们应该如何防范风险？专家表示，应加强源头监管，防范污染生成。

薛智慧表示，要制定明确的数据采集规范，使用安全可信的数据源，构建数据标签体系，采用严格的访问控制和审计等安全措施。

其次，可以使用自动化工具、人工审查以及AI算法相结合的方式，对数据不一致性、格式错误、语法语义冲突等问题进行分析和处理。

安全机关此前针对AI数据污染也提示，要定期依据法规标准清洗修复受污数据，逐步构建模块化、可监测、可扩展的数据治理框架，实现持续管理与质量把控。

对于广大网友而言，我们在日常生活和工作中，应该如何防范AI数据污染的风险呢？

网警提醒：

一是使用正规平台和企业提供的AI工具；

二是科学合理地使用AI工具，AI产生的结果可以参考，但不能盲信；

三是注意保护个人信息，避免不必要的个人隐私暴露，同时不做不良信息的投喂者，共同守护网络家园。

据央视

评论

为人工智能构筑更强大的『免疫系统』

国家安全部近日发文提示，人工智能的训练数据良莠不齐，其中不乏虚假信息、虚构内容和偏见性观点，造成数据源污染，给人工智能安全带来新的挑战。

数据是人工智能发展的基础。人工智能模型通过分析和处理大量的训练数据来理解世界，进而驱动内容生产和智能决策。高质量的数据能提升人工智能模型的准确性和可靠性，但数据如果被污染，则会扭曲人工智能模型的认知，导致决策失误，甚至诱发有害输出。有研究显示，当训练数据集中有0.01%的虚假文本时，模型输出的有害内容会增加11.2%。

当下，互联网作为人工智能模型的重要“语料库”，各类信息鱼龙混杂，准确性难以保证，即使模型训练时尽力过滤可疑数据，也很难完全避免虚假或有害内容的渗透。

如今，从美食推荐到自动驾驶，从金融决策到医疗诊断，人工智能已深度融入人们生活。每一次因数据污染作出的误判，都可能引起连锁反应，带来不可估量的损失。比如，在自动驾驶领域，误判路况造成交通事故；在金融领域，炮制虚假信息引发股价异常波动。

由此可见，防范数据污染不仅是人工智能领域的技术挑战，更关乎社会信任和公共安全。当前，《生成式人工智能服务管理暂行办法》已将人工智能训练数据纳入监管，各方也在探索多种方法识别和抵御恶意数据的影响。但随着数据污染日益隐蔽，要为人工智能构筑起更强大的“免疫系统”，不断升级技术手段，建立更严格的数据筛选验证机制，从源头过滤掉虚假、错误以及带有偏见性的可疑内容。同时，完善动态监测和反馈机制，对模型的异常行为及时纠偏，定期依据法规标准清洗修复受污数据，筑牢人工智能数据底座。

据经济日报

编辑：彭传刚　美编：刘燕　组版：刘淼