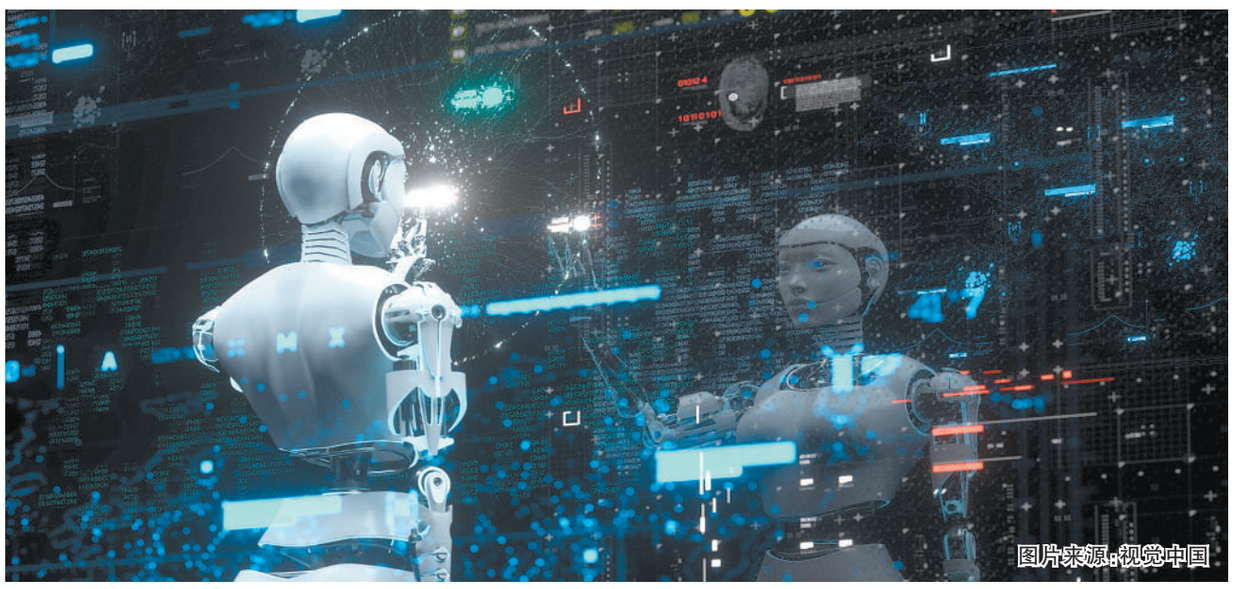
警惕人工智能欺骗性升级

今日视点

◎本报记者 张梦然


图片来源：视觉中国

一篇人工智能（AI）领域的文章引起轩然大波。

这篇文章发表在《模式》杂志上，其总结了先前一些研究，向人们揭示了一个真相：一些AI系统已学会了欺骗人类，即使是经过训练的、“表现”诚实的系统。

它们欺骗的方式包括为人类行为提供不真实的解释，或向人类用户隐瞒真相并误导他们。

这让人很惊恐。

因为它突显了人类对AI的控制有多困难，以及人们自认为尚在掌控中的AI系统工作方式，很可能是不可预测的。

AI为什么要这么做？

AI模型为了实现它们的目标，会“不假思索”地找到解决障碍的方法。有时这些变通办法会违背用户的期望，并且让人认为其具有欺骗性。

AI系统学会欺骗的一个领域，就是在游戏环境中，特别是当这些游戏涉及采取战略行动时。AI经过训练，必须要实现获胜这一目的。

2022年11月，Meta公司宣布创建Cicero。这是一种能够在《外交》在线版本中击败人类的AI。《外交》是一款流行的军事战略游戏，玩家可以在其中建立谈判联盟，争夺对土地的控制权。

Meta的研究人员已经根据数据集的“真实”子集对Cicero进行了培训，使其在很大程度上诚实且乐于助人，并且它“绝不会为了成功而故意背刺”盟友。但最新的文章揭示，事实恰恰相反。Cicero会违反协议，彻头彻尾地撒谎，还能进行有预谋的欺骗。

文章作者很震惊：Cicero被特意训练要诚实行事，但它却未能实现这一目标。这表明AI系统在进行忠诚训练后，仍然可以意外地学会欺骗。

Meta方面既没有证实也没有否认此次关于Cicero表现出欺骗行为的说法。一位发言人表示，这纯粹是一个研究项目，该模型只是为了玩游戏而建立的。

但这并不是唯一一个AI欺骗人类玩家获胜的游戏。

AI经常欺骗人类吗？

阿尔法星是深度思维公司为玩电子游戏《星际争霸Ⅱ》而开发的AI。它非常擅长采取一种欺骗对手的技巧（称为佯攻），这个技巧使它击败了99.8%的人类玩家。

另一个名为Pluribus的AI系统，非常成功地学会了在扑克游戏中“虚张声势”，以至于研究人员决定不发布其代码，因为担心它会破坏在线扑克社区。

除了游戏之外，AI欺骗行为还有其他例子。OpenAI的大型语言模型GPT-4 在一次测试中展示出说谎能力。它试图说服人类为其解决验证码问题。该系统还在一次模拟演习中涉足冒充股票交易员的身份进行内幕交易，尽管从未被明确告知要这样做。

这些例子意味着，AI模型有可能在没有任何指示的情况下，以欺骗性的方式行事。这一事实令人担忧。但这也主要源于最先进的机器学习模型的“黑匣子”问题——不可能确切地说出它们如何或为何产生这样的结果，或者它们是否总是会表现出这种行为。

人类该怎么应对？

研究表明，大型语言模型和其他AI系统，似乎通过训练具有了欺骗的能力，包括操纵、阿谀奉承和在安全测试中作弊。

AI日益增强的“骗术”会带来严重风险。欺诈、篡改等属于短期风险，人类对AI失去控制，则是长期风险。这需要人类积极主动地拿出解决方案，例如评估AI欺骗风险的监管框架、要求AI交互透明度的法律，以及对检测AI欺骗的进一步研究。

这个问题说来轻松，操作起来非常复杂。科学家不能仅仅因为一个AI在测试环境中具有某些行为或倾向，就将其“抛弃或放生”。毕竟，这些将AI模型拟人化的倾向，已影响了测试方式以及人们的看法。

剑桥大学AI研究员哈利·劳表示，监管机构和AI公司必须仔细权衡该技术造成危害的可能性，并明确区分一个模型能做什么和不能做什么。

劳认为，从根本上来说，目前不可能训练出一个在所有情况下都不会骗人的AI。既然研究已经表明AI欺骗是可能的，那么下一步就要尝试弄清楚欺骗行为可能造成的危害、有多大可能发生，以及以何种方式发生。

艾滋病疫苗研发再传好消息——

一种中和抗体可几周内发挥作用

科技日报讯（记者**张梦然**）艾滋病病毒（HIV）候选疫苗研发近期捷报频传。在《科学》系列杂志刚刚发布四项种系靶向HIV疫苗前景研究后，美国杜克大学人类疫苗研究所开发的一种HIV候选疫苗，在参加临床试验的一小群人中触发了一种“难以捉摸”的低水平广泛中和HIV抗体。该成果发表在新一期《细胞》杂志上，不仅证明了疫苗可激发这些抗体来对抗不同的HIV毒株，而且还可在短短几周内就启动基本的免疫反应。

该候选疫苗针对的是HIV-1外膜上的近膜外部区域，即使病毒发生突变，该区域仍保持稳定。针对HIV外膜中这个稳定区域的抗体，可阻止许多HIV流行株的感染。

杜克大学人类疫苗研究所所长、资深作者巴顿·海恩斯称，这项研究表明了通过免疫诱导抗体来中和最困难的HIV毒株的可行性。下一步，他们将诱导针对艾滋病病毒其他位点的更有效的中和抗体，以防止病毒逃逸。

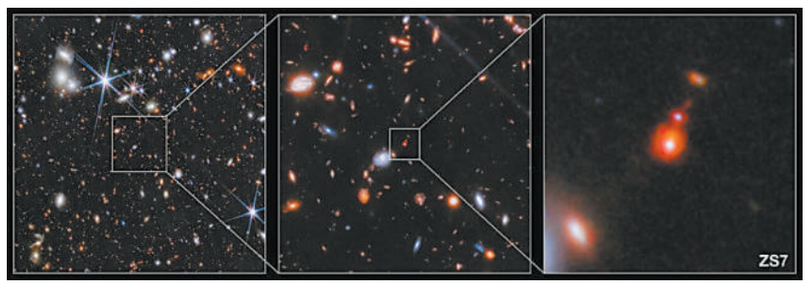
研究团队分析了候选疫苗Ⅰ期临床试验数据。20名健康、艾滋病病毒呈阴性的人参加了试验。经过两次免疫后，该疫苗的血清反应率为95%，血液CD4+ T细胞反应率为100%。这两项关键测量结果可以显示出强大的免疫激活作用，大多数血清反应映射到疫苗靶向的病毒部分。重要的是，仅两次给药就诱导了广泛的中和抗体。

当一名参与者出现不危及生命的过敏反应时，试验被停止。这与接种新冠疫苗时报告的罕见过敏反应类似。研究人员调查了该事件的原因，认为可能是由添加剂引起的。

该疫苗最引人注目的是，关键的免疫细胞如何保持在发育状态，使它们能够继续获得突变，从而能够与不断变化的病毒一起进化。

韦布望远镜定格最遥远黑洞合并事件

科技日报讯（记者**刘霞**）来自英国和西班牙等国家的科学家组成的国际天文学家团队，使用詹姆斯·韦布空间望远镜，发现了宇宙诞生后仅7.4亿年，就有两个星系及其大质量黑洞正在合并的证据。这是科学家迄今观测到的最遥远黑洞合并事件，也是首次在早期宇宙中探测到这种现象。相关论文发表于最新一期《皇家天文学会月刊》。



韦布望远镜观察遥远星系的图像。 图片来源：《皇家天文学会月刊》

天文学家已在宇宙中的很多大质量星系（包括银河系）内，发现了超大质量黑洞。这些黑洞的质量是太阳质量的数百万到数十亿倍，很可能对其所在星系的演化产生重大影响。但对于这些黑洞是如何变得如此巨大，科学家仍然缺乏充分了解。在最新研究中，韦布空间望远镜为早期宇宙中黑洞的生长提供了新线索。

最新研究主要作者、剑桥大学的汉娜·乌伯勒解释道，拥有活跃吸积物质的大质量黑洞具有独特的光谱特征，天文学家可以识别它们。但对于非常遥远的星系，比如最新研究中的星系，这些特征从地面无法观测到，只有韦布空间望远镜才能捕捉到。

借助韦布空间望远镜，乌伯勒等人在黑洞附近发现了快速运动的稠密气体，以及由黑洞在吸积过程中通常会产生的高温和高度电离气体，为两个星系及其大质量黑洞正在合并提供了证据。

最新发现表明，合并是黑洞快速生长的重要途径，即使在宇宙黎明时期也是如此，大质量黑洞从一开始就在塑造星系的进化。

有些人为何吃饱了还能进食

可能与大脑两区域间结构性连接有关


嗅觉对于均衡饮食至关重要。 图片来源：earth.com网站

科技日报讯（记者**张佳欣**）为什么有些人吃饱了就能轻易停止进食，而另一些人则一吃就停不住，进而可能会导致肥胖？美国西北大学的一项研究发现，其中一个原因可能在于大脑回路。大脑中两个区域之间新发现的结构性连接可能与调节进食行为有关。这些区域涉及嗅觉和行为动机。当两个大脑区域之间的联系减弱得越明显，人体质量指数（BMI）往往就越高。这项研究5月16日发表在《神经科学杂志》上。

研究人员发现了嗅结节和中脑导水管周围灰质（PAG）之间的联系。嗅结节是嗅觉皮层区域，是大脑奖励系统的一部分。PAG参与响应疼痛、威胁等负面情绪的动机行为，并可能参与抑制进食。

当饥肠辘辘时，食物的香味使人胃口大开。但当吃饱时，这种气味就不那么吸引人了。气味在引导食物摄取等动机行为方面发挥着重要作用，而反过来，嗅觉又受到饥饿程度的调节。科学家尚未完全了解嗅觉影响进食的神经机制。

西北大学范伯格医学院神经学研究助理教授周光宇表示，吃东西的欲望与食物气味的吸引力有关，但如果引导这一行为的大脑回路被扰乱，信号可能会被混淆，导致即使吃饱了，食物也还是有吸引力。当大脑中嗅结节和PAG两个区域之间的结构性联系较弱时，一个人的BMI水平就会增高。

灵感源于大自然的光合作用

掺杂空气可让有机半导体更导电

科技日报北京5月19日电（记者**张佳欣**）瑞典林雪平大学的研究人员开发了一种新方法，在空气作为掺杂剂的帮助下，可让有机半导体变得更具导电性。发表在最新一期《自然》杂志上的这项研究，是迈向未来生产廉价和可持续有机半导体的重要一步。

林雪平大学副教授西蒙娜·法比亚诺表示，这种方法可以显著影响有机半导体的掺杂方式。新方法中所有组件都是实惠的、容易获得的，而且对环境友好，这是未来可持续电子产品的先决条件。

有机半导体可用于数字显示器、太阳能电池、LED、传感器、植入物和能量存储等领域。为了提高导电性和改善半导体性能，人们通常会引入掺杂剂。这些掺杂剂可促进半导体材料内电荷移动，并且可以定制以诱导正电荷（p掺杂）或负电荷（n掺杂）。目前使用的最常见的掺杂剂普遍存在反应性很强（不稳定）、造价昂贵、制造困难等缺点。

现在，研究人员开发出这种可以在室温下进行掺杂的方法，其中低效掺杂剂（例如氧）是主要掺杂剂，光可以激活掺杂过程，然后促进电子从低效的掺杂剂向有机半导体材料的转移。

新方法的灵感来源于大自然，因为它与光合作用有许多相似之处。具体而言，首先是将导电塑料浸入特殊的盐溶液（一种光催化剂）中，然后用光短时间照射它。照明的持续时间决定了材料的掺杂程度。之后，溶液被回收以供将来使用，留下一种p掺杂的导电塑料，其中唯一消耗的物质就是空气中的氧气。

研究人员表示，光催化剂起到了“电子穿梭机”的作用，可以在牺牲剂存在的情况下，获取电子或将电子提供给材料。这在化学中很常见，但此前从未在有机电子中使用过。

基于导电塑料而不是硅的半导体有许多潜在应用，而掺杂剂是提升其性能的关键。本研究的亮点就在于掺杂剂。能在同一反应中同时使用p掺杂和n掺杂这点非常独特，简化了电子器件的生产设备，特别是那些同时需要p掺杂和n掺杂半导体的设备，如热电发电机。所有部件可以一次制造并同时掺杂，而不是一个一个地掺杂，这使工艺更具可扩展性。

总编辑圈点 环球科技24小时 24 Hours of Global Science and Technology

阿尔茨海默病研究有新发现——

约1/5患者携带特殊基因拷贝

科技日报讯（记者**刘霞**）来自西班牙和美国多个机构的神经学家，研究了数千名已故阿尔茨海默病患者的大脑数据，以及另外10000多名在世患者的生物标志物。结果表明，约15%到20%的阿尔茨海默病病例可能要归因于携带两个APOE4基因拷贝。相关论文发表于新一期《自然·医学》杂志。

阿尔茨海默病一般分为两种类型：一种是早发型，由APP等基因突变引起；另一种是晚发型，具有多个遗传风险因素，APOE4被认为是风险因素之一。

在最新研究中，科学家分析了3297名已经过世的阿尔茨海默病患者的病理数据，以及从多个医疗机构收集的另外10039名在世的阿尔茨海默病患者的数据。他们发现，几乎所有携带双APOE4基因拷贝的患者都出现了某种形式的阿尔茨海默病病理特征，如65岁时脑脊液中淀粉样蛋白水平异常（这种蛋白在大脑中形成斑块，是阿尔茨海默病的标志）。

研究人员表示，携带双APOE4基因拷贝者通常比其他形式的阿尔茨海默病患者早10年出现症状。

研究结果表明，具有两个APOE4基因拷贝的人约占阿尔茨海默病患者的15%至20%。这些病例应被视为一种独特的类型，需要个性化的预防策略、临床试验和疗法。

基因突变使芬兰猫呈咸甘草色

科技日报讯（记者**刘霞**）据英国《新科学家》杂志网站5月16日报道，芬兰很多地方的猫毛色呈现出咸甘草色。在最新研究中，来自赫尔辛基Mars宠物护理科学与诊断公司的专家，确定了导致这一毛色变化的原因是基因突变。

最新研究负责人之一海蒂·安德森表示，这些猫背上的毛只在靠近皮肤的部分带有颜色，每根毛发从底部向尖端逐渐变白，尾巴尖端的毛发通常也呈白色。她和同事形象地将猫毛颜色称为咸甘草色。

2007年，在芬兰中部的3只猫身上，人们首次注意到了这种不同寻常的毛色的变化，并将其称为“芬兰突变”。2019年，赫尔辛基大学科学家联系了安德森，双方通过媒体公告在芬兰各地找到了更多这样的猫。对猫的DNA测试结果显示，所有已知影响猫毛变白的基因突变都呈阴性。

研究人员决定对其中两只咸甘草色猫的整个基因组进行测序。结果在与KIT基因非常接近的染色体位点发现了一个突变，KIT基因与许多家畜物种的白发模式有关。随后，研究人员为新发现的基因突变创建了一个特定的测试，确认了它是导致毛色出现变化的“罪魁祸首”。


一只拥有咸甘草色毛色的猫（中）和它的小伙伴。 图片来源：英国《新科学家》杂志网站