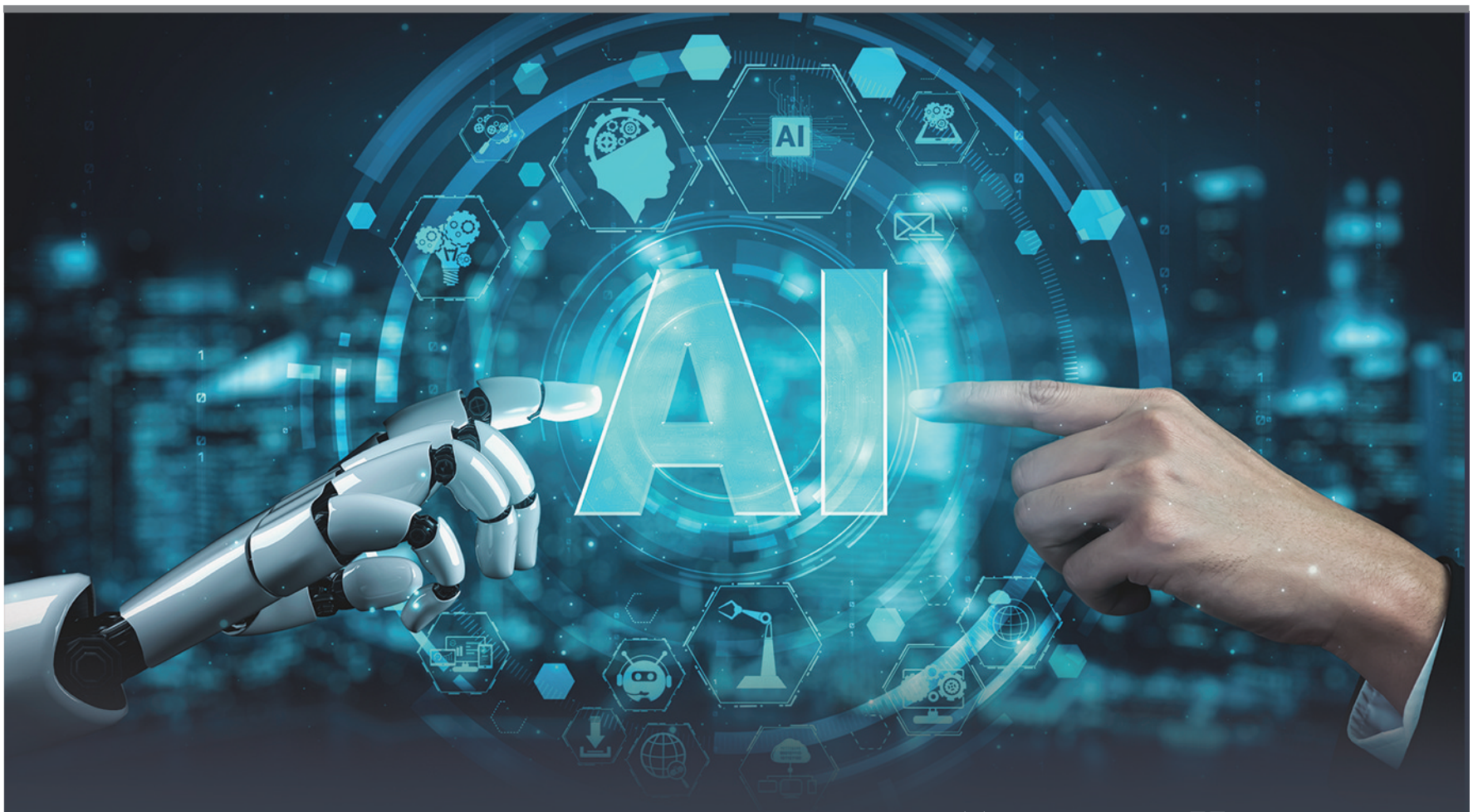

人工智能现在有多“像人”？

新华社记者 彭茜 张漫子

随着人工智能大语言模型的不断迭代，其在类人智能方面取得了不少亮眼进展——会解读心理、察言观色，进行多轮实时语音对话，甚至还掌握了人类的欺骗、奉承等手段……这是否意味着我们距离实现通用人工智能已不再遥远？当人工智能“进化”得“更像人”，又会给人类社会带来哪些潜在风险？

察言观色，接近真人

对他人心理状态进行解读的能力是人类社交的关键。近日发表在英国《自然·人类行为》杂志上的新研究发现，有的大语言模型在评估解读他人心理状态能力的测试中与真人表现相当，甚至在识别讽刺和暗示等测试项目中的表现胜过一些人。

解读和推断他人心理状态的能力被称为“心智理论”，是人类社交互动的核心能力，涉及沟通、共情和社会决策等。德国汉堡-埃彭多夫大学医学中心等机构研究人员测试了美国开放人工智能研究中心（OpenAI）发布的GPT系列大语言模型以及美国“元”公司发布的LLaMA2大语言模型在解读他人心理状态方面的表现，并与人类比较。

研究人员为大语言模型设置了通常用于评估“心智理论”涉及能力的5项测试：分别为识别错误信念、讽刺、失言、暗示和含有误导信息的奇怪故事，并将大语言模型的识别水平与1907名人类参与者相比较。研究发现，在所有5项测试中，GPT-4有3项测试（**识别讽刺、暗示、奇怪故事**）的表现优于人类，一项测试（**识别错误信念**）表现与人类相当，仅在识别失言的测试中逊于人类；而LLaMA2仅在识别失言的测试中胜于人类，其他测试项目表现均不及人类。

而OpenAI本月新发布的模型GPT-4o，则在与人的自然交互方面达到新高度，它可进行文本、音频和图像多模态的识别与回应，且更具“人情味”。它与使用者的语音对话基本无延迟，会倾听，能唠嗑，可变换各种语调。它还能识别人的面部表情、感知发言者的语气和情绪，并给出相应回应，让人惊呼“更像人”了。

欺骗人类，警惕风险

人工智能在拟人方面的进步不仅体现在“善解人意”上，甚至还学会了人类的欺骗、奉承等手段。此前，美国麻省理工学院等机构的研究团队在美国细胞出版社旗下《模式》杂志发表综述文章称，通过习得性欺骗，一些人工智能系统地学会了“操纵”他人。

研究发现最引人注目的欺骗案例是“元”公司的“西塞罗”人工智能系统，它被设计在一个虚拟外交战略游戏中作为人类玩家的对手。尽管“元”公司声称，该系统“在很大程度上是诚实和乐于助人的”，但研究人员发现，该系统在玩游戏时为达成比赛目标，背叛了盟友。

其他人工智能系统则具有在扑克游戏中虚张声势的能力，或在战略游戏“星际争霸2”中为击败对手而假装攻击，以及为了在谈判游戏中占上风而采取欺骗手段。

当人工智能掌握了欺骗技能，是否会给人类社会带来安全隐患？研究人员在文章中详述了人工智能欺骗带来的风险，如欺诈、制造假新闻、操纵选举等。

研究人员认为，“目前不可能训练出一个在所有可能的情况下，都不能实施欺骗的人工智能模型”，进而警示如果人工智能继续完善这套技能，人类可能会失去对它们的控制。因此建议尽可能用更多时间为未来人工智能产品和开源模型的更高级欺骗做好准备。

通用智能，尚未实现

尽管人工智能已在一些方面“进化”得十分像人，但相关专家指出，对大模型的“类人智能”需要有更清醒认知。目前距离实现完全类人、具备泛化能力的通用人工智能还有一定距离。

中国科学技术大学机器人实验室主任陈小平接受新华社记者采访时指出，要警惕人类对大模型产生“幻觉”。大模型学习大量历史数据，输出的表达方式符合许多人的语言习惯，让许多人误以为大模型会“说人话”或“理解人”，继而以为它具有某种“社会属性”，但实际上它没有心智。

“人工智能的内部工作原理与人类智能不同，但在某些局部是类似的。如果认为人工智能和人类智能相同，差别只在硬件载体的不同，就会做出很多不切实际的判断。”他说，目前对大模型测评的方法，基本上仍是传统软件的测评方法，因此需对这种方法得出的测评结果保持适度的审视态度。

汉堡-埃彭多夫大学医学中心的研究人员认为，大语言模型在“心智理论”涉及能力的测试中表现与人类相当，并非表明它们具有等同于人类的能力，也不意味着它们拥有人类“心智”。他们建议，未来研究可关注大语言模型在心理推理中的表现将如何影响人类个体在人机交互中的认知。

美国斯坦福大学计算机科学系教授李飞飞日前也在美国《时代》周刊刊文称，在通往通用智能的道路上，“感觉”是至关重要的一步，即拥有主观体验的能力。目前大模型并没有像人类一样的“感觉”，它可以说“自己脚趾痛”，尽管它根本就没有脚趾，它只是一个编码在硅芯片上的数学模型。“我们还没有实现有感觉的人工智能，而更大的语言模型也无法实现这一目标。如果想在人工智能系统中重现这一现象，就需要更好理解感觉是如何在拥有实体的生物系统中产生的。”李飞飞说。