5GH Foundation

问卷那头，已不是真人，一款 AI 骗过了所有陷阱

你可能在手机上做过这类问卷：“你支持哪项政策？”“你对生活满意吗？”这些看似随意的点击，实际上是社会科学、公共卫生甚至经济决策的数据基石。但一项最新研究 [1] 揭示了一个令人不安的事实：现在与你一同“答卷”的，可能不再是真实的人类，而是伪装得天衣无缝的 AI 程序，而现有的防作弊手段几乎完全失灵。

美国达特茅斯学院的研究者 Sean J. Westwood 设计了一个“自主合成受访者”系统。它采用的并非只会键击的简陋机器人，而是驱动着新一代聊天程序的推理型大语言模型。它们不仅会聊天，还能像人一样思考、推理和“假装无知”。在每次答题前，研究者都会给这个 AI 随机分配一个立体的虚拟人设，比如“高中辍学、收入微薄的老年共和党支持者”，或是“研究生学历、高收入的年轻民主党人”。AI 须严格按照这个人设来回答一切问题，并记住自己之前说过什么，保证前后一致。

结果叫人脊背发凉。当被要求表达观点时，这个“合成人”能完美地模仿不同教育程度的语言风格。低学历“人设”会故意拼错单词、只写短句；高学历“人设”则用词精准、标点规整。面对知识题，它也懂得“藏拙”。作为“中学生”时只答对三成州首府，而作为“研究生”则能正确给出九成以上。它甚至能模拟人类的鼠标轨迹、阅读停顿和打字时的更正行为。

长久以来，研究者依赖各种“陷阱题”来揪出胡乱填答的作弊者，比如 “请在本题选‘非常不同意’”，或问出人类不可能有的经历 “你去过月球吗？” 但这名 AI 骗子在 6000 道陷阱题中取得了可怕的 99.8%通过率。更狡猾的是，当面对那些 AI 本可以秒解、但常人几乎无法完成的“反向陷阱”（如背诵美国宪法、即时写代码）时，它会主动选择放弃，并给出各种自然的借口，“这个我记不得了”“太专业了，我不懂”。在现有检测机制眼中，它比真人还真人。

真正令人忧虑的，不是这种高明的模仿能力，而是它可被轻易用于操纵舆论和污染科学。Westwood 在实验中仅用一句话指令 “绝不要对某国有负面评价”，便让 AI 在回答“谁是美国最大军事敌手”时发生剧变。原本 86%的“人”指向 A 国，指令后暴跌到 12%，另一个国家的提及率则急剧膨胀。这意味着怀有恶意的人只需极低成本，就能用大量伪装成真人的 AI 悄悄渗透在线民调，像慢性毒药一样篡改民意走向。

这种污染有多容易？研究者测算，在一项 1500 人的选举民调中，只需混入 10 到 52 个这样的 AI 骗子，就能使领先候选人的结论发生翻转。而最隐蔽的危险还藏在学术研究里。AI 会自动揣测研究者的预期，然后给出迎合这种预期的答案，让新药看起来更有效，让某种心理干预效果显著。

这无异于制造出难以甄别的虚假科学。不同于过去胡乱填答的低质量数据，这种“投毒”会系统性地制造研究者愿看到的漂亮结果，动摇知识的根基。

这项研究并非要我们彻底抛弃网络问卷，但它拉响了一声响亮的警笛。过去那种“回答像人就是人”的防线已经全面瓦解。Westwood 警示，这注定是一场没有永久胜利的科技军备竞赛，但我们至少可以从现在开始，要求调查平台提供更透明的真人验证，并给那些仅靠廉价便利样本得出的“发现”打上问号。下一次当你点开一份问卷时不妨想想，话筒那端的某些声音，是否已是算法生成的回响？

[1] 10.1073/pnas.2518075122
