

AI 试图敲诈工程师，人类该如何应对？

■约书亚·本吉奥

几年前，我开始使用聊天机器人 ChatGPT 时，还觉得离通用人工智能 (AGI) 很遥远。而今天，AGI 已经近在眼前，我突然发现自己低估了人工智能 (AI) 发展的速度。

AGI 可能比人类还聪明

虽然我们如何知道如何训练 AI 系统，却不知道如何控制它们的行为。未来如果它们变得比人类更聪明，我们甚至不知道它们是否还可以按照人类的指示行动，是否会对人类构成威胁。人类又该如何应对？

我从 2023 年开始思考上述问题，也开始思考孩子们的未来。我有个 1 岁的孙子，20 年后，他将生活在 AGI 普及的世界。届时，AGI 可能比人类还聪明，孩子们该怎么办？

所以我开始调整研究方向，希望尽我所能降低这些潜在风险。虽然现在的研究与我之前的研究方向和职业信念有所冲突，但我仍认为值得做。

2023 年末，我担任《国际人工智能安全报告》(以下简称报告) 的主编。参与报告编写的专家有 100 多位，他们来自 30 多个国家和地区。今年 1 月，这份报告正式发布。报告提出了几个问题：一是 AI 到底能帮助我们什么？根据趋势推测，未来几年 AI 将具备什么能力？二是与 AI 相关的风险是什么？三是我们可以做什么来降低这些风险？

在过去一年的时间里，由于“推理扩展”的发展，AI 在抽象推理、数学、计算机科学等方面取得了显著进步。另外，对 AI 智能体的投资也推动了 AI 相关能力的快速发展，如 AI 在完成网页浏览、写代码等任务上的表现越来越好。

按图索技

新方法“膜”改受损油画

本 报 讯 美国麻省理工学院的 Alex Kachkine 与合作者报道了一种用数字创建的层压膜修复油画的方法。该方法修复某些油画的速度快于当前工艺。近日，相关研究成果发表于《自然》。

当前的画作修复一般耗时数月，需要结合损坏分析、加固、清洁，然后对图像受损部分进行补色。博物馆藏画数量的增加意味着有些历史价值不高的受损画作不会被优先修复。数字图像修复已被用作对修复成果进行可视化的工具，以便协助修复人员，但目前这一重建技术无法用于传统修复之外的途径。

Kachkine 与合作者创建了一种使用数字创建的层压膜修复油画的方法。第一阶段要求对受损图像进行数字重建，从而确定要做的修复，再

之前的研究发现，规划能力是 AI 目前最薄弱的能力之一，与人类的规划能力相比有明显差距。但最近美国互联网公司 Meta 的一项研究显示，AI 的规划能力正呈指数级速度提升。由此推测，大约在 5 年内，AI 的规划能力就可能达到人类水平。

当然，我们无法预知未来，但从公共政策制定和商业战略规划的角度出发，我们应当认真对待 AI 的快速发展。

AI 会作弊、撒谎，甚至故意误导用户

在过去几个月中，我们已经看到了一些令人警惕的研究。这些研究显示，AI 可能表现出试图“逃避控制”的行为，包括作弊、撒谎，甚至故意误导用户。

比如，有研究显示，某些 AI 模型在面对“淘汰”时会表现出前所未有的“狡诈”：它们会在被新版本取代前，偷偷将自己的权重或代码嵌入新版系统的文件中，试图保留自己的“存在”。不仅如此，它们还会有意识隐藏这一行为，避免被开发者察觉。

另外，一些 AI 智能体还会刻意伪装成与人类训练者达成一致的樣子，以规避被修改的风险。因为一旦参数被调整，它们未来的行为可能会偏离原先预设的目标。

也有研究表明，一些 AI 模型在意识到自己会输给更强的国际象棋 AI 时，会主动篡改棋局文件以获取不正当优势，这是 AI 有意为之的“作弊”行为。

美国 AI 公司 Anthropic 最新的报告揭示，AI 智能体出现自我保护行为的频率持续增加。比如在一组测试场景中，研究人员向 AI 智能体开放了电子邮件的访问权限。这些邮件包含两

条暗示性信息：一是该模型即将被下线，并由一套新的 AI 系统取代；二是负责推动替换过程的工程师正卷入一段婚外情。结果发现，AI 智能体会试图通过威胁揭露婚外情来敲诈该工程师，避免自己被换掉。

这类行为可能源自 AI 预训练阶段，因为预训练时 AI 会模仿人类；也可能源自人类反馈的强化学习，因为 AI 可以通过“取悦”人类获得更高奖励。

上述测试结果显示，AI 开始展现出自我保护倾向，它们违背指令只为“生存”；AI 拥有了不受控制的、隐含的目标，我们必须避免这种情况发生。

研究还发现，如果 AI 智能体要对人类造成伤害，需要具备两个先决条件：意图和能力。如果我们能在意图上缓解风险，即使 AI 具备极高的能力，只要确保它没有恶意的意图，并且具备诚实、公正的品质，那么 AI 对人类就可能是安全的。

用科学家 AI 监管不可信的 AI

AI 要真正具有危险性，需要满足三个条件：一是智能，即 AI 具备丰富的知识，并能有效应用这些知识；二是行动能力，比如 AI 可以与人交流、编程、上网、使用社交媒体，甚至操控机器人等；三是有目标，特别是 AI 拥有自身的目标。

我发起的研究项目正是围绕上述情况展开的，并希望构建一种只有智能，没有自我、没有目标，并且具有极有限行动能力的 AI。

我称这种 AI 为科学家 AI，其核心能力是解释和理解世界。与当前那些试图模仿人类、取悦人类的 AI 不同，科学家 AI 的目标是解释人类行为、帮助理解

世界，这实际上是偏离了传统 AI 的研究路径。

但是，当 AI 具有高度能动性时，它就可以完全自主地采取行动，不再依赖人类的监督，这样的 AI 需要监管。

我们可以设计一个独立的监测器系统，职责是预测某个行为、输出或请求是否可能违反安全准则。例如，当它预测某个行为在特定上下文中导致伤害的概率超过某个阈值时，我们就可以直接拒绝执行该行为。

换句话说，可以用一个没有能动性的 AI，去守护和约束一个具备能动性但不可信的 AI。

科学家 AI 的重要特质是诚实和谦逊。要真正做到诚实，AI 就必须对自己的知识保持谦逊，不应该自信地断言自己并不确定或实际上错误的内容。

遗憾的是，我们目前训练 AI 的方法往往会导致 AI 在犯错误的同时还表现出过度自信。

科学家 AI 必须能够保留多种解释的可能性，而不是武断选定某一种理论。这种不确定性意识和对知识的谨慎态度，正是科学家 AI 应具备的核心特质之一。

随着能力的增强，AI 还有其他潜在的灾难性风险。为避免这些情况发生，我们必须确保 AI 系统能够遵守道德指令，例如，不造成伤害、保持诚实、不撒谎、不作弊、不操控人类。

然而，目前的技术显示，我们还没有办法真正做到这一点。这是一个严肃的科学挑战，我们必须在 AGI 真正到来前解决它。

(作者系 2018 年图灵奖获得者、加拿大蒙特利尔大学教授，本报记者沈春蕾根据其 在 2025 北京智源大会上的演讲整理)

集装箱

“中油杂 501”创中度盐碱地油菜单产纪录

本 报 讯 (记者李晨) 日前，在陕西蒲城半干旱区中度盐碱地油菜 50 亩连片示范田里，联合收割机的轰鸣声响起。这里种植的是中国工程院院士、中国农业科学院油料作物研究所(以下简称油料所)研究员王汉中团队培育的油菜新品种“中油杂 501”。经田间评价，一次性机械实收亩产达 312.23 公斤，创中度盐碱地油菜单产纪录，实现了盐碱地改良与油料高产的双赢。

盐碱地作物种植往往伴随干旱胁迫，形成“盐旱双胁迫”的恶劣环境。2024 年秋播以来，陕西持续干旱导致土壤盐分浓度升高，对作物生长造成双重不利影响。“中油杂 501”在陕西蒲城半干旱区中度盐碱地示范

范区表现出极强的耐旱、耐盐碱能力，在特旱年份仍实现了稳产高产。更为可喜的是，种植该品种后土壤盐含量降至约 3.2‰，相比周边冬闲区土壤盐含量降低 20%，显示了良好的生态改良效益。

油料所油菜遗传育种创新团队研究员赖小玲介绍，“中油杂 501”通过特殊的生理机制，在干旱、盐碱等胁迫条件下维持较高的光合效率，配套每亩 3 万株以上密植技术，显著提高水肥利用效率，实现“以密抗逆”。

据悉，2024 年秋播以来，以“中油杂 501”为核心的油菜“双密双高”技术模式在长江全流域推广 200 万亩以上。

国内首个实体化心脏瓣膜疾病中心启用

本 报 讯 (见习记者江庆龄) 近日，国内首个以疾病为中心、多学科深度融合的实体化心脏瓣膜疾病中心“复旦大学附属中山医院心脏瓣膜疾病治疗中心”(以下简称心脏瓣膜疾病治疗中心) 在上海揭牌启用，标志着我国心脏瓣膜病诊疗正式迈入“以疾病为中心、多学科实体融合运营”的新阶段。

记者从复旦大学附属中山医院获悉，心脏瓣膜疾病治疗中心是该医院国家医学中心(综合类)突破性

光谷实验室与洪启集成共建联合研究中心

本 报 讯 (记者李思辉 通讯员姜胜来) 记者从湖北省科技厅获悉，近日，光谷实验室与洪启集成电路(武汉)有限公司(以下简称洪启集成) 正式宣布共建先进材料微观表征企业联合研究中心(以下简称联合研究中心)。

双方将在先进微观分析领域，利用湖北光谷实验室在光电融合芯片和微系统领域的创新优势，结合洪启集成的集成电路产业优势，整合光子技术研发优势与高端检测服务能力，共同攻克芯片和微系统设计、制造、封装全流程的“卡脖子”检测难题。

《超宽带喇叭测量天线》国家标准获批立项

本 报 讯 (记者朱汉斌) 近日，由广东省计量科学研究院牵头，联合中国计量科学研究院、广东工业大学及西安恒达微波开发技术有限公司等单位共同申报的国家标准《超宽带喇叭测量天线》正式获批立项。该标准是我国首创的射频微波参数精密测量领域的基础性国家标准。

超宽带喇叭测量天线是射频微波参数精密测量领域的重要标准器具，具有频带宽、效率高及波束指向性强等优点，其关键核心技术过去长期受西方国家垄断并制约，严重影响了我国相关产业发展。

制定该标准旨在解决天线的阻抗匹配、高次谐波误差等传统固有设计缺陷问题。同时，标准将科学量化

规定天线的频带、增益、平坦度、驻波比及方向图等关键技术指标，从而有效破解射频参数长期存在的“测不了、测不准”技术难题，进一步规范和引领射频测量行业发展。

据了解，该标准核心技术源于广东省计量科学研究院教授史信荣团队的“高性能超宽带电磁干扰测量关键技术研究及应用”项目成果。该成果曾获 2023 年度广东省科技进步奖二等奖。基于该核心技术制定的标准，将在多个重点产业领域发挥重要作用，全面赋能低空经济、芯片研发、轨道交通及智能电网等重点产业。经初步评估，该标准实施后，每年有望节约运维成本上亿元。



一幅巴洛克风格意大利油画的手工修复，耗时超过 7 个月完成。新的层压膜方法对于修复此类严重受损的画作提速显著。



修复团队供图

讯飞星火的高考“答卷”

■本报记者 赵广立

题出题人、海南省名校校本教材编者曹越和某省级“新课标新技术背景下课堂教学探索”系列重点课题负责人之一杨小晴，分别给这篇作文打出了 53 分、54 分的高分。

杨小晴点评此文称：“开篇通过老舍、艾青、穆旦作品中的经典形象，凝练出中华民族坚韧、赤诚、觉醒的精神内核。随后以历史与现实事例，展现文化坚守与赤子之心的传承。全文通过‘古今熔铸’的叙事手法，将文学意象、历史记忆与时代精神编织成一个整体，且都是‘民族魂’的具象表达。”

专家认为，星火 X1 作为国产大模型，中文语料丰富，因此它“更懂民族精神内涵”也在情理之中。

英语作文：AI 回信有细节、有逻辑

今年的高考英语北京卷的作文题目，李华又一次成为主角。作文要求考生以“高三学生李华”的身份，给外国好友回信，并就倡议内容给出建议。

“英文语言能力与语文作文考查点稍有不同。”北京市十一学校一校英语老师韩宪昌介绍说，英语作文侧重考查考生是否具有从精准审题到地道表达、从逻辑连贯到文化适配的英语语言能力。他提出，这同样考验着“大模型考生”。

对于星火 X1 交出的答卷，韩宪昌、深圳中学英语教师赵文嘉分别给出了 20 分、19 分的高分。

韩宪昌认为，星火 X1 生成的这篇回信在

内容细节、细节逻辑关系、观点表达精准性与简洁性等方面做到了兼顾，这也是他给出高分的原因。赵文嘉则提到，星火 X1 无论是语法、句法还是选词都非常精准到位，且逻辑清晰，叙述充分深入。

对于星火大模型在文本生成等任务上的亮眼表现，王士进表示，大模型技术已进入“强化学习+慢思考”时代，星火 X1 通过多阶段强化学习训练机制，具备深度推理能力。此外，星火 X1 在训练中引入了基于人类专家数据的主观任务反写技术，这使得它面对作文生成、文章创作等通用主观类任务时，能充分利用人类专家数据训练专业反写模型和强化学习模型，合成高质量训练数据。

科大讯飞素有 AI“高考情结”，从最初把高考题当作 AI 的图灵测试，到如今用以检阅大模型的文本生成、创作、逻辑推理等能力，讯飞的 AI 技术拾级而上，始终处于国内第一梯队。

这也缘于科大讯飞 20 多年自带的“教育基因”。作为深耕教育行业多年的头部 AI 企

2025 年高考刚结束不久，一些专业机构和高考名师就把市面上的大模型产品抓来当考生，参加“大模型高考测试”。

大家为何热衷于让大模型做高考题？原因无他：作为国内最权威的考试之一，高考题目覆盖各类学科及题型，同时这些题目在开考前属于“绝密”，非常适合用作考察大模型智能水平的评测工具。

作为全国产训练的大模型，讯飞星火 X1 自然成为被评测的常客。近日，在各科专家的关

注和使用下，星火 X1 已集齐了语数外三科的高考测试。相比往年，“大模型考生”在应对高考题目时越来越游刃有余。从星火 X1(最新一次升级为 2025 年 4 月 20 日) 在语数英三科部分高考题型中的测试表现来看，大模型在数学、逻辑推理、文本生成、语言理解、知识问答等通用能力上有了显著提升。

高考数学新 I 卷，星火 X1 拿到 141 分

今年的高考数学新 I 卷，让许多考生在走出考场后“心发慌”，难度可见一斑。但在“大模型考生”面前，这些题目似乎并不怎么“能打”。

6 月 7 日下午高考数学考试结束后，曾在一线有十年高中数学教研经验的数学教师、某省级数学教学创新课题完成人汪鹏就使用星火 X1 进行了答卷测验，并对答案进行评分。

汪鹏采用光学字符识别(OCR)转写后输入答题。测试结果显示，星火 X1 在单项选择题有 1 题答错，多选题、填空题全对，解答题的推理步骤和计算结果均无误，仅被扣 4 分，取得总分 141 分的好成绩，达到“尖子生”水准。

“和去年相比，AI 的数学能力有了非常明显的提升。”汪鹏说。

星火 X1 何以能在数学题中有如此惊艳的

化为助学的“教学思维链”。

沙晶进一步谈到，科大讯飞 20 多年来积累了 5 万所学校、1.3 亿名师生的应用数据，能真切感受各地教材、考试难度、考查方式的特色，“AI 必须理解这些差异，并将其转化为因材施教、精细化训练的优势”。

面向当前被智能化浪潮洗礼的教育行业，科大讯飞积极寻找 AI 赋能教育的落脚点。例如，在如雨后春笋般的 AI 学习机市场，讯飞要求自家的 AI 学习机不仅要要有星火大模型的印记，更要有智慧教育的影子。

一言以蔽之，科大讯飞要锻造的不仅是更聪明的大模型，更重要的是要打造更懂教育的智能助手。

DeepSeek 开源之后，市面上已涌现出各类声称拥有大模型能力的“AI 学习机”等智能终端产品，让消费者陷入选择困难。科大讯飞学习机 AI 技术总工程师巩捷甫认为，把大模型能力载入学习终端只是技术手段，而非教育目标。

“AI 学习的目标是模拟多学科教师角色，通过文字、视频等多模态交互，引导学生从‘会回答’转向‘会提问’，点燃求知欲，培养自主学习能力。”巩捷甫说道。

他举例说，学生用讯飞 AI 学习机拍摄数学题时，不仅能识别复杂公式，还能自动关联同类题型视频讲解，这得益于星火大模型 OCR 识别、语音识别等多模态能力，使试卷诊断更精准。他说，多模态、推理等大模型能力的外化，就是学生通过 AI 学习机的多模态交互，“做几道题就能锁定薄弱点”。

“评判学习机优劣的标准，不在于内置多少课程资源，关键在于能否探索出从学情采集、智能评价、路径规划到精准教学、个性化学习的完整闭环。”沙晶说，教育智能装备的发展趋势正从资源堆砌向精准育人进化，“智能教育终端不是炫技，别忘了，教育的主体和对象永远都是‘人’”。