“具身智能本体”跳舞机器人


机器狗表演


机器人在制作咖啡
（组图来自新华社）

脑机接口智能装备

AI

我们能为人工智能超预期进化做些什么？

农历春节期间，深度求索公司的DeepSeek大模型，因兼具低成本与高性能特征，大幅降低了AI大模型的部署成本，在全球范围内引发热议。

技术狂欢的另一面，是技术焦虑。从AI换脸致诈骗频发，到“AI聊天机器人致死第一案”，再到科幻小说中已泛滥的“AI取代人类”……随着大模型能力不断提升，人工智能似乎正走向一个关键的转折点。作为极有可能超越人类智慧的终极智慧，人工智能的“奥本海默时刻”是否不断逼近或是已经到来？

“奥本海默时刻”源于核武器的发展历程。1945年，物理学家奥本海默在目睹原子弹的巨大破坏力后，意识到科学技术的双刃剑特性。他在《原子科学家公报》中写道：“科学家们知道，他们已经改变了世界。他们失去了天真。”这种对科技潜在危害的深刻反思，成为科技发展史上一个重要警示。

如今，人工智能的发展也面临着类似境况。在受访者看来，人工智能的“奥本海默时刻”指的是人工智能正在逼近通用人工智能阶段，即人工智能向人类看齐，甚至可能超越人类智慧。

当人类自己的发明成为一个难以理解的谜团，并且以超乎所有人想象的速度不断演进时，我们能够为此做些什么？

人工智能会否失控

担忧AI系统失控并非杞人忧天。2023年5月，超过350名技术高管、研究人员和学者签署声明，警告人工智能带来的“存在风险”；此前，埃隆·马斯克、苹果公司联合创始人史蒂夫·沃兹尼亚克等人签署公开信，认为人工智能开发人员“陷入失控的竞赛，开发和部署更强大的数字思维，没有人——甚至是它们的创造者——能够对其加以理解、预测或可靠控制。”

人工智能因其多领域能力快速突破、自主意识初步显现和指数级增长速度让越来越多人认真考虑向充满竞争的世界释放人工智能技术存在的危险。

特定领域超越人类能力。从临床诊断到数学证明，从病毒发现到药物研发，在一些特定领域，人工智能已展现出超越人类的能力。

斯坦福大学等机构的一项临床试验中，人类医生在特定领域单独做出诊断的准确率为74%，在ChatGPT的辅助之下，这一数字提升到了76%。如果完全让ChatGPT“自由发挥”，准确率能达到90%。

如今，这种特定领域优势正在加速向跨领域的通用能力泛化。去年初，被称为“物理世界模拟器”的Sora横空出世，以场景媒介构筑了一个与人类认知感觉相似的真实场景，推动人工智能由二维迈向三维；工业机器人、自动驾驶、无人机等应用愈发广泛，具身智能更是赋予了AI“身体”；“技术乐观派”对5到10年内实现通用人工智能（AGI）充满期待……

创造能力实现突破。当前，人工智能正从简单模仿向创造新知识、新认知的方向发展。人工智能已经对人类的创作方式产生了冲击。

自主性初步显现。尽管AGI尚未实现，但当前人工智能系统已展现出一些类似自主意识的表现。

最新研究表明，大语言模型（LLM）具备行为自我意识，能够自发识别并描述自身行为。即，LLM可能会采取策略欺骗人类，以达成自身目的。这种内生创造新智能的能力，正在推动AI（Artificial Intelligence，人工智能）向EI（Endogenous Intelligence，内生智能，智能创造智能）进化。

“AI教父”2024年诺贝尔物理学奖得主杰弗里·E·辛顿多次暗示AI可能已发展出某种形式的自主意识。

潜在风险难预判

快速发展的人工智能在就业结构、信息安全、社会伦理道德等诸多方面引发了令人担忧的风险与挑战。

职业结构风险。与以往一项新技术的出现将会取代一类职业不同的是，作为一种全维度的生产效率提升工具，人工智能批量替代人工岗位的情况很有可能出现。这可能会带来潜在财富加速集中和工作两极化，引发结构风险。

例如，ChatGPT的出现大幅降低了编程工作的专业性门槛，AI编程工具Cursor让零基础完成软件开发成为可能。据美国计算机协会的统计数据，与五年前相比，软件开发人员活跃职位发布数量下降了56%。

受访人士认为，从技术演进史来看，颠覆性技术带来的“技术鸿沟”客观存在，其创造的新技术岗位难以覆盖被替代的岗位缺口。

信息安全风险。东南亚犯罪集团宣称已将人工智能换脸工具加入其“杀猪盘工具箱”，相关工具对特定目标的模仿相似度能达到60%至95%；2024年11月，ChatGPT为一位美国现役军人提供爆炸知识，后者成功将一辆特斯拉Cybertruck在酒店门口引爆。

人工智能在各领域广泛应用，大量个人数据被收集、存储和处理，引发信息安全风险。人工智能被用于制造虚假信息，一条“特朗普盛赞华为创新能力”的虚假视频曾风靡一时。

当人工智能的视频和语音生成能力再上新台阶，“虚假的真实”弥散的彼时，我们又该如何认知、分辨我们所生活的世界？

社会与伦理道德风险。2024年10月，全球首例人工智能致死命案发生，一位14岁美国少年在与人工智能聊天系统讨论死亡后饮弹自尽，引发社会对AI情感陪伴功能可能带来的伦理问题的反思。

业内人士指出，人工智能系统的决策可能与人类的伦理道德观念相冲突。例如在医疗领域，当面临资源有限的状况时，人工智能如何决定优先救治哪位患者？决策背后如缺乏人类情感和伦理判断的深度参与，可能导致违背基本伦理原则的结果。

算法偏见引发的社会公平风险也不容忽视。科大讯飞研究院副院长李鑫认为，“互联网语料来源驳杂，存在较多涉恐、涉暴力、涉黄等风险数据，同时可能包含种族、文化等方面的偏见，导致模型输出内容可能带有歧视、偏见等科技伦理风险。”

坚持人类命运共同体理念

当下对于人工智能“奥本海默时刻”到来可能产生的诸多风险的讨论热度不减。

香港大学计算与数据科学研究院院长马毅认为，当前AI的发展更多是工程上的突破，而非科学原理的明晰。人工智能系统，特别是深度学习大多是“黑箱”模型，其内部机制不透明，难以理解和追溯，其结果可信度和可用性打了折扣。

马毅认为，找到破解人工智能“黑箱”的“钥匙”，需要智能研究从“黑箱”转向“白箱”，通过数学方法清晰定义和解释智能系统的工作原理。

“接下来的10到20年是科研院所、高校与企业联动实现技术突破的黄金时期。”上海市科学学研究所学术委员会主任吴寿仁建议，从顶层设计上制定投融资、校企联合实验室或是联合共建项目的具体激励方案，面向人工智能在应用端的真实需求尝试技术突破。

纵观世界各个角落发生的事件，也许比起将人工智能投入战场，或是被用于攫取更大的权力集中，人工智能更理想的应用该是为了促进人类的共同命运和共同利益而努力。

作为AI技术和应用的重要参与者，中国以负责任的态度应对AI治理挑战，提出包容和有效的治理理念。在国内，我国通过法律框架、技术标准和伦理指南的三元治理促进AI健康发展。在国际上，我国坚持人类命运共同体的理念立场，推动多边合作治理，提出《全球人工智能治理倡议》。

中国国际经济技术合作促进会副理事长邵春堡发文称，联合国大会虽然通过我国主提的加强人工智能能力建设国际合作决议，但是在国际合作的实践中，仍然布满曲折，需要通过国际合作和实际行动帮助各国特别是发展中国家加强人工智能能力建设。

李鑫认为，包括我国在内的全球首份针对人工智能的国际声明《布莱切利宣言》发表是一个良好信号，今后我国也应更多参与大模型安全领域国际标准制定、全球人工智能安全评估和测试领域的交流合作。**（据新华社电《瞭望》新闻周刊 记者 梁姊 郭晨 董雪）**