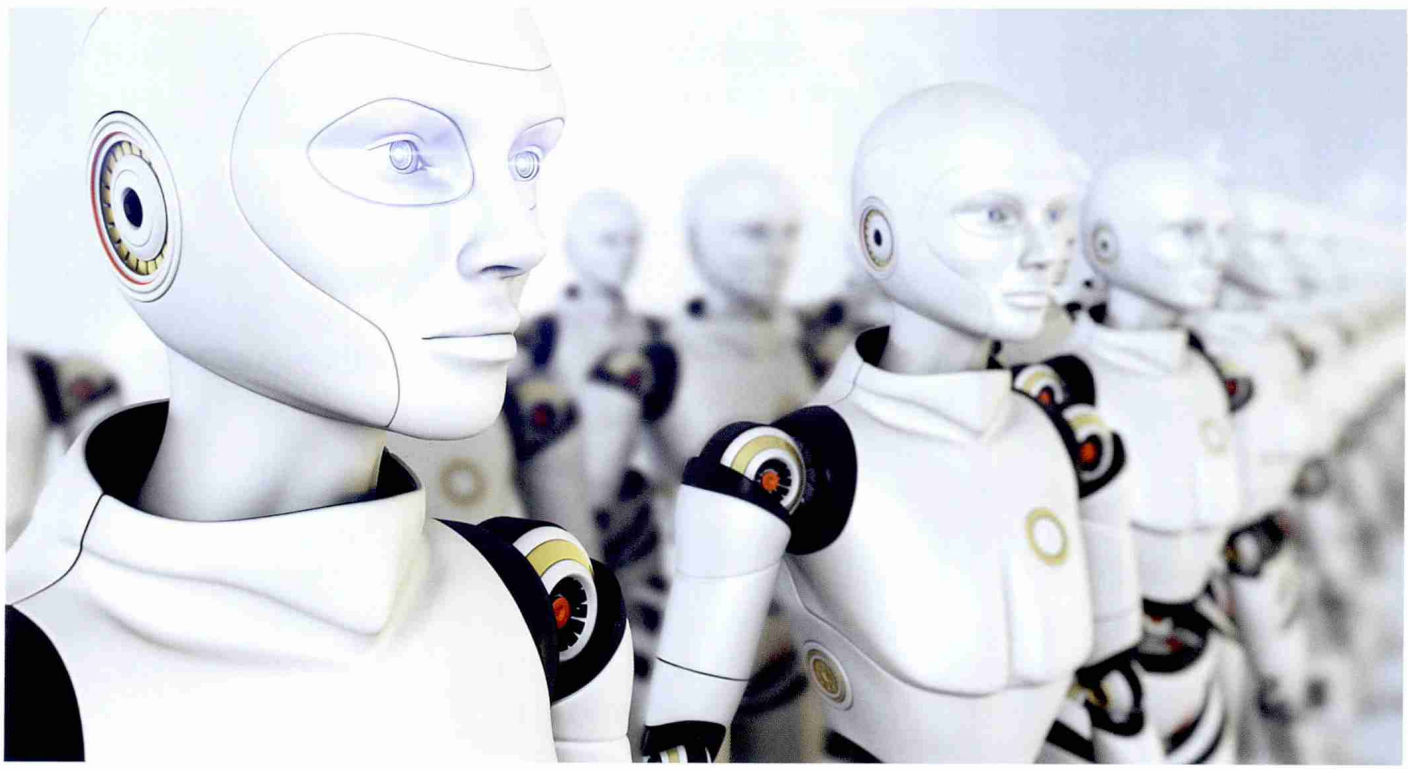
人工智能的真正风险

最好是期待人类的聪明才智，而非低估它；最好是承认风险的存在，而非否认它

□艾伦·达福（Allan Dafoe） 斯图尔特·罗素（Stuart Russell）



CFP供图

AI 对人类是威胁吗

有新闻报道称，人工智能（AI）研究尽管在未来可能取得成功且前途不可限量，但潜在的长期风险正逐渐显示出来。奥伦·埃齐奥尼（Oren Etzioni）是一位知名的人工智能研究人员，他对此感到不满（参见《专家并不认为超级智能对人类是一种威胁》）。在他直接公开指责牛津大学哲学家尼克·波斯特罗姆（Nick Bostrom）和他的最新著作《超级智能》（Superintelligence）后，埃齐奥尼指出，波斯特罗姆提出的“即将出现的人类层级智能的主要数据源”来自对 AI 研究人员意见的调查。在此之后，埃齐奥尼自己也对 AI 研究人员的意见进行了调查，声称他得出的结论推翻了波斯特罗姆的观点。

要知道，埃齐奥尼甚至没提到为什么他公开谴责《超级智能》所造成的影响，这点很重要。他也没有解释清楚为什么超级智能的 AI 可能会造成无法把控的负面影响，为什么提前开始解决 AI 的问题是如此重要。波斯特罗姆并没有把事实建立在超人类 AI 系统即将来临的这个预测上，他在著作中写道：“本书并不是在表明我们正处于人工智能重大突破的开端，也没有表明我们可以准确预测这一重大进展何时会发生。”

因此，在我们看来，埃齐奥尼的文章转移了读者的注意力，使读者没有关注该书的核心内容。埃齐奥尼以质疑调查结果为由，直接从个人偏好的角度出发攻击波斯特罗姆。我们觉得有必要更正一下。我们的一位同事（拉塞尔）甚至还参与了埃齐奥尼的调查，但却发现他的回答完全被埃齐奥尼曲解了。事实上，根据我们的详细分析，埃齐奥尼的调查结果与波斯特罗姆的观点完全一致。

那么，埃齐奥尼是如何得出他的新结论的呢？他自己设计了一份调查问卷，不过问卷的质量比波斯特罗姆

的差一些，又对结果进行了错误的理解，所以得出了不正确的结论。

文章的副标题是："如果你问那些真正应该了解的人，那你就会发现没有多少人认为人工智能对人类是一种威胁"。所以在此情况下，这让读者认为埃齐奥尼的确询问了真正了解AI的专业人士的看法，而波斯特罗姆并没有。但事实却正好相反。波斯特罗姆才是真正询问过的人，而埃齐奥尼谁也没问过。波斯特罗姆调查了最常受访的前100位AI研究人员。超过半数的受访者认为人类层级的机器智能对人类的影响将是"不太好的影响"或是"极其坏的影响（事关存亡的人类大灾难）",这样的几率很高（至少有15%的几率）。埃齐奥尼的调查则不像波斯特罗姆那样，他根本没有涵括任何有关人类威胁的问题。

相反，埃齐奥尼对此只问了一个问题，那就是我们何时将实现超级智能。正如波斯特罗姆的数据预测的那样，在埃齐奥尼的受访者中，超过一半的人（67.5%）选择了"至少25年"才能实现超级智能——毕竟，波斯特罗姆有超过一半的受访者给出的数据是"25年后，仅有50%的可能性达到人类层级的人工智能"。我们的同事（拉塞尔）在埃齐奥尼的调查中给出的回答是"至少25年"。而波斯特罗姆在自己的调查中写道："我本人的观点是，在较晚的实现日期方面，专家调查里的中间人群并没有足够的概率分布，所以时间上还无从推测。"

如何看待AI的潜在威胁

在设计出了让受访者可能选择"超过25年"这一选项的调查问卷后，现在埃齐奥尼又陷入了他自己的陷阱：他声称25年是一个"无法预见的将来"。所以可由此推测出，无论是拉塞尔还是波斯特罗姆，他们都不认为超级智能对人类是个威胁。这让拉塞尔和波斯特罗姆都很讶异，可能也会让调查中许多其他的受访者也感到惊讶。事实上，埃齐奥尼的文章标题本可以简单地起为《75%的专家认为走向超级智能是必然的趋势》。难道因为大部分专家认为超级智能离我们还有至少25年之远，我们就该忽视它的灾难性风险吗？按照埃齐奥尼的逻辑，我们还应该忽视气候变化的灾难性风险，顺便严惩提出这些问题的人。

75%的专家认为走向超级智能是必然的趋势，许多杰出的AI专家已经认识到AI具有威胁人类存续的可能性

还有一些人与埃齐奥尼和某些AI界人士的观点相反，他们认为AI的长期风险并不等同于表明超级智能及其伴随性风险"即将来临"。曾指出AI风险的人，包括一些杰出人物，比如阿兰·图灵、诺伯特·维纳、I.J.古德和马文·明斯基。甚至奥伦·埃齐奥尼自己也承认了这些风险。据我们所知，这些人中没有任何一个人曾断言超级智能即将到来，也还未有任何迹象表明超级智能即将来到我们的生活中。波斯特罗姆在《超级智能》中也没有任何的提及。

之后，埃齐奥尼再次重复了他那令人半信半疑的观点，"悲观的预测前景通常没有考虑到AI在预防医疗事故、减少交通事故等方面的巨大潜力"。对波斯特罗姆来说，埃齐奥尼的观点根本站不住脚。波斯特罗姆预计在控制AI方面的成功将使人类大范围地自发使用AI。埃齐奥尼的观点也十分荒谬。就像是在说分析核电站爆炸可能性的核工程师"没有考虑到廉价电力的巨大潜力"一样。因为可能某一天，核电站真的能够生产出便宜的电了，所以我们既不能提起核电站爆炸的可能性，也不能致力于解决核电站可能爆炸的问题。

切尔诺贝利事件表明，宣称某一强大技术不会引起任何风险是不明智的。宣称某一强大技术永远不会实现也是不明智的。1933年9月11日，卢瑟福勋爵（可能是世界上最杰出的核物理学家）认为通过原子裂变获得能量，简直是异想天开。然而在不到24小时之后，利奥·西拉德（Leo Szilard）便发现了中子诱发核链式反应。几年后，核反应堆与核武器的详细设计便出现了。所以说，最好是期待人类的聪明才智，而非低估它；最好是承认风险的存在，而非否认它。

许多杰出的AI专家已经认识到AI具有威胁人类存续的可能性。但与媒体报道中的失实陈述和误导说法不同，这种风险不该由自发的恶意引起，而是应来自促进AI发展过程中出现的不可预测性和潜在的不可逆性。早在1960年，诺伯特·维纳（Norbert Wiener）就已清楚地阐述过这个问题，但我们直到今天仍没能解决它。希望读者能够支持目前正在进行的努力。

（作者分别为耶鲁大学政治学系助理教授、加州大学伯克利分校计算机科学系教授）