人工智能威胁论溯因*

——技术奇点理论和对它的驳斥

李恒威　　王昊晟

提要：近些年来，史蒂芬·霍金、伊隆·马斯克、比尔·盖茨等科技意见领袖一再表达对人工智能威胁的担忧，许多从事人工智能研究的学者也多有附议。虽然以现有的路线是否能实现真正意义上的强人工智能仍存在一些根本的理论问题尚未澄清，但人工智能领域的发展可谓日新月异，一些智能程序或智能装置所展现出的炫目智力迫使人们不得不正视人有关工智能超越人类的论断和预言的可能性。2015 年，在波多黎各召开的人工智能安全会议上，诸多专家预测与人类同等水平的智能体将会在 2060 年之前诞生；而 2016 年横空出世的 AlphaGo 更是将这种预测演变为社会对于超级智能的忧虑乃至恐惧。然而，本文认为：当前这种所谓超越乃至控制和奴役人类的“人工智能威胁论”是一种理据不充分的过度忧虑，因为支持这种观点的主要理论基础或前提——即加速回报定律和作为其结果的技术奇点理论——是站不住脚的。为此，本文探讨和分析了加速回报定律和作为其结果的技术奇点理论的来龙去脉，概括了驳斥该理论的若干视角，从而形了一条釜底抽薪的反驳路线。

关键词：　人工智能威胁论　　加速回报定律　　技术奇点　　递归自我改进

DOI:10.16235/j.cnki.33-1005/c.2019.02.007

史蒂芬·霍金(Stephen Hawking)在其遗作《大问题简答》(Brief Answers to the Big Questions)的第九章“人工智能会比我们聪明吗”(Will Artificial Intelligence Outsmart Us？专题谈论了他关于人工智能，尤其是人工智能之于人类社会的看法。他认为，虽然创造人工智能是人类历史上最重大的事件之一，但如果我们不学会如何避免其带来的风险，那么人工智能可能将是人类的最后一次创造，“人工智能的出现将是人类有史以来最好或最糟糕的事情。”①“最糟糕的事情”就是他在很多场合所忧虑的“人工智能威胁”，也就是我们在《人工智能威胁与心智考古学》一文中所论述的人工智能带来的“IV 型：生存性威胁”。② 需要指出的是，霍金并非唯一一个认为人工智能会带来生存性威胁、并为之深感忧虑的人，但他的言论可以被视作这类观点的典型代表。从霍金散见于各处的陈说中，我们能够抽绎出其所说的生存性威胁意义上的人工智能威胁论的前提和基本内容。

对于霍金而言，人工智能之所以能够威胁人类，其核心的前提假设是——人脑以至于人等价于计

* 本研究得到国家社科规划基金重大项目(14ZDA029)、“中央高校基本科研业务费专项资金资助”的资助。

① Hawking. S. , *Brief answers to the big questions*. New York : Random House Large Print, 2018, p. 102.

② 李恒威、王昊晟：《人工智能威胁与心智考古学》，《西南民族大学学报(人文社科版)》2017 年第 12 期。

算机。在霍金看来,“蚯蚓的脑如何工作与计算机如何计算之间没有显著差异……因此,从原理上讲,计算机可以模拟人类智能。”[①]并且,就二者可以实现的目标而言,计算机甚至更胜一筹。

基于这一假设的前提,霍金认为,无论是从短期还是长期来看,人工智能都势必会对人类构成威胁。霍金指出:从近期看,各国军队已经开始了基于人工智能的自主武器系统(autonomous weapon system)的军备竞赛。如果这种拥有自主选择消灭目标能力的人工智能武器被研发并量产,那么一旦它们落入罪犯、恐怖分子或独裁者的手中,就势必会对公众、国防甚至全球安全构成严重威胁,这类武器系统将会成为“少数人压迫多数人的新手段”。[②] 而从中远期看,由于具有超级智能的机器可以不断地快速改进并完善自身的设计,而人类则受限于生物演化的缓慢节奏,这最终会使机器智能超越人类智能,而且它们还有可能以我们无法理解的方式征服或消灭人类。霍金在2014年12月接受BBC采访时就表示,“完整人工智能的发展可能意味着人类的终结……它会以不断增长的速度重新设计自己。而人类受制于生物演化的缓慢性,将无法与其竞争,并将会被取代。”[③]霍金的观点代表了当前诸多科技意见领袖和一些学者对于人工智能的看法,再加之媒体、影视和文学创作的渲染,人工智能似乎真的成了一种让人忧虑甚至恐惧的生存性威胁。

然而,从霍金的表述中不难看出,生存性人工智能威胁论的根源在于担忧人工智能的发展速度超过人类智能,以及随之而来的能力上的全面超越。因此本文尝试从支持该观点的主要前提和驱动力——技术奇点理论——入手,抽绎出它的立论基础,并从不同角度对其合理性进行驳斥,从而瓦解人工智能威胁的根源,由此阐明“人工智能威胁论”是一种理据不充分的论断,因此对其过度忧虑是不必要的。

一、人工智能的迷雾

回溯人工智能的发展历程,“人工智能”概念本身始终处于动荡变化之中。从1956年达特茅斯会议(Dartmouth Conference)召开伊始,人工智能既经历过辉煌的“黄金十年”,也曾遭遇过乏人问津的“AI凛冬”。近四分之三世纪的发展让人工智能领域产生了革命性变化,同时也使人们对“人工智能”概念本身的理解变得相当多样和含混。导致“人工智能”概念变得多样和含混的原因主要源自于两种转变:内涵的转变和外延的转变。

从内涵上看,“人工智能”概念与最初诞生时相比已出现了根本差异。20世纪50年代,当“人工智能”概念首次被提出时,研究者指的是通过软硬件来实现与人类智能相媲美的智能体(artificial agent)。人工智能可以分为弱人工智能(Weak artificial intelligence 或 Artificial Narrow Intelligence)与强人工智能(Strong artificial intelligence 或 Artificial General Intelligence)两类。[④] 弱人工智能是指在某些方面可以比人类更好地执行任务,而在其他任务方面存在严重缺陷或不足的人造智能。强人工智能则是指具有与人类一样的心智系统,而这也是当时研究人员力求实现的隐含目标。但在经历数十载的发展后,强人工智能仍只存在于理论假设和科幻作品之中。因此,今天的“人工智能”更多

① Hawking, S. , *Brief answers to the big questions*. New York : Random House Large Print, 2018, p. 100.

② “The best or worst thing to happen to humanity”—Stephen Hawking launches Centre for the Future of Intelligence. Retrieved from https://www.cam.ac.uk/research/news/the-best-or-worst-thing-to-happen-to-humanity-stephen-hawking-launches-centre-for-the-future-of , 2016.

③ Jones, R. , “Hawking: AI could end human race”. Retrieved from http://www.bbc.com/news/technology-30290540, 2014.

④ 事实上,“强人工智能”(strong AI)与“通用人工智能”(AGI)的含义是有所差别的。前者强调人工智能非功能主义或非模拟意义上的全同性,即具有与人类同样的感受、思维和意识等能力;后者侧重人工智能在功能上的全面性,即人工智能能模拟解决人类智能所能解决的任何问题,而不是某一类特定问题。

地转向了各个领域内相对“有限智能”的开发。正如伯克利大学教授迈克尔·乔丹(Michael Jordan)所言,“当前大部分所谓的‘人工智能’,尤其是在公众领域,实际上是指‘机器学习’”。[①] 按照亚瑟·塞缪尔(Arthur Samuel)的定义,机器学习是指“使计算机拥有在没有被明确编程的情况下学习的能力。”[②]这种学习是利用算法对数据进行分析加工,进而获得某种结果并以此进行推测或者判断。从“强人工智能”到“机器学习”,代表着人工智能领域研究企图和方向的转变。但遗憾的是,大众并没有区分这些概念,甚至在使用时将它们混为一谈,这无疑大大增加了人们在理解人工智能究竟能达到什么程度时的不确定性。从外延上看,“人工智能”概念在某种意义上是自相矛盾的。这种矛盾被称作“人工智能效应”(AI effect),即只要某个问题被人工智能成功解决,那么该问题就不再是人工智能的一部分。帕梅拉·麦考克(Pamela McCorduck)将其称作“奇怪的悖论”,她指出,“人工智能一旦成为实际上实现了具有某种智能行为的计算程序,它就很快被其他应用领域所吸收……人工智能的研究人员只负责处理‘失败’,即那些尚未被攻克的难题”。[③] Deep Blue 就是一个很好的例证。1997 年,IBM 的国际象棋程序 Deep Blue 成功击败国际象棋大师卡斯帕罗夫(Kasparov)。之后,当人们认识到该程序是用“暴力穷举法”实现这一点时,便批评它实际上并没有表现出“智能”。这也使得人工智能支持者经常面对这样一个问题,“当我们知道机器如何做一些‘聪明的事情’时,它就不再被认为是聪明的”。[④] 侯道仁(Douglas Hofstadter)也曾简洁地描述过人工智能效应:“人工智能是任何尚未完成的事情”。[⑤]

这两种转变导致人们既无法清晰地界定“人工智能”的内涵,也无法明确究竟有哪些应用属于“人工智能”,因此,无论是理论层面还是应用层面,人工智能都颇如一团迷雾。

另一个导致“人工智能”充满误导性的因素是企业、媒体以及文学作品。它们因其内在的行为方式或各自利益的考虑,都或多或少地夸大或扭曲了人工智能的实际能力,并想当然地赋予人工智能一些未经论证的特征和属性。由于无法清晰准确地理解人工智能的实现机理,加之人工智能带来的道德规范、安全威胁、失业、社会不平等之类的问题,这更加剧了公众对人工智能的某种负面看法,甚至使公众也产生了对“人工智能威胁论”的惶惑、忧虑乃至恐惧。

二、人工智能威胁论的根源:从加速回报定律到技术奇点理论

无论人工智能威胁的表现形式如何,当前人们对人工智能表现出紧张情绪的关键在于担忧人工智能的发展速度超过人类自身,甚至以人类无法理解的速度和方式继续发展,并最终取代人类。支持这种观点的主要驱动力是加速回报定律(The Law of Accelerating Returns)和作为其结果的技术奇点理论。

1. 加速回报定律

雷·库兹韦尔(Ray Kurzweil)提出的加速回报定律源自于半导体行业的摩尔定律(Moore's Law)。1965 年,英特尔公司创始人之一戈登·摩尔(Gordon Moore)提出了摩尔定律的原始版本,即

① Jordan, M., "Artificial Intelligence — The Revolution Hasn't Happened Yet". Retrieved from https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7, 2018.

② Samuel, A., "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal Of Research And Development, 3(3), 1959, pp. 210-229.

③ McCorduck, P., *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. Natick, Mass: A. K. Peters, 2004, p. 423.

④ "Promise of AI not so bright", Retrieved from https://www.washingtontimes.com/news/2006/apr/13/20060413-105217-7645r/, 2006.

⑤ Hofstadter, D. R., *Godel, Escher, Bach: An Eternal Golden Braid*. Vintage. 1989, p. 601.

半导体芯片上集成的晶体管和电阻数量将每年增加一倍，并预计这种增长速度将会持续至少十年。[①] 1975年，摩尔在IEEE国际电子组件大会上修正了摩尔定律，将“每年增加一倍”修改为“每两年增加一倍”。[②] 时任英特尔首席执行官的大卫·豪斯(David House)则以此预测，这些变化将导致计算机性能每18个月翻一番。

需要指出的是，摩尔定律是通过对历史数据进行观察和分析后的预测，并非物理或自然规律。但是从摩尔定律提出后半导体行业发展的实际情况来看，从20世纪70年代至今，该定律的预测大致都是正确的，并且该定律也被半导体行业作为指导长期规划和设定研究开发目标的参照。

库兹韦尔的“加速回报定律”则彻底改变了“摩尔定律”的适用范围，它将摩尔定律应用于所有形式的技术预测。1999年，他在《精神机器时代——当计算机超越人类智能》(The age of spiritual machines: When computers exceed human intelligence)一书中首次提出了“加速回报”(Accelerated Return)的概念，认为系统演化的变化速率(包括但不仅限于技术)呈现指数增长。[③] 2001年，他正式提出了摩尔定律的扩展定律，即“加速回报定律”[④]：

正反馈适用于演化，因为演化过程中某个阶段产生的更有力的方法将被用于创造下一个阶段。因此，随着时间的推移，演化过程的速率呈指数增长。并且，嵌入在演化过程中的信息的“秩序”(order)(例如，衡量信息如何与目的相匹配)同样随着时间增长。与上述观察相关的是，一个演化过程的“回报”(例如，速度、成本—效益或者某个过程的整体“力量”)亦随着时间的推移呈指数增长。在另一种正反馈回路中，随着某种特定的演化过程(例如，计算)变得更加有效(例如，成本效益)，更多的资源将会倾向于这种过程，促进其进一步发展。这将导致次级的指数增长(即指数增长率本身也呈指数增长)。生物演化就是这样一种演化过程。技术演化同样是这样一种演化过程。事实上，第一种技术创造物种的出现导致了技术的全新演化过程。因此，技术演化是生物演化的产物，同时也是生物演化的延续。某种特定范式将持续指数增长，直到该方法耗尽其潜力。当这种情况发生时，范式将发生转变，从而继续维持指数增长。[⑤]

库兹韦尔认为，基于错误的“直觉的线性观”(Intuitive Linear View)，而不是“历史的指数观”(Historical Exponential View)，我们对技术进步的理解出现了严重偏差。人们会直观地假设当前的技术发展速度会在未来的一段时间内持续下去，并以此推断未来10年或100年的技术进步。而事实上，技术变革是指数级的。并且，这种增长不仅仅是简单的指数增长，而是“双重”指数增长，也即是说，指数增长率本身也呈指数增长。库兹韦尔指出，这种指数增长不仅仅发生在适用于摩尔定律的半导体领域，而是发生在从电子技术到生物医学的各种技术领域。[⑥]

库兹韦尔强调，单一的技术无法维持这样的指数增长。当呈指数增长的技术接近某种障碍或瓶颈时，将会出现一种新的技术跨越这种障碍或瓶颈，从而保证整体上仍维持指数增长。这种范式转换的技术越迁在计算机领域得到了印证。从打卡式机械计算机、电磁继电式计算机，到真空管、晶体管计算机，再到早期集成电路计算机，以至现代超大规模集成电路计算机，每当某种计算机发展到达瓶颈时，总会有新的计算机设计与技术出现来保证这种指数级增长维持下去。

2. 技术奇点理论

技术的加速回报将带来技术奇点(Technological Singularity)。所谓技术奇点，简而言之是指能

① Moore, G. E., *Cramming more components onto integrated circuits*. New York: McGraw-Hill, 1965.
② Moore, G. E., "Progress in digital integrated electronics". *SPIE Milestone Series*, 178, 2004, pp. 179-181.
③ Kurzweil, R., *The age of spiritual machines: When computers exceed human intelligence*. New York: Viking, 1999.
④⑤ Kurzweil, R., "The Law of Accelerating Returns". Retrieved from http://www.kurzweilai.net/the-law-of-accelerating-returns, 2001.
⑥ 雷·库兹韦尔：《人工智能的未来：揭示人类思维的奥秘》，2016年，第243、245页。

够进行自我改进的人造智能体超越人类智能的时刻。这一概念可以追溯到人工智能发展的初期。1958年,斯坦尼斯瓦夫·乌拉姆(Stanislaw Ulam)在纪念冯·诺依曼的文中写道,"在(与冯·诺依曼)一次谈话中,我们集中讨论了技术的不断加速发展与人类生活方式的变化……这在人类历史中接近了一些关键的奇点,正如我们所知,人类事务将无法继续下去。"①

古德(I. G. Good)在1965年提出了"智能爆炸"(intelligence explosion)的观点,用更为具体的方式对"奇点"进行了描述,"我们将超智能机器定义为这样一种机器,它可以远超任何人类的任何智能活动。由于机器设计本身就是智能活动之一,那么一台超智能机器将可以设计更好的机器;毫无疑问,这将会是一次智能爆炸,人类智能将远远落后。因此,第一台超智能机器将会是人类的最后一项发明。"②

1986年,弗诺·文奇(Vernor Vinge)在其科幻小说《实时放逐》中首次描写了一种快速临近的技术"奇点"。1993年,文奇在NASA组织的"Vision-21"研讨会上撰文预测,"在三十年之内,我们将会具有创造超人智能的技术手段。随之,人类时代将会结束。"③

2006年,库兹韦尔在《奇点临近》一书中将"技术奇点"阐发为"技术奇点理论",他认为,依据加速回报定律,技术的范式转换将会变得越来越普遍,"对技术史的分析表明,技术变革是指数性的,与常识性的'直觉的线性观'相反。所以我们在21世纪将不会经历100年的进步——它将更像是2万年的进步(以今天的速度)。芯片速度和成本效益之类的'回报'也呈指数增长……几十年内,机器智能将超越人类智能,并导致技术奇点的来临——技术变化如此迅速而深刻,代表了人类历史结构的破裂。其含义包括了生物和非生物智能的合并,基于软件的不朽人类,以及以光速在宇宙中向外扩张的超高水平智能。"④根据库兹韦尔的观点,这种技术奇点将在21世纪中期,大约2045年左右发生。

基于技术奇点理论,许多评论家甚至科技界人士开始对人工智能可能带来的威胁表示担忧。除却前文所述的霍金之外,马斯克、盖茨等人也在公开场合表达过对人工智能相关问题的忧虑。诸多意见领袖的警告或者预言显然触动了大众恐惧的神经,人工智能威胁论逐渐上升为一种生存危机,被视作对人类的发展甚至存在构成了实质性的威胁。

三、对技术奇点理论的驳斥

事实上,技术奇点理论引发的人工智能威胁论的支持者大多来自于主流人工智能和计算机领域之外。即使不乏科技人士和意见领袖,但这些人工智能威胁论的拥趸通常默认了技术奇点理论的合理性和可实现性,认为技术确实将会以超越人类理解能力的速度进步,不能预警且无法避免。在这样的前提下,人工智能威胁的到来似乎已成定局,人们的担忧也看似符合情理。然而,当下的讨论大多忽视或回避了一个基本而重要的核心问题,即我们是否如技术奇点理论所言的那样,能设计并开发一种维持智能以指数方式快速增长,从而远超人类智能并威胁人类生存的机器。如果技术奇点理论这一前提不能成立,那么建立在其基础之上的人工智能威胁也将随之消弭。近年来,人工智能学界和其他领域学者对于技术奇点理论的核心论断提出了诸多批评和反驳。总结来看,以技术奇点理论的两

① Ulam, S., "John von Neumann 1903-1957". Bulletin of the American Mathematical Society, 64(3), 1958, pp. 1-50.

② Good, I. J., "Speculations concerning the first ultraintelligent machine". In F. Alt & M. Ruminoff (eds.), *Advances in Computers*, *volume* 6. Academic Press, 1965.

③ Vinge, V., "The coming technological singularity: How to survive in the post-human era". In Rheingold, H. ed., *Whole Earth Review*, 1993.

④ Kurzweil, R., "The Law of Accelerating Returns", Retrieved from http://www.kurzweilai.net/the-law-of-accelerating-returns, 2001.

个核心要素——技术进步与智能为出发点，这些批评和反驳可以分为以下几类：(1)否认技术进步与智能提升的相关性或二者仅弱相关；(2)承认技术进步与智能提升的相关性，否认后者可以满足加速回报定律；(3)承认技术进步与智能提升的相关性，否认前者可以满足加速回报定律；(4)承认技术进步与智能提升的相关性，否认智能可以形成威胁。这些批评和反驳不仅在于瓦解了技术奇点理论所推崇的指数增长假设，同时意在表明纯粹计算速度的增长并不能带来智能的提升，更无法让人工智能具备主动威胁人类的动机和意图。①

1. 计算速度与智能

托比·沃尔什(Toby Walsh)认为，支撑技术奇点理论的一个重要论点是，作为硬件的硅相较于人类脑的湿件(wetware)②具有显著的速度优势。并且，根据摩尔定律，这种优势随着时间的推移将呈指数增长。但在沃尔什看来，技术奇点论者忽略了十分重要的一点：计算速度的提升并不等同于智能的提升。沃尔什指出，“技术奇点论者最大的问题在于混淆了执行任务的能力与提升‘执行任务的能力’的能力之间的区别”，③前者对应于计算速度，而后者则是指智能。

如果只是单纯的提升计算速度，按照文奇的观点，这只不过类似于一只“快速思考”(fast thinking)的狗，“想象一下，如果狗的脑以非常高的速度运行，那么一千年的狗的经历是否能够比拟人类的洞见？”④文奇认为显然是不可能的，无论狗的思维有多么快速，它永远不会懂得下棋。这种计算速度的提升只不过是一种“弱”智能提升，与到达技术奇点所需要的全方位的“强”智能提升相去甚远。也正如史蒂芬·平克(Steven Pinker)所说，“我们没有任何理由相信奇点将会到来。你可以想象一种未来，但它并不一定具备实现的可能性。当我还是一个孩子时，人们就想象过圆顶城市、喷气式通勤工具、水下城市、超高建筑、核动力汽车等，这些未来式的幻想至今尚未实现。单纯的加工能力并非一种魔法尘埃，它无法神奇地解决你所有的问题。”⑤

对于智能的提升，大卫·查莫斯(David Chalmers)曾进行过如下阐述，“如果我们通过机器学习创造了AI，那么我们很可能可以改进学习算法并延长学习过程，进而创造AI+。”⑥查莫斯通过逻辑和数学方法推理认为，如果AI_0系统能够产生比起本身能力更强(哪怕只有很微小的提升)的AI1系统，那么经过n次迭代后的AIn系统就可以实现超级智能。在查莫斯的推理论证中，智能的提升以至超级智能的实现都取决于两个关键点，一是学习过程的延长，二是学习算法的改进。就目前技术而言，这两个问题的解决前者依赖于提升硬件处理能力，后者需要实现机器学习的自动化。但就实现超级智能这一目标而言，两者面临着各自的困难。因为延长学习过程并非是真正意义上智能的提升。基于深度学习算法的人工智能系统在近些年取得了令人瞩目的成就，在语音识别、计算机视觉、推理、自然语言处理等领域实现了突破性进展。但这些进步依赖于更大的数据和更深层次的神经网络，正如燕乐存(Yann LeCun)所指出，神经网络之所以能够打破连续语音识别的记录，只是因为它们变得足够大而已。⑦ 由此可见，在语音、图像等领域的进展并没有实现深度学习算法的改善和真正智能的提升，这只是硬件提升和数据量增大带来的规模效应而已。由于这些进步不是来自于对智能机制的理

① 李恒威、王昊晟：《人工智能威胁与心智考古学》，《西南民族大学学报(人文社科版)》2017年第12期。

② 湿件(wetware)，用于描述人体中与计算机的硬件(hardware)和软件(software)相对应的要素，尤其是指中枢神经系统和心智。之所以为“湿”(wet)，是因为与计算机软硬件需要“干燥”的环境不同，生物体中包含了大量水。这显然是一个对比的隐喻说法。

③ Walsh, T., “The Singularity May Never Be Near”. *AI Magazine*, 38(3), 2017, pp. 58－62.

④ Vinge, V., “The coming technological singularity: How to survive in the post－human era”. ? In Rheingold, H. ed., *Whole Earth Review*, 1993.

⑤ Pinker, S., “Tech luminaries address singularity”. *IEEE Spectrum*, 2008.

⑥ Chalmers, D. J., “The singularity: A philosophical analysis”. *Journal of Consciousness Studies*, 17, 2011.

⑦ Edwards, C., “Growing pains for deep learning”. *Communications of the ACM*, 58(7), 2015, pp. 14－16.

解，只是能力更强大的芯片和更加丰富的数据，因此它们总是会存在相对的极限。计算机科学家拉米兹·那姆(Ramez Naam)提出这样的思想假设，“想象你是一个运行在某种位处理器上、拥有超级智能的 AI。突然，你想要设计一个更快、更强大的处理器应用于你的运行……你现在必须生产这种处理器。而生产工厂的运转需要巨大的能量，它需要从周围环境中获取原材料，需要严格控制内部环境……所有这些工作都要花费时间和能量。现实世界是你螺旋式自我迭代的最大障碍。”①

亚姆波斯基(Roam Yampolskiy)认为，虽然几乎所有系统的性能都可以通过分配更多的内存、更快的处理器或更强大的网络来进行提升，但这种线性缩放显然与指数型增长不相符合。同时，单纯的硬件叠加并不能使系统更好地改进自身，“一般而言，硬件提升可能会使系统加速，但软件提升(新算法)才是实现元提升(meta－improvements)所必需的。”②鉴于算法对于智能提升的必要性，人们设想通过学习的自动化来实现算法的改进，即用一个算法指导另一个算法进行学习，让算法不断迭代升级，实现一种递归式的自我改进(Recursively Self－improving)。但是，这种自我改进在实现上却面临着通用性的难题。举例而言，AlphaGo 的终极版本 AlphaGo Zero 正是基于这样设想下的产物，它不依赖于人类棋谱和知识，仅仅凭借最简单的围棋规则，通过短时间(大约 40 天)自我学习就可以战胜人类顶尖棋手以及 AlphaGo 的其他任何版本。虽然 AlphaGo Zero 在学习自动化方面的结果令人欣喜，但需要指出的是，无论围棋看上去如何复杂，它始终是一种具有固定规则和有限可能构型的游戏，而真正的学习自动化所面临的开放式的领域和问题往往并不具有像围棋这样清晰的规则和条件。因此，通过 AlphaGo Zero 获得的学习自动化成果想要推广到更广泛的领域，实现由特殊向普遍的扩展仍存在着目前难以跨越的障碍。

2. 线性输出与收益递减

技术奇点理论预设了智力改进的速率是一个相对固定的乘数，从而会导致人工智能每一代都比上一代拥有显著的提升，使智能整体呈现指数增长，并最终导致“智能爆炸”。在这一点上，技术奇点理论支持者最有力的证据来自于计算机计算能力的增长。在过去的几十年中，计算机的计算能力基本保持着每 18 个月倍增的速率，这使得当前的计算机与最初相比，计算能力获得了大约百亿倍的提升。但正如前文所述，这种计算能力的提升只不过是一种基础输入能力(input)的指数增长，而作为最终结果的智能输出(output)仍呈现出线性增长的趋势。

例如，根据 20 世纪 80 年代至今计算机国际象棋的表现来看，虽然在最开始的一段时间中，计算机的 ELO 等级分③增长迅速，但就 30 余年的整体趋势而言，计算机的等级提升呈现出接近线性的趋势。换言之，计算机国际象棋输入能力的指数改进只带来了计算机国际象棋输出水平的线性改善。

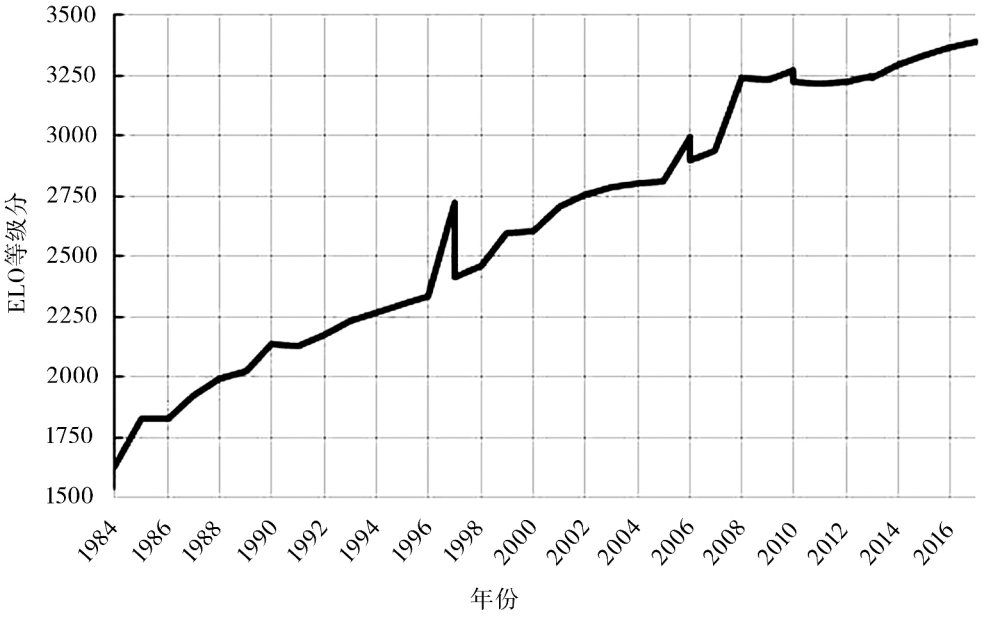
深度学习框架“Keras”的创建者弗朗索瓦·乔乐特(François Chollet)进一步阐明了这种指数输入与线性输出的关系。他认为，当前人工智能威胁论产生的技术前提智能爆炸根本不会发生，人工智能的发展是线性的，而非指数的。在乔乐特看来，智能是基于环境的，人脑是更广泛系统的一部分，后者包括人的身体、其所处的环境、他人以及整个环境本身。由于任何单独的智能体都受其所处环境的限定或限制，因此，人类智能发展并非单独受限于人脑，我们所在的环境才是智能发展的瓶颈。智能的进步来源于(生物的或数字的)脑、感知运动功能、环境和文化的共同演化，并不能通过简单地调整缸中脑的几个齿轮来独立完成。这种共同演化已经经过了无数时间，并将会随着智能转向数字化而

① Naam, R., “Top Five Reasons 'The Singularity' Is A Misnomer”. Retrieved from http://hplusmagazine.com/2010/11/11/top－five－reasons－singularity－misnomer/, 2010.

② Yampolskiy, R., “The Singularity May Be Near”. *Information*, 9(8), 2018, p. 190.

③ ELO，衡量各类对弈活动水平的评价方法，以其创始人阿帕德·埃洛(Arpad Elo)名字命名。它是一种以数值表示的评级系统，将等级差别转化为分数或取胜概率。

继续。“智能爆炸”不会发生，因为这个过程将以接近线性的速度前进。



计算机国际象棋ELO等级分①

由于输入的指数增长与输出的线性提升间的不对称关系，各种类型的人工智能系统在数十年来的实际发展中都经历着收益递减的过程。在研究初期，人工智能系统通常可以快速提升，甚至在某些时刻超越技术奇点理论所设想的指数增长速度，但随着完善度和复杂度的增加，人工智能系统往往会遭遇各类难以改进或跨越的瓶颈，导致无法维持固定的改进速率。

微软的联合创始人保罗·艾伦(Paul Allen)将这一现象称作“复杂性刹车”(complexity brake)，“随着我们对自然系统不断加深的理解，我们通常会发现，我们需要更多且更加专业的知识对其进行描述，并且不得不用愈加复杂的方法来持续扩展我们的科学理论……我们相信，我们对自然的理解正由于复杂性刹车而减慢。”②

乔乐特通过类比物理学、数学、医学等其他科学的发展速度指出，即使是在不断投入更多的科研人员并使用更快计算速度的计算机的情况下，上述各科学领域的发展远远达不到技术奇点理论所设想的指数增长，甚至无法维持稳定的线性提升。乔乐特认为，限制科学提升速度的主要原因包括：首先，既定领域内科学研究的难度会随着时间的推移而增加，任意领域的开创者都可以取得突破性的成果，而后续研究者想要实现相同的进展就需要付出成倍的努力；其次，研究人员之间的合作共享随着该领域研究的扩张而呈几何式增长，其结果导致前沿领域的论文发表愈发困难；再次，随着人类科学知识体系的不断增长，我们在教育培训上投入的时间和精力不断增加，对于每个研究者而言，可以探索的方向却变得越来越狭窄。③

3. 递归自我改进系统的限制

通常情况下，我们可以将人类自身视作一种递归自我改进系统(Recursively Self－improving

① Eckersley, P., “AI Progress Measurement”. Retrieved from https://www.eff.org/ai/metrics#Abstract－Strategy－Games, 2017.

② Allen, P., “Paul Allen: The Singularity Isn't Near”. Retrieved from https://www.technologyreview.com/s/425733/paul－allen－the－singularity－isnt－near/, 2011.

③ Chollet, F., “The impossibility of intelligence explosion”. Retrieved from https://medium.com/@francois.chollet/the－impossibility－of－intelligence－explosion－5be4a9eda6ec, 2017.

System)，即人们可以通过学习获得知识，再通过已有知识改进学习效率，如此循环，使人们不断变得更加聪明。这一点正如庄子所言："以其知之所知以养其知之所不知"。[①] 于是，从人工智能领域的最早期开始，就有研究者希望设计出一种自我改进的智能系统，以实现真正的人工智能。图灵(Alan Turing)在《计算机器与智能》(Computing Machinery and Intelligence)中就曾写道，"与其尝试编程模拟成年人的脑，为何不尝试模拟儿童的脑？只要儿童的脑接受适当的教育，那么就可以获得成年人的脑。儿童的脑大致就像刚刚买到的笔记本，只包含简单的机制和许多白纸。我们希望儿童的脑中有足够少的机制，以至于我们可以容易地编程。我们可以假设对机器进行教育的工作量与教育人类儿童的工作量大致相同。"[②]奇点支持论者认为，递归自我改进系统是实现技术奇点的有效途径。在上文中，我们已经述及这种自我改进系统在实现上存在的通用性难题。从更根本上而言，递归自我改进系统本身也存在着相当多的限制，通过这条路径实现技术奇点困难重重。

亚姆波斯基具体描述了这种限制，"任何软件系统的实现都依赖于存储、通信和信息处理等硬件，即使我们假设存在一种非冯·诺依曼(量子)结构可以运行这一软件。这就给计算带来了严格的理论限制，尽管摩尔定律预言了硬件的进步，但未来任何的硬件范式都无法克服这一限制。"[③]

就现实而言，当下人工智能系统的物理实现都依赖于冯·诺伊曼结构，超越该架构的量子结构尚有待进一步完善。而在冯·诺伊曼结构下，人工智能则面临着更多的限制。为了解决计算机的运行可靠性和运行效率，冯·诺依曼提出了将存储设备与中央处理器相分离的概念，而正是这一点，导致了冯·诺伊曼瓶颈(von Neumann bottleneck)出现：CPU 与存储器之间的数据传输能力远远小于存储器的容量，同时也远远小于 CPU 的计算能力，因此，数据传输能力在某些情况下就严重限制了计算机整体的效率。CPU 在数据输入或输出存储器时会处于闲置状态。同时，在技术改进层面，CPU 的提升要远远快于存储器的优化，这样一来，瓶颈问题随着时间的推移愈加严重。冯·诺伊曼瓶颈不仅限制着计算机效率的提升，同样掣肘着人工智能的发展极限。

除却硬件上的限制，递归自我改进系统在自我指涉(self－reference)方面也存在着严峻的挑战。亚姆波斯基认为，"很有可能的是，成为一个递归自我改进系统所需要的最低复杂度高于系统自身能够理解的最高复杂度。"[④]这种情况经常出现在较低等级智能水平的生物上。例如，一只松鼠并不具有理解其自身的脑如何运作的心智能力。此外，即使是一个能够进行完全自我分析的系统，随着自我改进的不断发生，该系统也可能会丧失这种自我分析能力。因为，系统复杂程度的不断提升，将导致理解自身所需要的智能也不断提升，并且前者者提升的速度可能比后者更快。

亚姆波斯基还提出，在系统递归自我改进的过程中会累积错误(errors)，这种错误类似于生物演化过程中的突变(mutations)。最初，这些错误(突变)对于系统(生物体)没有损害，并且难以被检测(察觉)。但随着其数量的不断累积，当这些错误到达临界值时，就会导致系统的运行出现错误，从而影响递归自我改进的质量，甚至导致整个系统完全崩溃。

4. 智能的工具性

对于许多技术奇点论者而言，"智能"是至关重要甚至唯一的标尺。人类智能被视作这个标尺上的一个里程碑式的节点，一旦人工智能跨越了这个节点，技术奇点就将到来，威胁也随之降临。正如博斯特罗姆(Nick Bostrom)所说，"人类级别的人工智能很快就会产生比人类更高级的智能……人类

① 庄子：《庄子·大宗师》，方勇译注，中华书局，2010 年，第 94 页。

② Turing, A., "Computing Machinery and Intelligence". Mind, LIX(236), 1950, p. 457.

③④ Yampolskiy, R., "From Seed AI to Technological Singularity via Recursively Self－Improving Software". CoRR, abs/1502.06512. 2015.

与机器智能相匹敌的时间可能很短暂。此后不久,人类将无法在智能上与人工智能相竞争。"[①]这种对智能的盲目推崇使技术起点论的支持者们错误地将智能提升等同于产生威胁。事实上,形成"威胁"的关键所在恰恰是智能之外的其他要素,即情绪、感受所表达的价值倾向。与人类迄今发明的所有解放人类体力和智力的装置一样,智能装置的本质同样是工具性的,它不具有内在的价值规范性,因此它不可能自主地(autonomously)具有价值表达和价值行为——譬如,意在威胁、奴役或灭绝人类。

杰夫·霍金斯(Jeffrey Hawkins)指出,"人们之所以有这样的担忧是因为他们将智能(即新皮层的算法)与古脑的情感因素(诸如恐惧、多疑和欲望等)归并起来了。智能机器是不具备这些能力的。它们既不会有野心和渴望财富,也不会寻求社会认同以及性满足。它们没有食欲、嗜好,也不会出现情绪不稳定的情况。智能机器不会有任何类似人类情感的东西,除非我们刻意把他们设计成那样。"[②]也就是说,实现智能并不等于同时赋予它以意图、动机、情感等价值成分和规范成分。事实上,智力或智能在生物的演化中始终是服务于生命内在的生物价值(biological value)。价值才是生存和演化的方向,才是驱动力。"生存的概念以及引申出来的生物价值的概念可适用于各种生物体,从分子和基因到整个有机体。"[③]平克同样认为,"人们对AI的误解在于,混淆了智能与动机——也即是对于欲望的感受、对于目标的追求、对于需求的满足——之间的区别。"[④]在平克看来,"智能是一种利用新工具来实现目标的能力,但目标本身是与智能无关的:聪明与想要获得某个东西并非是同一件事。"[⑤]

四、结 语

从上述的论证来看,技术奇点论在理论和实践上都存在严重的困难和局限,因此基于此推导出的人工智能威胁论也缺乏坚实可靠的根基。在现实中,人工智能学科当前的关注重点也往往是追求在特定问题上取得进步或突破,大多人工智能学者在其实际的研究工作中并非以强人工智能为目标开展其人工智能研究的。人工智能在通往"超级智能"的路途上尚存在诸多理论和实践上的鸿沟。综上而言,对由技术奇点引发的人工智能威胁论是一种有点过度的情绪性反应。当然,这并不意味着基于其他路线的人工智能不具有最终达到甚至超越人类智力水平的可能性。对于这一点,我们始终持一种开放的态度。

着眼当下,我们必须认识到,由于人工智能仍处在快速发展时期,出现种种问题和困难符合新事物的发展规律。因此,关注人工智能在经济、政治、文化、军事等领域内的发展,探讨其衍生出的法律、伦理问题,才是真正解决人工智能潜在"威胁"[⑥]的有效途径。我们始终坚信,人工智能的发展对于人类的自我提升,对人类文明的演进,都大有裨益。

〔作者李恒威,浙江大学哲学系、语言与认知研究中心教授;王昊晟,浙江大学哲学系、语言与认知研究中心博士研究生。杭州 310028〕

责任编辑:张东锋

① Bostrom, N., "When machines outsmart humans", Futures, 35(7), 2003, pp. 759－764.

② 杰夫·霍金斯:《人工智能的未来》,2016年,第224页。

③ Damasio, A., *Self Comes to Mind: Constructing the Conscious Brain*, Vitage Books. 2010, p. 45.

④⑤ Pinker, S., "We're told to fear robots. But why do we think they'll turn on us?", Retrieved from https://www.popsci.com/robot－uprising－enlightenment－now, 2018.

⑥ 即那些我们在《人工智能威胁与心智考古学》一文中提出的"工具性威胁""适应性威胁"和"观念性威胁"。参见李恒威、王昊晟:《人工智能威胁与心智考古学》,《西南民族大学学报(人文社科版)》2017年第12期。

● 本期新锐

李恒威，浙江大学哲学系、语言与认知研究中心教授、博士生导师；主要研究方向为认知科学哲学、意识科学、东方心学。出版专著 2 部、编著 2 部、译著 10 余部，在《中国社会科学》《哲学研究》《心理科学》《自然辩证法通讯》《自然辩证法研究》、*Constructivist Foundations*，*Open Journal of Philosophy*，*Frontiers of Philosophy in China* 等期刊发表论文 60 余篇，主持国家社会科学基金 2 项，主持国家社科基金重大项目子课题 2 项，主持教育部重大攻关项目子课题 1 项，2009 年获教育部“高等学校科学研究优秀成果奖”二等奖，2016 年入选“浙江省 151 人才工程”第一层次。

严俊，博士毕业于北京大学社会学系（中法联合培养博士生），2013–2014 年度美国杜克大学亚太研究中心（APSI）访问学者，目前就职于上海大学社会学院。他的主要研究兴趣包括经济社会学理论、艺术社会学、宗教社会学，已在《社会学研究》《社会学评论》《中国第三部门研究》《中国农业大学学报（社科版）》《上海财经大学学报》《南京农业大学学报（社科版）》发表多篇论文，在法国出版学术专著一部（Travail et migration : jeunesses chinoises à Shanghai et Paris, EDITIONS DE L′AUBE, 2017. 第二作者），并有一篇法语论文收入中法经济社会学文集（Sociologies économiques française et chinoise : regards croisés, ENS ÉDITIONS, 2014）。