关于强人工智能

周志华
南京大学

关键词：强人工智能　科技伦理

近来“人工智能”很热，关于“强人工智能”的一些讨论也见诸于网媒报端。作为一名人工智能研究者，在此谈点粗浅的看法仅供大家批评。

关于人工智能，长期存在两种不同的目标或者理念。一种是希望借鉴人类的智能行为，研制出更好的工具以减轻人类智力劳动，一般称为“弱人工智能”，类似于“高级仿生学”。另一种是希望研制出达到甚至超越人类智慧水平的人造物，具有心智和意识、能根据自己的意图开展行动，一般称为“强人工智能”，实则可谓“人造智能”。

人工智能技术现在所取得的进展和成功，是缘于“弱人工智能”而不是“强人工智能”的研究。正如国际人工智能联合会前主席、牛津大学计算机系主任迈克尔·伍德里奇 (Michael Wooldrige) 教授在 2016 年 CCF-GAIR 大会[1]报告中所说：强人工智能“几乎没有进展”，甚至“几乎没有严肃的活动”（“little progress, little serious activity”）。事实上，人工智能国际主流学界所持的目标是弱人工智能，也少有人致力于强人工智能。那么，这是不是因为强人工智能“太难”，所以大家“退而求其次”呢？不然。事实上，绝大多数人工智能研究者认为，不能做、不该做！

首先，从技术上来说，主流人工智能学界的努力从来就不是朝向强人工智能，现有技术的发展也不会自动地使强人工智能成为可能。

不妨看看现在人工智能技术所取得的成功。在图像识别、语音识别方面，机器已经达到甚至超过了普通人类的水平；在机器翻译方面，便携的实时翻译器已成为现实；在自动推理方面，机器很早就能进行定理自动证明；在棋类游戏方面，机器已经打败了最顶尖的人类棋手……可以看出，上述成功有一个共同的特点：它们都是在考虑某种特定类型的智能行为，而不是“完全智能”行为[2]。一方面，聚焦在特定类型的智能行为上，才使得任务成为可能而非空谈[3]；另一方面，如果目标是制造“工具”，那么考虑特定类型的智能行为就已足够，自主心智、独立意识、甚至情感[4]之类的东西，根本无须考虑。打个未必恰当的比方，如果人们的目标是造个工具砸东西，那么造出锤子来就好了，无须考虑让锤子有心智、意识，也不必考虑是否要让锤子自己感觉到“疼”。事实上，

[1] 全球人工智能与机器人峰会，是由中国计算机学会 (CCF) 主办，雷锋网承办的。此会于 2016 年 8 月 12~13 日在深圳举办。

[2] 人类始终在不断努力制造出在某个特定方面超越人类自身能力的工具，例如潜艇比人游得深、火箭比人飞得高，但似乎罕有人努力制造既是潜艇又是火箭的工具。类似地，人工智能研究也是在努力制造出在某种智能行为方面超越人类自身的工具。

[3] 事实上，“图灵测试”所考虑的也仅是机器能否“思考”(thinking)，而不是强人工智能语境下的“完全智能”。

[4] 人工智能中有关于“情感计算”的研究，但并非是研究如何让人造物“拥有情感”。

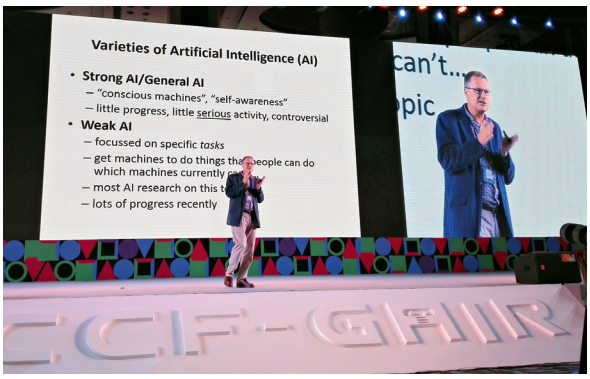
人工智能研究活跃的子领域，都是与制造智能“工具”直接相关的；而对“强人工智能”必不可少、却与“工具”不太相关的内容，如自主心智、独立意识、机器情感之类，罕有严肃的研究。所以，现有技术即便发展再快、发展再好，也不会直接使得强人工智能成为可能。

第二，即便想研究强人工智能，也不知道路在何方。

有一种说法，认为如果能够模拟出“人脑”，把其中的神经元、神经突触等全部同规模地仿制出来，那么强人工智能就会自然产生。然而，这种说法从来没有得到过一点点证明，严格说来甚至不能称其为“猜想”，因为猜想也应该有一些即便不够完备但尚能显示可能性的证据，例如通过仿制简单细胞，做出了单细胞智能生物。实际上，我们完全有更强烈的理由认为，即便能精确地观察和仿制出神经细胞的行为，也无法还原产生出智能行为。正如国际人工智能终身成就奖得主、多伦多大学赫克托·莱韦斯克 (Hector J. Levesque) 教授在他 2017 年的新著[5]中所说，即便在最理想的情况下，神经科学家也仅是能获得“目标代码”而已，没有理由认为获得了目标代码就能还原出源代码，因为这样的“反向工程”即便对软件程序来说也几乎是不可能的，更何况神经细胞内部还存在“分布式表示”[6]。

第三，即便强人工智能是可能的，也不应该去研究它。

任何一个科学研究领域或许都存在一些不该去触碰的东西。例如克隆人是被主流生命科学界所禁止的。强人工智能的造物将具有自主心智、独立意识，那么，它凭什么能“甘心”为人类服务、被人



牛津大学Michael Wooldrige教授

类“奴役”？有人把阿西莫夫的“机器人三定律”[7]奉为圭臬，但事实上这是行不通的。且不论三定律自身的矛盾和漏洞[8]，凭什么以为有自主心智和独立意识，且智能全面达到甚至超越人类水平的机器，就不会把这些约束改掉呢？即便它是善意的，人类又凭什么认为它会同意比它“愚蠢”的人类的判断？例如它会不会以为把人类全部关进监狱就可以避免人类互相残杀，这才是对人类整体最好的？至于说，到时候人类如果觉得危险了，可以把机器的电源断开……这只是开个玩笑吧，真到那个时候，机器恐怕早就能采用其他的方式摄入能源了。总之，强人工智能出现的那一天，恐怕真的就是人类面临最大生存危机的时候。所以，对严肃的人工智能研究者来说，如果真的相信自己的努力会产生结果，那就不该去触碰强人工智能。 ■

周志华

CCF 会士、常务理事、人工智能与模式识别专委主任。南京大学教授、计算机软件新技术国家重点实验室常务副主任。ACM/AAAS/AAAI/IEEE/IAPR Fellow, 欧洲科学院外籍院士。

zhouzh@nju.edu.cn

[5]《Common Sense, the Turing Test, and the Quest for Real AI》

[6]并非由单一神经细胞对应单一功能，而是诸多神经细胞共同发生作用。

[7]一、机器人不得伤害人，也不得见到人受伤害而袖手旁观；二、机器人应服从人的一切命令，但不得违反第一定律；三、机器人应保护自身的安全，但不得违反第一、第二定律。

[8]由于发现三定律有漏洞，阿西莫夫后来补充了第零定律：机器人不得伤害人类整体，或因不作为而使人类整体受到伤害。