“强人工智能”争论过程中的“态度转换”现象研究

王彦雨

(中国科学院 自然科学史研究所,北京 100190)

摘 要: 在分析“强人工智能(强AI)”概念的源起、类型及特征的基础上,对强AI争论过程中的“态度反转”现象进行了详细描述,即强AI争论过程中的两个主要参与主体——人工智能界及哲学社会科学界对于“强AI是否会实现”这一议题,均经历了态度上的反转(由赞同到反对抑或相反),并对“态度转换”现象背后的动力机制进行了解析。文章最后强调,传统的强AI概念过于科幻,并提出基于“自主性”理念的“强AI_2”,以期破解传统强AI争论过程中的“不落地”困境。

关键词: 强AI; 态度反转; 摩尔定律; 强AI_2

中图分类号: N02 文献标识码: A 文章编号: 1674-7062(2020)06-0026-08

在人工智能发展过程中,对“AI能否超越人类”(或是说“强AI能否实现”)这一议题,人工智能界及哲学社会科学界均经历了多次的态度转变,如人工智能界对于强AI的态度经历了由乐观到悲观这一大体进路,而哲学社会科学界则与之相反,从最初的质疑与批判,转变为支持与宣扬。我们将之称为强AI争论过程中的“态度转换”现象。研究关于强AI的“态度转换”现象、其背后的动力机制,对于厘清当前关于强AI的一些认识误区、预期未来的人机关系、反思智能社会的治理体系,均大有裨益。

一 “强AI议题”源起及演进

人工智能是建立在这样一种假说之上: 即人类思想可以被机器化、人工化,如玛格丽特·博登所言,“人工智能就是让计算机完成人类心智能做的各种事情”[1]。而关于“人工智能能够在何种程度上实现人脑功能”这一问题,出现两种观点: 弱AI观(weak AI)及强AI观(strong AI)。“弱AI观”认为,机器没有真正的智能、不拥有自主意识; 而“强AI观”则强调机器可以全面、综合地复现人类的所有思维能力,且聪明程度能够达到或超过人类。

实际上,‘强AI’概念源起于古德(Irving Good)提出的“智能爆炸”概念,“由于机器设计本身便是智能活动的一部分,因此一台聪明的机器本身便可以设计更好的机器; 毫无疑问,到那时,便会出现‘智能爆炸’,人类的智能将远远落后于机器”[2]。20世纪70年代,塞尔(John Searle)首次提出了“强人工智能”概念,认为“计算机不仅是用来研究人的思维的一种工具; 相反,只要运行适当的程序,计算机本身就是有思维的”[3]417-418,认为“人脑不过是一台数字计算机,人心不过是一种计算机程序……我把这种观点称为‘强人工智能’观点,简称‘强AI’观点’,‘机器不仅仅是模拟人的能力,它实际上可以说是能够‘理解故事’并‘为问题提供答案’”[3]417-457。从塞尔的定义看,“强AI”强调机器智能与人类思维的同一性,这种“同一性”不仅仅是基于信息学、机械学的,同时还是基于生物学、心理学的,它赋予AI“理解”“反思”等属人性品质。

塞尔之后,学界提出了许多类似于“强AI”的概念。如波斯特罗姆(Nick Bostrom)的“超级智能”

【收稿日期】 2019-03-27

【基金项目】 中国科学院自然科学史研究所重点培育方向项目“科技的社会风险”项目

【作者简介】 王彦雨(1982-),男,山东省菏泽市人,中国科学院自然科学史研究所副研究员,研究方向为STS、科技史。

(Superintelligence),即“在所有领域,包括科学创造力、社会技能等都远比最优秀的人类大脑聪明”[5];文奇(Vernor Vinge)则在1983年提出“技术奇点”(technological singularity),意指AI在某一临界点开始拥有自主目标、智能远超人类且拥有指数级的智能增长速度;库兹威尔(Raymond Kurzweil)则将AI与人的生物性融合在一起,“奇点将代表我们的生物思想与现存技术融合的顶点,它将导致人类超越自身的生物局限性”[6]。大体上,“强AI”可分为以下几个学派(如表1):

表1 “强AI”的不同学派及核心主张

学派	代表人物	核心主张
意向性学派	艾伦·纽厄尔、约翰·麦卡锡	AI具有类于人的自主性、意向性特征,与人类智慧已无异,如塞尔强调,“强AI的支持认为,一个适当的编程计算机不仅仅是思想模型的模拟,它们实际上本身便是思想”。
基于计算主义的“加速进化”学派	雷·库兹威尔、约翰·斯玛特	将智能还原为计算能力,将思维过程还原为单纯的计算过程,认为技术的加速进化及指数增长能够实现对人脑功能的超越,我们可以精确预测新技术何时出现、何时跨越关键技术门槛。
“智能爆炸”学派/递归改良学派	欧文·古德、埃利泽·尤德考斯基、弗诺·文奇	当超过某一临界点,AI可以创造新的、比自身更优秀的AI,AI能够对自身目标进行改进、修正、改变,或是创造自己的新目标,从而不再受控于设计者。
智能融合学派	雷·库兹威尔	生物智能与机器智能相互整合,这将创造一个比大多数科幻小说更为怪异的未来。

虽然不同学派对强AI的内在属性做了各自界定,但总体来讲,它具有如下特性:(1)超越性:须达到或超越人类智能;(2)自我目标形塑能力:指AI可实现“意识自醒”并形成“自己”独特的行动目标,这也是强AI的最关键属性;(3)具备较为完整的认知能力序列:既包括理性的逻辑推理、知识学习能力,也包括感性的情感感知、自我认知,关于他人、社会的理解能力,同时还包括类似于智慧的非线性思维能力(如策略设计)等;(4)情景适应性能力:针对环境变化进行自我反思,并进行适应性的行动变化能力;(5)通用性:可适用于多个问题域;(6)道德赋予属性:即强AI智能体不再被视为单纯的“物”范畴,而是被视为主体地位,被赋予社会属性(如权利、责任、道德等)。

二 “强AI议题”论争过程中的“态度反转”现象

虽然“强AI”这一概念在20世纪70年代末才提出来,但“强AI议题”实际上自20世纪50年代便已经产生,且伴随着一个比较有意思的现象——态度转换,即两个关于强AI的主要争论群体——AI界及哲学社会科学界,对于“强AI议题”的态度,均经历了一个转换过程:AI界从支持者逐渐转变为质疑者;而哲学社会科学界则从悲观论者转化为乐观论者。

(一)AI界的态度转换:从强AI的支持者转变为质疑者

AI界对强AI的态度,经历了“朴素担忧”(20世纪50年代初期—中期)、“极度乐观”(50年代中后期—70年代初)、“示微及路径分裂”(70年代中后期至80年代中后期)、“噤声”(80年代中期—2010年左右)、“谨慎发声”(近几年来)这几个阶段。具体来讲:

1.“朴素担忧”期。如维纳在1950年的《人有人的用处:控制论与社会》中强调,“机器,和生命体一样,是一种装置……从理论上说,如果我们能够造出一部机器,其机械结构就是人的生理结构的复制,那我们就可以有一部机器,其智能就是人的智能的复制”[7]43。维纳还设想未来会有一种具有“学习”能力的机器,并将之形容为“瓶装妖魔”,“像《一千零一夜》中阿拉伯渔翁在那只装有愤怒妖魔的瓶子上揭开所罗门的封印时所做的那样……不论该机器能够学习与否,都意味着他把自己的责任交给天风,任其吹逝”[7]153,(技术的)灾难性后果不仅会出现于童话世界,还会发生于现实世界”[8]。

2.“极度乐观”期。1956年达特茅斯会议的召开,使“人工智能”成为一个严肃的学科,并在定理证明、机器翻译等领域取得突破。这一时期,AI界对于“强AI”持极度乐观心态,“麦卡锡定义了制造匹敌人类能力的机器的目标。事实上,在人工智能历史上的第一个10年,这种乐观无处不在”[9]114,“他们试图建造用来展现具有真实心理特性的机

器:问题解决、思考、理解和推理,并且最终可能是意识、感觉和情绪"[10]。如1961年,西蒙(Herbert A. Simon)谈道"在20年内,机器可以做人类可以做的任何事情"[11];1966年,明斯基(Marvin Minsky)指出,"一旦我们设计出真正具有自我完善能力的机器,那么一个快速的进化过程便会开始,随着机器自身及其自身模型的改进,我们将会看到与'意识''直觉''智能'相关的所有现象"[12]。

3."示微及路径分裂"期。20世纪70年代中至80年代初,虽然AI界的主流观点依然是基于"强AI"的,但一种新的思潮正在兴起——"智能增强"。"智能增强"强调人工智能是辅助、扩展、服务,而非超越、取代人类,"智能增强基因几乎无处不在"[9]138。如:比尔·杜瓦尔(Bill Duvall)曾参与斯坦福研究所Shakey项目,但他后来转投恩格尔巴特实验室,"成了全世界第一个从'用计算机取代人类'领域的研究跳槽到'用计算机来增强人类智慧'领域的人"[9]8,其他加入这一阵营还包括艾伦·凯(Alan Kay)、拉里·泰斯勒(Larry Tesler)等。

4."噤声"期。20世纪80年代初,人工智能界凭借"专家系统"获得商业层面的短暂成功,但自1987年起,个人台式电脑迅速取代专用型的Lisp智能机[13];此外,日本"第五代计算机计划"在1991年宣告失败。无论是"专家系统"的兴起还是衰落,均代表着"强AI"观念在整个AI界的衰落,因为"专家系统"的目标是为了辅助人类在特定的专业领域内决策,而非取代人类。在20世纪90年代,人工智能界已经不再进行任何关于"强AI"的预言,以免被人贴上"白日梦"的标签[14],即使到2005年前后,AI研究者依然对"强AI"理念保持距离[15]。

5."谨慎发声"期。约2010年以后,虽然大部分的AI专家依然对强AI持怀疑态度,但也有一部分AI专家开始强调强AI最终会实现,如递归神经网络之父施米德胡贝(Jürgen Schmidhuber)认为,宇宙史上重大事件的发生间隔似乎在几何式地缩短,人工智能可能在2050年超过人类智商。鲍姆(Seth D. Baum)等对参加2008年"通用人工智能会议"的21位AI专家进行了采访[16],发现当时间定为2040年之前时,约有11人(约52%)认为AI会达到"超级人类水平"。

(二)哲学社会科学界:从强AI反对者到强AI支持者

对于强AI,哲学社会科学界在很长的一段时期内是坚定的反对者,但20世纪末以来,其逐渐转变态度,特别是当前,他们成为宣扬强AI及强AI风险的主要力量之一:

1.强AI的"批判者"(20世纪60年代—80年代中后期)。90年代以前,哲学社会科学界对强AI一直持反对意见。如20世纪60年代初,普林斯顿大学逻辑学教授卢卡斯(John Lucas)认为,"将一个新思想带到世界上来有两种方式,一是通过女性生儿育女,另一种是建构一个新的由阀门和继电器所构成的复杂系统。当我们谈到第二种方式时,我们必须强调尽管我们所创造的像一个机器,但实际上它不同,因为它已经不再是由各个部分所构成的累加系统"[17]。1965年,休伯特·德雷福斯(Hubert Dreyfus)分析了人工智能和人类智能的区别,认为前者不具有感知力,它习惯于数学/逻辑的严谨规则,"在理解事物时,我们往往是通过瞬间的感知"[18]1。德雷夫斯指出,"不出意外,同时也更为让人失望的是,人工智能游戏、问题解决、语言翻译依然需要等待模式识别领域有一个根本性飞跃"[18]45-52。

2.由强AI批判者转化为强AI坚定的支持者(20世纪初—2010年)。20世纪末21世纪初,哲学社会科学界由此前的悲观转变为乐观态度。尤德考斯基(Eliezer Yudkowsky)于2001年提出"种子人工智能"(seed AI)概念,意指具有自我完善功能的AI,其演变过程包括:①级联(Cascades),意指"一种发展会导致另一种发展";②循环(Cycles),意指过程的可重复性叠加,其中一种优化会导向另一种优化,并形成可重复过程;③洞悉(Insight),指大大提高优化能力的新信息、新知识;④递归(Recursion),是指AI能够重新设计认知算法[19]。波斯特罗姆则于2002年提出"生存性风险"概念,并将恶意编程的人工智能视为类型之一,"当我们创造出第一个超级智能实体时,我们可能会犯错误,并赋予它毁灭人类的目标"[20]。

3.强AI风险反思大潮的弄潮儿(2010年至今)。2010年以后,社会上出现了一股针对"强AI"的反思热潮,而哲学社会科学界成为这场运动的弄潮儿,在这一时期,他们更多关注"如何实现强AI",而不仅仅是"是否会实现"。如:①强AI生成机制研究,查尔默斯(David J. Chalmers)指出,如果AI0能够生产在能力上稍强于新一代AI1,那么如果这一过程持续下去,"衍生由AI+到AI++的演变"[21]。②强AI风险研究,巴雷特(Anthony Barrett)认为强AI风险的产生路径,如产生种子AI、限

制外壳失效等[22]；福克斯(Joshua Fox)则认为智能机器间由于相互竞争的需要，其行为可能会是“反道德的”[23]。③强AI风险治理研究，如索尔斯(NateSoares)认为，在设计智能体时应确保智能体具有容错能力、在进化过程中与设计者的“最初意图”保持一致[24]。

(三)新的参与者：科技企业家及非AI类科学家

当前这一轮强AI热潮中出现一个重要的参与者：科技类企业家和非AI界的科学家(如物理学家等)。与AI界摇摆不定态度不同，他们对强AI问题异常关注。如伊隆·马斯克(Elon Musk)表示，“随着人工智能发展，我们正在召唤恶魔”；盖茨(Bill Gates)在2015年谈道“我和那些担心超级智能的人同处一个阵营。”此外，非AI界的科学家也积极参与到最新一轮的强AI风险讨论之中，如霍金认为，“一个超级智能极端擅长达到其目标，当其目标与人类目标不相一致时，我们便会有麻烦”[25]。英国理论天体物理学家马丁－里斯(Martin Rees)也指出，25年之后人类将进入“无机后人类时代”，智能机器人将能够摧毁地球人类文明[26]。

三 “态度反转”的动力机制及具体过程分析

“强AI态度转换”现象背后的动力机制是什么？当我们分析这一现象时，需要看到三个层次：(1)实际能力层：即在特定时期，AI技术本身所具有的实际能力；(2)能力扭曲层(包括“技术潜力夸张层”和“技术局限性聚焦层”)：专业人士(如哲学家、一部分AI科学家等)将视角聚焦于对AI未知潜力的夸张性描述，或是单纯聚焦AI技术的局限性；(3)外围影响层：指非专业人士(如传媒界、公众等)对当时代AI技术发展的认知态度。在这三个层次中，“能力扭曲层”起着决定性作用，加之外围影响层，形成“强AI态度转换”现象的建构场(如图1)。当我们反思强AI态度转换现象时，应追溯其两条主导性技术范式——逻辑符号主义和连接主义——所蕴含的局限性及未知潜力之间的张力，AI界和哲学社会科学界通过对两种技术范式自身所蕴含的未来潜能及面临局限性的分析，形塑着各自对“强AI”的认可及质疑度。

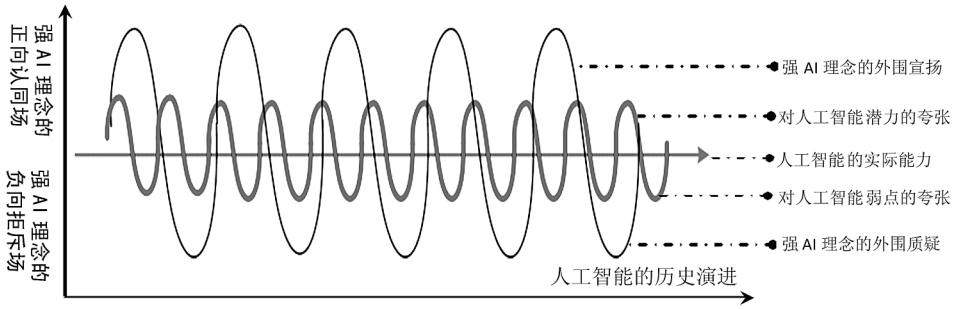

图1 强AI态度转换现象的影响因素

(一)初始态度阶段的动力机制分析：关于逻辑符号主义潜力及短板的争论

1956年至20世纪70年代中期，AI界对“强AI”持乐观态度，而哲学社会科学界则相反。造成这种差别的原因是两者对逻辑符号主义范式进行了不同视角的解读：AI界聚焦于其“技术潜力”，而后者则聚焦其局限性。

AI界之所以持乐观态度，是因为逻辑符号主义证明了人类思维可化约为可操作的符号/代码，机器可完成人类思维的特定功能，“他们乐观心态的基础是建立在这样一种假定之上：他们确信人类的信息处理过程与数字计算机的离散步骤是一样的”[18]47；且人工智能所能完成的任务属于高阶智慧，如：塞缪尔1952年所发明的跳棋程序既能打败州际冠军(1956)，同时也能打败设计者本人(1959)，纽厄尔(Allen Newell)和西蒙(Herbert A. Simon)的“逻辑理论家”(1955)可证明数学定理[27]123－125。基于这些进展，AI界开始聚焦“技术潜力夸张层”，认为一切关于人工智能/人类思维、逻辑推理/非逻辑直觉、高阶智慧/常识知识等之间的区分将被打破，人类思维最终将全部“人工化”。

与AI界不同，哲学界较早发现了逻辑符号主义路径的局限性：逻辑符号主义是基于逻辑、规则、信息的充分性，以及理想环境下的非容错性；而人脑则不仅于此，其信息处理过程同时还是非线性、非充分性、综合性的，且容错性强。如卢卡斯便证明了逻辑系统难以包纳“自由意志”：① 任何人都会拥有至少一个以上的逻辑系统L；② 对于任何逻辑系统L而言，一个足够熟练的逻辑学家H都可以构建一些为真但在L系统中无法证明的语句S；③ 因此，对于

类似 H 的人 M 来讲，单凭逻辑系统 L，无法可靠地预测 M 在所有情形下的行为；④ 因此 M 拥有“自由意志”[28]。德雷福斯则强调基于符号主义路径的人工智能只适应于理想环境，只能处理目标、需求明确的信息。

（二）AI 界第一次态度转换的动力机制分析：来自计算困境的挑战

自 20 世纪 70 年代中期起，AI 界关于强 AI 的态度由极端乐观开始转向悲观，之所以会发生这种态度转变，关键原因在于 AI 界发现：有限的计算能力难以应对组合爆炸等难题。如卡普（Richard Karp）1973 年指出，许多 AI 问题只能通过指数时基来解决，这需要不可想象的计算量[30]。人类对事物意义的确定是建立在大量背景信息基础之上，AI 在信息处理问题过程中亦如此，如在视觉识别过程中：AI 不仅仅要“看到”，还要解决“我要怎么看，看到了哪些，没有看到哪些”等背景信息问题，这便需要 AI 对图片中各个构成要素的像素代码进行复杂计算，包括目标对象辨识与计算、非目标对象辨识与计算、相似目标辨识与计算、背景辨识与计算等，而非目标的辨识则又需要进行回溯式分析，这一过程需要巨量计算，而当时尚无法建立如此巨大的数据库[27]300。机器翻译同样面临类似的困境，在面临稍微复杂的多重语义、代词、不太符合正统语言规则的句型及结构时，需要机器对巨量的、不同语境中的词语组合等进行解析，这远远超出了当时计算机所能够提供的计算量及存贮量。

（三）哲学社会科学界第一次态度转换的动力机制分析：来自他山之石的“助攻”及摩尔定律的推动

自 20 世纪 90 年代后期，哲学社会科学界对强 AI 的态度，开始由此前的批判者转化为支持者，这种态度转变与以下两种因素密切相关：

（1）直接思想来源：20 世纪 80 年代后期的基于奇点概念的“科幻小说浪潮”。当时，计算机科学界的一小部分人，通过科幻小说对未来强 AI 的实现场景进行文学式展示，如 AI 专家莫拉维克在《智力后裔：机器人和人类智能的未来》一书中认为，随着计算机性能的提升，2030—2040 年间，一种基于人工智慧的新物种将会出现[29]；文奇于 1971 年获得计算机博士学位，但其被人们所熟知却是其智能体小说，如《真名实姓》（1981）、《深渊上的火》（1992）等，其最大贡献在于 1983 年首次提出“技术奇点”概念[31]。这些科幻小说成为此后哲学社会科学界宣扬强 AI 理念的重要思想来源。

（2）技术基础条件：摩尔定律。20 世纪 80 年代起，摩尔定律不断被印证（如图 2），解决了此前 AI 发展过程中所面临的“计算能力”瓶颈，展示了 AI 在硬件层面不断接近人脑的一种可能性与发展潜力。摩尔定律下机器超越人类的一个成功案例便是 1997 年“深蓝”战胜世界国际象棋冠军卡斯帕罗夫。

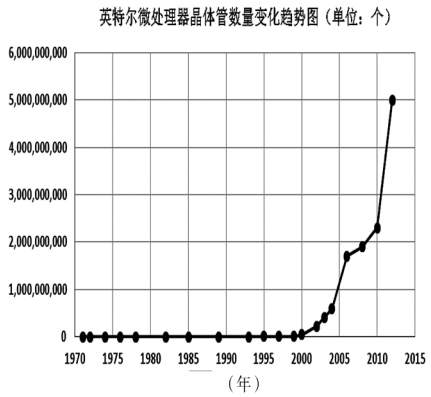

图 2（a） 英特尔微处理器晶体管数量的指数增长

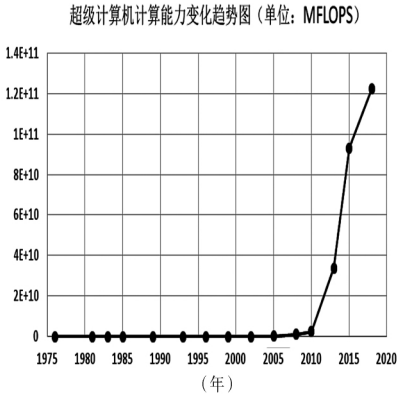

图 2（b） 超级计算机计算能力的指数增长

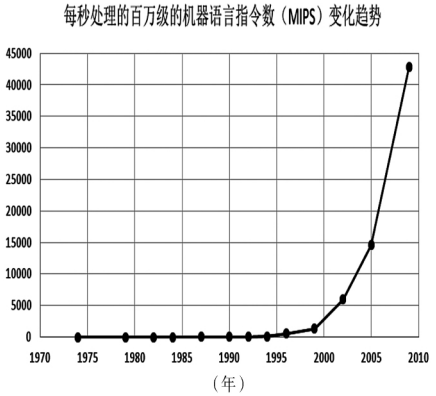

图 2（c） 每秒处理的百万级机器语言指令数的指数增长

（四）强 AI 议题成为 AI 界与哲学社会科学界的一种共同话题：逐渐走向现实的“类强 AI”技术（2010 年至今）

2010 年前后，无论是 AI 界还是哲学社会科学界，均出现了一股反思强 AI 的热潮，其背后的原因在于连接主义范式的兴起。连接主义与逻辑符号主义明显不同——它更像人脑，无论是结构层面还是功能层面，且具备了逻辑符号主义所无法实现的、类似于人脑的诸多特征。举几个简单的例子：

第一，类似于人的“学习能力”。可发现离散信息背后的规律，如张首晟所研发的人工智能系统，当把已知化合物的化学结构输入其中，它只用两个小时便“重新发现”了元素周期表。

第二，恐怖的学习效率。以围棋 AI 为例，呈现

出从以年为单位到以月，至以天为单位的功能倍升模式，这种恐怖的学习与进化效率，成为人们相信强AI能够到来的重要理由之一。

第三，“黑箱效应”——AI在一定摆脱人类控制。在逻辑符号主义范式下，“它们自身的有效性永远处于其创造者——程序员的独创性及霸道性所限制”[32]。而基于深度学习技术的AI在一定程度上实现了“自我掌控”，如2015年纽约西奈山医院的人工智能“深度病人”(Deep Patient)自我学习了70万份病历数据，便可预测精神分裂症，负责人表示就连他自己也不知道它是如何做到这一点的。

第四，一定程度的“自我改良”属性。如2007年，施米德胡贝建立的“哥德尔机器”成为第一个具有完全自我指涉能力元学习器，“当它一旦发现(对原有程序代码)的重写将非常有用时，便会迅速重写自身代码的任何一个部分”[33]。

四　从“态度反转”看未来的强AI：从“强AI_1”到“强AI_2”

学界关于“强AI”的争论，过于强调其“意向性”特征，将“具有社会性及意识性特征的人”作为参照，实际上这忽略了机器智能自身的特殊性及人的意识的复杂性。在这里我们提出了具有科学性特征的“强AI_2”概念。

(一)走向落地的强AI理念——强AI_2

传统的强AI理念(我们将之称为“强AI_1”)，往往过于科幻，强调其可实现“意识的自醒”。然而这并不具备科学层面的可行性，因为诸如“理解”“意志”等本质上是生物学、哲学或宗教学的；且，当前的生物学界很难对“意识/智慧的生成”等做出满意解释，如蒲慕明强调，当前神经科学研究主要限于对大脑结构的精细“解剖”，而对于整体性的“意识”的生物学形成机制，则根本无法攻克。

在这里，我们提出基于“自主性”(而非“意向性”)的强AI观——具备“目标自主性+行动自主性”能力的AI(强AI_2)。“自主性”与“意向性”不同(如表2)，它并非要求AI会产生某种自我意识，而意指它能够对设计者所赋予的初始目标进行自我微调、改变及生成新目标，或是通过与外界交互自主进化。如果说“强AI_1”的对应物是人脑，那么“强AI_2”则对应设计者，强调AI成为相对于设计者的异化物、异变物，这里的“强”不是指“机器智能会在整体功能层面超过人类大脑”，而是指“AI具有摆脱设计者控制、自主生成及实施新目标，或是自主进化的能力”。

表2　强AI_1与强AI_2之间的相异之处

属性	强AI_1	强AI_2
思考基点	意向性	自主性
临界点	形成基于自主意识的“新目标”及目标实施策略	形成区别于设计者的原初目标新目标(往往是被动性地)，并制定相应实施策略
意向性	有(能回答“我是谁”“他人是谁”“我要做什么”等问题)	无(“新目标”是通过与外部环境的交互/系统的递归性改良等来实现)
自主性程度	强(“新目标”往往超出已有的领域限制，与原初目标之间往往不具有逻辑一贯性)	弱(初始目标由设计者所赋予，且新目标与原初目标在很大程度上具有逻辑一贯性)
通用性	必需条件	并非绝对必需
参照物	人	设计者

(二)强AI_2的可能性

AI_2将关注点放之于当下的技术条件及未来的技术可能性，如在文奇、波斯特罗姆等人的论述中，拥有“递归性自我改良”能力是实现超级AI的关键性要素，而当前的AI已经在一定程度上实现了生成种子框架、自动改进已有框架等能力，“递归性自我改良”已不再是科幻小说里的猜想。如贝克(Bowen Baker)提出基于强化学习的元建模方法，它能够通过学习各种可能存在的架构，迭代发现具有改进性能的设计。此外，类似的“种子AI”也已产生，2019年初，哥伦比亚大学打造了一套能够“从零开始”认识、发展自己的AI：它可以通过自主训练、不断调整已有的自我模拟，实现自我进化，如通过35小时的训练，可在没有任何先验经验的情况下，自主学习“将小球放入水杯”，且在这一任务完成后自主学习了“写字”功能。

(三)强AI_2与未来的社会治理

在未来的“强AI_2”时代，传统的人与技术人工物之间的“主客关系”受到挑战：人们无法完全知晓“强AI_2”在与人的互动中会如何进化、会形成何种

目标，会在特定的情景做出何种决策，人类社会逐渐进入“泛伦理”及“泛风险”时代（“社会态3”）：

（1）“社会态3”的“泛伦理性”特征。AI_2 自带“伦理属性”，智能体引发伦理问题，不再被视为引发原有社会体系混乱的“异常”现象，而成为一种正常的社会态；智能体研究者、生产者面临更多的伦理诘难，传统的科学规范（如独创性）、企业规范（如技术先进性为上）已无法为自己的行为有效辩护，他们迫切希望获得更多的社会支持（伦理方面等），伦理优先逐渐取代科学主义成为主流意识形态。

（2）“社会态3”的“泛风险”特征。人类应警惕未来智能社会所可能引发的重大风险问题[34]，人工智能所具有的决策黑箱属性、自我改良与进化特征、情景中的目标异变现象，以及被恶意使用，使得未来的社会治理体系更为脆弱。人类一方面不得不越来越多地依赖于人工智能，另一方面，人类又不得不面临“风险的逐渐扩大”局面，对人工智能的“依赖与反依赖”两难困境将逐渐显现[35]。

表3　三种技术社会态及其特征

属性	社会态1：技术影响社会	社会态2：社会形塑技术	社会态3（强 AI_2 时代）：技术形塑社会
主客间的边界	边界清晰	边界清晰	边界模糊
人工物属性	单纯的工具性存在物	作为一种“社会风险/伦理负载”的工具性存在物	既非单纯的主体也非单纯客体
主导性技术文化	科学主义为主导	科学主义与伦理关注并存，且“技术风险/技术伦理”被视为是人机关系的“非正常态”	反思技术的伦理/社会风险成为主导性技术文化，智能体引发伦理问题被视为是一种正常的社会现象
科学家与社会科学家关系	科学共同体占主导地位，伦理学家等被无视	科学共同体占主导地位，且与伦理学家等处理一种紧张的关系态	社会科学家崛起，社会科学家或伦理学家与科学共同体的合作成为主流
技术伦理规范的形成模式	“精英规范”时代：伦理学界热衷建立一种理想性的规范体系	“精英规范”时代：伦理学界积极参与到相关讨论与互动之中，希望能够建立一种普适的规范体系	伦理规范的形成模式由“精英模式”向“平民模式”转换，消费者/公民介入规范制定过程
技术的风险强度	弱（人绝对主导技术）	中（人主导技术，但技术不可控性增强，如黑客、恐怖主义对技术的滥用）	强（智能体不再绝对受控于人）

【参　考　文　献】

[1] 博登. 人工智能的本质与未来 [M]. 孙诗惠，译. 北京：中国人民大学出版社，2017：3.

[2] GOOD I. Speculations concerning the first ultraintelligent machine [J]. Advances in computers，1966(6)：31－88.

[3] SEARLE J. Minds brains and programs [J]. The behavioral and brain sciences，1980，3(3)：417－457.

[4] 塞尔. 心、脑和科学 [M]. 杨音莱，译. 上海：上海译文出版社，1991：20.

[5] BOSTROM N. How long before superintelligence? [J] Linguistic and philosophical investigations，2006，5(1)：11－30.

[6] 库兹威尔. 奇点临近 [M]. 董振华，李庆诚，译. 北京：机械工业出版社，2011：2.

[7] 维纳. 人有人的用处：控制论与社会 [M]. 陈步，译. 北京：商务出版社，1989.

[8] WIENER N. Some moral and technical consequences of automation science [N]. New series，1960－05－06.

[9] 马尔科夫. 与机器人共舞 [M]. 郭雪，译. 杭州：浙江人民出版社，2015.

[10] PRESTON J，BISHOP M. Views into the Chinese Room：essays on Searle and artificial intelligence [M]. Oxford：Clarendon Press，2002：14－14.

[11] CREVIER D. Ai：the tumultuous history of the search for artificial intelligence [M]. New York：Basic Books，1993：109.

[12] MINSKY M. Artificial intelligence [J]. Scientific american，1966 (3)：251－257.

[13] HENDLER J. Avoiding another AI winter [J]. IEEE intelligent systems，2008，23(2)：2－4.

[14] MARKOFF J. Behind artificial intelligence，a squadron of bright real people [N]. The New York Times，2005－10－14.

[15] MENZIES T. 21st－century AI：proud，not smug [J]. IEEE intelligent systems，2003，18(3)：18－24.

[16] BAUM S，GOERTZEL B，GOERTZEL T. How long until human－level AI? results from an expert assessment [J].

Technological forecasting & social change, 2011, 78(1): 185 – 195.
[17] LUCAS J. Minds, machines and gödel [J]. Philosophy, 1961(36): 112 – 127.
[18] DREYFUS H. Alchemy and artificial intelligence [M]. Santa Monica: RAND corporation, 1965.
[19] YUDKOWSKY E. HANSON R The Hanson – Yudkowsky AI – foom debate [EB/OL]. (2013 – 04 – 01) [2019 – 01 – 03]. http://intelligence.org/files/AIFoomDebate.pdf.
[20] BOSTROM N. Existential risks: analyzing human extinction scenarios and related hazards [J]. Journal of evolution and technology, 2002, 9(1): 1 – 23.
[21] CHALMERS D. The singularity: a philosophical analysis [J]. Journal of consciousness studies, 2010, 17(9): 7 – 65.
[22] BARRETT A, BAUM S. A model of pathways to artificial superintelligence catastrophe for risk and decision analysis [EB/OL]. (2016 – 04 – 15) [2018 – 11 – 05]. https://arxiv.org/ftp/arxiv/papers/1607/1607.07730.pdf
[23] JOSHUA F, SHULMAN C. Superintelligence does not imply benevolence [EB/OL]. (2010 – 03 – 12) [2017 – 09 – 04] http://joshuafox.com/wp – content/uploads/2014/10/FoxShulman_SuperintelligenceBenevolence.pdf
[24] SOARES N, FALLENSTEIN B. Agent foundations for aligning machine intelligence with human interests: a technical research agenda [EB/OL]. (2017 – 06 – 23) [2017 – 08 – 05]. https://intelligence.org/files/TechnicalAgenda.pdf
[25] GRIFFIN A. Stephen Hawking: artificial intelligence could wipe out humanity when it gets too clever as humans will be like ants [N]. Independent, 2015 – 10 – 8.
[26] TEGMARK M. Life 3.0 [M]. New York: Random House, 2017: 280.
[27] McCorduck P. Machines who think: a personal inquiry into the history and prospects of artificial intelligence [M]. Norfolk County: A. K. Peters, 2004.
[28] LUCAS J. The godelian argument [EB/OL]. (1989 – 01 – 23) [2018 – 09 – 09]. http://www.leaderu.com/truth/2truth08.html
[29] BUCHANAN B. A (very) brief history of artificial intelligence [J]. AI magazine, 2005, 26(4): 53 – 60.
[30] MORAVEC H. Robot: mere machine to transcendent mind [M]. Oxford: Oxford University Press, 1999: 127 – 155.
[31] VINGE V. The coming technological singularity [EB/OL]. (1993 – 04 – 09) [2019 – 01 – 01]. http://www.frc.ri.cmu.edu/~hpm/book98/com.ch1/vinge.singularity.html
[32] FEIGENBAUM E, FELDMAN J. Computer and thought [M]. New York: McGraw – Hill Book Company, 1963: 250.
[33] SCHMIDHUBER J. Gödel machines: fully self – referential optimal universal self – improvers [M]//GOERTZEL B, PENNACHIN C. Artificial general intelligence. Berlin: Springer, 2007: 200.
[34] 刘益东. 科技巨风险与可持续创新及发展研究导论: 以致毁知识为中心的战略研究与开拓 [J]. 未来与发展, 2017(12): 5 – 7.
[35] 刘益东. 智业革命: 致毁知识不可逆增长逼迫下的科技转型产业转型与社会转型 [M]. 北京: 当代中国出版社, 2007: 204 – 206.

A Study on the Phenomenon of "Attitude Reversal" in the Debate of Strong Artificial Intelligence

WANG Yan – yu
(*The Institute For History of Natural Science, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: Based on the analysis of the origin, types and characteristics of the concept of "strong AI", the paper makes a detailed description of the phenomenon of "attitude reversal" in the process of "strong AI" debate, that is, the two main participants including the artificial intelligence research circle and the social sciences circle, both have experienced an attitude reversal on the issue of "whether strong AI will be realized". The dynamic mechanism behind the phenomenon of "attitude reversal" is analyzed. Finally, the paper emphasizes that the traditional concept of strong AI is too sci – fi, and puts forward the concept "strong AI_2" in order to solve the "landless" dilemma in the process of traditional strong AI debate.

Key words: strong AI; attitude reversal; Moore's Law; strong AI_2

(责任编辑　许玉俊)