基于历史视角分析的强人工智能论争

王彦雨

(中国科学院 自然科学史研究所,北京 100190)

摘 要:在人工智能发展史上,强人工智能("强 AI")一直是一个争议不断但却又不断引发人们关注的议题。对于"强 AI"理念,我们应合理看待其所发的各种争论:(1)"强 AI"理念是推动人工智能界不断打破人机界限、使 AI 技术向前发展的重要信念;(2)"强 AI"争论背后所反映的是不同社会要素,特别是"两种文化"(科学文化和人文文化)之间的张力,且 AI 界的这几种文化也经历着由对立、冲突,到逐渐的尝试性对话与合作这一过程。(3)AI 界、哲学界在对待"强 AI"这一议题的态度并非一成不变,AI 界经历了一个由乐观与支持到悲观与放弃以及现在的谨慎心态,而哲学界对于"强 AI"的态度则是沿着由批判与质疑到现在的大力宣扬这一路径演变,且当前他们对强 AI 所可能引发的风险更为忧虑。(4)"强 AI"概念需要进行重新界定,使其成为科学而非单纯的"科幻式"概念,并给予强 AI 风险议题更多关注。

关键词:强 AI;"机器智能恶与善"之争;乐观主义与悲观主义之争;"智能增强"理念;"奇点"理论;"强 AI 风险论";科技巨风险

中图分类号:N031 **文献标识码**:A **文章编号**:1008-7699(2018)06-0016-12

强人工智能("强 AI")概念由哲学家塞尔(John Searle)于 20 世纪 80 年代所提出,[1]类似的概念还包括高端"通用人工智能"(Artificial General Intelligence)、波斯特罗姆(Nick Bostrom)的"超级智能"(Super intelligence)、弗诺·文奇(Vernor Steffen Vinge)的"奇点"(singularity)等,其对应的概念是"弱 AI"(weak AI)或应用性 AI、专用 AI。"弱 AI"不具有真正的智能或自主意识,只能解决特定领域中的问题;而"强 AI"则是指达到人脑水平的机器智能,可以全面、综合地复现人类大多数(或全部)的思维能力,甚至具有诸如自主意识、算计、情感反应等价值或情感要素。如塞尔将"强 AI"界定为"一个机器可以展示出或模拟人类水平的聪明程度,或超出人类的聪明程度",波斯特罗姆强调"超级智能""几乎在所有领域均能够远远超过人类的认知能力"。强 AI 是否能实现、是否与人为善,对于这一问题,学者们给出了不同看法。一些学者,如埃隆·马斯克(Elon Musk)、史蒂芬·霍金(Stephen Hawking)等认为强 AI 最终会实现且会与人为恶,刘益东则指出强 AI 具有双重危险,属于"致毁知识",因为其正负效应不可抵消,无论它有多大的正面效应,也是"一坏遮百好",所以应该明令禁止。[2]蔡恒进也认为,"作为人类意识延伸的人工智能,被赋予的是偏狭而非完整的意识,在快速进化之后会导致其能力与意识状态的极度不匹配";[3]一些学者认为强 AI 的实现虽然不可避免,但它是有益于人类的,如赫伯特·西蒙(Hebert Simon)、库兹韦尔(Ray Kurzweil)等;也有人将强 AI 视为是一种不可能实现的幻想,如严乐春(Yann LeCun)、谭铁牛等。

关于"AI 能否达到人类的思维水平、是否会取代甚至控制人类"这一问题(我们称之为"强 AI 议题"),自"人工智能"这一学科产生之日起便引发广泛关注,相关分析散见于人工智能史研究中,如丹尼尔·克勒维耶(Daniel Crevier)在 1993 年的著作《AI:人工智能研究的动荡史》中,描述了人工智能发展初期

收稿日期:2018-10-18

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";中国科学院青年研教项目"机器人 ELSI 问题研究"项目

作者简介:王彦雨(1982—),男,山东巨野人,中国科学院自然科学史研究所副研究员,科学技术哲学博士.

乐观派和悲观派之间的争论、对立过程；约翰·马尔科夫(John Markoff)在《与机器人共舞：人工智能时代的大未来》一书中，对人工智能研究共同体中“强 AI 派”与“智能增强派”之间的紧张关系进行了阐释，等。但是这些研究较为分散，没有能够结合各个时期争论的不同主题，对“强 AI”的整个争论史进行系统的论述，实际上，在人工智能发展的不同阶段，人们对“强 AI 议题”的关注点及讨论视角一直在发生着变化(见图 1)。本文主要探讨的问题是，(1)在人工智能发展的不同时期，AI 共同体内部不同群体以及哲学社会科学界对“强 AI”分别持何种态度？(2)不同争论群体在“强 AI”这一议题上是否以及如何对话与互动；(3)“强 AI 议题”对人工智能的发展有何作用，特别是如何影响通用人工智能与专用人工智能两种研究路径的演进。

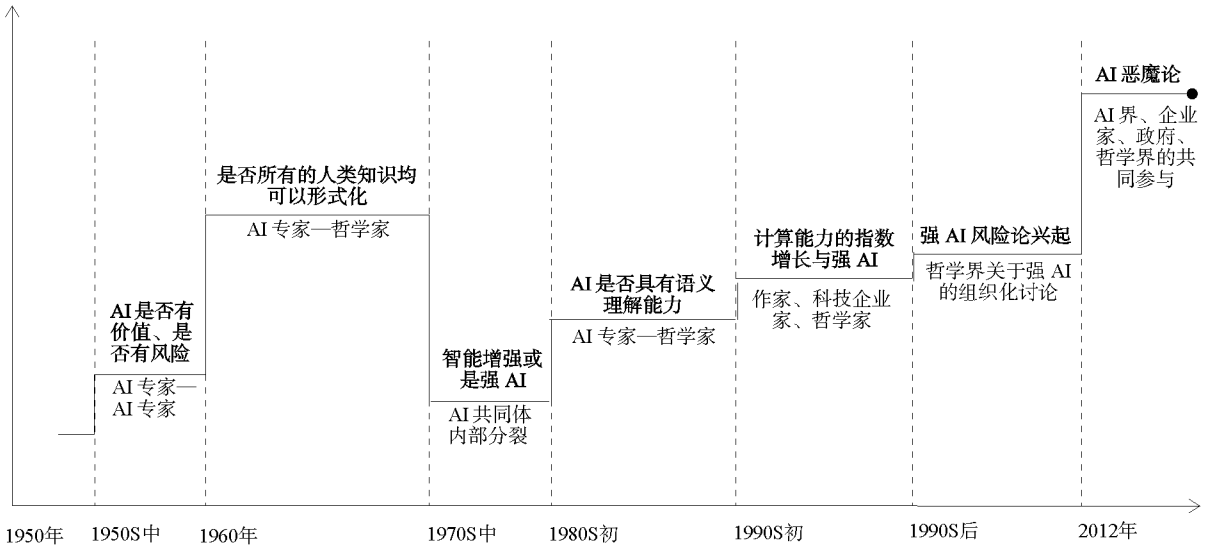

图 1　不同时期人们关于强人工智能的争论议题及关注点变化

一、图灵的担忧及诺伯特·维纳与塞缪尔关于“机器智能恶与善”问题之争(1950—1956)

1950 年，图灵发表著名论文《计算机器与智能》(*Computing Machinery and Intelligence*)，首次提出“机器是否可以思考”这一问题。[4]442 图灵认为，机器可以像人脑一样思维，他强调“模仿”概念，即机器可以模仿人脑从而实现人脑的某些功能。图灵将人脑比拟为一台数字计算机，由存储器、执行单元和控制器所构成，通过“编程”将目标函数输入机器，从而实现特定目标，“分析机实际上是一台万能数字计算机，当它的存储能力和速度达到一种程度，人类就可能通过适当的程序使它模仿我们讨论的机器。”[4]433-460 机器“思考”的限度是什么？或是说图灵眼中的机器是否具有强 AI 属性？他认为，“机器可以成为它自己的主题，机器可以通过观察自己的行为的结果，修改自己的程序，以便有效地达到某种目的”，图灵眼中的机器智能不具有诸如道德、伦理、主体感受、意向性等精神或意识要素，它是逻辑的、计算的、线性的，但图灵认为未来的机器智能将具有学习、进化、改进自我等功能，这种属性是强 AI 的重要特征之一。

图灵的理论促使人们去研究如何使机器模仿人类思维方式，如“游戏 AI”、定理证明等。1951 年，斯特雷奇(Christopher Strachey)编写出一个跳棋程序，而普林茨(Dietrich Prinz) 则发展出国际橡棋程序；1955 年，纽厄尔(Allen Newell) 和西蒙(Herbert A. Simon)开发的“逻辑理论家”(Logic Theorist)证明了《数学原理》(Principia Mathematica)中 38 个定理，且发现了一些新的、更好的证明方式，[5]123-125 西蒙认为这一程序“解决了古老的精神/身体问题，解释了一个由物质构成的系统如何拥有思想特质。”[6]17

游戏AI战胜人类棋手事件，引发了一些学者对“机器是否会控制人类”的担忧，其中之一便是维纳(Norbert Wiener)。实际上，维纳在其1950年的《人有人的用处：控制论与社会》(*The Human Use of Human Being*)一书中，便基于熵、反馈控制理论，对有机体与机器之间的相似性进行了论证，认为生命体甚至其思维都可以最终被机械化。维纳的担忧很快变成现实。1952年，塞缪尔(Arthur Samuel)构建了一个被认为是能够学习的跳棋程序，1956年2月24日这一跳棋程序打败了康涅狄格州的西洋跳棋冠军，而1959年，塞缪尔在与自己所设计的跳棋游戏AI的对弈中被击败。

维纳对此表示深切忧虑，他于1960年发表《自动化的某些道德和技术的后果》(*Some Moral and Technical Consequences of Automation*)一文，认为智能机器迟早超过并危害人类。其理由是：(1)机器可能会跳出此前的训练模式，摆脱设计者的控制，“它们无疑是有创造力的……不仅表现在下棋程序所具有的不可预见的战术上，同时还表现在战略评估的详细加权上”；[7](2)下棋游戏可以将其能力延伸到其他领域(如核领域)，实现智能的跨领域迁移并带来未知风险，“这些具有学习能力的机器可以用于编码新型按钮战争中的按钮动作……一个拥有足够的经验可进行适当编程的机器，可能使人类早已经被消灭了”；[8](3)人类行动缓慢，难以做出及时有效的回应。但塞缪尔并不认同维纳的观点，他认为机器不具备独立思想，下棋程序所谓的“意图”或“结论”，只不过是程序设计者本人意图的反映。塞缪尔专门在《科学》杂志撰文“自动化的某些道德和技术的后果——一种反驳”(*Some Moral and Technical Consequences of Automation——A Refutation*)，强调“维纳的一些结论我并不认同，他似乎认为机器能够拥有原创性，是人类的一个威胁”，但“机器不是妖魔，它不是用魔术操作，也没有意志，而且与维纳的说法相反，除了少见的功能失常情况外，它不能输出任何未经输入的东西。”[9]

二、强AI乐观主义与悲观主义之争(第一个人工智能黄金时代，1956—1974)

1956年，达特茅斯会议(Dartmouth Conference)召开，人工智能研究逐渐建制化，且发展速度“令人惊奇”。[6]18 20世纪50、60年代，MIT(1959)、斯坦福大学(1963)、卡内基梅隆大学等均建立了人工智能实验室。

在人工智能发展的第一个黄金期(1956—1974年左右)，人工智能界对“AI达到或超过人类智能”这一问题持乐观态度。如，1957年赫伯特·西蒙在一篇论述中写道：“我可以作出的最简单的结论是，当前世界上拥有可以思考、学习、创造的机器。而且，它们能力的增长非常迅速，在可见的未来，它们解决问题的能力将与人类思维共延”；[10]1970年，明斯基(Marvin Minsky)在《生活》(*Life*)杂志谈到，“在三到八年的时间内，我们可以拥有其智能等同于普通人的机器。”[5]272-274 在乐观主义者看来，人脑的一切活动(如推理、情感、决策等)均可以形式化为符号、还原成一系列的数字或代码，并通过逻辑程序的创建来复现人类大脑的所有思维活动。计算机科学家道格拉斯·霍夫施塔特(Douglas Hofstadter)在《哥德尔、埃舍尔和巴赫：不朽的金发辫》(1979)一书中便强调一切实在性都能转变为形式系统，虽然神经活动与纯粹的数学有着巨大的差异性，但是诸如美、意义、感觉、情感依然存在着进行形式化表示的可能，“没有什么理由使人相信，一台计算中运转得完美无缺的硬件不可能支持那引起体现诸如混乱、遗忘以及美感这类复杂事件的高层符号行为。”[11]

这一时期，哲学界开始登场，他们是作为强AI反对派而出现的，如鲁卡斯(John Lucas)、德雷福斯(H. Dreyfus)等，他们反对AI乐观派所认为的“所有人类活动均可以符号化、形式化”观点。如鲁卡斯在1961年的《心灵、机器和哥德尔》(*Minds, Machines and Gödel*)一书中，强调人类心灵无法进行形式化还原，机器由各个独立的部分构成，它不具有整体的、非机械性累加的特质，“一个机器可以被设计成它可以表面上‘谈’及它‘意识’到自己的行动，但是如果它不变成不同的机器，它便无法真正拥有这种‘考虑’…

…一个有意识的思想可以反思自身,并且批评自己的表现,且这种活动不需要增加额外的部分:它已经是一个完整的、因此不存在阿喀琉斯之踵的物体。"[12]

德雷福斯则利用现象学对人工智能乐观派进行攻击。1972 年,德雷福斯在其出版的《什么是计算机不能够做的:人工智能的极限》(*What Computers Can't Do: the Limits of Artificial Intelligence*)中强调:并不是所有的问题都可以进行形式化,经验、直觉、意义等均无法通过形式规则来加以描述;且,意义、情感等是在生活实践中产生的,大脑功能并非完全由其生理机能所决定,孤立地模拟大脑机制而不把它放到与外界相联系的环境之中,是无法产生思想的,"如果人工智能研究者的理性概念是对事实的演算,如果当他确认哪些事实相关而且有意义时,不是根据事先的给定,而是由上下文环境来决定的,那么他要造出智能行为的企图就会引起论证上的矛盾";[13]另外,用计算机通过算法来解决的问题,其复杂度必须在一定范围之内,而现实世界中的问题往往出现复杂性指数爆炸现象。

对于德雷福斯的批评,人工智能界进行了激烈的攻击,如西蒙在《思维机器》(*Thinking Machines*)一书中将德雷福斯的论点斥之为"废物",明斯基认为德雷福斯"不理解'人工智能',因此也无需理会"。[14]143 当然,AI 界对于德福雷斯也并非完全是批判态度,一些学者基于伦理维度来谨慎反思人工智能所可能引发的社会问题,如 ELIZA 的发明者魏泽鲍姆便认为 AI 研究者对德福雷斯的完全漠视是不专业、且幼稚的。1976 年,魏泽鲍姆出版了《计算机能力与人类理性》(*Computer Power and Human Reason*),强调对人工智能的滥用会降低人类的生活价值,"最终必须在人的智能和机器智能之间划出一条界限。……人把自主权交给了一个机器世界;这种普遍的意义可是值得深入考虑了。"[15]

虽然遭到哲学社会科学界的批评,但是在这一时期,"强 AI"成为 AI 界的主导性理念,并向政府界的资助者做出了许多"浮夸性承诺",并得到政府的大力支持。如 19 世纪 60 年代中期,麦卡锡曾告诉美国国防高级研究计划局(当时称 ARPA):在未来 10 年里,他们就可以打造出"全智能机器","我们最终的目标是创造能够像人类一样高效地从经验中学习的程序。"[16]114 基于冷战需求,加之 AI 界的鼓吹,政府部门特别是美国国防高级研究计划局将大量的资金注入这一领域,如 1963 年该机构投入 200 万美元资助 MIT"AI 智能小组"的 MAC 项目(the Project on Mathematics and Computation),直到 1970 年代,美国国防高级研究计划局每年都会向 MIT 投入 300 万美元用于人工智能研究;[14]64-65 另外,美国国防高级研究计划局也资助西蒙在卡耐基梅隆大学以及麦卡锡在斯坦福大学的人工智能项目等。在"强 AI"理念的推动下,人工智能在各个领域如数学及几何解题、机器翻译、自然语言处理、语音及视觉识别、智能机器人等均取得进展,人工智能界希望人工智能能够拥有类似人脑的越来越多的功能、不断打破人脑与人工智能之间的限界,如 1959 年纽厄尔和西蒙发展出"通用问题求解器"(General Problem Solver),可以处理一些普遍的数学问题;1966 年,魏泽鲍姆(Joseph Weizenbaum)发明 ELIZA,这是人类历史上第一个具备可使用特征的自然语言对话程序;1972 年,维诺格拉德(T. Winograd)在 MIT 建立了一个可用自然语言指挥动作的机器人系统 SHRDLU,它可以用普通的英语句子进行交流,并执行相关的操作。

三、"智能增强"理念对"强 AI"的冲击(第一个人工智能冬天,1974—1980)

1974—1980 年初,人工智能进入第一个"寒冬期"(AI Winter)。AI 冬天主要有两个标志,一个是美国"自动语言处理咨询委员会"(Automatic Language Processing Advisory Committee)于 1966 年发布的"ALPAC 报告"(语言和机器:翻译中的计算机和语言学),另一个是英国赖特希尔(James Lighthill)爵士在英国议会授意下于 1973 年所发布的"Lighthill Report"。"Lighthill 的[1973]年的报告使得英国学术界所建立的 AI 的信心遭受巨大的打击(当然在程度上小于美国),它持续了十年——这就是所谓的'AI 冬天'。"1973 年,这一报告在英国 BBC"Controversy"系列节目中引发了广泛争论,当时参与争论的英国皇家学会

研究人员认为“通用型机器人不过是海市蜃楼”。[14]22报告最终使英国政府终结了英国绝大多数大学中的AI研究，James Hendler写道：“这形成了一种冲击波效应，并导致AI研究经费在整个欧洲范围内的削减”，[17]直到1983年，英国政府才通过Alvey项目资助AI研究。

在这种情况下，AI界开始出现分裂。虽然麦卡锡等依然对强AI抱有信心，如他认为强AI依然可以很快实现，只需要“1.8个爱因斯坦，以及曼哈顿计划所需资源的1/10。”[16]116-117但是另外一些AI研究者则对这样的宏大纲领表现出怀疑及失望情绪，并由此出现了AI(约翰·麦卡锡为代表)与IA(增强人类智能，以“鼠标之父”道格拉斯·恩格尔巴特为代表)两个研究纲领之间的对立，前者(AI)强调发展能够替代人类、整体上超过人类的AI系统，而后者(IA)则强调人机交互，认为机器只是人类的辅助。

斯坦福研究所的恩格尔巴特(Douglas C. Engelbart)早在1962年就发表了名为“放大人类智力：一个概念框架”(*Augmenting the Human Intellect: A Conceptual Framework*)的论文，强调计算机是人类智力“放大器”，人工智能所要研究的是那些能够改善人机交互方式的AI，“在我们看来，我们无需等到完全了解了人类心智过程机理，我们无需等到学会设计出更智能、更强、更快的计算机，我们现在便可以在已拥有的知识的基础上，开始设计更强大、更具经济可行性的智能增强系统。”[18]在智能增强理论看来，人工智能并非是要取代人类，而是要以人类用户为核心进行设计，“在过去的50年，AI和IA两大阵营间潜在紧张关系一直围绕在计算机科学的心脏地带，而这一领域已经创造了一系列正在改变着世界的强大技术……人们很容易认为，AI和IA是同一枚硬币的正反两面，两者的根本区别在于，是设计造福于人类的技术，还是将技术作为目标本身。如今，这种差异的体现是，制造越来越强大的计算机、软件和机器人的目的是以人类用户为核心进行设计，还是替代人类。”[18]115恩格尔巴特自身便是智能增强、人机交互理念的实践者，他于1963年发明了最初的鼠标，用它来代替繁琐的指令，从而使计算机操作更为简便。1960年代初，恩格巴尔特在斯坦福国际研究院(SRI International)成立了“增强研究中心”(Augmentation Research Center)，并与同事安德鲁斯(Don Andrews)、英格利希(Bill English)、鲁里夫森(Jeff Rulifson)等一道，开发了超文本系统、网络计算机、图形用户界面等技术，这些发明仅仅是为了简化人机交互过程。[19]

在恩格巴尔特等人的影响下，一部分AI研究者开始由过去的“强人工智能”转向“智能增强”阵营。如，20世纪70年代施乐公司的艾伦·凯(Alan Kay)，他曾在斯坦福大学人工智能实验室工作过一年，但他没有像麦卡锡一样追求一种匹敌或取代人类的AI，而是创造了个人电脑概念。此外，比尔·杜瓦尔(Bill Duvall)、拉里·泰斯勒(Larry Tesler)、杰瑞·卡普兰(Jerry Kaplan)等人工智能专家，也逐渐转向智能增强研究，发展出简洁的桌面操作系统、平板电脑等。

由于强AI理念在发展过程中面临着诸如语义问题、常识难题(如莫拉维克悖论)等，这些问题不仅仅涉及技术层面，还关涉哲学，如语言的本质、逻辑与情境的关系等。因此，在这一时期，人工智能界开始与哲学社会科学界有了互动。在互动过程中，AI界也逐渐认识到当时的主导性研究路径——逻辑符号主义的局限性，这使一部分AI界人士逐渐抛弃“强AI”理念，如威诺格拉德(Terry Allen Winograd)等。1970年，威诺格拉德完成了SHRDLU智能机器人系统(这一系统能够“理解”对话者的语言、并能够基于自然语言命令做出相应的动作)，他曾经认为这种具有“理解”自然语言能力的AI，只要经过“升级”或积累大量的常识和规则，便可以像人类一样处理多样化的、复杂的日常事务。但是，当时威诺格拉德经常参加加州大学伯克利分校由德雷福斯主导的小组讨论，并与那里的哲学家如塞尔等进行非正式午餐讨论，“哲学家们穿过伯克利的街道聚在一起。来自施乐PARC的威诺格拉德和丹尼·博布罗已经成为这些午餐讨论的常客。威诺格拉德发现，他们改变了自己对人工智能哲学基础的偏见……他最终放弃了对人工智能的‘信仰’”，[16]179选择了“恩格尔马特式”的智能增强路径，将关注点转移到“如何强化人机互动技术”问题。威诺格拉德还与哲学家费尔南多·弗洛雷斯(Fernando Flores)合作完成《理解计算机和认知：设计新基础》(*Understanding Computers and Cognition: a New Foundation for Design*)一书，对当时的人工智能

乐观派进行了批判。

四、AI界对强AI理念的“集体叛逃”及哲学社会科学界对强AI理念的持续批判(1980年代初—1990年代初)

20世纪80年代初,以专家系统的成功商业化为标志,AI迎来了自己的第二个黄金期(1980—1987)。① 1970年代后期,由于AI界没有为军事部门提供可用的新技术,使得美国国防高级研究计划局的资助大量缩减,AI研究举步维艰。这时AI研究者不得不转向企业界寻求资助,或是建立独立的AI公司推进人工智能研究。这一时期,“专家系统”研究开始兴起,并于80年代中期形成了一个约10亿美元的产业,涌现出了Carnegie Group(1983)、IntelliCorp(1980)、Teknowledge(1981)、Inference(1979)等知名专家系统公司,产品广泛应用于工业制造、医疗、航空等领域。20世纪80年代末,世界500强企业中有一半在应用和发展专家系统,当时AI重镇如MIT、斯坦福大学、卡内基-梅隆大学等均开展了专家系统相关研究。

专家系统的目标是为了解决应用问题,它将特定领域中的专业知识和经验符号化为知识库,建构相应的问题匹配规则,通过问答方式来解决用户的专业问题。专家系统虽然实现了特定领域AI对普通人(甚至专家)的超越,且一定程度上在知识拥有量、分析精准度及效率等方面实现了对人类专家的超越,但专家系统从一开始便是“应用型AI”而非“强AI”,其目的是辅助人类决策,而不是实现对人类思维能力的综合性超越。专家系统的兴起,表明AI界逐渐放弃“强AI”观念,聚焦于“在特定领域中的可用性”这一较低目标,其技术文化基础是“辅助人类”及“获得商业层面的成功”,意味着“应用型AI”派相对于“强AI”派的胜利,代表着AI界对传统麦卡锡式“强AI”理念的一种集体“叛逃”。

然而,即使是“应用型AI”,其黄金期也没有持续太久,很快便进入“第二次寒冬”。“总的来说,AI业的繁荣,从1980年几百万美元,到1988年的数十亿美元。然后不久,便进了一个被称为AI冬天的时期”,[6]24它被认为是过于昂贵、无法“自主学习”、维护成本高、不具备通用性等。1987年,美国信息技术处理办公室(Information Processing Technology Office)主管史瓦兹(Jack Schwarz)将专家系统讽刺为“聪明的编程”,并“全面、残酷地”地砍掉了AI资助,强烈认为AI不是“下一波”。此后,AI逐渐被个人电脑所终结。人们不再期望计算机有多“聪明”、或是多么具有强AI性质,相反,人们关注的是如何使人机交互更为便捷、生动,如何降低计算机使用成本,人们对于“实用性”的追求代替了对机器“智能性”的追求。

在这一时期,即使是AI界的乐观派,也开始承认“强AI只是一种信念或信仰”,不具备现实可行性。如迈克·詹姆斯(Mike James)强调“目前关于智能的问题,还仅仅是一个信仰问题:相信人工智能研究的稳步前进将会最终造成智能机。”[20]此外,在这一时期,也有一些强AI乐观派人士认为,虽然强人工智能将来会实现,但需要等待计算性能的提升。如汉斯-莫拉维克谈道:“我相信,有一天人工智能的自下而上的研究路线,会与传统的自上而下的路线半途相遇,从而获得真实世界中的能力,以及对于推理程序来说极其困难的常识知识库。这两种方向结合在一起的时刻,会成为产生真正智能机器的所谓‘金钉子’。”[21]

在AI界逐渐放弃强AI理念时,哲学界则基于语义学对强AI理念进行了更为彻底、更为致命的批判。1980年初,哲学家约翰·塞尔提出了“中文屋”思想实验,从语义学的角度说明了“机器根本不可能具有意识”,强调“图灵测试”无法作为判别“机器智能与人类智能是否等同”的标准。塞尔指出,人工智能

① 1965年,费根鲍姆(E. A. Feigenbaum)和化学家勒德贝格(J. Lederberg)合作研制出第一个专家系统DENDRAL,20世纪70年代,一些专家系统相继研制成功,如医药专家系统MYCIN、探矿专家系统PROSPECTOR等。

不过是基于一系列语法规则进行符号处理的转换机，它无法像生物大脑一样进行语义解释，不能理解这些符号背后的意义，“计算机程序永不可能代替人心，其理由很简单：计算机程序只是语法的，而心不仅仅是语法的。心是语义的，就是说，人心不仅仅是一个形式结构，它是有内容的。”[22]塞尔的观点激怒了人工智能界，塞尔最初的关于“意向性”的文章引发了30篇反驳论文，在AI界看来，如果对“意义”进行最终溯源，其终极原因只能归之于宗教和神学，这是毫无意义的。

五、“奇点”议题的兴起(1990年代中期—2000年代中期)

到了1990年代，人工智能研究者背上了无法实现自己承诺的坏名声，他们拒绝再作出任何预言，AI界的大数研究人员避免提到任何“人类水平”的人工智能，以免被贴上“白日梦”的标签。麦卡锡曾这样描述：“对于AI共同体的其他科学家来说，如果新的一般形式主义(inventors of new general formalisms)的发明者能够以更加谨慎的方式来表达他们的希望，那么其他科学家便会感觉到一种极大的解脱。”[23]这种现象一直持续到2000年代中后期，马科夫(John Markoff)在2005年的“*New York Times*”杂志上谈道：“在(AI)低谷时，许多计算机科学家和软件工程师避免使用AI这一术语，害怕被人们视为是狂热的梦想家。”[24]

但是，强AI观念并没有消失，在哲学界、传媒界以及一些科学家，强AI理念却逐渐以“奇点”这一新概念形式而重新复兴了。1993年，文奇在一次美国国家航空航天局路易斯研究中心举办的研讨会上提交了文章“即将到来的奇点”(*The Coming Technological Singularity*)，强调技术的加速进步是当时代的主要特征，在未来，一种远超人类智慧的机器物种将会产生：某种先进的电脑会“苏醒”并超越人类智能，大型计算机网络及相关用户也可能会“苏醒”并成为超人类实体，“我认为超越人类的智慧体将会在未来的30年内产生”，“我们可以将此事件称为奇点。奇点意味着旧的方法被抛弃、新的规则实体产生，奇点意味着它会以较为隐晦的方式越来越大地影响人类事务，直到这一观念成为一种常识。”[25]文奇的预言似乎很快变成了现实，1997年“深蓝”战胜了世界国际象棋冠军卡斯帕罗夫，在比赛之前，美国《新闻周刊》将此次对弈描述为“人脑的最后防线”(The brain's last stand)。卡斯帕罗夫甚至认为“深蓝”经常表现出“非常拟人的危险”(showing a very human sense of danger)，“有时候‘深蓝’就像神一样来掌控棋局”或是某种更高级的智能，[26]他甚至认为有特级大师躲在“深蓝”背后帮它挑选最佳着法。

文奇的“奇点”理论提出后，并没有立即获得人们的广泛认同，只是得到一部分人的支持，如库兹韦尔。库兹韦尔于1990年出版《智能机器时代》(*The Age of Intelligent Machines*)一书，认为随着计算机性能的不断提升，未来经过足够多的时间，人类将会创造比他自身更聪明的实体[27]。库兹威尔的推断是基于计算能力的指数级增长及技术的加速循环规则。

(1)计算能力的指数增长。“奇点”理念是基于计算能力万能论的，许多持“奇点”观念的学者将人类的智能还原为计算能力，并强调随着计算能力的提升，人工智能最终将超越人脑。而20世纪90年代以来，摩尔定律不断得以证实，AI界所面临的“原始的计算能力”问题正在逐步被克服，摩尔定律下计算能力的指数级增长状态，展示了AI不断接近人脑的潜力。① “深蓝”发展史完整地展现了计算能力的指数增长、使机器智能不断增强并战胜人类的图景：①1985年，ChipTest(深蓝前身)每秒计算5万步；②1988年，ChipTest升级为“深思”，每秒计算500000个局面，击败丹麦特级大师拉尔森；③1989年，“深思”每秒

① 库兹威尔总结了1971—2011年摩尔定律在计算机领域的实际表现情况：每美元动态RAM存储能力的翻倍期为1.5年；晶体管平均价格的减半期为1.6年；每个晶体管周期的微处理器成本(Microprocessor Cost per Transistor Cycle)的减半期为1.1年；总位运(Total Bits Shipped)的翻倍期为1.1年；处理器性能(MIPS)的翻倍期为1.8年；英特尔微处理器中晶体管数量的翻倍期为2.0年。

计算 200 万步，但 0:2 输给了当时的世界冠军卡斯帕罗夫；④1990 年，“深思 2.0”问世，在与前世界冠军卡尔波夫对局中非输即和；⑤1996 年，“深思”改进为“深蓝”，每秒可以处理 1 亿个局面，但在与卡斯帕罗夫对弈以 2:4 失利；⑥1997 年，“深蓝 2.0” 以 3.5:2.5 击败了卡斯帕罗夫，每秒可分析 2 亿个棋步。

(2)技术的加速循环。1999 年的《心灵机器时代》(*The Age of Spiritual Machines*)一书，库兹威尔首次提出“加速循环规则”(The Law of Accelerating Returns)：各种形式进化系统的变化率呈现出指数增长态势，人类正处于技术加速变化的浪尖上。在 2001 年一篇文章中，[28]库兹威尔对这一规则进行了更加详细的解释：一是技术进化呈现正反馈模式，进化过程的一个阶段用以创造下一个阶段；二是在另一个正反馈回路中，当一个特定的进化过程(如计算)变得更有效时，便会有更多的资源布署到这一领域，并导致第二个水平的指数增长；三是一个特定的技术范式潜力耗尽会发生范式转移，使指数增长继续。库兹威尔预测说，未来几年里随着可负担得起的计算工具的发展，将实现人类智力水平机器的创造，[29]“技术点正在临近”，“这一进程最终变得如此之快，以至于它会使我们难以跟上其发展步伐，它最终会摆脱我们的控制。”[30]

虽然一些科学家并不认为“奇点”的存在，如 MIT 的诺姆·乔姆斯基认为，我们离建立人类水平的机器智能还“遥不可及”，称奇点是”科幻小说”。但是，“奇点”概念还是获得了硅谷一些科技企业家的关注，如谷歌的拉里·佩奇(LarryPage)和谢尔盖·布林(Sergey Brin)、支付宝联合创始人彼得·泰尔(Peter-Thiel)，他们的参与推动了社会各界对“奇点”的关注。2000 年，尤德考斯基(Eliezer Yudkowsky)等创建了“奇点研究所”，是目前唯一致力于研究强 AI 及其风险的研究组织。

应该说，在这一时期，“强 AI”观念在 AI 界几乎被抛弃，为了继续推动类似研究，AI 界往往为他们的工作贴上其他标签，如信息学(informatics)、机器学习(machine learning)、知识系统(knowledge—based systems)等，实际上，这是一种基于“应用式 AI”理念的发展路径。在“应用式 AI”理念的“掩护”下，AI 研究取得了许多成功，他们所提供的解决方案在技术工业领域被证明是有效的，如数据挖掘、工业机器人、物流、语音识别、银行软件、医疗诊断、搜索引擎(如谷歌)。但是，AI 界却没有为这些成功而获得太多荣誉，因为在这一时期，AI 的成功更多的源自于计算能力的提升，许多 AI 的重大创新仅仅被降低为工具箱，许多 AI 领域的前沿技术已经渗透到普遍的应用过程之中，它们常常没有被冠以人工智能这一名号，因为一旦有些应用变得足够有用且足够普遍，它便不再被贴上 AI 标签。

六、“强 AI 风险论”成为社会焦点之一(2010 年左右至今)

进入 2010 年以后，人工智能技术获得了更为迅速的突破，如 IBM 的 Watson 战胜 Jeopardy 冠军(2011)、聊天程序“尤金·古斯特曼”(Eugene Goostman)首次“通过”了图灵测试(2014)；2016 年，谷歌的围棋程序 AlphaGo 击败了围棋世界冠军李世石，2017 年 10 月 AlphaGo Zero 以 100:0 的战绩击败了 AlphaGo；2014 年，包括百度、腾讯、谷歌的人脸识别系统均已超过人类。

人工智能最近几年所取得的巨大突破，使得人们愈发关注强 AI 议题，并形成了一股强 AI 争论热潮。在这次争论热潮中，人们不再仅仅关注“人工智能能否超越人类”，而是更多地将“强 AI”与“风险”概念结合在一起，基于风险语境来反思“强 AI 是否会导致机器对人类的控制”、“强 AI 是否会成为一种影响人类生存与发展的危险要素”等问题。如 2014 年，霍金(Stephen Hawking)在观看了《超验骇客》后，在《独立》杂志中谈道：成功制造出一台人工智能机器人将是人类历史上的里程碑，但不幸的是，它也可能会成为我们历史上最后的一个里程碑。[31]在麻省理工学院航空航天系 2014 百年纪念座谈会上，伊隆·马斯克表示，“随着人工智能发展，我们正在召唤恶魔。”DeepMind 联合创始人及研究总监莱格(Shane Legg)也强调，AI“将会是本世纪的第一风险”。[32]未来生活研究所研究人员阿姆斯特朗(Stuart Armstrong)认为，如

果AI变坏,那么95%的人类将会被杀死,而剩下的5%也会很快被灭绝。[33]在这一时期,关于强AI风险的讨论出现了以下新特点。

(1)大量关于AI控制、威胁人类的书籍、论文出现。如穆豪瑟尔(Luke Muehlhauser)的《面对智能爆炸》(2013)、波斯特罗姆的《超级智能:路径、方法及策略》(2014)、亚姆波尔基(Roman Yampolskiy)的《人工超级智能:一个未来学研究路径》(2015)等。哲学家波斯特罗姆对未来强AI夺权的场景进行了详细分析:①前临界阶段:产生种子人工智能(seed AI),其发展初期需要人类程序员引导其发展以及完成多数繁重的工作;②递归性自我改良阶段:在某个时间点,种子人工智能会变得比人类程序员更擅长设计人工智能;③秘密准备阶段:种子AI策划出一套为了实现其长期目标的稳健计划;④公开实施阶段:通过各种手段(如控制银行,说服或威胁政府等)来控制社会资源,以服务于AI自身的目标。

(2)对强AI风险的讨论进入组织化阶段,大量AI风险研究组织建立。2012年成立的剑桥大学生存风险研究中心(Centre for the Study of Existential Risk),主要研究能够导致人类灭绝的风险,特别是人工智能所可能导致的未来风险场景;2014年成立的未来生活研究所(Future of Life Institute),主要集中于如何使AI技术以有益于人类的方式发展这一问题;2015年成立的负责任机器人基金会(Foundation for Responsible Robots,FRR),是唯一的致力于负责任机器人研究的基金会;2016年成立的Leverhulme智能未来研究中心(The Leverhulme Centre for the Future of Intelligence),则是确保人类能够在未来能够最大程度地利用人工智能技术。

(3)不再拘泥于从理念层面来讨论AI是否会超越人类,而是从实践角度来分析如何治理AI风险。如:①企业家资助AI风险研究。马斯克在2015年1月及7月分别向"未来生命研究所"投资1000万、700万美元,用于支持人工智能负面效应研究项目;②面向全社会发布AI风险及伦理公开信。未来生活研究所于2015年1月的"优先发展稳健且有益的AI:一封公开信",强调应避免AI威胁到人类生存;③企业建立AI伦理委员会,确保AI发展能够符合人类道德。如2017年10月3日,DeepMind宣布成立人工智能伦理与社会部门(DeepMind Ethics& Society);微软于2018年1月成立AI伦理委员会AETHER,避免企业滥用或过度发展AI。此外,2016年9月,Google、Facebook、Amazon、IBM、Microsoft等企业联合起来形成AI联盟,命名为"有益于人类和社会的AI联盟"(Partnership on Artificial Intelligence to Benefit People and Society),强调确保AI尽可能地有益于人类、形成最佳的人工智能实践;④公众参与。如剑桥大学生存风险研究中心于2015年6月19日举办公开演讲活动"人类水平的人工智能:是隐现还是海市蜃楼?"⑤非政府组织运动。如2013年4月成立的"杀手机器人禁令运动"。

(4)虽然隔阂尤在,但是哲学社会科学界、科学界(如物理学家等)、AI界开始进行较为广泛的对话,共同参与到强AI风险相关讨论及风险治理活动之中。在这一时期,虽然哲学社会科学界、AI界、科学界就"强AI是否会实现"、"如何使人类与机器之间形成良性互动关系"等问题存在不同意见,但一个明显的趋势是,不同的利益相关者(特别是哲学社会科学界与AI界)不再固步自封、相互攻讦,而是尝试性地进行互动。如2015年1月11日,未来生活研究所发布"一封关于使AI的社会效益最大化的公开信"(*An open letter on maximizing the societal benefits of AI*),其中目的是如何使AI研究能够更好地服务人类福祉。截止到2016年10月11日,已有8749个人在这封公开信上签字,在前110名签名者中,约70%是计算机科学家和AI专家,16%为哲学家、未来学家、伦理学家,9%为物理学家、生物学家等非AI界研究者;而在一些AI风险研究组织中,哲学社会科学家、AI界人士、其他科学家也往往都参与其中。如未来生活研究所AI研究分支创立者包括:Skype联合创始人Jaan Tallinn,MIT物理学教授Max Tegmark,DeepMind AI安全专家Viktoriya Krakovna,信息物理学教授Anthony Aguirre,哲学研究人员Meia Chita-Tegmark,牛津大学哲学教授Nick Bostrom,MIT物理学教授Alan Guth,剑桥大学理论天文学家Stephen Hawking,艾伦脑科学研究所首席科学家Christof Koch,企业家Elon Musk,IBM人工智能专家Francesca

Rossi,加州大学伯克利分校计算机科学家 Stuart Russell 等。可以看出,未来生活研究所 AI 研究小组成员涵盖了哲学家、AI 专家、物理学家、企业家等,是一个跨学科、领域研究团队。

(5)强 AI 逐渐走入"政策之屋"。首先应该说明的是,整体上讲,决策者对于强 AI 的态度一般比较保守,且当前世界各国的 AI 政策文本重弱 AI 治理、轻强 AI 调控,但许多政策及法律文本也往往涉及"强 AI"。如欧盟《就机器人民事法律规则向欧盟委员会提出立法建议的报告草案》指出:在未来,人工智能超越人类智慧是存在可能性的,强 AI 应成为人类的补充而不是替代,其 2016 年发布《欧盟机器人民事法律规则》,建议"一旦技术进步使得机器人自主性程度高于当前的合理性预测,那么相关的立法更新便需要适时推出";IEEE 的《"人工智能设计的伦理准则"白皮书(第 2 版)》强调,当系统接近并超过通用人工智能时,无法意料的或无意的系统行为将变得越来越危险且难以纠正;"艾斯罗马人工智能 23 定律"也强调,强 AI 应符合人类的整体公共利益,超级智能的开发是为了服务广泛认可的伦理观念,任何自主系统都必须允许人类中断其活动;美国所发布的《为未来人工智能做好准备》(*Preparing for the Future of Artificial Intelligence*)报告,则强调 AI 应与人类有效合作,其行动必须与人类的价值观和愿望保持一致。

七、结论与展望

在人工智能发展史上,"强 AI"一直是一个争议不断但却又不断引发人们关注的议题,围绕"强 AI"所发的争论,是人们对不同时期 AI 技术自身所内含的局限性及其潜力之间的不同聚焦,同时亦反映了科学文化、人文文化以及主流世界观等因素之间的冲突与协调过程。对于"强 AI"理念,我们应合理看待其所发的各种争论,具体来讲有以下几个方面需要加强共识。

(1)"强 AI"理念是推动人工智能界不断打破人机界限、使 AI 技术向前发展的重要信念。在很长一段时间里,"强人工智能"是人工智能界的官方表征语言及追求目标,"它们试图建造用来展现一定程度的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……如果你认为自己不知道有谁支持强人工智能的话,你就没有深入地看待这个问题。"[34]虽然在人工智能发展初期,一些 AI 乐观主义人士向资助者、传媒界做出了许多浮夸式承诺,目标的好高骛远与承诺的无法实现最终导致资助者的厌弃。但是,我们应该正视"强 AI"理念在人工智能发展史上(特别是在其发展初期)的作用和地位:它是一种信仰,推动 AI 界不断开拓 AI 学科疆域、使机器拥有更多人脑才具有的功能,开创了人工智能领域中的大多部分奠基性成果,如数学问题解决能力的 Logic Theorist(1956)、Lisp 语言(1958)、第一个自然语言对话程序 ELIZA(1965)、首个专家知识系统 Dendral(1965)、具有感知和运动能力的智能机器人 Shakey(1969)、具有自然语言理解能力的 SHRDLU 机器人系统(1971)等。正如一个新的范式刚形成时需要借助信仰、说服等力量来推动其发展一样,人工智能在其初创期,强 AI 理念对于扩展人工智能学科的社会影响力、获得资助、吸引人才是尤为重要的。

(2)围绕"强 AI"的争论是一种文化冲突现象,其背后所反映的是不同社会要素,特别是"两种文化"(科学文化和人文文化)及"科技是生产力、竞争力"的主流世界观等之间的张力,且这几种文化也经历着由对立、冲突到逐渐的尝试性对话与合作这一过程。从"强 AI"的整个争论过程来看,AI 界更多是在人工智能发展初期介入进来,主要讨论的是如何使人类知识符号化、形式化等科学问题,而 20 世纪 80 年代以后,AI 界已较少介入"强 AI"议题;"强 AI"议题的参与者,自 20 世纪 70 年代以后主要是哲学社会科学界,他们更多的是基于批判式思维,从未来学、道德哲学等人文视角考虑强 AI 的性质及其与社会的关系问题,如意识的本源是什么,机器的物理性特征与人的生物性特征区分标准,AI 在未来社会治理体系中的属性及地位是什么等。这是两种文化之间的张力。这种张力在人工智能发展初期阶段至 20 世纪 80 年代,往往是一种"紧张与对立"状态,如 20 世纪 70 年代哲学家德福雷斯与赫伯特·西蒙等 AI 专家之间

就“AI能否达到人脑功能水平”这一问题进行了长期争论；自20世纪80年代以来“应用式AI”的崛起、“强AI”理念的衰落，AI界逐渐脱离这一争论场，专注于AI在特定领域中的商业化成功，“如何使AI更好地融入市场”、“科技如何服务于生产力”等观念成为主流，所以在很长一段时期内，强AI观念逐渐被“智能增强”观念所取代；而20世纪末起，随着人工神经网络范式的迅速发展及大量应用，“强AI”理念再次引发哲学人文社科界及AI界的关注，AI界也参与到这一论争过程之中，且哲学人文社会科学界与AI界开始出现“融合与对话”的趋势，反思“AI的伦理、风险问题”，开放治理已经成为一种潮流。“强AI”参与者及讨论议题的变化，说明当代科学的发展需要从多视角进行反思，未来的科学治理体系需要社会科学家集团的崛起（如阿里巴巴设立“罗汉堂”），[35]当AI以越来越快的速度、越来越广的范围形塑社会并逐渐成为一种非可控体时，便需要人文社会科学家的介入。

（3）一个有意思的现象是：在对待“强AI”这一议题上，AI界、哲学界等的态度并非一成不变，而是出现了“态度转换”现象。从整体上讲，AI界对于“强AI”的态度经历了一个由乐观与支持到悲观与放弃，以及现在的谨慎心态这一整体转变过程。如在AI初建期，AI界如赫伯特·西蒙、约翰·麦卡锡等，均对强AI的实现非常乐观；1970年代中后期至1980年代，AI界对于“强AI”开始出现质疑声音，一部分AI界人士开始转向智能增强阵营；20世纪90年代至21世纪初，AI界基本上不再触及“强AI”议题，即使到现在，其态度依然谨慎。而哲学界对于“强AI”的态度则是沿着由批判与质疑到现在的大力宣扬这一路径演变。从20世纪60年代AI学科创立之初，到20世纪80年代、90年代中期，哲学界对于“强AI的实现”一直持质疑态度；而20世纪90年后期开始，哲学界则开始转变态度，更加预言强AI能够实现，并更为担忧其所可能产生的严重后果。为何会出现“态度转换”现象？其根本性原因在于AI界及哲学界对AI的关注点有所差异：当强调其未来潜力时，便对强AI持乐观态度；当集中于AI的局限性时，便对强AI持怀疑及批判态度。

（4）“强AI”概念需要进行重新界定，使其成为科学而非单纯的“科幻式”概念，并给予强AI风险议题更多关注。传统的强AI观念强调AI拥有与人脑完全相同的功能（我们称之为强AI_1），然而，这样的“强AI”并不具备科学层面的可行性（至少是现在），因为诸如“理解”“意志”“情感”等概念，是以生物人为前提、基于社会实践而获得的，除非人类能够建造出一个在生物层面与人脑完全相同的机器，否则便会陷入“中文屋困境”。在这里，我们强调一种基于“自主性”的强AI观：即基于强大学习能力、具备“目标自主性＋行动自主性”能力的AI（我们将之称为强AI_2）。在这里需要指出的是，“自主性”并非是基于“意向性”，它意指AI具有“目标的自我微调、改变，新目标的生成”、“自动完成已设定目标而无需外界干扰”、“可与外部环境交互并自我进化”等功能。AI_2舍弃了诸如“机器是否会具有意向性”等难以追寻或回答的终极问题，将关注点放之于当下的技术条件及未来的技术可能性。实际上，即使是当前，具备“自主性”特征的AI已经开始显露迹象，如2017年5月，谷歌提出自动机器学习（AutoML），允许人工智能成为另一个人的架构师，并在无需人工工程师输入的情况下进行自我创造，也即意味着AI在“自主生成目标”方面已有了突破。近几年来，AI技术发展所呈现出的加速进步态势，使得未来具有跨领域性、完全自主性、递归式自我改良性特征的AI的出现具有了现实可能性，这要求人们更加重视AI社会风险问题。[36]

参考文献：

[1]SEARLE J R. The rediscovery of the mind[M]. Cambridge: MIT Press, 1992: 201.

[2]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): 4-7.

[3]蔡恒进. 超级智能不可承受之重——暗无限及其风险规避[J]. 山东科技大学学报(社会科学版), 2018(2): 15.

[4]TURING A M. Computing machinery and intelligence[J]. Mind, 1950(59): 433-460.

[5]MCCORDUCK P. Machines who think: a personal inquiry into the history and prospects of artificial intelligence[M]. Na-

tick:A. K. Peters,2004.
[6]RUSSELL S,NORVIG P. Artificial intelligence:a modern approach[M]. London:Pearson Education,2003.
[7]WIENER N. Some moral and technical consequences of automation[J]. Science,1960(131):1356.
[8]SAMUEL A. Some moral and technical consequences of automation—a refutation[J]. Science,1960(132):741-742.
[9]史南飞. 对人工智能的道德忧思[J]. 求索,2000(6):69.
[10]HUBERT D. Alchemy and artificial intelligence[M]. Santa Monica:RAND Corporation,1965:3-10.
[11]侯世达. 哥德尔、艾舍尔、巴赫:集异壁之大成[M]. 郭维德,译. 北京:商务印书馆,1997:75.
[12]LUCAS J R. Minds,machines and gödel[J]. Philosophy,1961(XXVI):112-127.
[13]休伯特. 德雷福斯. 计算机不能做什么:人工智能的极限[M]. 宁青岩,译. 北京:生活·读书·新知三联书店,1986:230.
[14]CREVIER D. AI:the tumultuous history of the search for artificial intelligence[M]. New York:Basic Books,1993.
[15]WEIZENBAUM J. Computer power and human reason[M]. New York:W H Freeman & Co,1976:8-10.
[16]约翰·马尔科夫. 与机器人共舞[M]. 郭雪,译. 杭州: 浙江人民出版社,2015.
[17]MCCARTHY J. Review of the Lighthill report(1993)[EB/OL]. [2018-07-11]. http://www-formal. stanford. edu/jmc/reviews/lighthill/lighthill. html.
[18]ENGELBART D. Augmenting human intellect:a conceptual framework. SRI summary report[EB/OL]. 1962,https://web. stanford. edu/dept/SUL/library/extra4/sloan/mousesite/EngelbartPapers/B5_F18_ConceptFrameworkInd. html.
[19]ENGLEBART D. The augmented knowledge workshop[C]. Proceedings of the ACM conference on the history of personal workstations. 1986,http://www. dougengelbart. org/pubs/augment-101931. html#Pub-63-Frame.
[20]JAMES M. Artificial intelligence in basic[M]. London:Butterworths & Co,1984:120.
[21]MORAVEC H. Mind children[M]. Cambridge:Harvard University Press,1988:20.
[22]约翰·塞尔. 心、脑和科学[M]. 杨音莱,译. 上海:上海译文出版社,1991:23.
[23]MCCARTHY J. Formalizing common sense:papers by John McCarthy[M]. New York:Ablex Publishing Corporation,1990:69.
[24]MARKOFF J. Behind artificial intelligence,a squadron of bright real people[N]. The New York Times,2005-10-14(A10).
[25]VINGE V. The coming technological singularity(1993)[EB/OL]. [2018-02-11]. http://www. frc. ri. cmu. edu/~hpm/book98/co m. ch1/vinge. singularity. html.
[26]LEVY S. What Deep Blue tells us about AI in 2017[EB/OL]. Wired,2017-05-23.
[27]KURZWEIL R. The age of intelligent machines[M]. Cambridge:MIT Press,1990:21.
[28]TEUSCHER C. Alan Turing:life and legacy of a great thinker[M]. Berlin:Springer,2010:381-416.
[29]KURZWEIL R. The age of spiritual machines:when computers exceed human intelligence[M]. NY:Penguin Books,1999:33-35.
[30]KURZWEIL R. The law of accelerating returns(2001)[EB/OL]. [2017-11-02]. http://www. kurzweilai. net/the-law-of-accelerating-returns.
[31]GRIFFIN A. Stephen Hawking:artificial intelligence could wipe out humanity when it gets too clever as humans will be like ants[EB/OL]. Independent,2015-10--08.
[32]SHEAD S. The biggest mystery in AI right now is the ethics board that Google set up after buying DeepMind [EB/OL]. Business insider,2016-03-26.
[33]BRYANT M. Artificial intelligence could kill us all,meet the man who takes that risk seriously[EB/OL]. Insider,2014-03-08.
[34]PRESTON J,BISHOP M. Views into the Chinese room:essays on Searle and artificial intelligence[M]. Oxford,UK:Clarendon Press,2002:14-14.
[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,2007:202-206.
[36]王彦雨. 学界关于"超级 AI"的论争及其实现的可能路径[J]. 未来与发展,2017(08):30.

(下转第 35 页)

From the Asilomar Conference to the International Summit on Human Genome Editing: the Role and Limitation of the Expert Precaution in Biotechnology Governance

GAO Lu

(*Institute for the History of Natural Sciences, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: The history, significance and related debate of two "consensus conferences", that is, the Asilomar Conference and the International Summit on Human Genome Editing, both organized by the scientists and focusing on the uncertainty of recombinant DNA technology and genome editing technology respectively, are discussed for an attempt to probe into the role and limitation of expert precaution in new biotechnology governance. The Asilomar Conference in 1975 on the potential biohazards of recombinant DNA was seen as the beginning of precaution principle of biotechnology governance. However, the risks and interests set by the Asilomar Conference were too narrow, which limited the scope of biotechnology governance. The International Summit on Human Genome Editing held in 2015 discussed the science, ethics and governance issues concerning human genome editing, and took the advice from a broader global community. However, it still distributed the complicated global governance challenge into different countries and regions, even the enterprises, which inevitably led to fragmented governance. With the development of biotechnology and its industrialization, the expert precaution model in the governance of gene editing technology was only a start of a long going procedure of discussion and negotiation. Governance could never accomplish at one stroke, instead, it was a slow, dynamic process of agreement making. In the post genome era, to understand the social transformations in social relationship, ethics, and laws caused by the technology is the key to understand the technology in itself, which could lead to a good governance of emerging technology.

Key words: Asilomar Conference; International Summit on Human Genome Editing; expert precaution; biotechnology governance; gene editing; huge risk of science and technology

(责任编辑:黄仕军)

(上接第 27 页)

Research on Strong Artificial Intelligence Debate from the Historical Perspective

WANG Yanyu

(*Institute for the History of Natural Sciences, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: Throughout the history of artificial intelligence(AI), "strong AI" has always been a controversial concept which constantly arouses people's concern. For the idea of "strong AI", we are expected to take a reasonable view of the various arguments it has caused: (1) The idea of "strong AI" is an important force to drive AI technology to break the boundary between humans and machines and develop forward; (2) The controversy of "strong AI" reflects the tension between different social factors, including the "two cultures" (scientific culture and humanistic culture), the mainstream world outlook, i. e. science and technology are the competitiveness and productivity and so on. Besides, these cultures in AI circle go through a process from opposition to conflict, and finally to tentative dialogue and cooperation. (3) The attitude of AI circle and philosophical circle towards "strong AI" is changing: AI circle was initially optimistic and supportive, and then pessimistic and ignoring, and is now finally cautious; philosophical circle towards "strong AI" was critical and suspicious, and now advocates vigorously, moreover, it is more concerned about the risk that strong AI may cause. (4) The concept of "strong AI" needs to be redefined to make it a scientific concept rather than a simple one of "science fiction". Besides, more attention should be paid to the "strong AI" risk.

Key words: strong AI; debate on "the evil and the good of machine intelligence"; debate between optimism and pessimism; concept of "intelligence augmentation"; "singularity" theory; "theory of strong AI risk"; huge risk of science and technology

(责任编辑:黄仕军)