机器意识最新进展的哲学反思

The Philosophical Reflections on the Latest Developments of Machine Consciousness

游均 /YOU Jun　周昌乐 /ZHOU Changle

（厦门大学哲学系，福建厦门，361005）
(Department of Philosophy, Xiamen University, Xiamen, Fujian, 361005)

摘　要：机器意识研究在近十年取得了瞩目的成就，尤其在脑智功能技术实现方面取得了长足的进步，但相应地，其在意识觉知机制方面却依然没有起色。本文分析了机器意识研究在可行性、判定标准和解释上所遭遇到的困难，并指出了出路所在。

关键词：机器意识　脑智外现　觉知内显

Abstract: In the past ten years, the research on machine consciousness has made considerable progress in the function simulation technology of explicit intelligence, but relatively speaking, the research on the awareness mechanism is insufficient. This essay has analyzed the triple dilemma of machine consciousness research, and points out the solution to the dilemma.

Key Words: Machine consciousness; Explicit intelligence; Inexplicit awareness

中图分类号：N0　　**文献标识码：**A　　DOI:10.15994/j.1000-0763.2018.06.015

近十年，随着脑科学研究和人工智能研究的不断突破，通过机器实现意识能力的相关研究渐渐引起人们关注，这项研究通常被称为“机器意识”（Machine Consciousness）或是“人工意识”（Artificial Consciousness）。机器意识研究不仅推动并深化了人工智能方面的研究，在探索神秘的意识现象方面也有着卓越的贡献。就发展状况来说，加梅斯（D. Gamez）回顾了截止2008年的机器意识的研究进展，[1]随后在2013年，瑞杰（J. Reggia）再次对当年的机器意识研究做出了精彩的总结。[2]近几年，机器意识在技术上又有突破性进展，佼佼者如AlphaGo，已能在围棋领域彻底战胜人类的世界冠军。因而，本文将对机器意识的最新进展做简略性回顾。

一、机器意识的最新进展

近十年的机器意识研究，在实现方式上，往往采用编程算法或是神经网络模拟的方法，在实现方向上，则主要有机器单独实现和脑机融合实现两大方向。

1. 机器单独实现

机器单独实现即只依靠机器自身来实现脑智特征能力。在其表现形式方面，除了传统计算机所拥有的的思维、计算能力之外，同时也偏重于语言能力、想象能力、情感能力和自我反思能力等脑智特征的实现。亚历山大（I. Aleksander）等人甚至提出将表述能力、想象能力、注意能力、计划能力和情感能力五项特征作为测试机器是否具有意识的特征公理。[3]具体而言，（1）在情感表现方面，皮卡德（R. W. Picard）、派珀德（S. Papert）、本德（W. Bender）等人对MIT媒体实验室的各种情感机器人研究作出过总结。[4]胡得利卡（E. Hudlicka）提出了关于情感的计算模型。[5]

基金项目：国家自然科学基金项目“针对涉身行为的自我意识机器人构建方法及其实现”（项目编号：61273338）。
收稿时间：2017年3月30日
作者简介：游　均（1986–）男，福建福州人，厦门大学人文学院博士研究生，主要研究方向为意识哲学、科学哲学。Email: yjcfss@qq.com
周昌乐（1959–）男，江苏太仓人，厦门大学信息学院教授，主要研究方向为机器意识及其哲学问题。Email: dozero@xmu.edu.cn

日本研究者林（E. Hayashi）和则开发了基于模拟多巴胺系统的会趋利避害的情感机器人，并给出了一些模仿行为的实现。[6]（2）在语言能力方面，安吉尔（L. Angel）早在1989年，就提出过基于语言与主体的意识机器的体系结构。[7]奥古斯汀（V. Agustin）和费诺（M. Ferno）等人则提交了一份关于内部言语的研究综述。[8]雷舍（E. Lesser）和海克能（T. Haikonen）等人给出了基于言语和感知、认知的机器意识架构。[9]斯蒂尔（L. Steel）开发了可以根据给定场景中的对象进行互相对话的机器人，[10]在其后续研究中，在没有实现预设语法程序的情况下，斯蒂尔的机器人还可以在不断对话中自行掌握语法。[11]（3）在想象能力方面，克劳斯（R. Clowes）和克里斯理（R. Chrisley）等人做出了关于机器意识的涉身和想象方面的综述。[12]斯图亚特（S. Stuart）则在躯体想象方面作出了研究，[13]他强调，要在真正意义上实现机器意识，与世界的涉身交互是必须的，因而躯体想象是必须研究的重点问题。[14]（4）在计算能力方面，DeepMind公司的AlphaGo围棋机器人通过深度学习，将蒙特卡罗搜索树（MCTS）和价值网络相结合，从而计划出在棋盘上下一个落子的位置权重，借此打败了人类棋手中的世界冠军。[15]

除了上述几个只实现了单一或少数脑智特征能力的机器人，值得注意的还有综合了情感、想象、环境认知等多方面脑智特征的机器人。如加梅斯等人的CRONOS机器人、布鲁克斯（R. Brooks）和布里吉尔（C. Breazeal）等人研发的COG机器人、以及契里亚（A. Chella）等人研发的CiceRobot，还有亚历山大等人开发的仿脑机器人核心架构。CRONOS通过躯体和环境的相关关系，建立内部模型进行环境认知；[16]COG机器人则侧重于关节注意机制、情感表现等方面的研究；[17]CiceRobot机器人可以通过想象能力与外部环境输入的视觉感知材料相比较来引导行为；[18]亚历山大的仿脑机器人核心架构基于神经表征模组Neural Representation Modeller（NRM）而运行，可以同时满足亚历山大提出的五条机器意识特征公理。[19]

2. 脑机融合实现

脑机融合研究是位于机器意识和神经科学交叉领域的一个比较重要的前沿研究方向。其兴起主要是基于以下几个方面的考量：第一，机器智能和人类意识彼此都具有对方所不擅长的优势。机器擅长于快速精确计算、海量记忆存储、快速检索信息等，而且在高速运动、飞行、深海探索和宇宙探索等对人类身体有所限制的环境下也可以进行自如行动，而人类意识则擅长快速学习、抽象想象、创造性思维等，脑机融合研究则可以将双方的优势互相结合，从而实现人类智力的进一步开发，以及让机器在某种程度上实现意识能力；第二，脑机融合研究可以为某些残障人士提供更为接近其原本器官的功能上的替代物。例如，直接链接到脑部的反馈电路使得肢体残障人士可以对机器义肢拥有和原本肢体类似的感受，也可以借助摄像摄影设备让盲人恢复部分视觉，随着研究水平的提高，甚至可以让患有癫痫、帕金森病等脑部疾病的患者更换上相关脑区的功能替代物，从而得到康复。

脑机融合的可行性主要体现在：首先，机器通常使用电信号来处理信息，而大脑的神经信号主要采用的也是电脉冲信号的方式，这一点二者存在着共同点；其次，大脑的各个区域存在功能分工，如负责语言、负责视觉、负责运动等，因而可以针对性地采集特定功能区域的神经信号并将其与人工设备进行信号对接，在掌握其具体功能后，便可以研发该部位的功能替代物；最后，大脑本身具有可塑性，可塑性在将来自人工设备的信号和神经信号建立起联系的过程中起着重要的作用。

通常，脑机融合主要分为两个方向，一是从机到脑，即通过各种人工设备产生电子信号来刺激大脑，以传输某些特殊的感觉信息，或是模拟某些特殊的神经功能，这方面的典型代表是人工耳蜗，随着其技术发展，还可以恢复盲人的视力，使机器义肢拥有触觉，以及治愈帕金森病和癫痫等。二是从脑到机，即通过大脑的原生信号来操控人工设备，这方面比较著名的应用是利用运动皮层的神经信号来实时控制机器手臂的运动，其未来可以发展为直接通过脑信号进行虚拟现实的交互，远程操控机器替身等。

在脑机接触手段上，脑机融合主要分为侵入式和非侵入式两种。侵入式指的是需要通过外科手术，直接在大脑皮层植入电极，这种方式可以最高效地采集和传输神经信号，但风险较大，目前主要以动物实验为主，非侵入式则采取风险相对较小的传统的表层信息采集技术，常用的有EEG

（脑电图）、fMRI（核磁共振功能成像）、MEG（脑磁图）、NIRS（近红外光谱）、PET（正电子成像术）等。

在研究文献方面，巴赫（P. Bach）、克赛尔（S. W. Kercel）等人对于大脑可塑性和感觉替代方面做了相关研究。[20] 尼尔斯（B. Niels）、科恩（L. G. Cohen）针对非侵入式的脑机融合所使用的各项技术在中风患者和肌肉萎缩症的患者的临床应用方面也做了详细的研究。[21] 克里斯坦（H. Christian）和汤加（S. Tanja）在他们对于神经信号的语音识别技术所做的综述中评述分析了不同的脑成像技术使用自动语音识别技术来识别神经信号中语音的潜力。[22] 值得注意的是列别德夫（M. A. Lebedev）和尼克列利斯（M. A. L. Nicolelis）等人的研究，他们对脑机结合做了比较全面的综述，并且给出了分类路径图和操作原则。[23]

在具体的技术实现方面上，蔡平（J. K. Chapin）等人用人工神经网络算法将实验鼠运动皮层神经集群电信号转换为按压水泵的机械臂控制指令，首次实现了大脑对外部设备的直接控制。[24] 霍赫贝格（L. R. Hochberg），巴彻（D. Bacher）等人则成功地使得因中风而四肢瘫痪的患者通过意念控制完成了通过机械臂抓取杯子的喝水行为。[25] 类似的，梵思汀赛尔（M. J. Vansteensel）等人的团队成功地使得患有ALS（Amyotrophic Lateral Sclerosis）晚期肌萎缩侧索硬化的荷兰女患者使用意念进行拼写，从而实现对外交流。[26] 卡波格罗索（M. Capogrosso）最近发表在《*Nature*》上的研究更是成功地将神经信号解码并且通过机械装置中转后对脊椎发放信号从而直接让猕猴瘫痪的后肢恢复运动能力。[27]

二、机器意识的困难

尽管机器意识研究的成果斐然，上述研究中的机器人在各方面表现得越来越接近于人类，甚至在某些方面超越了人类的水准，但当前的机器意识研究依然有着三个难以回避的困难。

1. 机器意识可行性的怀疑

由于意识的神秘性和高度复杂性，关于意识运行的具体机制众说纷纭，仍未有一个可以通过科学实验证实的定论。因而，宣称机器可以实现人类同等意识的机器意识研究在诞生之初就遭遇到了可行性上的怀疑。人们对可行性的怀疑包括了两个方面——理论层面的可行性和技术层面的可行性。在理论层面的可行性方面，机器意识难以回避这样的怀疑[28]：我们的神经生物系统是特殊的，意识只能通过神经系统产生，因而机器无法产生意识。这种怀疑基于这样一个事实：迄今为止，我们尚未发现在生物神经系统之外能产生意识的存在物，有可能我们就处于必须通过神经系统才能实现意识的可能世界。如此一来，通过机器实现意识也就成了无源之水，无本之木，从根本上就不可能实现。而在技术层面的可行性方面，机器意识的工程师们指出：“大部分现有的意识理论往往来自哲学或是心理学，并不提供关于有意识的存在是什么，以及意识是如何在机器中产生的解释。他们只是提供或多或少的关于意识的隐喻式描述，而不是能直接通过计算术语实现的模型。”[29] 然而，试图通过计算术语实现意识的前提有三个，第一，必须将意识形式化；第二，这个形式化必须是可计算的，得有合适的算法；第三，这个算法必须有着合理的复杂度。[30] 而就目前的技术水平而言，这三个条件都无法满足。首先，就形式化方面来说，目前机器意识所流行的实现方法都是基于神经网络或是符号编程的，雷多闻（M. Radovan）早在1997年就已经证明，神经网络的这种神经连接主义方法的表达能力与传统的符号逻辑主义方法是等价的，[31] 而对于符号逻辑主义方法，卡普兰（C. Caplain）也已经证明其不可能描述意识现象，[32] 换言之，如果不突破目前逻辑系统的范式，我们可能始终都无法通过计算术语实现真正的意识；其次，目前的计算算法也无法通过计算术语来完整地描述意识现象，仅仅是实现视觉的最佳算法就会被一个像素干扰从而识别错误；[33] 最后，人脑约由10^{12}个神经元组成，而每个神经元都有大约10^3个突触，而意识正是由这些数量巨大且相互连接的神经细胞之间不可预测的，非线性作用的结果。如果通过计算机来模拟，每个突触需要4字节的内存空间，总计就需要4×10^{15}字节的空间，但目前的科技水平还远远未达到这个硬件条件，[34] 而且随着计算量的增加，计算复杂度也将呈现指数级的增长，随之而来的能耗将大大超出我们所能供应的程度。因此，机器意识在技术的可行性上也具有相当大的难度。

2. 机器意识判定标准的困难

就算机器意识的两方面可行性都顺利达成，在判定机器是否真正实现了意识能力方面，也存在着问题。由于意识的第一人称的私密性，除了觉知主体，他者是无法直观地觉知到意识是否产生的，通常采用的方法是从外部言行、输入输出等方面进行推断，这种行为主义和功能主义的判定方式遭到众多学者的反对。他们认为我们无法仅仅通过言语和行为就判定机器具有意识，赛尔（J. Searle）就通过中文屋的思想实验指出，就算中文屋可以在外部表现地和你对答如流，但其却根本没有理解中文的真正含义。[35] 而查尔默斯（D. Chalmers）关于哲学行尸（zombie）的思想实验进一步地强调了——功能主义和行为主义判定方式无法区别外表言行相似，却没有内部觉知体验的行尸和真正有意识的个体。[36] 因此仅凭借外部言行是无法判断机器是否真的实现意识的。

鉴于功能主义与行为主义判定方式的这一弊端，霍兰德（O. Holland）[37] 和塞斯（A. Seth）[38] 将意识的判定标准分为两个等级，强人工意识和弱人工意识，弱人工意识仅仅致力于外部行为表现，或是达成相对应功能的输入和输出，并不要求机器具备真正的意识，而只有实现与人类同等意义上的意识，即主体觉知意义上的意识能力，才能被称为强人工意识，换言之，我们可以将弱人工意识相对应于脑智外现，强人工意识相对应于觉知内显。如此一来，不难看出，上述机器意识研究的成果，基本上都是脑智外现的研究，而鲜有觉知内显的研究。值得注意的是脑机融合研究，其中的从脑到机方面的研究只需要读取相应的脑电特征的输入输出来操控机器，依然是属于脑智外现方面的研究，并不涉及觉知内显，因而直接对应于弱人工意识。而从机到脑，并产生相应感受方面的研究，就目前来说，虽然是通过机器让我们产生了相关的主体感受，但这种产生感受的机制依然是借用人类大脑的原生机制，研究者既没有探究机器如何独立产生主观体验，也没有借此探究人类的觉知机制，而只是止步于我们感受体验的输入端，因而并不属于强人工意识，也要归入弱人工意识。

如此一来，我们就可以确定机器意识的界限。普林茨（J. Prinz）指出，外部言行表现是意识判定的必要条件，觉知则构成了意识判定的充分条件，机器意识可以满足意识判定必要条件，但当前我们没有任何手段确定机器意识是否满足意识判定的充分条件。[39] 目前学界针对机器意识脑智外现的诸多判定标准比较直观清晰，也容易获得认可，这些构成了弱人工意识的下确界。就是说，目前有一类明确的判定标准来区分人工意识和非人工意识，满足这些判定标准的都可以归为人工意识，起码是弱人工意识。而相对应的，满足意识判定的充分条件的觉知机制则构成了弱人工意识的上界，同时也是弱人工意识和强人工意识的分界岭。但困难之处在于，我们可以从定性的角度找到弱人工意识和强人工意识的分界，但却无法从定量的角度找到这个分界的确界。即是说，我们目前无法给出一系列定量的、明确的判定标准来区分弱人工意识和强人工意识，其原因在于，意识觉知在其本质上是通过第一人称视角直接把握到的，而第一人称视角的验证是无法定量的，所有得到公认的定量的标准都必须经过第三人称视角的验证。由此一来，就形成了一个两难：我们要么放弃通过第三人称视角的方式来判定机器的意识觉知是否实现，仅保留第一人称视角来验证的方式，要么保留第三人称视角的验证方式，从而放弃给出强人工意识和弱人工意识之间的确界标准。这就是当前机器意识判定标准的困难所在。

3. 机器意识中的解释鸿沟

即便工程师们成功构造出了有意识的机器，在解释为何这种构造机制会产生意识的方面，也存在着困难，查尔默斯（D. Chalmers）称之为意识的难问题：“我们不只是要知道哪些过程引起了经验，我们还必须看到关于为什么和是怎样的说明。”（[40]，p.373）哈尔纳德（S. Harnad）同样表示，就算我们完成了有关意识的所有正向和逆向工程，我们也仅仅只是掌握了意识的相关运行机制，并未真正解释意识是如何产生的，他认为“就算所有的认知科学方面的工作都完成了，机器中的幽灵却依然困扰着我们。”[41] 可以说，仅仅通过物理方面的解释对于意识是不充分的，工程师和神经科学家们可能在无数次实验中发现与意识相关的物理构造，但这种构造无法给予意识以充分的说明。莱布尼茨（G. W. Leibniz）就曾指出，若我们走进有意识的机器的内部，所能看到的也就只是相互作用的零部件，其中并没有可以解释意识的东西。[42] 用列文（J. Levine）的话来说，在物理的神经过程与心灵的意识现象之间存在着难

以逾越的解释鸿沟（The Explanatory Gap）。[43]麦吉恩（C. McGinn）则采用了一个更为诗意的比喻："物质脑之水以某种方式转化为意识之酒，但我们对这种转化的本性却一无所知……所谓心–身问题就是理解关于这个奇迹是如何发生的问题。"[44]

三、机器意识的出路

1. 寻求技术上的创新

即便存在着种种困难，但总体而言，这些困难并不足以阻挡机器意识研究的进程。神经科学和人工智能科学的新进展正逐渐使得机器意识的诞生具备越来越高的可能，主要体现在：首先，大脑采用有规律的电信号/化学信号作为信息传递的载体及信号运算的物理手段，这一点机器也可以通过对电信号编码来进行模拟。卡波格罗索的研究就是通过电信号模拟神经信号的典型。其次，神经科学发现大脑各皮层存在着模块化的功能特征，通常来说，前额叶负责理性思维，原始脑负责情感处理。具体而言，单就语言功能就在大脑中就分别有S区（运动语言中枢），W区（书写语言中枢）、V区（视觉语言中枢）、H区（听觉语言中枢）等几个模块，其中，S区受损的表现为听得懂也看得懂语言但无法说话，W区受损表现为听得懂也看得懂但无法写字，V区受损表现为看不懂文字但是听得懂，H区受损的表现为看得懂文字，也能读写就是听不懂。更为重要的在于，不仅脑智能是模块化的，意识觉知也是如此。例如我们视觉觉知的过程，就是从视网膜转化信号，一路经过初级视觉区（V1区）、纹外皮层（V2–V5）等脑区进行的，在这一过程中，任何一个模块发生问题，都可能导致最终无法产生视觉觉知，初级视皮层受损的患者会出现盲视现象，纹外皮层受损则可能出现偏盲现象。这种模块化处理信息的特点在机器系统中亦是非常普遍与常见的。随着对意识的研究的深入，人们也可能在理解觉知机制后研发出意识觉知模块，从而在真正意义上实现和人类感同心受的机器意识。最后，量子计算机和意识的量子模型的相关研究使得机器可以突破传统计算机和传统逻辑模型的种种限制，进一步符合意识的原生生成机制。

2. 寻求解释上的革新

不难发现，意识本身的私密性和神秘性使得人们对意识的第三人称的观察和第一人称的体验难以有机地结合在一个统一的解释中，人们对于意识的认识，在第一人称视角和第三人称视角之间存在着一种缺失。侯世达（D. R. Hofstadter）在其名著《歌德尔、艾舍尔、巴赫》中将其归纳为"为阐明大脑中发生的思维过程，我们还剩下两个基本问题：一个是解释低层次的神经发射通讯是如何导致高层次符号激活通讯的；另一个是自足地解释高层次的符号激活通讯——建立一个不涉及低层神经事件的事论。"[45]

换言之，在当前对意识的解释中，起码存在着如下两种解释框架：

A. 物理状态/神经状态→意识状态

B. 觉知结构→意识状态

其中，我们把解释A看作是因果解释，而把解释B看作是生成解释。对解释A来说，也许在物理状态中我们把握到了某些意识得以产生的重要的相关参数，但仅凭这些物理参数仍然无法解释为何意识会得以产生，物理状态和意识状态之间存在着一条难以逾越的解释鸿沟。相应地，解释B则可以充分地说明意识状态的产生，查尔默斯就在其对意识难问题的思考中指出"用来解释觉知的过程，就是意识之基础的组成部分。"（[40], p.386）而我们所需要的，一个科学的，令人信服的解释理论则是一个沟通物理状态、觉知结构、意识状态三者的解释理论，这样的解释理论必须满足解释框架C：

C. 物理状态→觉知结构→意识状态

对于解释C而言，通过构造觉知结构作为物理状态与意识状态的中介，弥补了直接通过物理状态生硬地解释意识状态所造成的解释鸿沟，物理状态在功能上构成了觉知结构，觉知结构则生成了意识状态，如此便能弥补从物理状态到意识状态的解释鸿沟。一个意识的合理解释应该是按照物理状态–觉知结构–意识状态三者对应的方式表达的。在这一点上，麦吉恩也有着类似的洞见，他指出，在意识状态和物理状态之间存在着一种隐藏结构："我设想的这种隐藏结构不会处于内格尔所认为的那两个层次：它大概位于两者之间的地方。这个中间层既非现象学的，也非物理性的，它不会按照这一分裂的任何一方的模型来塑造，因此也就不会发现自己无法通到另一边。它的刻画要求概念的彻底革新"[46]同时，这种方式也为

判定标准的困难提供了一种出路。尽管无法直接把握到机器人的第一人称的觉知内容，我们依然可以通过寻找其所对应的第三人称相关物来解决，这个第三人称相关物不同于弱人工意识通常采用的外部言行标准，而是物理状态——觉知结构——意识状态的一体化理论中，与觉知结构相对应的物理状态。

但需要注意的是，根据多重可实现理论，任何具备意识觉知结构的系统都将拥有对应的意识，而却可以有多种不同的物理状态对应到觉知结构中来。因此，解释C也就可以看作是两个部分的组合：

解释C1：物理状态——觉知结构

解释C2：觉知结构——意识状态

其中，解释C1是因果解释，解释C2是生成解释，C1中的物理状态既可以是神经细胞的神经状态，也可以是机器的计算状态，在形成觉知结构这一点上，神经状态或是计算状态在多重可实现理论中是等价的。而承认多重可实现假说的这一论断，正是承认机器实现意识的理论前提。综上所述，就当前研究来说，我们可以根据对解释C1和解释C2两部分的侧重不同，分为侧重C1的信息加工解释进路和侧重C2的现象结构解释进路。信息加工解释进路基于因果性解释，主要致力于说明意识发生过程中的因果关系。而现象结构解释进路则基于结构性解释，其强调意识觉知内容的结构构成。相较而言，信息加工解释进路有利于机器实现，但缺乏针对现象方面的深入探究——对其来说，在提供了关于信息加工的某些细节之后，意识现象就突然出现了，而这一出现的机制并没有真正被信息加工解释进路所阐明。而相对应的，现象结构解释进路则从意识觉知的结构着手，说明了意识觉知是如何从这种结构之中产生的，但其往往缺乏对应这种结构的信息加工的因果过程的说明，因而难以在机器之中得以实现。

3. 寻求研究上的整合

我们的意识是一个复杂庞大的整体，而研究者们为了研究方便，往往针对意识的不同方面进行各自的研究，因此，将意识的不同方面的研究进行整合，也是机器意识研究的出路所在。

根据上文，我们已经将机器意识研究分为了脑智外现研究和觉知内显研究，其中，脑智外现研究分别从情感、语言、想象、计算、认知等多方面展开了机器实现，而觉知内显研究则分别从信息加工解释进路和现象结构解释进路进行了探索性的研究。信息加工进路中比较有代表性的有巴尔斯（B. J. Baars）提出、德阿纳（S. Dehaene）与尚热（J. P. Changeux）以及沙纳瀚（M. Shanahan）等人不断发展完善的全局工作空间理论及其模型；海客能（P. O. Haikonen）的机器意识理论；托诺尼（G. Tononi）的信息集成理论所罗门（A. Sloman）的机器意识理论等，而现象结构进路中著名的有阿姆斯特朗（D. Armstrong）和利康（W. Lycan）等人的高阶感知理论；卫斯伯格（J. Weisberg）和罗森赛尔（D. M. Rosenthal）以及卡拉瑟斯（P. Carruthers）等人的高阶思维理论；克里格尔（U. Kriegel）、金那罗（R. Gennaro）等人的同阶理论；玄奘、窥基等人的唯识论等。

因此，机器意识的研究整合也可分为三部分：

D1脑智外现研究和觉知内显研究的整合

D2脑智外现研究中各个具体表现方面的整合

D3觉知内显研究中信息加工进路和现象结构进路的整合

针对D1来说，当前的研究主流相对侧重于脑智外现的机器实现，觉知内显机制的研究则相对薄弱地多。针对这种工程师们往往只关注脑智外现的机器实现的现状，亚历山大不无担忧地表示：“完全使用人工智能那种功能的方法的那些人至少必须解释在什么程度上它们的模型可以说包括了一个现象的世界，不然他们的工作就不能被视为是对机器意识的目标做出贡献”。[47]因此，在当前研究的基础上，整合脑智外现研究和觉知内显研究也就成为机器意识研究所必须要达成的目标。在这一方面，巴尔斯、德阿纳与尚热、沙纳瀚以及所罗门、海客能等人都提出了各自的整合模型。但这些模型全都基于信息加工进路，并没有针对现象结构做出深入研究，其理论都还有待补充与加强。除此之外，脑机融合技术的发展也使得这方面的整合有着广阔的前景。

对D2而言，我们的意识可以从事多种复杂的任务，因此那些只能从事简单任务的机器人很难被认同其拥有意识。因此，复杂环境下处理多种任务的综合机器人也就应运而生，诸如加梅斯等人的CRONOS机器人、布鲁克斯等人的COG机器人、以及切里亚等人的CiceRobot等。

至于D3，通过整合C1和C2进路，使其能具

体解释在呈现现象结构的过程中发生了哪些功能机制的因果关系变化，这是真正触及到众人皆感棘手的意识觉知的核心问题，需要同时整合第一人称视角方法和第三人称视角方法，在目前也尚未有人对此方面做出过令人信服的研究。

综上所述，在机器意识的研究工作中，需要注意加强脑智外现研究中各项表现能力的整合，并以实现信息加工进路和现象结构进路的整合为主要突破口，最终实现脑智外现和觉知内显的整合，只有这样才能够找到机器意识的真正出路。

［参 考 文 献］

[1] Gamez, D. 'Progress in Machine Consciousness'[J]. *Consciousness & Cognition*, 2008, 17(3): 887–910.

[2] Reggia, J. A. 'The Rise of Machine Consciousness: Studying Consciousness with Computational Models'[J]. *Neural Networks*, 2013, 44(8): 112–131.

[3] Aleksander, I., Dunmall, B. 'Axioms and Tests for the Presence of Minimal Consciousness in Agents I: Preamble'[J]. *Journal of Consciousness Studies*, 2003, 10(4–5): 7–18.

[4] Picard, R. W., Papert, S., Bender, W., et al. 'Affective Learning—A Manifesto'[J]. *Bt Technology Journal*, 2004, 22(4): 253–269.

[5] Hudlicka, E. 'Challenges in Developing Computational Models of Emotion and Consciousness'[J]. *International Journal of Machine Consciousness*, 2012, 1(1): 131–153.

[6] Hayashi, E., Shimono, M. 'Design of Robotic Behavior that Imitates Animal Consciousness'[J]. *Artificial Life & Robotics*, 2008, 13(1): 203–208.

[7] Angel, L. *How to Build a Conscious Machine*[M]. Boulder: Westview Press, 1989, 320–322.

[8] Agustin, V., Ferno, M. 'Inner Speech: Nature and Functions'[J]. *Philosophy Compass*, 2011, 6(3): 209–219.

[9] Lesser, E., Schaeps, T. Haikonen, P. et al. 'Associative Neural Networks for Machine Consciousness: Improving Existing AI Technologies'[A], *IEEE Convention of Electrical & Electronics Engineers in Israel*[C], New York: IEEE, 2008, 11–15.

[10] Steels, L. 'Language Games for Autonomous Robots'[J]. *IEEE Intelligent Systems*, 2001, 16(5): 16–22.

[11] Steels, L. 'Language Re-Entrance and the 'Inner Voice''[J]. *Journal of Consciousness Studies*, 2002, 10(4–5): 173–185.

[12] Clowes, R., Torrance, S. Chrisley, R. 'Machine Consciousness: Embodiment and Imagination'[J]. *Journal of Consciousness Studies*, 2007, 14(7): 7–14.

[13] Stuart, S. A. J. 'Machine Consciousness: Cognitive and Kinaesthetic Imagination'[J]. *Journal of Consciousness Studies*, 2007, 14(7): 141–153(13).

[14] Stuart, S. A. J. 'Conscious Machines: Memory, Melody and Muscular Imagination'[J]. *Phenomenology & the Cognitive Sciences*, 2009, 9(1): 37–51.

[15] Chouard, T. The Go Files: AI Computer Clinches Victory against Go Champion[EB/OL]. https://www.nature.com/news/the-go-files-ai-computer-clinches-victory-against-go-champion-1.19553. 2016–03–12.

[16] Gamez, D., Newcombe, R., Holland, O., et al. 'Two Simulation Tools for Biologically Inspired Virtual Robotics'[A], *Proceedings of the IEEE 5th Chapter Conference on Advances in Cybernetic Systems*[C], New York: IEEE, 2006, 85–90.

[17] Brooks, R. A., Breazeal, C., Marjanović, M., et al. *The Cog Project: Building a Humanoid Robot*[M]. Computation for Metaphors, Analogy, and Agents. Berlin: Springer–Verlag, 1998, 52–87.

[18] Chella, A., Liotta, M., Macaluso, I. 'CiceRobot: a Cognitive Robot for Interactive Museum Tours'[J]. *Industrial Robot*, 2007, 34(6): 503–511.

[19] Aleksander, I. 'Why Axiomatic Models of Being Conscious?'[J]. *Journal of Consciousness Studies*, 2007, 14(7): 15–27.

[20] Bach-Y-Rita, P., Kercel, S. W. 'Sensory Substitution and the Human–Machine Interface'[J]. *Trends in Cognitive Sciences*, 2003, 7(12): 541–516.

[21] Niels, B., Cohen, L. G. 'Brain-Computer Interfaces: Communication and Restoration of Movement in Paralysis'[J]. *The Journal of Physiology*, 2007, 579(3): 621–636.

[22] Christian, H., Tanja, S. 'Automatic Speech Recognition from Neural Signals: A Focused Review'[J]. *Frontiers in Neuroscience*, 2016, 10(429): 1–7.

[23] Lebedev, M. A., Nicolelis, M. A. L. 'Brain–Machine Interfaces: Past, Present and Future'[J]. *Trends in Neurosciences*, 2006, 29(9): 536–546.

[24] Chapin, J. K., Moxon, K. A., Markowitz, R. S., et al. 'Real-Time Control of a Robot Arm Using Simultaneously Recorded Neurons in The Motor Cortex'[J]. *Nature Neuroscience*, 1999, 2(7): 664–670.

[25] Hochberg, L. R., Bacher, D., Jarosiewicz, B., et al. 'Reach

and Grasp By People With Tetraplegia Using a Neurally Controlled Robotic Arm'[J]. *Nature*, 2012, 485(7398): 372–375.

[26]Vansteensel, M. J., Pels, E. G., Bleichner, M. G., et al. 'Fully Implanted Brain-Computer Interface in a Locked-In Patient With ALS'[J]. *New England Journal of Medicine*, 2016, 375(21): 2060–2066.

[27]Capogrosso, M., Milekovic, T., Borton, D., et al. 'A Brain-Spine Interface Alleviating Gait Deficits After Spinal Cord Injury in Primates'[J]. *Nature*, 2016, 539(7628): 284–288.

[28]Kiverstein, J. 'Could A Robot Have A Subjective Point Of View?'[J]. *Journal of Consciousness Studies*, 2006, 14(7): 127–139.

[29]Arrabales, R., Ledezma, A., Sanchis, A., 'CERA-CRANIUM: A Test Bed for Machine Consciousness Research'[J]. *International Workshop on Machine Consciousness, Towards a Science of Consciousness*, 2009, 1: 1–20.

[30]德雷福斯. 计算机不能做什么[M]. 宁春岩译, 上海: 三联书店, 1986, 4–7.

[31]Radovan, M. *Computation and understanding*[M]. Mind Versus Computer. IOS Press, 1997, 211–223.

[32]Caplain, G. *Is Consciousness a Computational Property?*[M]. Mind Versus Computer. IOS Press, 1997, 190–194.

[33]Su, J., Vargas, D. V., Kouichi, S. 'One Pixel Attack for Fooling Deep Neural Networks'[J/OL]. https://arxiv.org/abs/1710.08864. 2018–02–22.

[34]Buttazzo, G. 'Artificial Consciousness: Hazardous Questions (And Answers)'[J]. *Artificial Intelligence in Medicine*, 2008, 44(2): 143.

[35]Searle, J. 'Minds, Brains and Programs'[J]. *Behavioral & Brain Sciences*, 1980, 3(3): 417–424

[36]Chalmers, D. J. *The conscious mind: in search of a fundamental theory*[M]. Oxford: Oxford University Press, 1997, 251.

[37]Holland, O., Goodman, R. 'Robots With Internal Models A Route to Machine Consciousness?'[J]. *Journal of Consciousness Studies*, 2002, 10(4–5): 77–109.

[38]Seth, A. 'The Strength of Weak Artificial Consciousness'[J]. *International Journal of Machine Consciousness*, 2012, 1(1): 71–82.

[39]Prinz, J. 'Level-Headed Mysterianism and Artificial Experience'[J]. *Journal of Consciousness Studies*, 2003, 10(4–5): 111–132.

[40]高新民、储昭华. 心灵哲学[M]. 北京: 商务印书馆, 2002, 373–386.

[41]Harnad, S. 'Can a Machine Be Conscious? How?'[J]. *Journal of Consciousness Studies*, 2003, 10: 69–75.

[42]莱布尼茨. 神义论[M]. 朱雁冰译, 上海: 三联书店, 2007, 479.

[43]Levine, J. 'Materialism and Qualia: The Explanatory Gap'[J]. *Pacific Philosophical Quarterly*, 1983, 64(4): 354–361.

[44]McGinn, C. 'Can We Solve the Mind-Body Problem?'[J]. *Mind*, 1989, 98(391): 349.

[45]侯世达. 歌德尔、艾舍尔、巴赫: 集异璧之大成[M]. 本书翻译组译, 北京: 商务印书馆, 1997, 467.

[46]丹尼尔·丹尼特. 意识的解释[M]. 苏德超、李涤非、陈虎平译, 北京: 北京理工大学出版社, 2008, 498.

[47]Aleksander, I. 'Designing Conscious Systems'[J]. *Cognitive Computation*, 2009, 1(1): 22–28.

[责任编辑 李斌 赵超]