

作为人工智能下一个关口的意识研究

——从加扎尼加的意识学说切入

李恒威 阮泽楠

【摘要】随着近年来人工智能(AI)的惊人进步,AI是否具有意识以及如何构建有意识的AI系统越来越受到学界和公众的关注。要回答这类问题,我们仍然需要或者必须从意识理论中寻找灵感和答案。立足AI意识问题的当前背景,我们考察了加扎尼加在《意识本能:揭秘脑是如何形成心智的》中提出的意识学说。透过脑异常这个窗口,加扎尼加得到了意识现象学的两个重要论断:意识与体验内容不同并且是可以分开的;意识首先是一种感受。为此,他提出了一个关于意识神经机制的模块—层级架构的主张,并用“开水冒泡”隐喻来形象地说明他的意识观。根据这一思路,构建有意识的AI系统需要我们在底层逻辑上辨明“意识”与“智能”的差异和关系,否则,我们就会在AI系统是否有意识以及AI是否会造成对人类的生存性威胁等一系列理论问题上产生困扰。

【关键词】人工智能;加扎尼加;脑异常;模块—层级架构;“开水冒泡”隐喻;意识

【作者简介】李恒威,浙江大学哲学学院教授,研究方向:意识认知科学、认识科学与东方心学比较、人工智能;阮泽楠,浙江大学哲学学院博士生,专业方向:认知科学哲学、意识科学(杭州 310058)。

【原文出处】《社会科学战线》(长春),2024.3.16~25

【基金项目】国家社会科学基金重大项目(22&ZD034);国家社会科学基金项目(20BZX045);科技部科技创新2030“脑科学与人类脑研究”重大项目(2021ZD0200409)。

引言

人工智能(AI)的惊人进步,尤其是在智能水平、通用性和自主性(autonomy)上的颠覆性表现,促使学界和公众越来越关心诸如AI的下一个关口等AI未来发展的深层问题。其中一个关键问题是:AI最终是否会具有意识,以及人类能否创造出有意识的AI?前者是一个理论问题,后者则是一个工程问题。围绕这两个问题,当前主要有三种观点。

第一种观点。在最近发布的一篇名为《人工智能中的意识:来自意识科学的启示》(Consciousness in Artificial Intelligence: Insights from the Science of Consciousness)长篇论文中,图灵奖得主约书亚·本吉奥(Yoshua Bengio)联合来自哲学、神经科学以及人工智能三大领域的19位学者组建跨界团队,系统回顾了当代主流的意识理论以及构建有意识的人工智能

系统的可能性。^①

他们认为,让AI拥有意识在科学上是可行的。他们从当代意识理论中提炼出了若干意识的指标(indicators)属性,并以此作为评估AI具有意识的标准。其基本观点是,尽管目前还不能确切地判定任何一个AI系统真的具有意识,但通过现有技术可以在AI系统中实现许多意识的指标属性。促使本吉奥等人对当前AI模型最终能否发展出意识能力持积极态度的原因在于,他们从当代意识理论中提炼出的指标属性在计算功能主义的意义上确实是能够逐步在AI系统中实现的。因此,他们设想的工程策略非常直接,就是不断地增加系统功能以使其满足具有意识的相关指标属性。事实上,在如何让机器拥有意识的问题上,全局神经工作空间理论(global neural workspace theory, GNWT)的倡导者斯坦尼斯拉斯·迪

昂(Stanislas Dehaene)就提出过,当我们将正常成年人脑特有的两种信息计算功能——全局广播C1和自我监控C2——赋予机器时,其行为表现就好像它是有意识的。^②

第二种观点。与秉持计算功能主义判定AI系统是否有意识的立场不同,以整合信息理论(integrated information theory, IIT)为代表的一类观点首先将意识的关键锚定为意识体验的主观方面或现象意识(phenomenal consciousness)。IIT认为,意识体验的主观性或现象性(phenomenality)是由系统的内在因果力决定的,而不是由外在观察者看到的功能表现所决定的。^③IIT的支持者克里斯托夫·科赫(Christof Koch)强调意识与智能(intelligence)是不同的,他将意识与智能看作两个正交的维度:当前AI系统有很高的智能,但却没有意识;与之相反,一个基于生物工程的类脑器官会具有一定程度的意识,但却没有智能。^④

第三种观点。认知神经科学家安东尼奥·达马西奥(Antonio Damasio)、雅克·潘克赛普(Jaak Panksepp)等人主张从生命演化的角度来看待和审视意识。他们认为生命与心智具有连续性,唯有与现实生命活动联系起来,属于心智范畴的诸多概念才能被深入理解。达马西奥和潘克赛普都强调感受之于意识的核心和基础地位。达马西奥指出,“内稳态感受(homeostatic feelings)本身自发和自动地就是有意识的”^⑤,潘克赛普认为,意识并不取决于高级认知功能(或高级智能),而是产生自更基础的情感心智(affective mind)和感受。^⑥达马西奥认为,AI的一些先驱为了走捷径采用了更为“经济”的方式,即他们仅仅致力于让机器模拟更容易客观地加以评估的智能,而回避了那些让他们觉得过于主观从而难以科学介入的情绪感受。但依据心智的生命观,^⑦恰恰是那些根植于生命自我调节的情感要素才是AI实现意识的关键。因此,要让AI系统拥有意识,首先要构建出该系统的“内稳态感受”,而“为了做到这一点需要为机器人提供一个满足存续所需规范和调整的身体”^⑧。

总的看来,要回答“AI是否具有意识以及如何构建有意识的AI”这类问题,我们仍然需要或者必须从意识理论中寻找灵感和答案。正因如此,我们注意到“认知神经科学之父”加扎尼加(Michael S. Gazzaniga)

在其专论意识的著作《意识本能:揭秘脑是如何形成心智的》中提出的意识学说。^⑨本文将讨论加扎尼加意识学说的关键思想,即他关于意识的现象学和神经机制——模块—层级架构。加扎尼加的意识学说将有助于澄清“AI是否具有意识以及如何构建有意识的AI”这类问题。

一、脑异常与意识现象学

通常,机器的故障反而能让人们更好地认识它的工作原理。同样,脑异常(脑损伤和脑疾病)也是人们间接窥视意识现象学以及意识现象学表现与脑的神经结构—动力学关系的一个极佳窗口。透过脑损伤和脑疾病这个窗口,加扎尼加得到的一个基本观察是,意识是有韧性的(resilient),它很难被彻底抹除。由于意识在神经上的韧性(resilience),使得脑异常与意识现象学之间存在复杂的关系,而加扎尼加关于意识的本质、功能和机制的学说在很大程度上源于对这种复杂关系的认识。

1. 意识与体验内容是有区别的

脑异常案例给加扎尼加带来的一个关键启示是,不能把意识与体验内容等同起来。换言之,不同的脑异常会给患者带来各种异乎寻常的体验内容,但意识一直持存着(persisting)。加扎尼加举了很多病例来说明这个区别,以及这个区别对理解意识的本质和构建意识的机制的重要性。

阿尔茨海默病会缓慢地在全脑各处造成严重的神经损伤,例如当内嗅皮层和海马神经元发生病变后,患者会出现短时记忆丧失。随着病情的加重,一个人在知、情、意三个方面的能力都会严重下降,患者神情木讷,寡言少语,言辞不清,行动迟缓,甚至认不出自己至亲的人。尽管患者“体验到的世界越来越局限,体验的内容也很古怪,与过去正常时自我体验到的完全不同”,但只要他的心脏还在跳动,“尽管内容如棋局的盘面一样变化不定,但他的意识仍然会在受退化蚕食的脑中幸存下来”^⑩。

精神分裂症患者存在感知、思维、情感、意志等多方面的障碍,其中最突出的症状是幻想(hallucination)和妄想(delusion),例如,患者总觉得别人在背后指点和监视自己。加扎尼加在书中描述了一个被称为“B先生”的精神分裂症患者,他总认为自己正被联邦调

查局(FBI)特别关注,自己每时每刻的行为都被监视,而且联邦调查局还将他每日的生活录制为“B先生秀”的节目向公众播放。尽管B先生对世界的体验内容与正常人相去甚远,但这就是他真实的主观世界,并且他总是努力在认知上做一番解释,以使他的体验世界变得合情合理。援引精神分裂症,加扎尼加想提出的观点是:尽管B先生的脑功能出现了异常(他的体验内容与正常人相去甚远),但B先生的意识并没有消失,他依然能够觉知到自身的存在。

人类意识体验的内容很大部分是语言性的。我们现在知道,人类的语言能力依赖特定的生物结构,语言的学习、理解和表达由不同的脑区负责,不同脑区的损伤会导致不同的失语症。例如,威尔尼克区(语言的听觉中枢)受损会导致感觉性失语,患者以口语理解障碍为突出特点,对他人和自己的言语均不能理解,严重时答非所问,但患者自发语言呈流利性,常滔滔不绝地独自说个不停,却用词混乱、语法紊乱,使人无法理解。尽管不同语言区的损伤会导致不同的意识体验,但没有一个会摧毁意识本身。再如狼孩,毫无疑问他具有意识,也有意识体验,但他未习得语言时的体验内容与假设他学会说话时的体验内容将非常不同。

菲尼亚斯·盖奇(Phineas Gage)可能是脑损伤案例中最有名的。他的左侧前额叶在一次铁路施工现场的爆炸事故中因一根钢钎穿过头骨而遭到严重损伤。在事故后他的个性发生巨大变化,他过去对待同事那种富有同理心的人格已经被暴躁好斗的意识体验取代,但他的意识毫无疑问地还在。“不管额叶综合症患者情绪失控的潜在原因是什么,有一个事实在所有病例中都是一致的:患者仍然有意识。”^⑩

2. 意识与感受

脑异常案例给加扎尼加带来的另一个关键启示是,意识首先是一种感受(feeling)。事实上,让意识成为“难问题”的正是感受,也正因为感受,哲学和科学才不得不面对如芒在背的“主观性”。意识研究领域提倡、接受和支持这个观点的人不在少数,像内格尔(Thomas Nagel)、查默斯(David Chalmers)、托诺尼(Giulio Tononi)、科赫、汉弗莱(Nicholas Humphrey)、潘克赛普、达马西奥、索姆斯(Mark Solms)等都是典型。例

如,达马西奥将感受置于意识(乃至心智)的中心地位,认为“感受是意识之源”^⑪，“归根结底,意识是作为一种感受开始的,它固然是一种特殊感受,但仍然是一种感受”^⑫。

长期以来,存在着这样一种信念,即认为所有形式的意识都来源于大脑皮层,没有大脑皮层,人们不仅没有意识,而且也无法在任何层面上产生意识。但脑异常病例提供了质疑和反对这种信念的证据。先天无脑畸形儿患者或积水性无脑畸形儿患者因为遗传和发育异常或疾病导致大脑皮层缺失或只残余少量大脑皮层。对这些患儿的研究发现,这类孩子“不仅是清醒的,时常表现出警觉性,而且能对周围环境做出情绪反应,或是对环境事件做出有针对性的反应”^⑬。与这些患儿熟悉的人尽管一眼就能看出患儿与大脑皮层健全的孩子之间的差别,但关键是,他们从不否认这些患儿有意识,能对刺激做出适当的反应,能与正常的孩子一样体验种种感受,诸如疼痛、冷暖、快乐。

加扎尼加援引积水性无脑畸形的病例等证据是要证明,皮层下结构可以独立地产生有意识的主观感受,大脑皮层对这类原始的主观感受来说并不是必需的,但大脑皮层却是一系列增强认知能力和丰富体验内容的神经基础。加扎尼加将负责情绪和情绪感受(emotional feelings)^⑭的皮层下结构比喻为意识的引擎,而大脑皮层的演化和发育则极大地扩展和丰富了体验的内容。只要支持情绪感受的“皮层下引擎”完好运作,大脑皮层中的任意部分都能够产生意识体验。

皮层下结构中的边缘系统、间脑、中脑等区域在演化上先于大脑皮层,与生命的维持和存续所涉及的情绪反应和情绪感受密切相关。“如果我们将意识看作对事物的主观体验,那么情绪必然是意识的基本组件。”^⑮据此,加扎尼加提出笛卡尔“我思,故我在”的基础是“我感受,故我在”(sentio ergo sum)^⑯:笛卡尔实际上只需借由皮层下结构产生的感受就能锚定他作为一个“我”的存在,其中可以不涉及大脑皮层的高级认知功能,不需要高度发达的概念性的“我思”就能“我在”。相比成年人的认知控制能力,额叶还未发育完全的婴幼儿的控制能力会更加依赖皮层

下结构。

二、意识的神经架构

在意识科学的时代,止步于描述各种脑异常(脑损伤和脑疾病)导致的引人入胜的意识现象学是不够的,加扎尼加的目标是透过现象学在认知神经科学上推进意识的自然科学研究。人脑中每时每刻都进行着无数电、化学和激素过程,但我们的体验却好像是一个丝滑统一的整体,脑是如何将这些过程组织起来产生出这种统一性的意识?加扎尼加的科学目标就是试图阐明精心演化的脑是如何曲尽其妙形成奇迹的,即物质是如何形成心智的。

脑异常的意识现象学是我们推断意识神经机制极富价值的线索和指引。如上所述,加扎尼加特别重视从脑异常病例中得到的意识现象学的这两个重要论断:论断1是意识与体验内容不同,并且是可以分开的,某种体验内容的异常或消失并不代表意识本身的丧失;论断2是意识首先是一种感受。托诺尼等人的整合信息理论,从意识的五条现象学公理(axiom)出发去构想能够支撑这些公理的物理系统必须满足的神经结构—动力学条件的假设,即相应的五条物理学公设(postulate)^⑧。加扎尼加依照同样的方法也构建了一个神经架构,以便解释上述两个重要论断以及一系列意识现象学。

加扎尼加认为,要充分地理解意识的本质和特点就必须同时考虑皮层下结构和大脑皮层在意识中的地位和作用。依据论断2,加扎尼加为其意识的神经架构添加的第一个结构是皮层下结构,因为单独的皮层下结构也能生产有限类型的意识体验(尤其是情绪感受)。依据论断1,加扎尼加增加的第二个结构是皮层结构,因为尽管大脑皮层不是产生意识所必需的,但只要皮层下加工过程的支持,大脑皮层就能为有机体提供类型多样的丰富的认知内容,为有机体获得更高级的心智能力(概念思维和反思)铺平道路。论断1尤其表明,体验内容不可能由脑内某个“超级中枢的”(Grand Central)单一结构产生,因此要解释体验内容的神经韧性就必须采取另一种方式来思考。为此加扎尼加认为,只有以模块—层级方式构造的脑才能解释有关意识现象学的种种脑异常的事实。在他的学说中,意识的神经架构依赖的

关键概念包括模块、层级(或分层)、协议(protocol)、整合等。

1. 模块—层级架构

简单地说,脑是由模块构成的神经网络。其中,每个确定的模块本身也是一个专门的、局部的神经网络,负责行使某个特定功能。一方面,为实现专门的功能,相对于模块之间较为松散的连接,模块内部的连接要更为致密;另一方面,脑是由大量相对独立的运行模块形成的一个功能整体,因此,模块之间的通信和协调是至关重要的。再者,一些模块是分层的,也就是说,模块本身由亚模块组成,亚模块又由亚—亚模块构成。加扎尼加认为,从神经科学上讲,意识就是由成千上万个相对独立运作的模块相互作用形成的某种整体功能。所以,他一再强调意识不是一个事物(thing),而是一种功能(或能力),“它是一个嵌入架构中的过程的产物,就好比民主并不是一个事物,而是一个[社会体系运行]过程的结果”^⑨,是一种功能(或能力)。

对模块—层级架构而言,协议是理解其运作的一个基本概念。每个层级的模块之所以能相对独立地工作,是因为每个层级都拥有自己的工作协议。所谓“协议”,是指一组规定层级内和层级间的接口或交互作用的规则或规范明细。波音777飞机是典型的模块—层级架构的复杂系统,其中飞机座椅就是处于其中一个层级的模块。加扎尼加以座椅层级的模块为例来说明协议的特点。在建造波音777飞机时,对制造座椅模块的工程师来说,他所需要的信息只是一套有关固定于该飞机机舱内座椅的尺寸标准,而无须知道也不用关心发动机模块的尺寸标准,譬如空气动力学、合金材质、燃油等的参数。只要满足这套标准,座椅制造工程师可以尽情发挥他在设计和制造上的创意和灵活性。显然,“协议既是对该层级的约束,但也允许在那些约束范围内有灵活性”^⑩。由于模块—层级架构系统中不同层级模块的形态完全不同,因此在它们之间传输的必要信息必须针对特定的层级进行抽象,即不同层级的模块之间必须要有接口参数,例如,舱内座椅与整个机舱的接口参数主要涉及座椅的空间布局、固定强度等。在一个模块—层级架构的系统中,其层级“堆栈”

(stack)中的每一层级的模块会依据相应的协议加工从下一层模块接收输出信息,并将结果传递给上一层模块或返回给下一层模块。每个层级的模块都不“知道”前一个层级模块接收的输入信息是什么样的,也不知道前一个层级模块对信息做了哪些加工,它也不需要知道这样的信息。协议使得各层只能加工那些接收自邻层的信息。

一个模块可以单独作为一个层级,但更高层级的模块是如何产生的呢?这时我们就需要“整合”概念了。“低级功能活动部分无法单独创建‘高级’体验,但在层级架构中,新的功能活动层级可以从低的功能活动层级中产生”^③,导致新层级涌现(emergence)的就是整合。可以说整合是宇宙间最神奇的机制,原子模块因整合形成了分子模块,分子模块因整合形成细胞模块,神经细胞模块因整合形成功能脑区模块,功能脑区模块因整合形成整个脑系统。“人脑拥有很多高度整合的模块,使得我们能够将不同模块的信息结合在一起,形成抽象的思维”^④,并最终诞生有意识的、反思的心智。

2. 模块一层级架构的优势

脑为什么以模块一层级架构的方式构造和运行呢?加扎尼加在书中论述了这种架构存在的一些显著优势。首先,模块一层级架构的脑能够减少能量消耗。脑被分成各个单元,因此在完成特定任务时,只需一个给定模块中的区域保持活跃就可以了。很显然,这种模块一层级架构与一体架构,即实现任何一个功能都要调用全局相比,肯定是一种更节能的构造方式。“这好比比你夏天住在凤凰城(Phoenix),如果你在晚上只开卧室空调,就比给整个房子制冷要省钱。”^⑤其次,模块一层级架构的脑在功能上也更高效。专业分工是古今中外所有社会系统运行中的普遍现象,因为一个事务由专业人士来处理肯定会更高质高效。如果社会系统没有分工,造成的结果不仅是没有社会合作,而且每个人都被迫成为多面手,最终的后果肯定是他哪一项职责都无法胜任。同样,如果脑是一个浑然一体的单元,它就无法应对环境对有机体发出的多样的、性质不同的挑战。幸好有相对独立运行的模块分工,脑才能通过协调这些“专家”完成种种超级挑战。“比起事事每个人都做一

点,诸专家同时做带来的经济产出更高。这样看来,为了能有效地同时处理多种类型的信息,脑以模块化的方式演化就显得很合理了。”^⑥最后,模块一层级架构的脑蕴含着一个也许对有机体而言最重要的优势:灵活性。有机体能对不断变化的环境做出更快的适应或演化。原因很简单,由于模块是相对独立的,一个模块的修改、复制或升级不会危及或触动其他运转正常的模块,这种方式使得有机体在应对新问题和新的挑战时能更灵活地做出调整。同时,当系统出现故障时,也更容易辨识出问题的源头,我们不需要修理或报废整个系统,只要处理出问题的层级或层级部件即可。此外,加扎尼加认为,即使不考虑模块一层级架构之于有机体演化的优势,这种架构也有助于有机体更好地学习新技能,因为倘若每次学习一门新技能,整个脑都大范围地调整一次,那么有机体就很可能因为学习新技能而丢失了已有的技能。

总之,模块一层级架构的好处是,“它在资源稀缺时节省能量,在时间有限时允许专门的并行认知(parallel cognitive)加工,在新的生存压力出现时更容易改变功能,并允许我们学习各种新技能”^⑦。加扎尼加认为,如果我们深思其理,就会发现脑根本不可能以模块一层级架构以外的其他任何方式来进行构造。

3. 小世界网络

大胆一点,我们可以说,模块一层级架构是一个普遍的宇宙学事实。神经科学家格奥尔格·施特里特(Georg Striedter)的研究发现,模块一层级架构的脑在神经元的连接上遵循三条法则:第一,一个神经元通常连接的神经元数量不会随着脑尺寸的增加而改变,也就是说,随着脑的扩容,尽管脑的整体连接性会下降,但脑的局部加工更为独立;第二,脑的整体连接长度尽可能保持最短,这样,大多数神经元只与临近神经元相连,使得局部神经元之间能够建立高效的通信,从而自然形成一些具有专门功能的局部区域,如此一来,模块就出现了;第三,脑中神经元之间不是所有的连接都是最小化的,因为为了协调各模块完成更复杂的加工和实现功能,脑中相对独立的不同层级的模块之间需要形成通讯连接,这就出现了一些间隔较远的模块之间的“快捷”(short cut)连接。

脑依循上述法则形成的架构也就是所谓的“小世界”(small-world)架构。脑的小世界网络一方面需要模块(神经元聚类)实现专门功能,另一方面它又需要模块之间通过连接形成更高层级的模块,以实现更复杂的功能。如此一来,就出现了模块内相对致密的连接与模块间相对稀疏的连接,如图1^③。

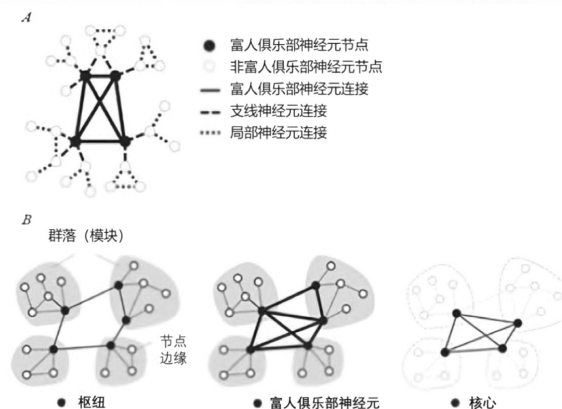


图1 小世界网络

(A)富人俱乐部神经元与非富人俱乐部神经元之间的关系。由局部互连形成的群体中的神经元可以通过与富人俱乐部通道搭上关系取得其他局部群体的信息。(B)神经网络(模块)是由致密互连的神经元形成的群体构成的。模块网络的典型特征就是存在若干神经群落。模块内部的连接往往要比模块之间的连接短。这样一来,受制于空间的模块就有助于节约与连接和沟通有关的成本。模块之间的功能整合要求增加高成本长距离的轴突投射以连接在空间上远距离的脑区,这就产生了连接器枢纽,它可以接入数量多到不成比例的长途的模块间连接,它具有一个高参与度指数,并且在网络中占据着一个在拓扑上更为中心或者“潜在瓶颈”的角色。

三、脑心演化的嵌套层级理论

在对意识本性的理解上,加扎尼加与潘克赛普的观点相当接近。潘克赛普等人从生命演化的视角提出了一个脑心(Brain Mind)或心脑(Mind Brain)^④演化的嵌套层级理论(图2^⑤)。对于哺乳动物,其脑心由三个层次的过程组成,分别是初级过程(primary-process)、次级过程(secondary process)和三级过程(tertiary process)。

初级过程是一些位于新皮层下的(sub-neocortical)脑结构,实现对哺乳动物生命状态的调控,是哺乳动物脑心中古老且普遍的价值结构,它们以原生情感(raw affect)的形式表达对环境的价值评价。哺乳动物的原生情感有三个一般类型:情绪性情感(emotional affect),源于情绪行动系统,体现为行动的驱力、动机

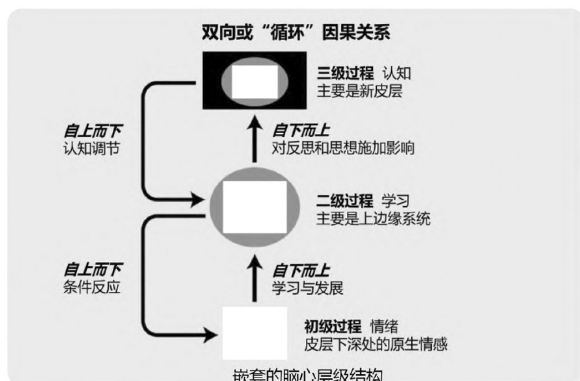


图2 脑的每一个系统中进行的自下而上和自上而下(循环)因果作用

该图概述了一种假设,即高级脑心功能要发挥作用就必须与低级脑心功能进行整合,其中初级过程被描绘成正方形,次级学习过程被描绘成圆形,在顶端的三级过程被描绘成矩形。

和意图;体内平衡情感(homeostatic affect),源于脑——身体的内感受器,表现为饥渴、口渴、性欲等;感官性情感,源自外感受感官,表现为被触发的令人愉悦或厌恶的情感。

次级过程是基于基底神经节的情绪学习和记忆机制,包括:经典条件作用(例如,通过杏仁基底外侧核和中央核实现的恐惧情绪);工具的和操作的条件作用(例如,通过伏隔核实现的探索行为);行为习惯和情绪习惯(由背侧纹状体实现,很大程度上是无意识的)。通过将原生情感与感知相结合,次级过程赋予哺乳动物广泛的条件性学习(conditioned learning)能力。大鼠对猫的铃铛声的恐惧采用的就是次级情绪过程的基本认知策略。

三级过程是基于新皮层实现的高级心智过程,新皮层是在人类身上特别发达的脑结构,它促成了丰富的认知和思想,尤其是在人类生活中极为显著的反思和慎思能力,包括:认知执行功能(诸如由额叶皮层实现的可能和规划);对情绪的反思和调节(主要涉及内侧额叶区);做出选择和采取行动的“自由意志”。^⑥

潘克赛普认为,心智存在需要区分开的两个方面,即心智的情感方面(affective aspect)与认知方面(cognitive aspect),而他的心智理论特别强调情感在心智中的基础性,“我们的观点是,理解情感对于我们理解人类本质而言至关重要”^⑦,“从根本上说,情感

是我们心理存在的基石……原生情感感受是整个心智装置的原始基础——它们是核心—自我的原初生物基质——也许是‘灵魂’概念的神经基础”^⑧。为此,潘克赛普将意识分为两种类型或形式:情感意识(affective consciousness)与认知意识(cognitive consciousness)。首先出现的是演化上较为古老的情感意识,这是一种对原生情绪感受的意识;认知意识在演化上则是一个相对的后来者,它不仅丰富了对世界的认知,还使得个体可以对自己的情绪感受进行思考。这种区分在意识研究中有相当的普遍性,例如,现象意识与通达意识(access consciousness)的区分^⑨、现象意识与认知意识的区分^⑩等。

潘克赛普提出,当初级过程这个演化上古老的系统跟划定“自我”与外部世界之间界限的这一最初的神经“身体地图”联系在一起时,主观的情感体验就诞生了。因此原生情感首先反映的是生物体与环境之间的一种价值关系或“态度”,尽管这种价值关系肯定与生物体对环境的感知和认知判断有关。原生情感本身既不聪明也不智能,但它是人类和动物的脑心共享的演化上同源的结构—功能,它为不同物种的高级心智活动提供了许多相似的生物“价值结构”和生存指引。不同哺乳动物新皮层的大小和复杂性千差万别,致使它们呈现出不同类型和水平的认知能力和智能,但是在解剖学和神经化学上与新皮层存在显著区别,深藏在大脑皮层之下的同源情感基质在所有哺乳动物中是极为相似的。由于脑的模块—层级架构既嵌套又彼此循环连接,因此在人类高度发达的意识体验——反思意识(reflective consciousness)或自知意识(auto-noetic consciousness)中,情感与各种相当抽象的言语思想和观念混为一体,以至于我们很难将意识的认知方面与情感方面剥离开,从而以最纯粹的形式来理解情感和感受,但原生情感即使在人类成人身上仍然以一种充满活力的意识形式表现出来,“大部分人都会记得他们在愤怒时握紧拳头,脸色通红,非常害怕,并且感受到深深的悲伤和喜悦”^⑪。

脑心演化的嵌套层级理论显然符合加扎尼加的模块—层级架构的思想,即模块—层级既在结构—功能上相对独立,又在结构上双向连接(循环因果关

系),从而实现结构—功能的协调和整合。在脑心的嵌套层级中,初级过程情感指引和推动次级过程的学习机制,之后,这两者又与三级过程的认知能力相结合,从而形成了尤其呈现于人类水平的心智景观。理解这一点会极大地帮助我们更全面地认识脑心复杂性。

四、“开水冒泡”隐喻

现在,我们可以结合模块—层级架构和脑心演化的嵌套层级理论来审视加扎尼加提出的有关意识的“开水冒泡”隐喻。

情感意识与认知意识的区分与加扎尼加得出的“论断2”和“论断1”正好是对应的。能够恰好同时说明这两点的脑异常病例就是加扎尼加做过精深研究的裂脑人现象。当连接两侧脑半球的胼胝体被神经外科医生切断后,一个人的脑中会出现两套意识体验。一方面,由于皮层下结构被两个脑半球所共有,因此裂脑患者有着相同的情绪驱动力和感受,这表明情感意识(感受)的基础性(论断2);另一方面,由于负责高级认知的新皮层不同,因此两套意识体验在同一时间拥有不同的认知—情感内容,这表明认知意识的神经韧性,也就是说脑损伤会造成特定的认知—情感缺陷,却无法完全消弭意识(论断1)。这使得加扎尼加认识到,“意识体验表面看上去是一个统一的整体,但实际上是协同(in concert)发生的,其中多个系统彼此并行运行,它们分别输出各自的加工结果”^⑫。再者,裂脑研究还让加扎尼加认识到,“意识不是某个使得我们所有心智事件都变得有意识的特殊网络的产物。相反,每个心智事件都由一些有能力使我们意识到它们加工结果的脑模块管理”^⑬。这些加工结果(即体验内容)由不同的模块产生,就像一壶开水中沸腾的泡泡,它们一个接一个出现,并在一段时间后破裂,被其他泡泡取代,这是一个在时间上无缝衔接的持续的动态过程。为此,加扎尼加将意识出现的过程比喻为“开水冒泡”。

通过意识如何出现的开水冒泡隐喻,加扎尼加认为,人们通常关于“意识体验是一部连贯的电影或者是一段连续的流(a continuous stream)”的直觉是一种错觉(illusion)。事实上,意识体验是由许多相

对独立的“认知”泡泡组成,通过嵌套层级中的循环连接与皮层下的“情感”泡泡紧紧地交织在一起,最终合成为一连串(a stream)的认知—情感泡泡,随着它们腾跃到表面的时间而被连接为一个看上去连续的整体。

究竟是哪个意识体验的泡泡最终冒出来呢?换言之,究竟是哪个模块的加工结果最终胜出呢?这就涉及控制的问题。“泡泡是胡乱腾跃出来的,还是某种动态控制系统的产物?是否存在一个控制层级,负责给一些泡泡放行,同时抑制其他泡泡?”^⑧某一时刻某个过程占据了你的意识高峰,其他一些过程则被挤了出去,竞争失败的信息加工不会被意识到,这部分结果只表现在无意识水平上。

对模块加工的控制受两个方面的影响。简单地说,一种是由外部刺激自下而上形成的结果,另一种是由生物体内部的选择机制自上而下造成的结果。就自下而上来说,例如,如果你吃一块糖,那么胜出的自然是负责加工甜味感觉的模块,而不可能是加工苦味感觉的模块。就自上而下来说,“从螃蟹到鸟类再到灵长动物,我们在许多动物中都发现了某种形式的选择性信号增强机制,表明这是一种被我们最后一个共同祖先所共有的能力”^⑨。之所以会出现这种信号增强机制,根本原因在于,生物体不可能关注刺激其各个感官的所有信息,而是必须有选择地考虑那些与其生存最相关的信息,最好是优先考虑那些迫切的关于危险、食物、交配的信息,而不是其他的。这种选择机制在所有生命形式中保存着。最终,这种选择性信号增强就演化为现在所说的“注意”。不论是自下而上还是自上而下的控制,模块加工结果之间的竞争和最终的胜出归根到底是受生物体原生情感——追求最优生存状态是所有生命的最终目的或目标——的支配。

结语

1938年,弗洛伊德在回应行为主义者的强烈攻击时写道,意识是“一个无可比拟的事实,它藐视所有的解释和描述。尽管如此,如果有人谈到意识,我们就能立刻并且通过我们最个人化的体验知道它意味着什么。一个极端的思路——比如美国行为主义学——认为,可以构建一门无视这一事实的心理

学!”^⑩相比于弗洛伊德的时代,在当今科学界,意识的地位早已今非昔比,无视意识的科学肯定是狭隘的和不完美的。但构建有意识的AI系统,这首先不是一个工程问题,而是一个理论问题,因为一个系统有意识与有智能可能在底层逻辑上完全不同:“也许对我来说最令人惊讶的发现是,我现在认为我们人类永远不会造出一个模仿我们个人意识的机器。无生命的硅基(silicon-based)机器以一种方式工作,而有生命的碳基(carbon-based)系统则以另一种方式工作。”^⑪当我们澄清了“意识”与“智能”的差异和关系,我们就会克服在“AI系统是否有意识以及AI是否会造成对人类的生存性威胁”等一系列理论问题上的困扰。

注释:

①<https://arxiv.org/pdf/2308.08708>.

②斯塔尼斯拉斯·迪昂:《机器会拥有意识吗?》,李恒威、蔡诗灵译,《新疆师范大学学报》(哲学社会科学版)2019年第1期。

③李恒威、蔡诗灵、阮泽楠:《人工意识与整合信息理论》,《浙江学刊》2021年第4期。

④C. Koch, *The Feeling of Life Itself: Why Consciousness is Widespread but Can't be Computed*, Cambridge: MIT Press, 2019, pp.151-152.

⑤A. Damasio, H. Damasio, "Feelings Are the Source of Consciousness," *Neural Computation*, Vol.35, No.3, 2023, pp.277-287.

⑥J. Panksepp, L. Biven, *The Archaeology of Mind: Neuro-evolutionary Origins of Human Emotions*, New York: W. W. Norton Company, Inc., 2012.

⑦李恒威:《心智的生命观与后人类社会议题》,北京:中国社会科学出版社,2023年。

⑧A. Damasio, *Feeling Knowing: Making Minds Conscious*, New York: Pantheon Books, 2021, p.104.

⑨M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018.

⑩M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.136.

⑪M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.150.

⑫A. Damasio, H. Damasio, "Feelings are the Source of Consciousness," *Neural Computation*, Vol.35, No.3, 2023, pp.277-287.

⑬A. Damasio, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, London: William Heinemann, 1999, p.312.

⑭M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.144.

⑮情绪和情绪感受在加扎尼加那里是两个递进的概念。情绪表现为基本的“奖惩系统”，而情绪感受意指由价值引导和奖惩进一步带来的具有情绪性的体验，它们都在皮层下系统中开始加工。

⑯M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.146.

⑰N. Humphrey, *A History of the Mind: Evolution and the Birth of Consciousness*, Dordrecht: Springer, 1999, p.132.

⑱李恒威、蔡诗灵、阮泽楠：《人工意识与整合信息理论》，《浙江学刊》2021年第4期。

⑲M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.106.

⑳M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.115.

㉑M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.106.

㉒M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.105.

㉓M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.93.

㉔M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.93.

㉕M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.95.

㉖丘奇兰德：《触摸神经》，李恒熙译，北京：机械工业出版社，2015年，第202页。

社，2015年，第202页。

㉗这两个术语的表达和用法参见 J. Panksepp, L. Biven, *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*, New York: W. W. Norton Company, Inc., 2012, p.7.

㉘J. Panksepp, L. Biven, *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*, New York: W. W. Norton Company, Inc. 2012, p.7.

㉙J. Panksepp, L. Biven, *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*, New York: W. W. Norton Company, Inc. 2012.

㉚J. Panksepp, L. Biven, *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*, New York: W. W. Norton Company, Inc., 2012, p.45.

㉛J. Panksepp, L. Biven, *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*, New York: W. W. Norton Company, Inc., 2012, p.46.

㉜N. Block, "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences*, Vol.18, No.2, 1995, pp. 227-247.

㉝N. Humphrey, *Sentience: The Invention of Consciousness*, Oxford: Oxford University Press, 2022.

㉞J. Panksepp, L. Biven, *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*, New York: W. W. Norton Company, Inc. 2012, p.13.

㉟M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.202.

㊱M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.207.

㊲M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.218.

㊳M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.219.

㊴M. Solms, *The Feeling Brain: Selected Papers on Neuropsychanalysis*, London: Karnac Books, 2015, p.177.

㊵M. S. Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, New York: Farrar, Straus and Giroux, 2018, p.236.