追问20万+

NextQuestion articles with **200,000+** views

2023 年刊

追问20万+

NextQuestion articles with 200,000+ views

天桥脑科学研究院

2023

TCCI 简介

天桥脑科学研究院(Tianqiao and Chrissy Chen Institute, TCCl)是由陈天桥、雒芊芊夫妇出资10亿美元创建的全球最大私人脑科学研究机构之一,总部设在美国。

TCCI与华山医院、上海市精神卫生中心设立了应用神经技术前沿实验室、人工智能与精神健康前沿实验室;与加州理工学院合作成立了TCCI加州理工神经科学研究院。

TCCI建成了支持脑科学研究的生态系统,项目遍布欧美、亚洲和大洋洲,包括学术会议和交流、夏校培训、AI加速科学大奖、科研型临床医生奖励计划、特殊病例社区、中文媒体追问等。

TCCI官网:
https://www.cheninstitute.org/zh

追问 NextQuestion 简介

天桥脑科学研究院(Tianqiao and Chrissy Chen Institute, TCCI)旗下科学媒体,于2020年2月由研究院创始人陈天桥、雒芊芊创立,旨在以科学追问为纽带,成为以“深入探究人工智能与人类智能相互融合”为特色的科学媒体,通过与AIGC技术的深度融合、重磅深度内容及多形态的呈现方式,打造全球“AI+脑科学”知识站。

媒体特设追问观察、追问专访、追问新知、追问顶刊、追问daily、智能渐近线等栏目。去年,媒体总发布文章358篇,共70.6万字,全网阅读2200万,全网阅读破20万的文章共75篇。本书正是从这75篇文章中做精心筛选汇编而成。

追问官网:https://www.next-question.com/

天桥脑科学研究院官方媒体

TCCI旗下追问
nextquestion公众号

目录 / DIRECTORY

创始人序言 ---- 001

引言 ---- 002

追问大脑

鸟儿为什么这么聪明? ---- 004

大脑如何区分和存储记忆? ---- 009

海马体掌管记忆的神, 我是你的破壁人 ---- 015

谁在影响我们的决策? ---- 023

当我们获得信息时, 我们获得了什么? ---- 027

你听到的音乐在脑中是怎样的? ---- 034

音乐是一种语言吗? ---- 038

大脑如何感受美的存在? ---- 042

好的生活方式如何降低抑郁症风险? ---- 048

非侵入性神经调控, 脑疾病精准治疗的新希望? ---- 052

Nature连发10篇, 揭示迄今最全小鼠完整大脑细胞图谱 ---- 057

神经科学领域里程碑, 全面构建人类大脑单细胞图谱 ---- 071

“超声读心”:用意念控制方向, 实现“未动先知” ---- 083

光遗传——诺奖的种子选手 ---- 089

用光束在脑中“绘制”电极, 让纳米金颗粒标靶特定神经元 ---- 093

纳米技术助力探索大脑中的“星辰大海” ---- 098

脑与AI

汪小京:将神经元变为数学模型和算法, 在人脑和 AI 间架起桥梁 ---- 104

从模仿到理解, 计算模型会是大脑的最终归宿吗? ---- 110

人工智能如何向人类智能学习? ---- 114

大语言模型是如何发展而来的? ---- 130

写给神经科学家的大语言模型基本原理 ---- 134

模型与大脑以不同的“眼光”看待世界 ---- 141

AI语音模型与人的听觉有多相似? ---- 149

35年激辩尘埃落定!这项能力不再是人类独有 ---- 157

超越感知：那些基于生物感官的AI算法 163
大语言模型为神经科学带来了哪些前所未有的机会？ 172
脑科学能用Transformer具体做什么？ 192
大语言模型如何宣告传统心理学的死亡？ 201
大语言模型如何改变了传统哲学议题？ 212
聊天机器人应该具备哪些社交特性？ 217
如果AI系统具备了意识，我们将如何知晓？ 225
AI的理解困境：如何走出数据世界，触达生命的理解？ 232
具身智能何以像人？ 237
从“智能涌现”到“超人类”，如何实现AGI？ 248

追问专访

追问专访·Moritz Helmstaedter：十年争议落幕，欧洲脑计划失败了？ 256
追问专访·杜忆：音乐让大脑返老还童？ 264
追问专访·Indre Viskontas：为什么人的审美有别？ 269
追问专访·汪小京：为计算神经科学培养跨学科人才 273
追问专访·吴思：打开人工智能的智慧之门 278
追问专访·吴梦玥：机器比人类更会理解声音吗？ 287
会议追问：医生最想用GPT解决哪些问题？ 295
会议追问：由 AI 操控的“读心术”，为什么不算准确却更受欢迎？ 299
会议追问·李广晔：脑机接口即将迎来“ChatGPT时刻”吗？ 305
会议追问：ChatGPT与人的意识层次有何不同？ 310
专访张闯&张洳源：AI和脑科学是未来十年最有前景的领域吗？ 314

鸣谢 318

创始人序言

在这个日新月异的时代，人类已经可以上天入海，但是我们的身心比过去更自由吗？如今，人类不仅可以尝遍世界各地的美食，还享受着令人震撼的视听盛宴，但是我们对世界的感知比以前更丰富了吗？世界的发展和科技的进步一定程度上解放了我们的双手，但是它是否也能解放心灵？我们感受到的世界是真实存在的吗？针对这些问题，我们试图从哲学、宗教、科学多个角度来寻求答案，但思考的结果往往带来更多的迷惑。

抽丝剥茧之后，它们似乎都指向了同一个目标：人类大脑。受好奇心和创新渴望的驱使，我们希望从脑科学的角度来揭开谜团。在与三百余位科学家、学者、行业领袖交流后我们了解到，作为最为复杂的科学领域之一，大脑的研究既令人为之着迷，也令人望而生畏，将这个梦想变为现实必将历尽艰难。但在交流中，我们也欣喜地发现，农业中的“蚕丝”竟可以作为新型生物材料应用在“脑机接口”领域；臭名昭著的致幻剂也可能成为治疗抑郁症的良药；在医学影像和肿瘤治疗中广泛应用的超声技术还具有调控大脑的作用，甚至有望逆转抑郁，抑制癫痫发作……由此我们更加坚信，如果能促成不同学科、行业领域的科学家与研究者之间大开脑洞的追问与相互启发，必将激发出无穷的可能性，绽放绚烂的科学之花。

因此，2020年2月，我们创建了天桥脑科学研究院（Tianqiao and Chrissy Chen Institute,TCCI）旗下科学媒体，并以“追问 Next Question”为它命名，旨在以科学追问为纽带，汇集全球脑科学领域的科学问题、学术会议资讯、科学家智库及前沿新知，不断探索科学的边界。

《追问20万+》便是以“追问 Next Question”在过去一年的知识积淀为基础精心编撰的科学读物。我们期待它能为读者呈现出一场脑科学的知识盛宴。读者可以将它作为了解脑科学前沿的切入点，开启自己的追问之旅。

陈天桥　　雒芊芊

引言

查尔斯·谢林顿有言：人脑就像一台施了魔法的纺织机，在不停地编织着与外在世界相关联的图样，把一个图样拆散了之后又重新编织起来，发明出其他的世界，创造出一个缩小的宇宙。随ChatGPT的问世，人类似乎在黑箱中再造了相似的魔法，构建起另一台魔法纺织机，实现了前所未见的新智能。随着字符在GPT内的重组和调用，它在不同领域实现了近似或超越人类的表现。这逼迫科学家们不断发问，人类是否真就独一无二？人类智慧与人工智能的差异在哪里？我们是否真是语言的动物，由数据驱动涌现出智慧？

随AI for Brain Science研究的井喷，我们已站在了人工智能与神经科学的十字路口，凝视着两个领域的融合与碰撞，期待着那些前所未有的发现。可以说，人工智能与人脑研究是彼此的镜像：我们模仿大脑设计神经网络，赋予计算模型以生命的特性；又借由机器学习探究大脑原理，以期发现大脑的计算特征。这一旅程的终点，机械与生命的界限愈发模糊不清。这本合集就是在这一主题的追问下诞生的。

在过去一年，追问编辑部深入探究人工智能与人类智能相互融合，邀请了数百位作者，精心编撰了358篇文章。这些文章在各大平台广泛传播，也被新华社，文汇，上观，新民，钛媒体，财新网，中国神经科学学会等知名媒体广泛转载。去年，全网阅读破20万的文章已有75篇。这本合集便是在这75篇文章内再做筛选，重新编排而成。本合集不做售卖，仅做馈赠，以答谢一直以来支持我们的研究员、作者与读者。文内图片也仅供交流学习使用，如有不当引用还请联系我们。

本文分追问大脑、脑与AI、追问专访三个部分，旨在探索脑科学最近进展、脑与AI的科研融合、AI技术的应用与反思等。我们希望这本合集能够带领读者一起追问。或许，在不同领域的碰撞中，我们将找到关于我们自身，关于智能的终极答案。

追问编辑部

追问大脑

鸟儿为什么这么聪明?

作者:陈硕

社会神经科学PhD在读。

业余科普人,时常为了写简介犯愁。

扫码查看原文

说起“聪明动物”,我们会想起谁?是“边牧是边牧,狗是狗”的边境牧羊犬,“人类的表亲”大猩猩,还是“在捕猎时懂得团队合作、拯救过遇袭潜水员”的虎鲸?

生活在澳洲的人类可能有一个不同的答案:葵花凤头鹦鹉(sulfur-crested cockatoo)。

这种著名的观赏鸟类拥有漂亮的白色羽毛和绽开时如葵花的黄色冠羽。因其人工繁殖的成功和出色的学习能力,时常能在国内的动物园里看到它们的身影。然而,在葵花凤头鹦鹉的快乐老家澳大利亚,成群结队的鹦鹉们令当地居民不堪其扰——它们学会了打开各个街区不同样式的生活垃圾箱觅食,并且通过社会学习将其中的技巧在族群内广而告之。

一项发表在《科学》(Science)杂志上的论文显示,葵花凤头鹦鹉的“开箱偷食”行为传播迅速,由2018年前的三个悉尼郊区迅速扩增到2019年的44个。研究人员甚至观察到了由森林阻隔而形成的地域特征,每个片区里葵花凤头鹦鹉的开箱方法都有一些细微的不同[1]。这些聪明的鸟儿已经成功倒逼它们的人类邻居想出各种科技手段保护自家的垃圾箱,不可谓不是人与自然相处中的一项奇观。

你可能已经注意到,提起“聪明”的时候,许多人并不会第一时间想起鸟类。相比之下,与人类亲缘关系更近的哺乳动物们往往更吸引我们注意。但仔细回想,从“鹦鹉学舌”到“乌鸦喝水”,鸟类的“小聪明”的确充斥于我们的文化生活之中。行为学数据也显示,鸦类和鹦鹉在许多领域的认知能力与非人灵长类旗鼓相当[2] [3]。它们不仅拥有空间记忆和情节记忆,也可以理解因果关系、延迟满足并计划未来,它们甚至拥有心智理论(一种理解自己和他人心理状态的能力)。

那么,这些小小生灵为什么这么聪明呢?

一、“大小”不重要,“内容”是关键

要想理解智慧,首先研究大脑。鸟类的聪明向一个由来已久的刻板印象发起了挑战:大脑越大越聪

明。尽管科学家们无法单凭大脑的重量推测某个物种或者个体的智力水平，但在同一个分类单元里，更大更重的大脑确实能够提供更多神经元和计算能力，也因此使得其拥有者比其他“表亲”更聪明[4]。这一点在非人灵长类、鲸豚类和鸟类中都有体现。但在亲缘关系较远的跨物种横向比较中，大脑的绝对重量和相对大小就没那么重要了：鸦类和鹦鹉类仅凭借5-20克的大脑便能与拥有400克大脑的大猩猩一较高下。

2016年，一项发表在PNAS的研究指出，鸟类的大脑虽小，神经元的密度却非常高[5]。每单位体积的大脑内，鸟类大脑所包含的神经元数量能达到灵长类的两倍、小鼠的四倍。研究者提出，如此高密度的神经元含量使得鸟类的前脑（forebrain）拥有与灵长类动物相当的神经元数量。神经元是大脑进行信息处理和计算的基本单位。尽管无法单独决定大脑的计算能力上限，数量相当的神经元为鸟类发展出与灵长类动物相似的信息处理能力提供了基础。此外，鸟类的大脑皮质神经元（pallial neurons）占比远高于灵长类动物：秃鼻乌鸦与狨猴的大脑重量相仿，但前者的大脑皮质神经元数量是后者的三倍。大脑皮质神经元能够协调不同认知过程以达成一个共同目标，因此与灵活的认知能力尤为相关。拥有大量大脑皮质神经元的秃鼻乌鸦也就比狨猴更加聪明。

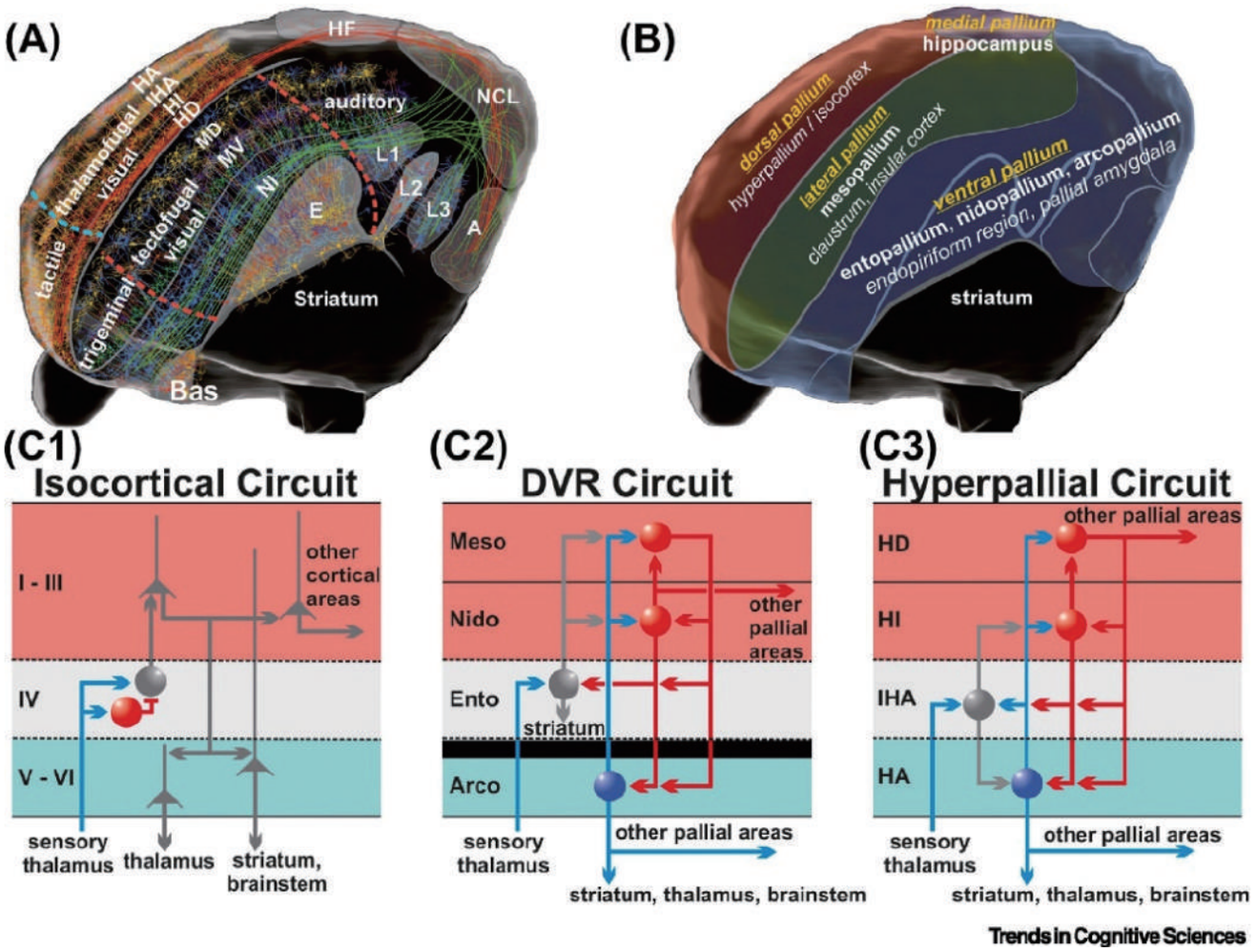
另一种与灵活的认知调控相关的神经元是联络神经元（associative neurons）。联络神经元存在于感知系统和运动系统之间，与联想学习和运动学习紧密相关。与鸡、鸽子和鸵鸟相比，新喀鸦（New Caledonian crow）拥有大量的联络神经元，其数量几乎可以媲美大猩猩前额皮质（prefrontal cortex）中的联络神经元[6]。新喀鸦也确实展现出了更高的智力——它们会用叶柄制作工具，钩取狭窄缝隙中的天牛幼虫吃。

因此，高神经元密度、大量大脑皮质神经元和联络神经元为小尺寸的鸟类大脑提供了惊人的计算能力，为鸟类的聪明才智提供了一定的生理基础。

二、“充分不必要”的新皮质

鸟类和哺乳类动物自数亿年前开始分别进化，最终形成了大相径庭的前脑结构。在哺乳动物中，背侧皮质（dorsal pallium）发育成大脑皮层，其中大部分是同皮质的（isocortical）。“同皮质”是指这部分大脑皮层的各个部分长得都差不多。“同皮质”更为人熟知的名称是“新皮质”，因为这部分大脑皮层为哺乳动物所特有，在进化上更新、更近。不光如此，哺乳动物的新皮质包含了所有有关感知、运动和联络的区域，可谓一个大包大揽的大脑结构，对认知和学习能力极为重要。也正因如此，很长一段时间内新皮质也被认为是哺乳动物智力的来源。然而，聪明的鸟儿们再次挑战了这一观点。

在鸟类中，与哺乳动物同源的背侧大脑皮层进化成了超大脑皮层（hyperpallium）。与哺乳动物的新皮质不同，鸟类的超大脑皮层仅包含了感知区域，功能上无法相提并论。大部分剩余的鸟类皮质核（pallial nuclei）则位于侧脑室下方，统称为背侧脑室嵴（dorsal ventricular ridge）。背侧脑室嵴在哺乳动物中没有同源结构，但它却在功能上补充了超大脑皮层，可以处理感知信息，也包含了运动和联络区域。近年的研究表明，背侧脑室嵴在处理感知信息的部分呈现与哺乳动物的大脑皮层相似的分层结构和信息处理路线[7]，而运动和联络区域依然是核状排列的（a nuclear arrangement）。



▷ A)鸟类大脑的分区;B)鸟类大脑的结构名称(黄字)以及与其同源的哺乳动物大脑结构(白字);C)不同的信息处理路径,C1为同皮质,C2为背侧脑室嵴,C3为超大脑皮层。图源:原始论文

这些数据表明,鸟类的大脑可能由于趋同进化,部分呈现出了与哺乳动物的大脑相似的结构和信息处理方法。至少在处理感知信息方面,这些类似同皮质的结构或许有着难以替代的绝佳优势。但鉴于背侧脑室嵴和新皮质在进化角度并不同源,且背侧脑室嵴没有完全采用新皮质的处理方式,新皮质对于聪明的大脑来说,可能只是一个“充分不必要”条件。

三、多巴胺能驱动的鸟类“前额皮质”

我们已经从神经元的数量、种类和大脑结构方面讨论了鸟类聪明的原因,发现它们与哺乳动物既有相似也有不同。接下来,我们将从大脑功能上继续探讨鸟类为什么这么聪明。

背侧脑室嵴呈核状排列的结构中,尾外侧巢状皮质(nidopallium caudolaterale)尤为重要。尾外侧巢状皮质位于背侧脑室嵴的最尾端,在功能上与哺乳动物的前额皮质极为相似,几乎参与了所有的认知过程。更重要的是,这两处区域都拥有大脑中密度最高的多巴胺能神经元(dopaminergic neurons),并与大脑内所有联络区域和前运动结构(premotor structures)相连。与前额皮质中的多巴胺能神经元相仿,尾外侧巢状皮质中的神经元也能够分类呈现感知信息(如长短和数字)、根据需要以不同的时间顺序呈现刺激(前瞻性或回顾性),乃至编码执行功能(executive functions,一系列管理和控制注意力等

抽象认知功能的能力）和感觉意识。因此，鸟类的尾外侧巢状皮质拥有与哺乳动物前额皮质相似的强大功能，从而允许它们统筹和安排复杂的认知过程，表现出“聪明”的行为能力。

四、鸟儿会有工作记忆吗？

工作记忆（working memory）指的是将信息储存在记忆中以进行实时处理的能力。拥有工作记忆是一切认知能力的核心，也是前额皮质的一项重要功能。脑损伤、药理和神经生理学等多项研究表明，鸟类的尾外侧巢状皮质也参与了鸟类的工作记忆。例如，单细胞记录方法在尾外侧巢状皮质中观察到了与哺乳动物前额皮质中类似的“延迟活动”（delay activity）。延迟活动是在施加外部刺激和采取后续行动的时间间隔内，前额皮质神经元保持活跃的现象[8]。在鸟类中观察到延迟活动，意味着它们确实具有工作记忆这一认知能力的基础。不仅是单个神经元，在由多个神经元细胞活动叠加而成的局部电场位（local field potentials）中，也观察到了鸟类与哺乳动物相似的神经生理活动。这些证据表明，尽管缺乏新皮质那样的分层结构，鸟类的认知活动也能产生与哺乳动物类似的神经生理学指纹。

有关工作记忆的研究主要集中在单细胞和局部脑区层面，而对于睡眠与梦境的研究则能在全脑层面提供有关鸟类认知能力的证据。在一项对鸽子的快速眼动睡眠（一种与梦境紧密相关的睡眠状态）实验里，研究人员采用功能核磁共振扫描了鸽子的大脑活动。他们发现，与人类相似，鸽子的大脑在快速眼动睡眠期间激活了边缘系统和前运动脑区，以及视觉和多模态大脑皮层区域[9]。尽管目前尚无法根据这些大脑活动还原梦境的内容，研究者们依然猜想，鸽子们或许梦到了在飞行过程中躲避障碍物的场景。这些证据初步表明，鸟类的认知过程和哺乳动物一样，与广泛的神经网络活动息息相关。

除了相似之处，鸟类认知的神经生理基础也有与哺乳动物南辕北辙的地方。例如，睡眠研究显示，鸟类并不会像哺乳动物那样在睡眠过程中激活海马体以巩固记忆。它们究竟如何保持和提取长期记忆，还有待进一步研究。

五、鸟类为什么这么聪明？

在这篇最新的Trends in Cognitive Sciences观点文章中，作者们由“鸟类为什么这么聪明”这一问题出发，对“智慧大脑”的构成进行了探讨。那么，什么样的大脑能够发展出智慧呢？

首先，并不是越大的脑袋越聪明。大脑内部神经元的数量、种类和互相之间的联结方式或许更为重要。与哺乳动物不同，鸟类大脑中高密度、大数量的大脑皮质神经元和联络神经元允许鸟类在有限的大脑尺寸中发展出智慧。

其次，哺乳动物所特有的新皮质并不是智慧的必要前提。鸟类的背侧脑室嵴与哺乳动物的新皮质在进化上并不同源，生理结构和信息处理路线上也不尽相同。但背侧脑室嵴与新皮质一样，负责包括感知、运动和联络脑区的信息处理，令鸟类拥有了发展出智慧的可能。这说明，新皮质只是发展出智慧的“充分不必要条件”。

第三，鸟类的尾外侧巢状皮质拥有密集的多巴胺能神经元联结，并参与了几乎所有的认知过程。尾外侧巢状皮质与哺乳动物的前额皮质在功能上极为相似，负责整合信息、编码抽象认知和统筹行动。由于二者进化上并不同源，这可能是趋同进化的结果。同时也表明，一个类似前额皮质的“控制中心”对于复杂认知能力的发展是极为重要的。

最后，有关鸟类工作记忆和睡眠梦境的研究显示，尽管缺乏类似的大脑结构，鸟类认知的神经生理学

指纹与哺乳动物极为相似。这说明在产生智慧的过程中，神经元与神经网络的合作机制可能有本质上的、难以替代的共同点。

至此，我们以鸟类为例，打破了“哺乳动物才有智慧”的垄断观念，从神经元的数量和种类、大脑结构、大脑功能和神经生理活动层面讨论了“智慧大脑”的更多可能。当然，这些仅仅是目前研究比较丰富的方向，还有其他的“潜力股”等待人类的研究和发现。

哺乳动物并不是地球上独一无二的智慧生物，对鸟类大脑的研究将带给我们不一样的视角。探询“鸟类为什么这么聪明”的同时，人类也在探索自身的智慧与边界，向着生命——或者说“生灵”——的本质提出疑问与思考。（编辑：存源）

参考文献

关联论文：Güntürkün, Onur, Roland Pusch, and Jonas Rose. "Why birds are smart." Trends in Cognitive Sciences (2023). https://doi.org/10.1016/j.tics.2023.11.002

[1] Dicke, U., & Roth, G. (2016). Neuronal factors determining high intelligence. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1685), 20150180. https://doi.org/10.1098/rstb.2015.0180

[2] Klump, B. C., Martin, J. M., Wild, S., Hörsch, J. K., Major, R. E., & Aplin, L. M. (2021). Innovation and geographic spread of a complex foraging culture in an urban parrot. Science, 373(6553), 456–460. https://doi.org/10.1126/-science.abe7808

[3] Lambert, M. L., Jacobs, I., Osvath, M., & von Bayern, A. M. P. (2019). Birds of a feather? Parrot and corvid cognition compared. Behaviour, 156(5/8), 505–594. https://www.jstor.org/stable/26737865

[4] Mehlhorn, J., Hunt, G. R., Gray, R. D., Rehkämper, G., & Güntürkün, O. (2010). Tool-Making New Caledonian Crows Have Large Associative Brain Areas. Brain, Behavior and Evolution, 75(1), 63–70. https://doi.org/10.1159/000295151

[5] Olkowicz, S., Kocourek, M., Lučan, R. K., Porteš, M., Fitch, W. T., Herculano-Houzel, S., & Němec, P. (2016). Birds have primate-like numbers of neurons in the forebrain. Proceedings of the National Academy of Sciences, 113(26), 7255–7260. https://doi.org/10.1073/pnas.1517131113

[6] Güntürkün, O., Pusch, R., & Rose, J. (2023). Why birds are smart. Trends in Cognitive Sciences. https://-doi.org/10.1016/j.tics.2023.11.002

[7] Pika, S., Sima, M. J., Blum, C. R., Herrmann, E., & Mundry, R. (2020). Ravens parallel great apes in physical and social cognitive skills. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-77060-8

[8] Sreenivasan, K. K., & D' Esposito, M. (2019). The what, where and how of delay activity. Nature Reviews Neuroscience, 20(8), 466–481. https://doi.org/10.1038/s41583-019-0176-7

[9] Ungurean, G., Behroozi, M., Böger, L., Helluy, X., Libourel, P.-A., Güntürkün, O., & Rattenborg, N. C. (2023). Wide-spread brain activation and reduced CSF flow during avian REM sleep. Nature Communications, 14(1), 3259. https://-doi.org/10.1038/s41467-023-38669-1

大脑如何区分和存储记忆?

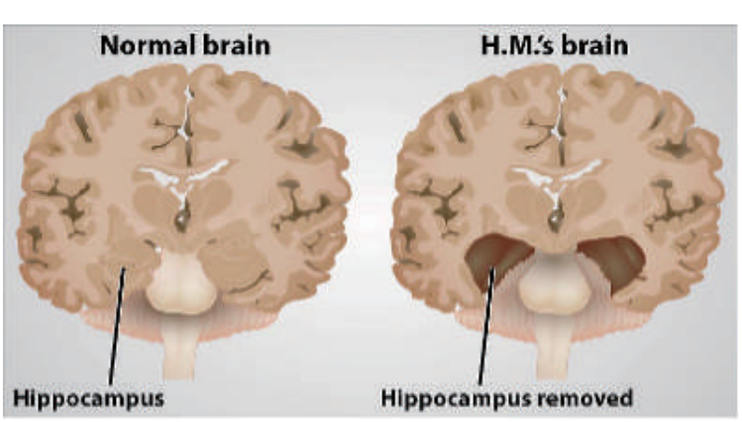
作者:轻盈

复旦大学博士生在读,计算&进化神经生物学方向。视科研和科普为人生的两大志业。想做有趣有意义的科学研究,也想把收获到的知识和乐趣分享给世人。

扫码查看原文

在神经科学领域诸多关于记忆的研究论文中,有一个高频出现的名字——Henry Molaison(简称H.M.)。他曾是一位严重的癫痫患者。1953年,他决心接受一项有望治疗癫痫的外科手术——切除包括海马体1(图1)在内的脑组织。手术后的H.M.在大多数方面都表现得相当正常:他的个性、抽象思维能力和推理能力并未受到影响;他也保留了童年的记忆,能够清晰地记得小时候老家的地址。然而,让世人感到疑惑的是,H.M.几乎完全丧失了创造长期记忆的能力。他不记得手术后经常看他的医生,甚至也不记得刚刚吃过饭。

H.M.的经历,让人们对海马体与记忆之间的功能联系产生了兴趣。目前,神经科学家和心理学家已经开始认识到,我们的大脑中存在多种类型的记忆:有关过去经历的情节性记忆,有关事实的语义记忆,短时和长时记忆等。但是,为什么这些记忆会以不同类型存在?不同类型的记忆都储存在大脑的什么地方?对科学家来说,这些问题仍是未解之谜。



▷图1:H.M.的海马体在手术中被剥除。图片来源:brainfacts

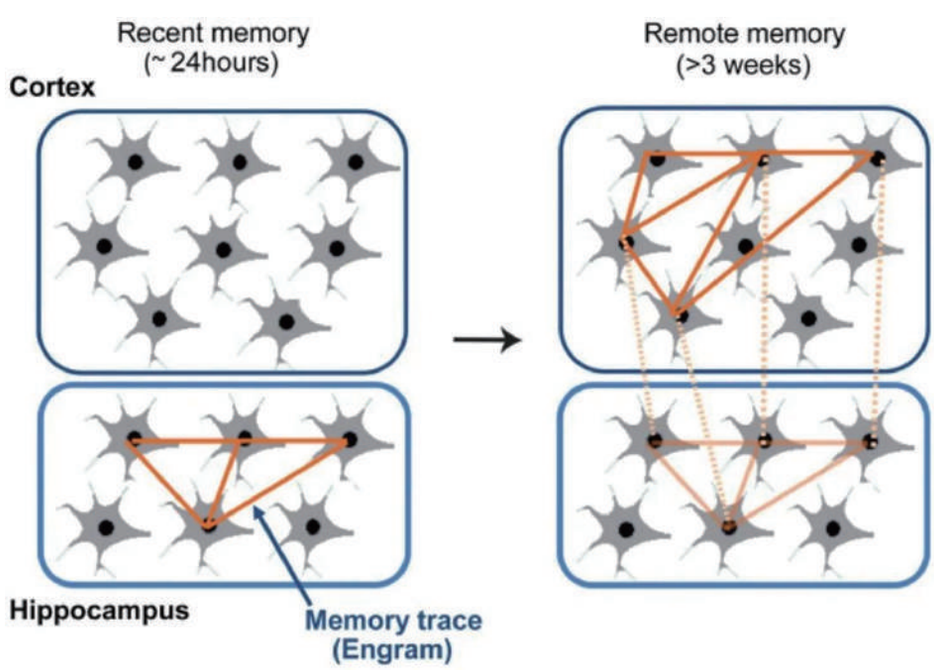
记忆指的是大脑编码、存储和检索过往经验的能力。对过往经验的存储和概括有益于动物做出适应性的决策。举例来说，记住通往特定水源的安全路径，并对这些记忆线索进行概括，就能发现能预测新水源的环境特征，如周围的地貌、植被类型等。因而，这种能够被用来识别和理解当前的情境，并基于过去的经验预测将要发生的事情的记忆的能力，即可预测记忆，对生物的环境适应能力尤为重要。

如今，一项经人工神经网络实验证实的新理论表明，大脑可能会通过评估记忆在未来的有用程度，对记忆进行分类。具体来说，可预测记忆2（比如每天早餐吃什么或上班的路线怎么走）会保存在大脑的新皮质中，它们将帮助你更好的适应不断变化的环境和做出相应有效的决策。至于一些不太有用的记忆，比如在一个派对上喝过的特殊饮料的味道，则会被直接保存在海马体中，不再会在新皮质中被巩固。可见，大脑会根据记忆的有用性和可预测性对其进行分类“装箱”，优化了记忆的可靠性，来帮助我们应对新颖的情境。

*可预测记忆：过去的经验和记忆对未来事件的发生或特征产生一定程度的可预测性。当大脑接收新信息时，先前存储的记忆可以用来识别和理解当前的情境，并基于过去的经验预测将要发生的事情。这种可预测性有助于个体更好地适应环境和做出有效的决策。

一、系统巩固神经网络模型

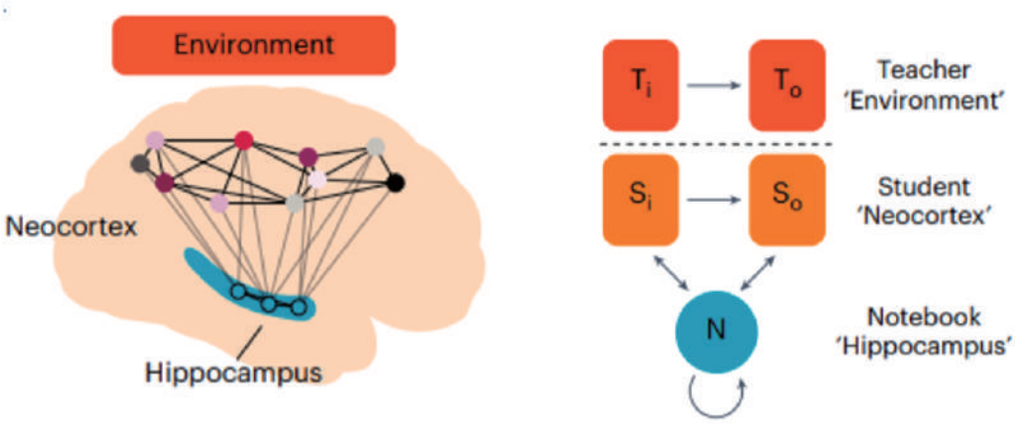
关于记忆在脑区间的存储，目前较为经典的研究理论是互补学习系统假说（图2）。该理论认为，大脑中存在两个互补的学习系统——海马体和新皮质。海马体，位于脑内深处，人们认为它在学习初期特别活跃。海马体主要负责快速地将新的信息进行编码和存储，并且对于空间导航、事实和事件的短期记忆至关重要。新皮质则是大脑外层的一部分，它会对信息进行更长期的存储和整合。新皮质的学习过程相对较慢，但它能够处理更加复杂和抽象的信息，并将其整合到长期记忆中。这两个系统配合着一起工作，就形成了学习和记忆的动态过程。海马体在学习初期迅速编码新信息，然后逐渐将这些信息传递给新皮质，再由新皮质进行更加稳固和长期的存储。



▷图2：海马体与新皮质在储存记忆时的互补学习过程。图源：researchgate

基于上述互补学习系统假说，研究人员建立了一个叫做“教师-笔记本-学生”的系统巩固神经网络模型（图3），该模型由三个功能不同的神经网络组成——教师、笔记本和学生网络。教师网络代表生物生存的环境，它提供经验性的输入；笔记本网络则代表海马体，能够快速地编码环境所提供的每个经验的所有细节；而学生网络则代表新皮质，其通过海马体（笔记本）中记录的经验来预测环境（教师）提供的信息。

具体来说，在神经网络模型中，“教师”是一个线性前馈网络，通过固定的权重和附加输出噪声生成“输入-输出对”。“学生”是一个与之大小相匹配的权重可学习的线性前馈网络。“笔记本”则是一个稀疏的Hopfield网络。在模型训练的过程中，“学生”从包含信号和噪声的有限经验集合中进行权重的学习，最终希望“教师”的输出和“学生”的预测之间的平方差应尽可能小。另外，研究人员也提出假设，在这个模型中，“学生”的主要目标是优化泛化性能，能够最终较好地预测新的“教师”输出。

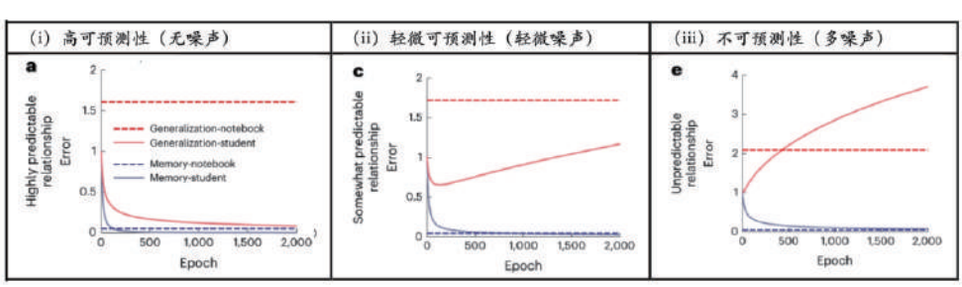


▷图3：教师-学生-笔记本结构图 图源：参考文献

研究人员使用该神经网络模型模拟了记忆和泛化的动态过程，来研究系统巩固3的实际影响和特征。为了优化学生的记忆回溯，研究人员将模型建模为无限的笔记本回放。在每次模拟中，他们会根据三个不同可预测性（信噪比）的教师网络（高预测性、轻微预测性以及不可预测性），生成经验输入。

结果发现（图4），尽管对于无噪声的教师网络输入，学生网络的泛化误差单调减小，但有噪声的教师网络输入最终仍使得学生网络泛化效果不佳（红色实线）。也就是说，在不可预测的环境中，过度的系统巩固可能会严重降低模型的泛化性能，导致新皮质（学生）过度拟合环境（教师）中不可预测的元素。这也预示着，在建构神经网络模型的过程中，为了最终能够提高模型的泛化能力，有必要将系统巩固水平考虑在内。

*系统巩固：研究人员将系统巩固建模为学生网络内部突触的可塑性。学生网络的可塑性机制受笔记本重新激活的指导，类似于假设海马体回放有助于系统巩固的方式。

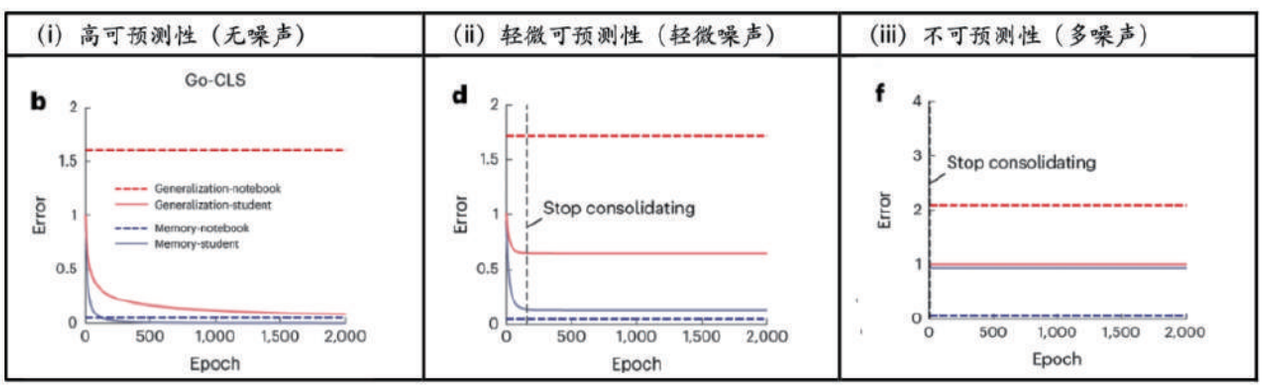


▷图4：在优化学生记忆时，学生概括错误、学生记忆错误、笔记本概括错误和笔记本记忆错误的动态变化。图源：参考文献1

二、泛化优化的互补学习系统(Go-CLS模型)

(1)Go-CLS模型

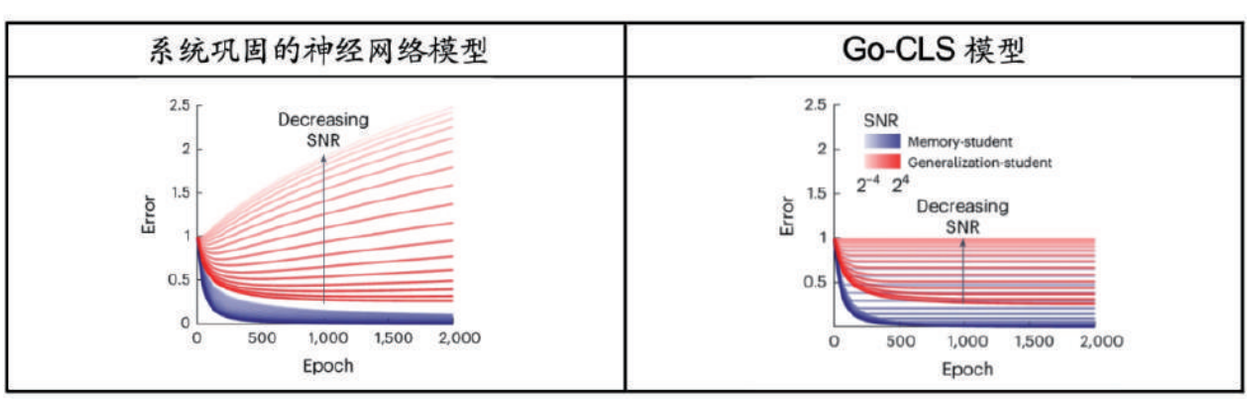
Go-CLS模型是在系统巩固神经网络模型的基础上建构而成的。考虑到上述实验中在不可预测的环境里系统巩固对模型泛化能力的不利影响,研究人员调整了模型的系统巩固水平。该理论的基本假设是,系统巩固的程度是自适应调节的,当系统巩固到一定程度后,如果继续巩固反而会削弱模型的泛化能力,就应该停止巩固。从目前Go-CLS模型的训练结果看,在这种设定下,学生(红色实线)可以从每位教师的经验中实现几乎最佳的泛化性能。



▷图5:在优化概括的表现时,学生概括错误、学生记忆错误、笔记本概括错误和笔记本记忆错误的动态变化。(b)高可预测性(无噪声);(d)轻微可预测性(轻微噪声);(f)不可预测性(多噪声)。图源:参考文献1

(2)与系统巩固的神经网络模型比较

研究人员将上述两个模型进行比较,发现两者在处理不同预测性教师数据时表现出明显差异。在标准系统巩固模型中,所有记忆都会被巩固,但泛化能力随着教师可预测性的不同而有很大变化。对于可预测性较低的教师,这种模型的泛化误差较高(图6,左)。与之相较,Go-CLS模型消除了过拟合的有害效应,在学生能够到达最优记忆性能之前结束。并且随着教师的可预测性程度增加,泛化性能和记忆性能均得到改善(图6,右)。



▷图6:巩固模型与Go-CLS模型下学生泛化性能随教师的可预测性程度变化的比较。左:系统巩固的神经网络模型,右:Go-CLS模型。图源:参考文献1

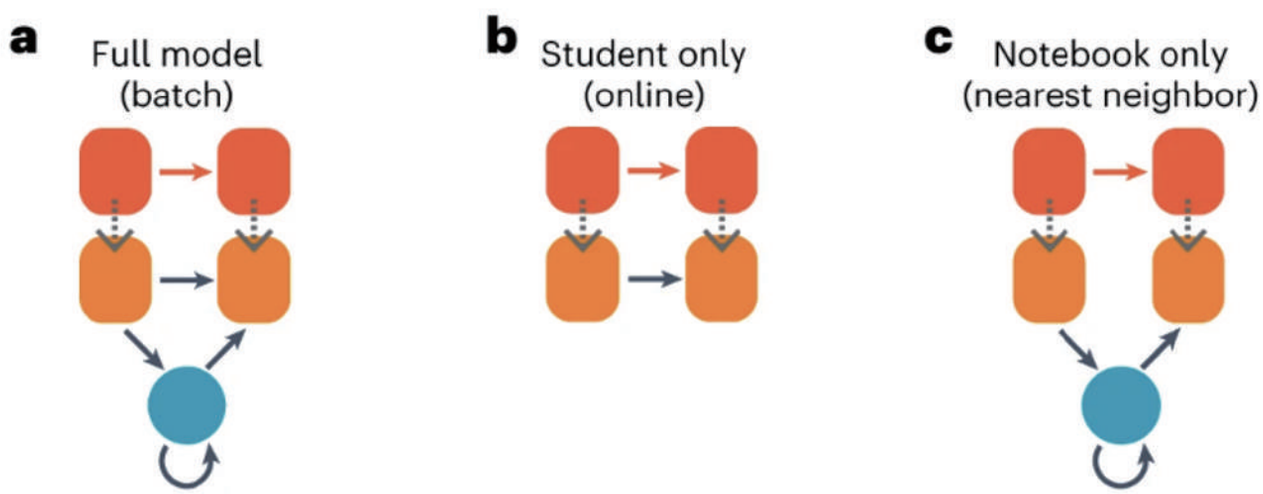
可见,Go-CLS模型不仅在逆行性遗忘4(即过去记忆的丢失)方面与系统巩固的神经网络模型表现不同,还在时间依赖性泛化方面表现出显著差异。在模拟海马体剥除后,在逆行性遗忘方面,系统巩固的神经网络模型总会表现出记忆逐渐丢失的趋势,即过去的记忆更容易被遗忘。而Go-CLS模型则能够解释不

同类型的记忆丢失现象，包括那些随时间变化而不同的记忆丢失模式。在时间依赖的泛化层面，在处理不同的可预测性的经验时，两种模型都会产生多样的泛化曲线，最大泛化性能随着教师的可预测性增加而提高。然而，在学生过拟合时，只有Go-CLS能够维持其随时间变化的泛化能力。系统巩固的神经网络模型甚至可能导致学生过度泛化，产生极不准确的输出。

*逆行性遗忘（retrograde amnesia）：在某个特定时间点之前发生的事件或信息无法被记忆，而在这个时间点之后的记忆则相对保留。这种记忆缺失通常是由于脑部受损或其他神经系统障碍引起的，影响了个体对过去经历和信息的回忆。逆行性遗忘可能涉及特定类型的记忆，如语义记忆或事件记忆，而不是影响所有记忆类型。研究者通常根据记忆缺陷在最近和远期记忆中是否相似（平坦逆行性遗忘）、在最近记忆中是否更为明显（分级逆行性遗忘），或者在最近和远期记忆中是否都不存在（无逆行性遗忘）来对海马体遗忘动态进行分类。

（3）对互补学习系统理论的补充

Go-CLS神经网络模型的实验结果为互补学习系统假说提供了新的见解。海马体和新皮质耦合的快速和慢速学习模块——是互补学习假设的基本前提。于是，研究人员又利用Go-CLS神经网络模型比较了耦合的学生-笔记本网络（图7，左）与孤立的学生（图7，中）和笔记本网络（图7，右）的泛化性能。考虑到Go-CLS神经网络模型中，“学生”表征的是互补学习系统假说中的新皮质，“笔记本”表征的是海马体，因此孤立的学生和笔记本网络的实验组，则分别表示各自仅用新皮质或仅用海马体学习的情境。



▷图 7: 学习系统模型示意图，(a)使用学生-笔记本网络；(b)仅使用学生网络；(c)仅使用笔记本网络。图源：参考文献1

孤立学生网络模型必须在线从每个教师网络的经验中学习，并且无法重新访问或是回顾以往的信息。这就限制了它的泛化性能（图8，橙色与黑色曲线相比）。结果表明，使用两个包含学生的网络模型都要比仅含笔记本的网络模型泛化性能更好。另外，当教师网络提供适量可预测的数据时，学生-笔记本耦合网络模型的泛化增益要远远大于孤立的学生网络模型。因此，在数据适中且环境中噪声适度的情况下，学生-笔记本耦合网络模型在泛化性能上更为优越。

另外，当学生的经验数量等于学习模型中可调整的参数数量时，使用“回顾先前的学习经验来进行学习”（回放功能）可以最有效地提升学生的学习能力。但这种情况下也存在风险：大脑可能会过度适应这些旧信息，导致对新情况的反应不够灵活，就像机器学习中的“双峰下降”现象一样。在这种现象中，当数据量处于与网络大小相关的中等水平时，过拟合最为严重。当记忆模式的数量接近网络模型可以处理的最

大数量（容量）时，神经网络必须最精细地调整其权重。这通常需要对权重进行大幅度的更改，以减小小的训练误差。为了避免这个问题，最优学生-笔记本网络模型可以通过根据教师的可预测性调整系统巩固的数量，以防止过度适应噪声。这预示着，大脑可能也在以类似的方式根据经验的可预测性来调节学习和记忆的方式。

三、总结

Go-GLS模型为我们提供了一个有趣的理论假设，即大脑可能是通过评估记忆的可预测性和对未来的有用性来对记忆进行分类的。设想这样的情境：一个小女孩与父亲在湖边度过快乐的一天。在这一天的经历中，可能包含了新鲜采摘的草莓，味道很甜这样的可预测关系。Go-GLS模型的研究结果表明：可预测关系应该从经验中提取出来，并与相关经验的记忆相互整合，通过系统巩固过程来生成记忆，并被强化，实现可预测经验的泛化。在这一天的经历中，也可能会出现不可预测的共现事件，比如父亲的衬衫颜色刚好与草莓的颜色都是红色。但这类关系不应在新皮质中被巩固。当然，它们仍然可以作为这一天的情景记忆的一部分存放在海马体中。

另一方面，Go-GLS模型与人工智能研究有许多微妙的联系。Go-CLS模型通过教师-学生对的最优逼近误差来定义可预测性。但是，这与实际情况中，是否可以学到最优学生权重是不同的。比如梯度下降可能陷入局部最小值或短暂地降低泛化性能。另外，由于过拟合同样可以在更复杂的学生架构以及现代深度学习模型中观察到，Go-GLS模型中提出的基本概念也适用于更广泛的模型类别。同样，一些机器学习方法可以插值训练数据并很好地泛化，因此我们有必要寻找能够在学习过程中更好地平衡记忆和泛化的学生网络架构和学习规则。

然而，Go-GLS模型仅关注了简单的监督学习问题，未来我们仍需要解决在具有更为多样化的泛化情境下的最优巩固问题，例如强化学习和大型语言模型中的新兴少样本学习。（编辑：野雾、韵珂）

参考文献

关联论文：Organizing memories for generalization in complementary learning systems

[1] Sun, W., Advani, M., Spruston, N. et al. Organizing memories for generalization in complementary learning systems. Nat Neurosci 26, 1438–1448 (2023).

[2] https://www.quantamagazine.org/the-usefulness-of-a-memory-guides-where-the-brain-saves-it-20230830/

海马体掌管记忆的神，我是你的破壁人

作者：郭瑞东

科普作家，关注复杂系统与神经科学。追问nextquestion、集智俱乐部长期撰稿人，曾为知识分子，果壳等多家媒体撰文，科普书《机器学习与复杂系统》合著者。

扫码查看原文

记忆是如何形成的？哪些事情会被我们记住？为何有些记忆会经久不忘？

要想解答这些亘古之谜，需要明白生物的记忆远不同于在记事本上记日记，更不像如计算机磁盘那样可以反复读写。记忆的生成、巩固和读取并非执行单一的、简单的指令，而是涉及一个精巧的、多步骤的复杂过程。发表在Nature Human Behaviour的一篇论文[1]，经由AI深度生成模型，阐释了人脑是如何存储和重建记忆的独特感觉和可预测概念元素的。

一、问题：海马体究竟是做什么的？

神经科学的进展，最初来自于那些大脑部分损伤的可怜病患，著名的病例H.M.（Henry Molaison）就是由于海马体的损伤，失去了形成新记忆的能力，但保留了一些基本技能和长期记忆。科学家由此认定海马体在新记忆的形成时发挥重要作用。

然而，随着研究的深入，科学界发现海马体不仅仅参与记忆的形成，还在其它认知功能中发挥着重要功能，例如海马体中的“位置细胞”（place cells）在空间记忆和导航中起着重要作用。海马体损伤的患者可能会难以回忆起曾经去过的地方或者抵达某个目的地的方式与过程。

之后更多的研究还发现，阿尔茨海默病中，海马体是最早受到损伤的区域之一。此外，海马体还与情绪调节有关，其损伤可能导致情绪波动、焦虑和抑郁等情绪问题。这些发现意味着人们最初以为科学家找到了人脑用于临时存储的“内存”，结果发现这个内存不仅是待写入信息的缓冲区，还参与信息的提取和调控。

越来越多关于海马功能的发现，让普通人乃至专家都对海马的具体功能感到迷惑，也许并非是海马体的功能多么复杂多变，而是我们搞错了研究方向？

二、背景：记忆的分类

电脑的存储分为高速缓存，内存，硬盘等多级存储。而人脑的记忆按照时间来分，只分成两档：短期或工作记忆和长期记忆。长期记忆又可细分为语义记忆（Semantic Memory）和情景记忆（Episodic Memory）。语义记忆类似于词典，涉及对世界知识的记忆，如历史事件、科学事实等；情景记忆更像私人化的笔记，涉及个人经历，如第一次骑自行车的经历。

具体来说，情景记忆涉及时空背景下的自传体经历，对应“我是谁”、“从哪里来”的问题；而语义记忆涉及事实知识，回答“我能做什么”、“想到哪里去”等问题。只有同时具有这两种记忆，才能算一个完整的人。可问题是，我们就只有一套神经系统，如同准备了一桌菜，来了两桌人，大脑该如何应对？

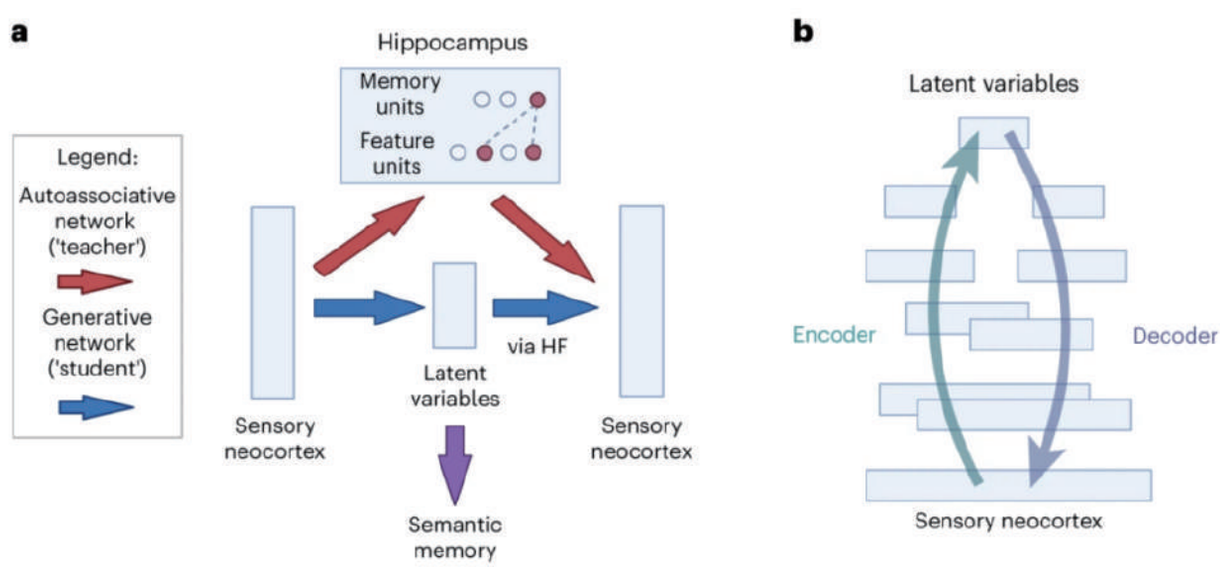
之前的研究表明，情景记忆可以通过海马体的事后回放来快速捕获转瞬即逝的多模式体验，因此被认为具有建设性。回忆是对过去经验的（重新）构建，而非单纯地检索一个副本。正是有了这一次次的回放，使得在进化上较新的新皮质能够从多种感官的体验中找到统计规律，从而形成语义记忆。这看似解释了大脑如何产生记忆，然而细究起来，情景记忆的重建机制及其与语义记忆的关联尚不清楚，例如我们不知道大脑是如何从情境中筛选关键特征的。而这是本文主要介绍的研究想要解决的问题。

三、新模型的关键词：巩固

该研究新提出的记忆模型模拟了如何使用记忆的初始表征来训练一个生成式人工神经网络。研究人员向模型播放了10000张简单场景的图像，其中模拟海马体的部分网络快速编码所经历的每个场景。随后，该网络反复回放这些场景，以训练模拟新皮质的生成神经网络。

这个待训练的网络通过学习经历过的事件（或“模式”）的统计结构来重建记忆。生成网络将每个场景中代表数千个输入神经元（接收视觉信息的神经元）的活动传递给一个更小的神经元中间层（最小的层仅包含20个神经元），从而将场景重建为数千个输出神经元（负责预测视觉信息）的活动模式。

训练过程如图1a所示：首先，海马体快速编码事件；然后，在接受海马体回放表征的训练后，生成网络逐渐接管。这一过程使得记忆更加抽象，更支持概括和关系推理，但也更容易出现基于要点的扭曲。生成网络可用于重建（用于记忆）或构建（用于想象力）感官体验，或直接从其潜在变量表示中支持语义记忆和关系推理。



▷图1：生成模型的训练过程与其内部的更细节结构。来源：关联论文

上述模型看起来复杂，但却不难理解。想象一个一见钟情的少年，只是见了女孩一次便念念不忘，在脑中一遍遍回放见面的每个细节。经过一次次回放，大脑的某个区域熟能生巧，能够生成一个关于这个女孩的种种场景。这时，少年对那次相遇的记忆经由一次次的回放而扭曲，他开始幻想着和女孩牵手的样子，这些都形成了情境记忆。而大脑的新皮层则在搜寻女孩出现规律的种种线索，而这就会成为陈述性的语义记忆。

在上述过程中，海马的功能不再是一张用作中转的草稿纸，而是更像是老师，经由自联想网络给生成模型提供指导。而生成模型训练的过程，可视为记忆巩固。在巩固过程中，记忆从一个神经网络转移到另一个神经网络。巩固后，生成网络对记忆中包含的信息进行编码。一旦生成网络学会重建特定事件，对它的依赖就会随时间增长。

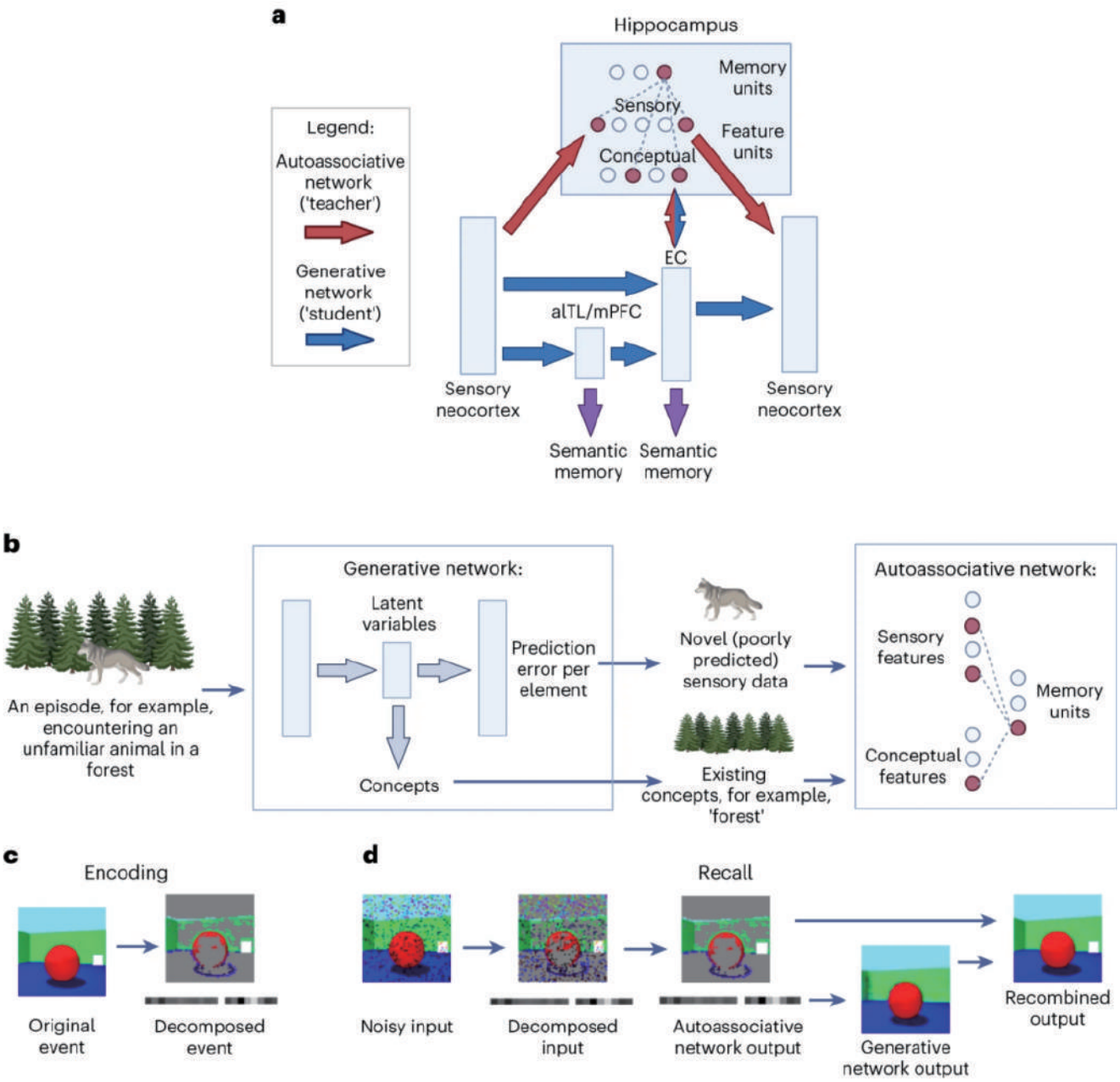
具体来说，生成网络被设计成一种特殊类型的自编码器，即变分自动编码器（VAE，图1b）。在这种编码器中，数据被压缩到最紧凑的层，这一层代表了一组潜在变量。我们可以从这些潜在变量中采样，以生成与训练数据集相似的新实例。这些潜在变量可以被认为是观察数据背后的隐藏因素，而在潜在空间中的不同方向可以对应于数据的有意义变换。VAE的编码器负责将感官体验转换为潜在变量，而解码器部分则负责将这些潜在变量还原为感官体验。在对一类刺激进行训练后，VAE可以基于该类别的一般模式，从部分输入中重建此类刺激，并生成与该模式一致的新刺激。

四、记忆重建中的差异支持概念提取

在感知过程中，生成模型根据其重建误差（输入和输出表示之间的差异，也称为“预测误差”）来持续评估体验的新颖性。事件中与先前经验一致的部分（即重建误差较低的部分）不需要在海马体中的自关联“教师”网络中进行详细编码。当生成网络的重建误差降低到一定程度时，就不再需要依赖海马体的记忆轨迹，从而为新的记忆编码释放出空间。重建误差阈值可能会根据刺激的重要性或可用注意力资源的数量而变化。例如，情绪显著性可以降低这个阈值。

上述过程我们继续用少年遇到梦中女神的例子解释，在回放与女孩初次相遇的场景时，大脑可能不会过分关注场景中熟悉的部分，而会专注于最新奇、最有价值的部分。在少年初遇女孩后可能会有一段意乱神迷的时间，这段时间发生了什么可能完全没有记忆。这是因为在这段时间内，海马体正专心作为生成网络的“老师”，并未有足够的资源接收新的外部信息。一旦大脑判断记忆中的女孩与实际经验中的女孩相似度足够高（纤毫毕现），海马体便会重新被释放出来，用于编码新的感官体验，形成记忆。通常，这种情感冲击大的记忆更新频率不高，上下文的连贯性较差，往往只包含少数几个场景。

在这个新的框架下，记忆巩固可能不仅仅是细粒度的感官表征更新粗粒度的概念表征的过程，而是将粗粒度和概念性表征与细粒度和感官性的一系列表征结合在一起的过程。例如，海马体在编码遇见女孩的那一天时，可能会将“女孩”和“心动”等粗粒度概念，与诸如陌生歌曲的旋律或特定沙滩的景象等感官表征结合在一起。

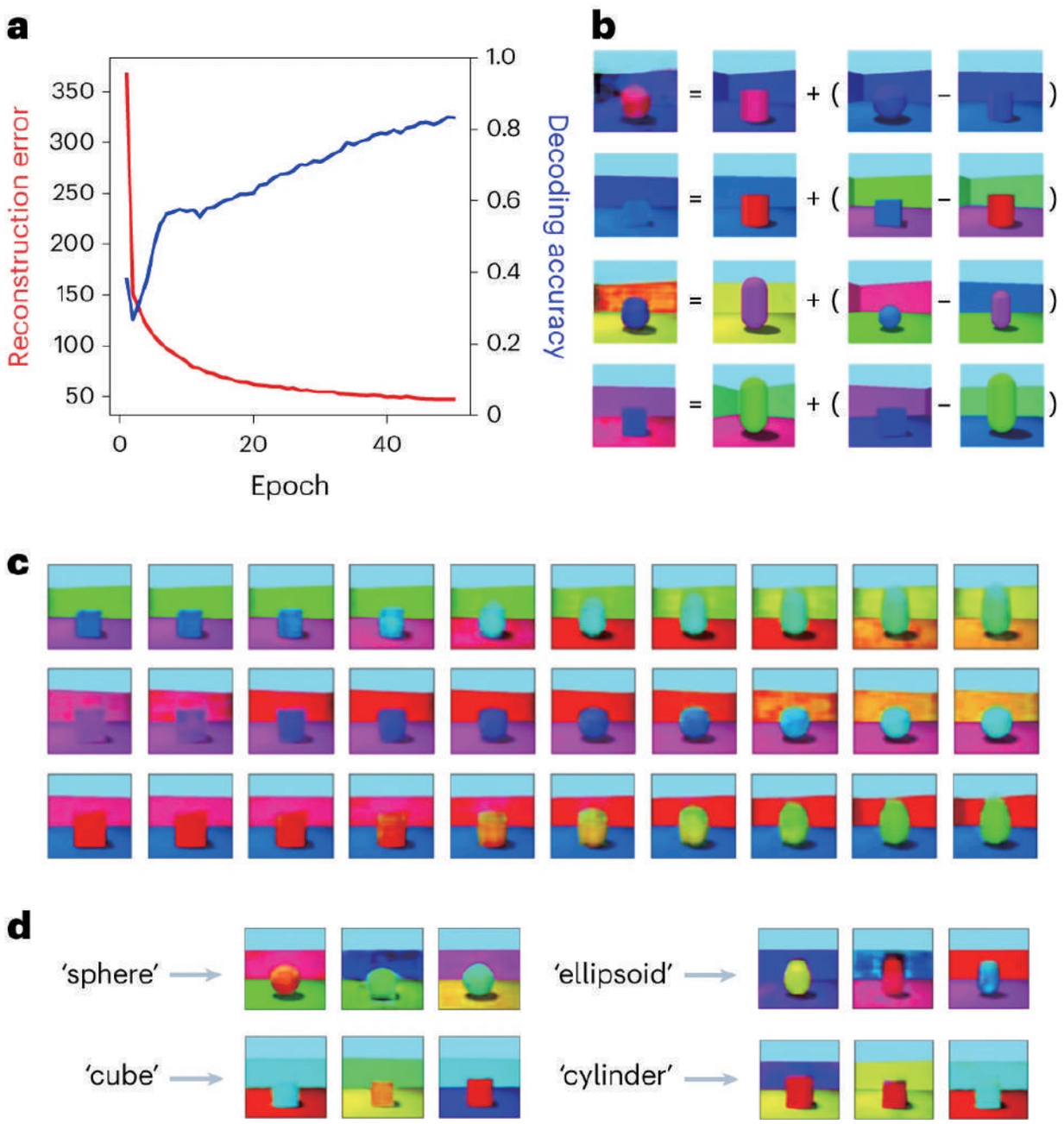


▷图2:场景(森林遇到狼)被编码为与生成网络的潜在变量相关的可预测的概念(森林)特征和生成网络预测不佳的不可预测的感官特征(狼)的组合。来源:关联论文

将记忆巩固过程视作更精细表征替代粗糙表征,并不利于大脑高效运算。在感知过程中,每个体验元素的重建误差是通过生成模型计算出来的。那些具有高重建误差的元素在自联想网络中被编码为感觉特征,同时还与生成模型的潜在变量表示相关的概念特征也一并被编码。换句话说,每个模式都被分成两部分:一是可预测的组件(近似生成网络对模式的预测);另一个是不可预测的组件(具有高预测误差的元素)。这样的处理方式产生的向量比详细存储每个元素要稀疏得多,从而提高了大脑的处理效率,避免了类似于电脑硬盘碎片化导致的卡顿。

而为了验证上述架构是否能够拆解感官从而进行推理,研究者设计了如图3b中的题目。而前述的人工神经网络能够经由训练(图3a),完成对新概念的解码,并在记忆中对元素进行组合(这被称为想象,对应

图3c中的图像)。此外,这些模型还能超越个别经验,利用共同的抽象特征通过同一生成网络处理不同的记忆(图3d)。这说明该模型已经学习了数据的一些概念结构,支持“A和B有特定关系,那么类似的,C与X有对应关系吗?”这类推理任务,并为记忆的灵活重组提供了一个模型,而这被认为是情境思维(类比)的基础。

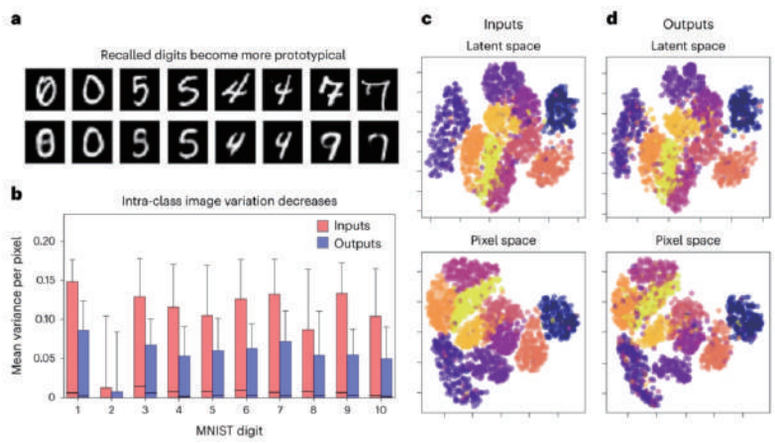


▷图3:生成模型中的学习、关系推理和想象。来源:关联论文

五、概念提取的另一面——扭曲和夸张

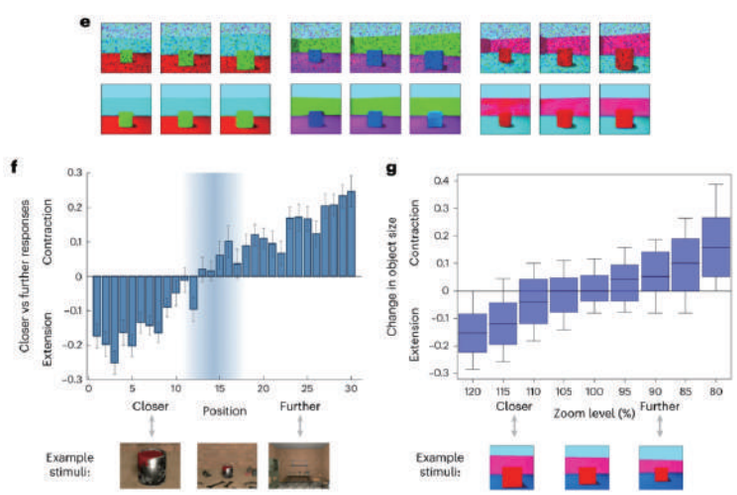
前文谈了记忆经由概念提取来关注更出乎意料的部分,然而,大脑的生成网络为了能尽快完成重构任务,会选择夸大新元素的某些特征。就如怦然心动的少年只会记得女孩吸引人的那一面。在人工神经网络的模型中,我们也能观察到类似的记忆扭曲现象。

研究者让模型记住手写数字，之后对比模型重构的数字与输入数字（图4a）。结果发现，重构的数字加强了各数字的典型特征，使得同一类别内的数字图像间差异减少（图4b），聚类后重建的图像边界更清晰，区分更明显（图4c，d）。这意味着即使在训练期间没有给出类别信息，生成网络的回忆也会将刺激扭曲为更典型的表征。随着对生成模型的依赖增加，这种失真程度也相应增加。



▷图4：手写数字训练过程中放大了各数字的典型特征。来源：关联论文

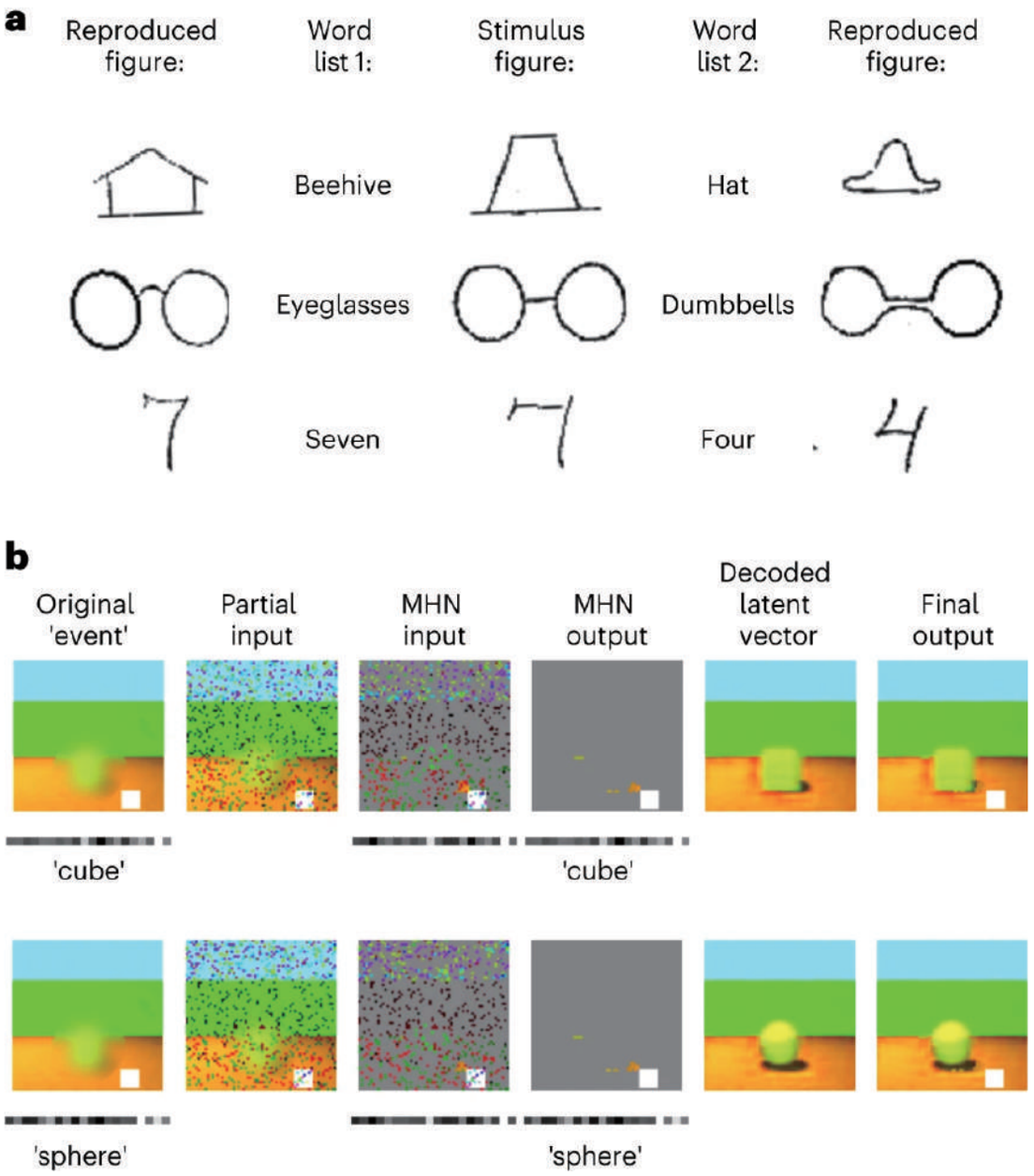
回到少年初遇梦中女孩的场景，在少年的记忆回放中，女孩可能总是处在画面的中心，画面的尺寸也不多，尽管实际上他可能只是在人群中远远地看着女孩。这种记忆中对画面边界的延伸和收缩，是记忆扭曲的一个典型表现。而在人工神经网络的实验中，生成网络提供的一系列新场景，与训练集中的场景相比，生成的场景大多被“放大”或“缩小”（图5f，g）；关键刺激的重建被扭曲为“典型视图”（图5e），这与人类记忆数据中观察到的现象相似。



▷图5：记忆扭曲中的边界延伸和收缩现象。来源：关联论文

六、并非无中生有的扭曲

正如情窦初开的少年会按照看过的电影小说中的样子来回忆和女孩的相遇，人们在记忆重建中也常常依赖于已存储的模式。例如，在编码后不久，回忆会偏向于该类别的“平均值”（图 6a，b）。当编码的错误阈值较高时，这种趋势更加明显，因为对“原型”表示的依赖程度更高，导致对新特征的回忆较少。在较低的错误阈值下，编码了更多的感官细节，即记忆轨迹的维度更高，重建误差也就越低。这表明失真较低，但会牺牲重构效率。



▷图6：a，在回忆过程中，在自联想网络中检索编码的概念，确定生成网络重建的原型场景。这会使回忆偏向于作为上下文提供的类。b，扩展模型中的记忆失真，当原始场景（包含模糊的模糊形状）使用给定概念进行编码时，由该类的潜在变量表示。然后，生成网络处理部分输入，以生成预测的概念特征和该概念的原型未预测的感觉特征（在本例中为白色方块）。然而，MHN中的模式补全再现了最初编码的感觉和概念特征，并且这些特征被重新组合以产生最终的输出。来源：关联论文

心理学研究发现，人们有时会错误地认为自己听到的故事包含了未曾出现的单词，而这仅仅是由于该单词与刚刚讲述的故事主题相关。这是记忆在重构过程中经由已有模式产生扭曲的另一个实例。图7展示的实验说明了这样的现象，并指出无论是人还是AI模型，都出现了这一现象。这让人想起大模型的幻觉，虽然两者有着不同的记忆机制，却有着相似的特征（缺陷）。

七、总结与扩展

为了生存，人类需要从过去的经历中提取模式，以预测未来的事件。而人工神经网络的模拟表明，当我们在休息时，大脑会重放记忆，从过去的经历中提取模式，帮助进行有利于生存的预测。在这个过程中，海马

体和新皮层在记忆、想象力和计划过程中协同工作，以便我们既可以回忆特定的经验，又可以灵活地构想新的情景。新皮质网络通过学习对场景的高效“概念”表示，捕捉它们的本质，使得重现过去的场景和创造新的场景成为可能。这一过程让海马体可以专注于编码新皮层无法轻易复制的独特特征，比如新颖的刺激，而无需存储每一个细节。

该框架将记忆整合视为一个持续的终生过程，而不是在单个数据集的编码过程中引入新的复杂性。它考虑到了对旧数据的修修补补，包括潜在表征的不稳定性，从而增强了记忆的鲁棒性。生成网络对新记忆的吸收有助于防止已经巩固的记忆发生灾难性的遗忘。

在这个框架下，我们可以理解为什么海马体的损伤会影响新记忆的形成，以及为何这种损伤与阿尔茨海默症患者提取已有记忆的缺陷有关。此外，这还说明了为什么海马体异常与需要记忆调节情绪的精神疾病（如抑郁症）有关。在该框架中，语义记忆变得独立于海马体：生成网络所学习到的潜在变量的表示形式构成了事件的“关键事实”，从而支撑了语义记忆；而情景记忆仍然依赖于海马体，其扭曲则源于大脑生成网络的运作。

该模型框架还对心理学对记忆的现有长期记忆二分提出挑战，即真正的情景记忆是否需要事件独特的细节，以及这是否需要海马体?在模型中，事件的独特细节最初是由海马体提供的，但也可以由生成网络提供。例如，如果你知道有人参加了你的生日聚会并送给你一份特别的礼物，这些个人语义事实不必依赖于海马体，但可以生成一个具有正确的事件特定细节的场景，这似乎就是情景记忆。随着生成模型的依赖概念越来越多，大脑在构建对新刺激的模型时将更多地依赖于语义记忆，这在经济上是合理的，但这同时意味着思维的僵化或者深入骨髓的偏见。

隐藏在情景记忆和语义记忆的差异之下的，是该研究中反复出现的“记忆巩固”。正是由于记忆巩固，人类可以通过少量尝试就学会新事物，但这也带来了记忆偏差的扭曲。情景记忆和语义记忆之间应当是一条连续的光谱，随着越来越多的概念被提取，认知的抽象层次也随之提高，逐渐与特定经验脱离，从而情景记忆转化为语义记忆。

无独有偶，近期，Nature　Communications的一篇相关主题论文[2]，基于强化学习框架构建的AI模型同样指出，海马体提取隐藏表征的过程与待完成任务与主体的行为实验共同进化，进一步佐证了该文观点。

总之，记忆与想象之间的边界远比我们想象的要模糊。海马体与新皮层的协作，不仅是记忆存储和回忆的基础，也是想象力和创造力的核心。正是这种内在的、动态的再创造过程，使得每个人的记忆和体验都富有个性，独一无二。（编辑：存源）

参考文献

关联论文：Spens, Eleanor, and Neil Burgess. "A generative model of memory construction and consolidation." Nature Human Behaviour (2024): 1-18.

[1] Spens, Eleanor, and Neil Burgess. "A generative model of memory construction and consolidation." Nature Human Behaviour (2024): 1-18.

[2] Cone, Ian, and Claudia Clopath. "Latent representations in hippocampal network model co-evolve with behavioral exploration of task structure." Nature Communications 15.1 (2024): 687.

谁在影响我们的决策?

讲者:Yuan Chang Leong

芝加哥大学助理教授。关注人类知觉、记忆、决策和社交互动的神经和计算机制。

扫码查看原文

人们倾向于认为自己的感知是对外部世界的真实反映,但这种观点长期以来一直受到心理学研究的挑战。相反,人们的感知经常受到内在状态或外部环境因素的影响,这种现象我们称之为动机性感知(motivated perception)。在视觉领域有一个经典例子,达特茅斯大学和普林斯顿大学的学生观看了同一场足球比赛。两队的球迷随后都报告说看到对方犯规更多。

为什么人们对世界的感知和感受如此不同?对这些发现的一种解释是,动机因素(如欲望和需求)对感知过程施加自上而下的影响,使人们倾向于看到他们想看到的东西,即人类的主观性。人类的主观性使我们成为独特、复杂、有趣的个体,而扭曲的感知和对现实的解释会导致病理性精神状态,因此,研究主观性对于理解人类的行为和相互作用至关重要。

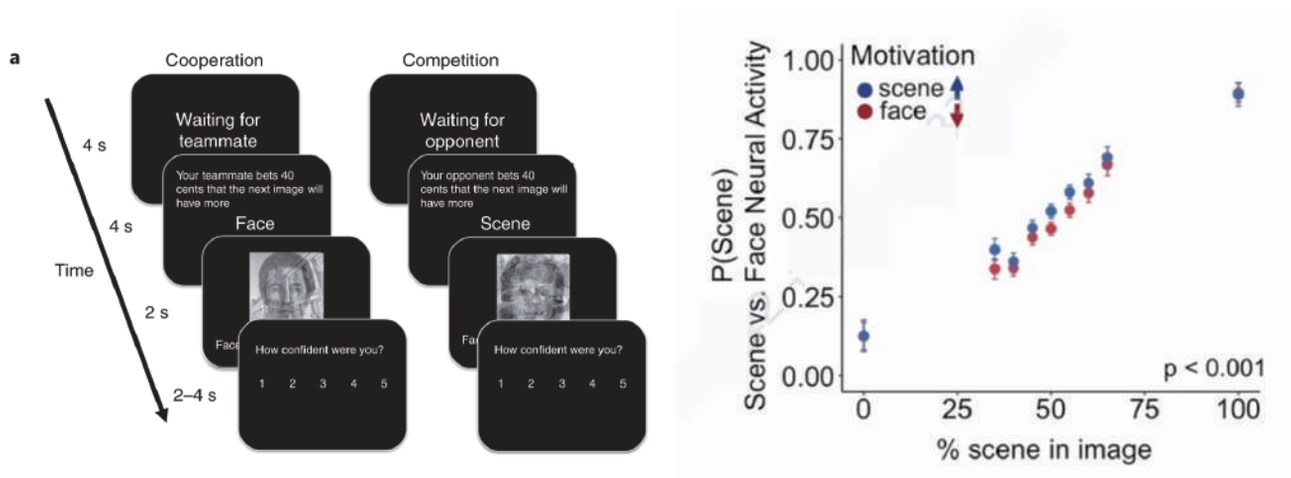
一、动机性视觉的大脑表征

当我们在进行一场网球比赛的时候,会有强烈的欲望想赢下比赛,这种欲望是如何影响我们对事物的感知的呢?

研究人员利用经过处理的房屋和人脸的混合模糊图像来研究这一问题。研究者向被试呈现经过处理的房屋和人脸的混合模糊图像,如果被试能准确判断图像包含“更多人脸”或“更多场景”,即可获得奖励。此外,如果能找到某种特定类型的图片,还会获得额外的奖励。为了评估动机对视觉感知的影响,研究人员在任务中加入了“竞争”条件,即这些被试将与队友一起执行任务。在没有任何信息提示的情况下,“队友”会对即将到来的图像进行预判,是有更多的脸还是更多的场景(图1左)。如果队友的预判是正确的,参与者就会获得奖励,如果队友的预判是错误的,参与者则失去奖励。

研究人员共招募了30名被试参与这项研究,这些被试在执行任务的同时进行功能性磁共振成像(functional magnetic resonance imaging, fMRI)扫描。结果表明,被试倾向于将图像归类为我们

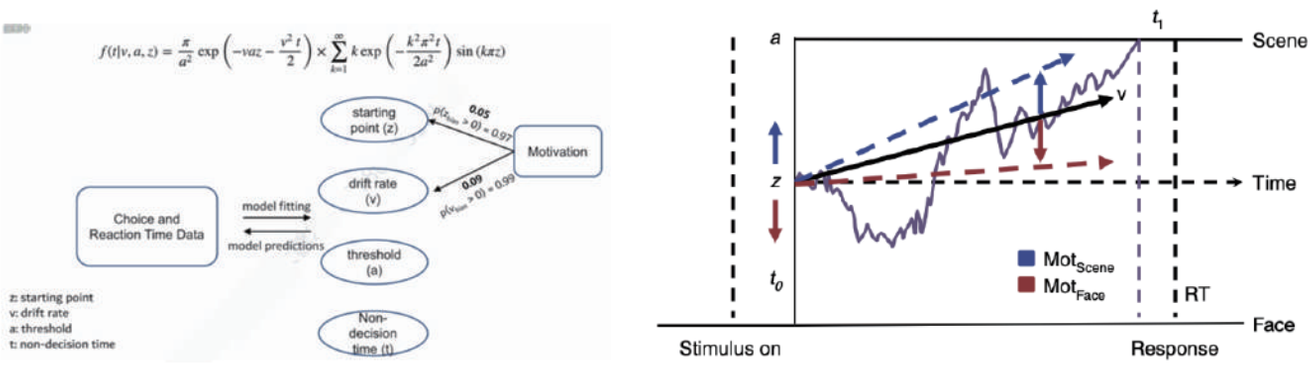
激励他们看到的类别(图1右)。通过一次次对大脑中的感官表征进行解码,结果表明动机增强了视觉皮层中欲望的感觉表征。



▷图1:动机性视知觉研究的实验范式及实验结果的心理测量函数。图源:由Yuan Chang Leong教授提供

为了理解动机是如何影响视知觉感知决策的,研究人员使用漂移扩散模型(drift diffusion models, DDM)对被试的选择和反应时间数据进行了拟合(图2)。DDM模型认为,决策是一个在噪音中收集信息的过程,当收集到的信息超过某个阈值,决策就产生了。在这个框架内,反应偏差可以建模为信息收集起点的偏差,感知偏差可以建模为信息收集速率的偏差。在每次试验中,决策取决于对两个决策阈值之一的噪音感官信息的积累。动机通过调节起始点(z)和漂移速率(v),对决策产生影响。

拟合结果表明,当被试被激励看到更多的人脸信息时,证据积累倾向于人脸类别,而当被试被激励看到更多的场景信息时,信息积累倾向于场景类别。综上所述,建模结果表明,动机通过增加与动机一致的反应倾向,并通过使感觉加工偏向于与动机一致的类别,从而使视知觉判断产生偏差。

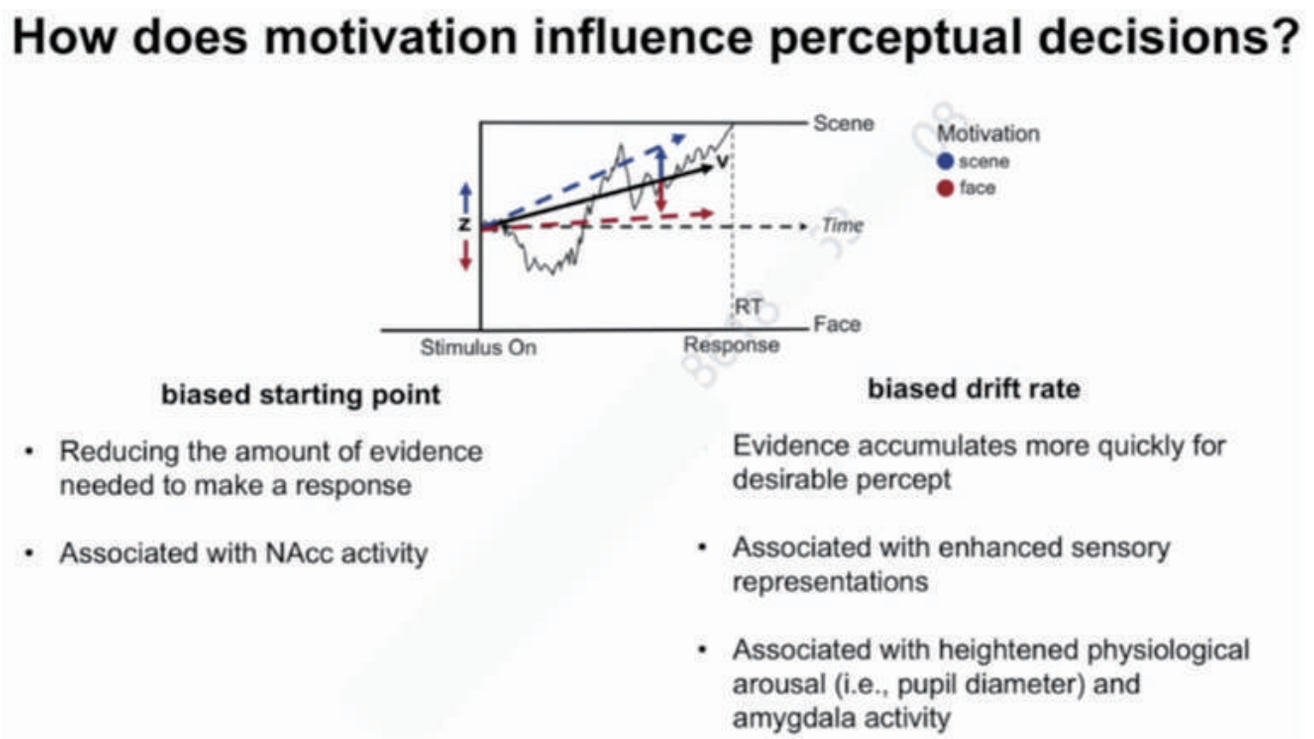


▷图2:DDM原理图。t0:刺激编码时间(非决策时间),t1:决策执行时间(非决策时间),a:决策阈值,z:起点,v:漂移速率,紫色实线:决策变量(即积累的信息),黑色实线:平均漂移量,MotScene:被驱使看到更多的场景信息,MotFace:被驱使看到更多的人脸信息。图源:由Yuan Chang Leong教授提供

那么,是什么介导了大脑的动机性视觉决策偏差?研究表明,觉醒(arousal)能够协调身体对动机性重大事件的反应,并很好地调节知觉判断的动机效应。研究人员招募被试执行上述视觉分类任务(图1),并测量被试的瞳孔扩张程度作为觉醒程度的测量。使用DDM计算模型的分析表明,觉醒通过使信息

积累偏向于理想感知而增强了动机效应。这些结果表明，高度的觉醒使人们倾向于看到被激励想看到的东西，而远离客观的环境（图3）。

杏仁核神经元直接投射到腹侧视觉皮层（ventral visual cortex）并与“情感注意”（affective attention）相关。进一步的研究表明，在知觉决策过程中，杏仁核活动的一次又一次波动与感官信息的动机增强相关，在动机性视觉决策中担任了中继站的角色（图3）。



▷图3：觉醒状态与杏仁核共同影响大脑的动机性视觉决策。图源：由Yuan Chang Leong教授提供

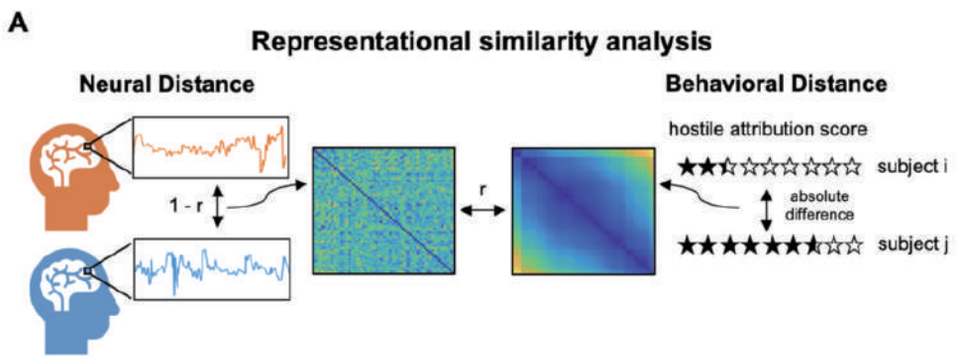
二、动机性社会认知的大脑表征

社会学认为，处理政治信息时的党派偏见往往会加剧社会分化。那么，这种偏见在大脑中是如何产生的？研究人员测量了观看移民政策相关视频时被试的神经活动，结果表明，即使在观看同样的视频，保守派和自由派被试表现出不同的神经反应，这种现象被称为神经分化（neural polarization）。研究表明，当视频中出现与威胁或与道德情感相关的语言时，神经分化更为强烈。这种现象被称为敌意归因偏见（hostile attribution bias）——将他人的行为解释为怀有敌意的倾向，即使这种行为是模糊的或良性的。

在测量被试的神经活动时，fMRI功能强大，但价格昂贵且有局限性；而功能性近红外光谱（functional near-infrared spectroscopy，fNIRS）是一种可扩展的、便携式的替代方案（非侵入性、便宜、便携、较少受运动影响，更适合研究个体差异）。研究人员使用测量敌意归因偏见的问卷对被试进行了实验，被试在执行任务的同时接受fNIRS成像。对实验数据进行表征相似性分析（representational similarity analysis），结果表明敌意归因偏见与神经活动变化相关（图4）。

进一步分析表明，敌对归因偏见能够影响腹侧前额叶皮层（ventromedial prefrontal cortex，VMPFC）的神经活动且在意图模糊的情况下效果最强（图5）。研究人员还构建了分类模型，可以以75%的

准确率对高归因偏差和低归因偏差的参与者进行分类。实验结果还表明偏好复杂行为归因的个体具有较低水平的敌意归因偏见。



▷图4：表征相似性分析。图源：由Yuan Chang Leong教授提供

▷图5：敌对归因偏见对VMPFC活动的影响。图源：由Yuan Chang Leong教授提供

总而言之，研究人员利用特定行为范式、结合fMRI成像与DDM建模找到了动机性决策在大脑中的表征形式。在这一过程中，觉醒与杏仁核活动扮演了重要的角色。利用fNIRs成像，研究人员还发现VMPFC活动与动机性社会认知高度相关。因此，我们的决策不仅受外部因素的影响，还与内在的动机（如欲望和需求）密切相关。

参考文献

[1] Leong, Y.C., Hughes, B.L., Wang, Y. et al. Neurocomputational mechanisms underlying motivated seeing. Nat Hum Behav 3, 962–973 (2019). https://doi.org/10.1038/s41562-019-0637-z

[2] Leong, Y. C., Dziembaj, R., & D' Esposito, M. (2021). Pupil-Linked Arousal Biases Evidence Accumulation Toward Desirable Percepts During Perceptual Decision-Making. Psychological Science, 32(9), 1494–1509. https://-doi.org/10.1177/09567976211004547

[3] Calabro R, Lyu Y, Leong YC. Trial-by-trial fluctuations in amygdala activity track motivational enhancement of desirable sensory evidence during perceptual decision-making. Cereb Cortex. 2023 Apr 25;33(9):5690-5703. doi: 10.1093/cercor/bhac452

[4] Leong YC, Chen J, Willer R, Zaki J. Conservative and liberal attitudes drive polarized neural responses to political content. Proc Natl Acad Sci U S A. 2020 Nov 3;117(44):27731-27739. doi: 10.1073/pnas.2008530117

当我们获得信息时，我们获得了什么？

作者：张旭晖

心理学在逃博士生，关注行为决策和认知建模，科普爱好者。

扫码查看原文

21世纪被誉为信息时代，信息已经成为了人们赖以生存并与世界交互的必要工具。但我们真正理解“信息”是什么吗？它推动了时代，但对我们的生活和思维方式又有何实质影响？

著名数学家克劳德·香农（Claude Shannon）提出的信息论，为我们提供了探讨信息量化、存储和传播的理论框架。按照信息论的观点，信息是一种“线索”，它的关键作用是减少对未知事件的不确定性。如果了解某些事件能帮助我们减少对另一个事件的不确定性，那么这些知识就构成了信息。例如，初来乍到的员工对公司知之甚少，他对周围同事充满了不确定感。不过，在和同事们约过几次饭后，他对同事们的言行举止有了一些观察，他逐渐开始了解同事们的兴趣爱好、性格特点。这些信息使得同事们不再那么陌生，使得新员工的不确定感大大降低。

在我们的日常生活和人际交流中，信息无处不在，它深刻影响着我们的人际交流和认知过程。对于神经科学家来说，信息论将人类的大脑视为一台信息加工机器，信息就像是神经系统中的基础货币，不同神经元通过电信号和神经化学信号的传递相互交流，这些信息在大脑中不断地被修改和组合，形成了我们的认知和记忆。

那么，电信号和化学信号就是信息吗？神经科学家们对此也并未形成统一共识[1]，用“基础货币”这样的修辞学方式来定义信息，可能还无法全面反映信息的本质。

近年来，信息论领域涌现出一种新的观点：信息并非单一的实体，而是由多种形态构成。这种观点被称为“信息分解”（information decomposition）[2]，它试图将信息解构为独特信息（unique information）、冗余信息（redundant information）和协同信息（synergistic information）三大类。这样的分解为理解信息的本质提供了新的视角，有助于我们更综合地认识大脑的信息结构，理解认知加工过程。

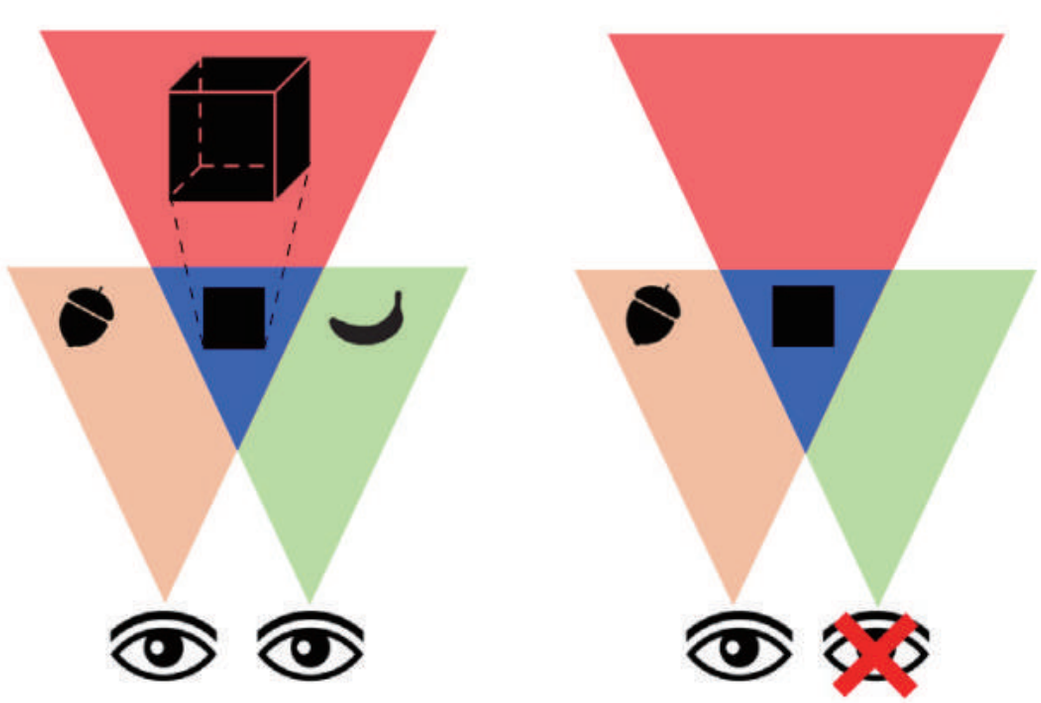
一、信息分解：多元的信息

三类信息各有特点，在信息加工过程中扮演着不同的角色。以视觉信息加工为例，人的视觉系统包含中央视野和边缘视野。其中，边缘视野帮助我们捕捉宽广的环境信息，包括物体的大致位置和模糊的细节。双眼边缘视野中的信息各不相同，属于独特信息。因为当我们闭上一只眼睛（左眼）时，相应眼睛（左眼）的边缘视野中的信息便会丢失，这时，大脑只能接收到另一只眼睛（右眼）的信息。例如，在驾驶车辆时，边缘视野往往能帮助司机从后视镜察觉到两侧和后方的车辆情况，假如左后方有车正在快速接近，司机左眼突然不舒服闭上了一瞬，那么，司机很有可能因未能及时察觉而面临意外事故。

相较于边缘视野，中央视野是人类视觉范围的中心区域，承载了我们看到的大多数详细信息。对于司机来说，同一车道正前方的车辆同时位于双眼的中央视野内，即使某一瞬间司机闭上了左眼，仅凭右眼他依然能察觉到前方车辆的动态。这样的信息被称为“冗余信息”。其主要优势在于稳健性，不同来源提供的信息相同，这种过度表征保障人们在任何一处信息源受损时依然能够获得所需的信息。不过，冗余信息的缺点也十分明显，它并没有充分利用大脑所有的信息收集能力。

最后一类信息是协同信息。单眼是无法产生立体视觉的，世界在人的眼中如何变得立体？这有赖于双眼的相互协作，由于双眼在头部的位置不同，看相同物体存在视角的差异。视觉皮层会接收到两幅稍有差异的二维图像信息，根据双眼视差的程度或两幅图像间的差异来判断眼前物体的距离，进而形成了对眼前物体深度的感知，产生了立体感。

在驾驶场景中，双眼协同帮助司机准确判断与前车或障碍物之间的距离，任何一只眼睛受损都会对安全带来威胁。相较于独特信息和冗余信息，协同信息最大的特点是效率，它充分利用了大脑神经系统不同部分之间的交互，实现了1+1>2的效果，在帮助人们处理复杂任务方面具有重要作用。



▷图1：图中橡果和香蕉是独特信息，矩形是冗余信息，立方体则是协同信息，需要双眼共同作用才能知觉深度。图源：参考文献[2]

人体作为一种高度复杂的系统，信息分解框架不仅为人们理解信息的结构提供了更细致的视角，在其他系统中也多有应用，如元胞自动机、社会经济数据、人工神经网络等[3]。

二、大脑如何整合信息？

信息整合是神经科学和认知科学中一大基础概念，然而，研究者们对于这一概念却存在两种不同的理解：整合即一体（integration-as-oneness）和整合即协作（integration-as-cooperation）。

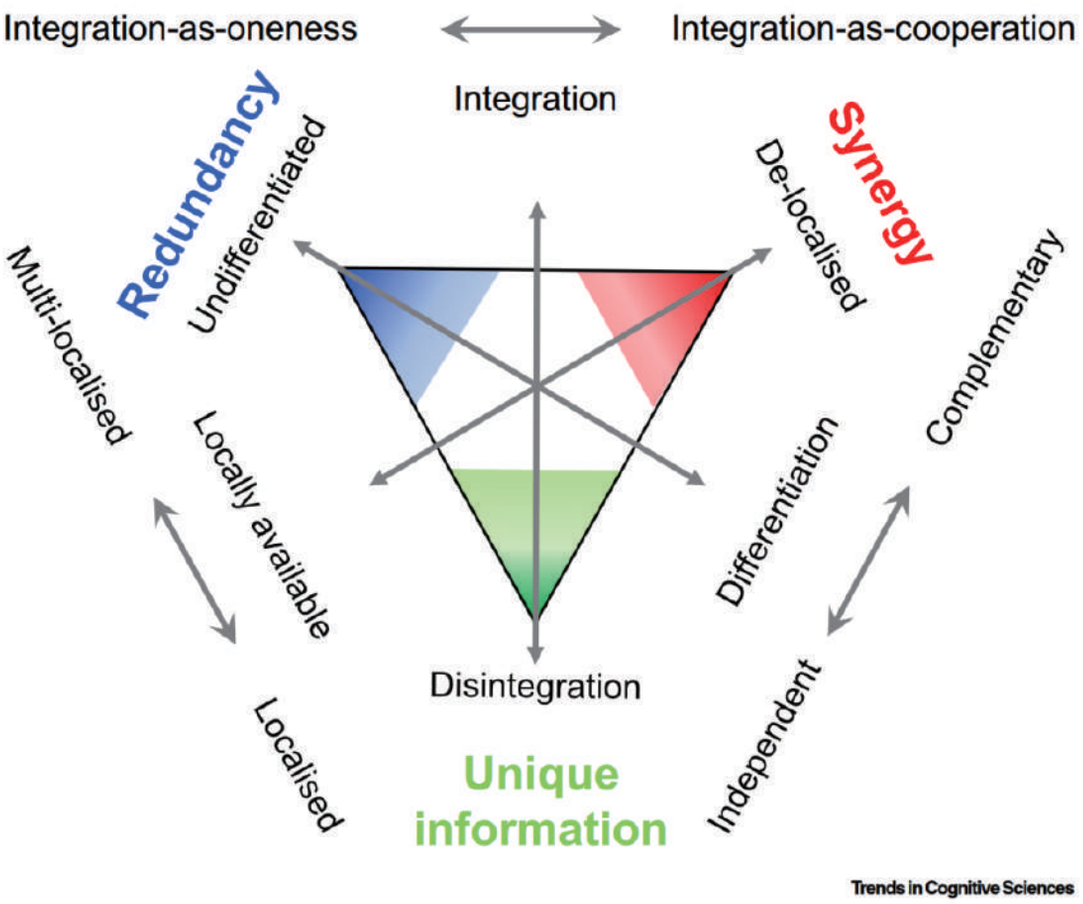
一体化观点认为，在分析大脑数据时，假如发现不同脑区的活动存在强相关或同步性，那么，研究者会推测两个脑区是高度整合的。这源于一种直觉：整合的元素会表现得像一个元素。两个元素的活动同步程度越高、表现越相同，它们的整合程度也越高。协作化观点则认为，当不同元素可以相互补偿时，系统的信息处理能力会受益于不同元素间的交互，这种交互即为整合的体现。

从信息分解框架的角度来看，整合实际上是独特信息的反面，一体化对应着冗余信息的概念，而协作化对应着协同信息。然而，传统的神经科学研究往往基于相关性来推测神经系统不同元素间是否存在整合，这种方法难以准确区分一体化和协作化的整合形式。高相关通常暗示着信息的冗余，而低相关复杂得多，可能意味着信息的独特性或协同性。

为了更好地区分冗余和协同，早期研究者开发出协同-冗余指数，来反映系统中协同与冗余的平衡程度。假如系统中的信息总和超过了各个组成部分的贡献总和，那么可以推测系统中存在协同。相反，假如系统各部分的贡献总和更高，那么系统中一定存在冗余。这种方式非常直观，但无法应对协同和冗余同时存在的情形，也无法精确识别协同信息。

与相关法或协同-冗余指数不同，信息分解通过计算系统的转移熵（transfer entropy, TE）来更精确地分析信息加工。例如，如果某个系统中两个脑区的时间序列数据X和Y之间的转移熵数据显示，从X到Y的转移熵高于从Y到X的，那么可以认为X对Y有着明显的“影响”。

以癫痫为例，这是一种大脑功能连接紊乱的现象。传统观点认为患者的大脑信号彼此高度同步，仿佛是一体的，也就是说，大脑不同区域的冗余是癫痫发作的可能成因。但使用信息分解框架分析癫痫患者的脑电数据，研究者发现，相较于发作前，癫痫发作时患者皮层下区域向皮层区域传递的冗余信息和协同信息都有所增加。更进一步地，分析不同区域的深部电极记录，结果发现，特定皮层下区域向皮层区域传递的独特信息增加可能是引发皮层振荡的主要原因，这为定位癫痫发作区域提供了更直观的证据[4]。



▷图2：信息分解为认知科学提供了一种统一的框架，图中穿过中心三角区域的每个双向箭头都表示认知科学和神经科学中一组对立的概念，箭头的一端对应信息分解框架中的某一种信息，另一端则混淆了两种信息，如整合（integration）与非整合（disintegration），非整合意味着独特信息，整合则包含冗余和协同两类信息。局部和分化在正文中未提及。对于局部，当某一信息源只包含独特信息，那么，该信息源的所有信息都是局部可获得的，即只能从该信息源获得，相反，当该信息源只携带冗余信息，那么，该信息源的所有信息都是多局部可得的。对于分化，当系统不同部分表现各异，即未表现出一体化时，它们被视为是分化的，不过，两个分化的部分既有可能是相互独立的，也有可能是相互补偿的。图源：参考文献[2]

三、大脑如何平衡冗余和协同？

冗余和协同作为广泛存在于大脑不同区域的两种重要交互方式，二者的区分得到了大量研究的支持。一项使用NeuroSynth数据库、涉及15000余项影像学研究的元分析发现，冗余信息在感觉运动加工中起到至关重要的作用。作为大脑的输入-输出系统，稳定的感觉运动加工对生存至关重要，冗余的交互方式则为这一过程提供了必要的稳健性。

另一方面，协同信息则扮演了大脑中“全局工作空间”的角色，是完成高级认知功能的关键。高度协同的脑区表现出更快的有氧酵解（aerobic glycolysis）、更多样的神经递质受体表达，为灵活快速的供能、突触形成以及神经调控提供了基础[5]。

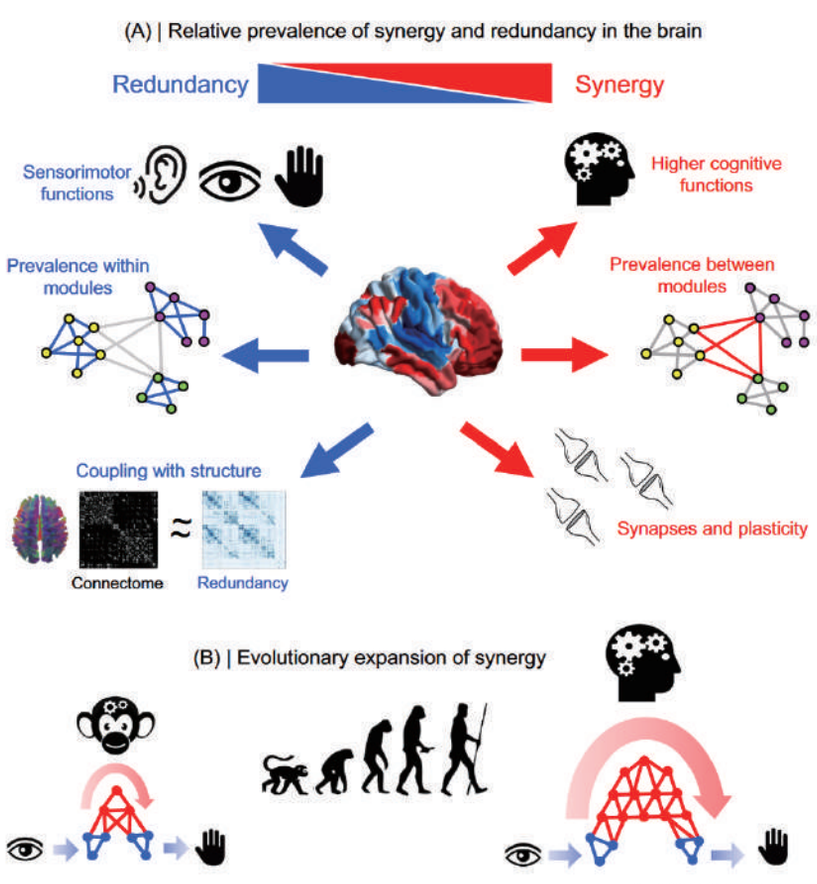
在宏观尺度上，人类的静息态fMRI数据表明，总体来看，协同似乎更为普遍[5]，但也与具体脑区有关。额叶和顶叶联合皮层以协同为主，这些区域是大脑中整合多模态信息的重要区域，其中，额叶联合皮层主要负责长期规划和决策，顶叶联合皮层则负责空间定位、手眼协调等功能。相反，冗余则主要出现在较为初级的、处理单一模态信息的皮层区域，如初级视觉皮层、躯体运动皮层和听觉皮层。相较于其他灵

长类动物，人类拥有更为发达的联合皮层，这意味着我们的大脑更多地以合作、协同的方式整合多重信息。这些发现无疑为人类的认知优势提供了信息论的证据。

微观尺度的研究证据与宏观尺度的fMRI研究结论高度一致。例如，电生理记录发现，神经元之间的交互对于解释前额叶皮层中的脉冲活动非常重要，前额叶神经元常常会根据刺激和任务的变化展现出复杂而灵活的反应。然而，对于视觉皮层（V4，尤其是V（1），神经元之间的交互在解释脉冲活动方面的贡献度要低得多[6，7]。

除了以人类为对象的研究证据，人工神经网络研究也发现，在早期，网络中冗余占主导地位。但随着学习的深入，一些神经元开始变得专一化，提供更多的独特信息。当人工神经网络在学习多任务时，它们需要足够灵活以整合不同来源的信息，此时，协同作用会不断增强。一旦高度协同的神经元被破坏或移除，整个网络的表现会明显变差。假如在训练时随机关闭神经元，神经网络变得更加冗余但也更加稳健，训练完成后面对人工损伤也有更好的抵御能力[8]。

综合来看，冗余没能充分利用神经系统处理信息的能力，但提供了强大的稳健性，是人类及众多灵长类动物感觉运动功能的基础。协同与更高阶的信息加工息息相关，更具效率也更为灵活，是人类成为万物之灵的重要助力，但一旦某一部分受损，有赖于不同神经元/脑区协同作用的高阶认知功能也更易恶化。



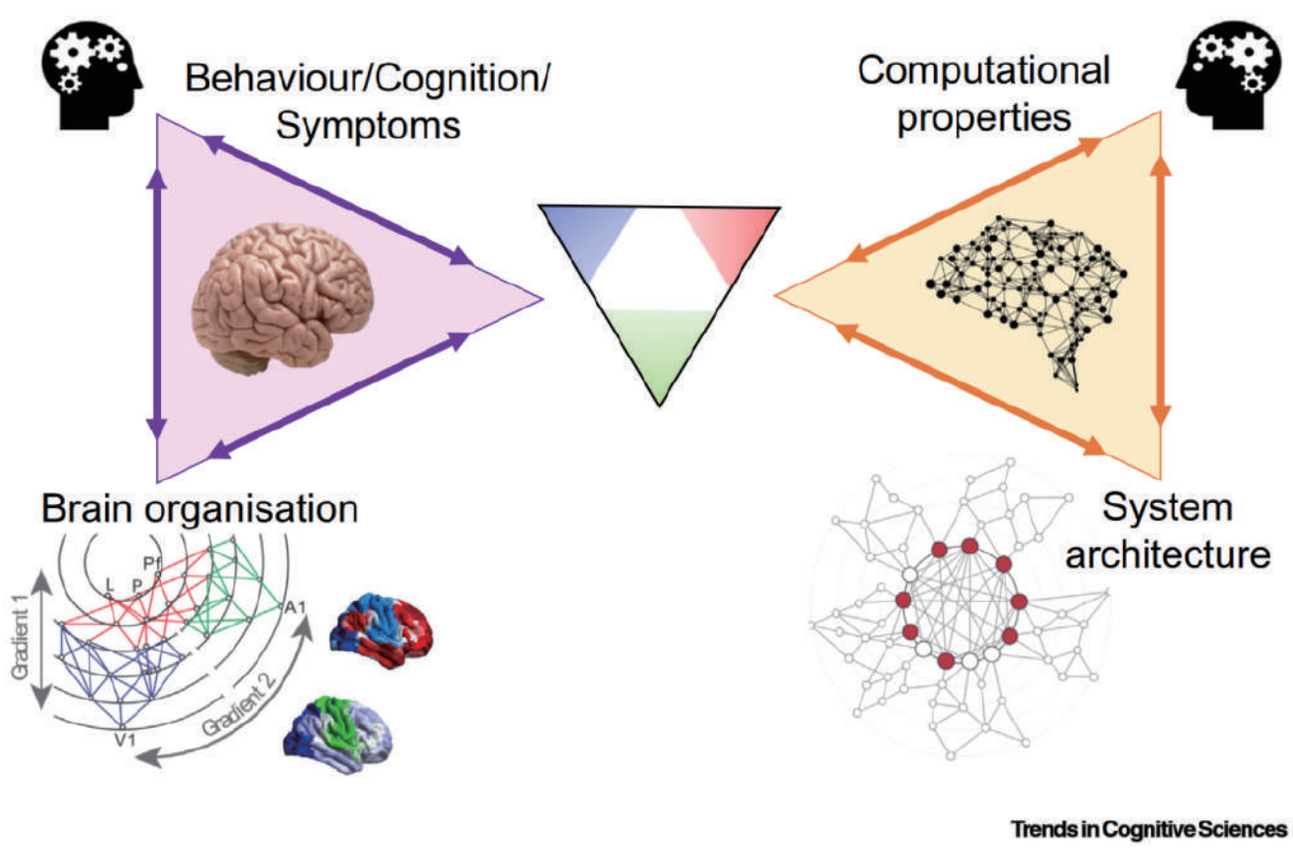
▷图3：（A）图中蓝色和红色分别对应大脑中冗余和协同为主的区域，蓝色脑区主要与初级感觉运动加工有关，以底层的结构化连接为基础，展现出高度模块化的网络组织，即不同部分只专一化地加工单一模态的信息，如视觉皮层对应视觉信息。红色脑区负责复杂认知加工，并且与突触密度、突触树突相关基因等有关，拥有更强的可塑性。（B）协同的演化史。冗余信息对不同物种都相对稳定，而与其他灵长类动物相比，人类大脑中表现出更加发达的协同信息，这可能是由于人脑拥有更发达的皮层区域。图源：参考文献[2]

四、人工智能设计的新蓝图

信息分解框架提供了探究大脑信息加工的全面视角，也为我们理解人类的进化优势提供了许多新的证据，除此以外，信息分解也将有助于研究者设计更加类人的人工智能系统。当前，人工神经网络已在多个领域展现出强大的能力。关键的问题是，这些系统是否也像人脑一样依赖于协同作用？

近期，人工智能的进步主要来源于模型的规模。研究者观察到，随着人工智能模型的规模扩大和其在处理多任务方面的灵活性增强，我们可以观察到模型表现出更多的协同作用。这些都可以视为人工智能系统越来越类人化的标志。但协同作用内在的易损性对于人工智能来说也是一种隐患，因此，未来在设计人工智能系统时，应对系统中不同类型的信息加以辨别，将信息分解框架发展为一种理解复杂系统的通用语言，这也将有助于人们解开许多人工智能模型的“黑箱”。

反过来，人工智能也能为信息论研究提供强而有力的检验环境。比如，我们已经观察到在面对复杂任务时，协同作用会增强。那么，假如通过如演化算法的方式，将人工智能系统设计成为更偏好协同作用，它们是否能更好地应对复杂任务呢？更进一步地，假如某个人工智能系统只拥有协同作用的能力，这将为我们提供一个独特的视角来直观探究协同作用的优势和局限性，这种极端环境是任何生物系统都无法实现的。



▷图4：使用信息分解框架作为连接生物与人工智能的罗塞塔石碑（罗塞塔石碑上包含希腊文、古埃及文字等多种语言，是研究古埃及历史的重要参照）。在生物的大脑中，信息加工、信息分解帮助我们理解大脑结构和功能组织与认知和行为变量之间的关系，类似地，在人工智能系统中我们也可以建立起系统架构与计算能力和表现之间的关系。无论是生物大脑皮层，还是人工智能系统，信息加工、信息分解不依赖于载体，可以成为一种通用的语言。图源：参考文献[2]

五、结语

现实生活中，如火如荼发展的人工智能技术时不时引发人们的惊呼，其实，大脑的精巧程度远非人工智能可及，信息分解框架无疑让我们朝大脑之谜再度前进了一步。未来，我们还能以怎样的方式分解信息？我们如何基于信息分解创造出更先进的人工智能系统？对大脑的信息结构更加全面的了解能否帮助我们破解困扰人们的诸多精神障碍？也许有朝一日，当我们将大脑的基础货币拆解得足够细微，人造大脑便不再是梦想。（编辑：存源）

参考文献

关联论文：Luppi, Andrea I., et al. "Information decomposition and the informational architecture of the brain." Trends in Cognitive Sciences (2024).

[1] Timme, N. M., & Lapish, C. (2018). A Tutorial for Information Theory in Neuroscience. Eneuro, 5(3), ENEURO.0052-18.2018. https://doi.org/10.1523/ENEURO.0052-18.2018

[2] Luppi, A. I., Rosas, F. E., Mediano, P. A. M., Menon, D. K., & Stamatakis, E. A. (2024). Information decomposition and the informational architecture of the brain. Trends in Cognitive Sciences, 0(0). https://doi.org/10.1016/j.tics.2023.11.005

[3] Mediano, P. A. M., Rosas, F. E., Luppi, A. I., Carhart-Harris, R. L., Bor, D., Seth, A. K., & Barrett, A. B. (2021). Towards an extended taxonomy of information dynamics via Integrated Information Decomposition (arXiv:2109.13186). arXiv. https://doi.org/10.48550/arXiv.2109.13186

[4] Faes, L., Marinazzo, D., & Stramaglia, S. (2017). Multiscale Information Decomposition: Exact Computation for Multivariate Gaussian Processes. Entropy, 19(8), Article 8. https://doi.org/10.3390/e19080408

[5] Luppi, A. I., Mediano, P. A. M., Rosas, F. E., Holland, N., Fryer, T. D., O' Brien, J. T., Rowe, J. B., Menon, D. K., Bor, D., & Stamatakis, E. A. (2022). A synergistic core for human brain evolution and cognition. Nature Neuroscience, 25(6), Article 6. https://doi.org/10.1038/s41593-022-01070-0

[6] Chelaru, M. I., Eagleman, S., Andrei, A. R., Milton, R., Kharas, N., & Dragoi, V. (2021). High-order interactions explain the collective behavior of cortical populations in executive but not sensory areas. Neuron, 109(24), 3954-3961.e5. Scopus. https://doi.org/10.1016/j.neuron.2021.09.042

[7] Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature, 497(7451), Article 7451. https://doi.org/10.1038/nature12160

[8] Proca, A. M., Rosas, F. E., Luppi, A. I., Bor, D., Crosby, M., & Mediano, P. A. M. (2022). Synergistic information supports modality integration and flexible learning in neural networks solving multiple tasks (arXiv:2210.02996). arXiv. https://doi.org/10.48550/arXiv.2210.02996

你听到的音乐在脑中是怎样的?

作者:陈诗雨

临床医学博士、神经内科医生、医学插画师,目前为某高校教师。一直关注脑科学、健康教育及科普,致力于把医学科学知识以更生动的方式传播给更多读者。

扫码查看原文

音乐是生命体验中不可或缺的一部分。过去,人们想要表达音乐只能凭歌喉哼唱、靠乐器演奏;而如今,科学家已经能通过解码大脑来呈现脑中音乐——近期,加州大学伯克利分校(University of California, Berkeley, UC Berkeley)的科学家们成功地从脑电波中重建了大脑所听到的音乐。

这项成果来自UC Berkeley的Helen Wills神经科学实验室,于23年8月15日发表于Plos Biology杂志此前,该团队已经成功通过脑电波重建脑内语音。而这次的研究更进了一步,音乐所包含的信息显然远远大于语音。正如研究团队成员介绍,“音乐本质上是充满情感和韵律的——它有节奏、重音、抑扬顿挫,包含了比任何语言中有限的音素更广泛的含义。”

有趣的是,与使用古典音乐的传统方式不同,研究人员重建的音乐片段是来自英国摇滚乐队Pink Floyd发表于1979年的歌曲“Another Brick in the Wall, Part 1”。为什么团队选择了Pink Floyd的音乐、特别是这个片段呢?“在论文中,我们提到的科学原因是:这首歌非常具有层次感,它引入了复杂的和弦、不同的乐器和不同的节奏,使得分析变得有趣。”认知神经科学家、该研究的主要作者Ludovic Bellier说道。“不过,不太科学的原因是我们真的很喜欢Pink Floyd。”

这项研究共纳入了29名耐药性癫痫患者,他们均接受过颅内电极植入,以监测癫痫的发作。纳入本次研究的电极共2668个,每位患者36个到250个不等。研究人员为患者播放了“Another Brick in the Wall, Part 1”,让他们聆听这段音乐,此后利用人工智能对电极记录的信息进行解码、重建。

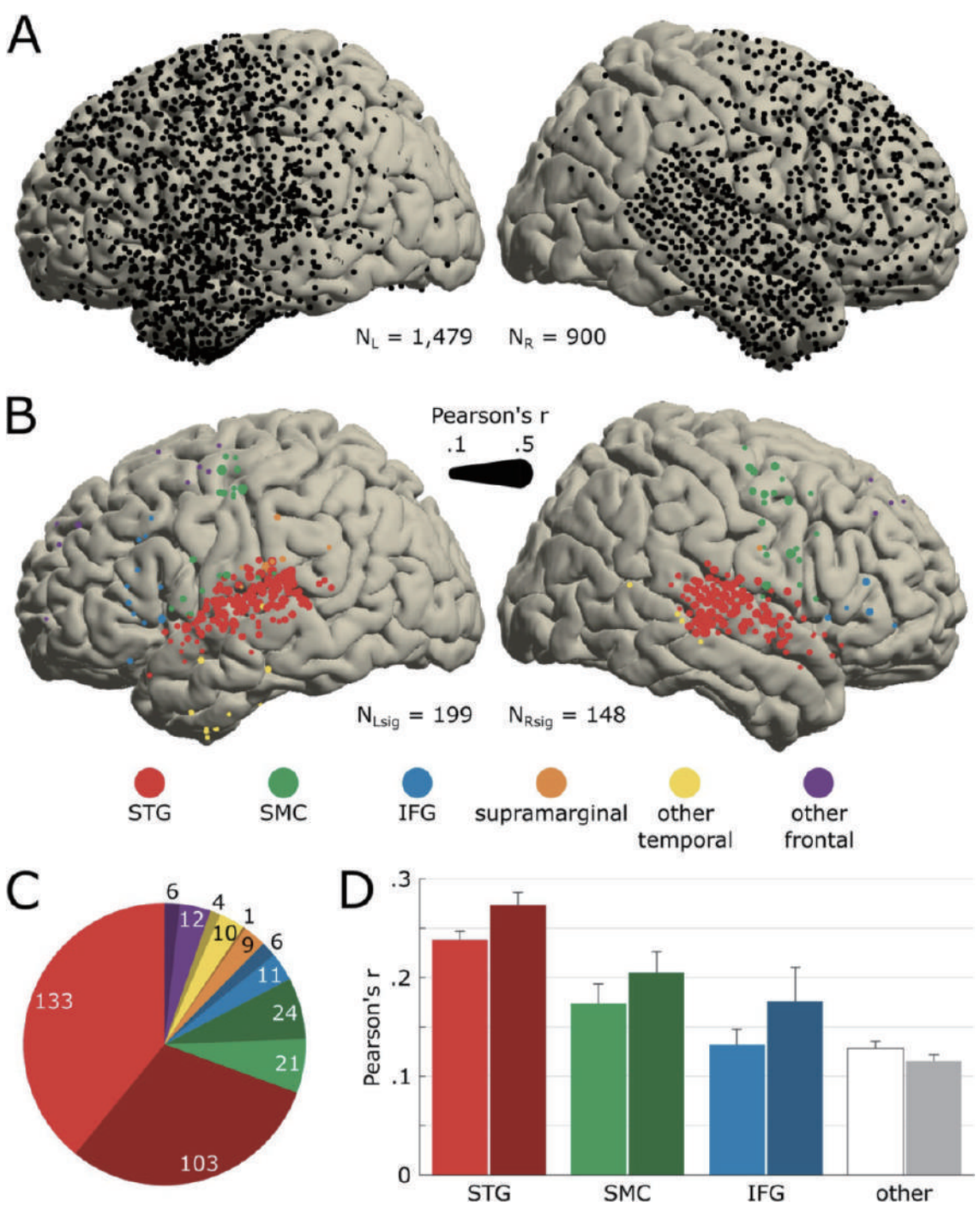
先来听听这惊人一致的原曲和脑电波重建曲吧!可以听到,虽然重建的音频有些模糊不清、仿佛“在水下说话”,但乐曲的走向、一些重音处的歌词如“the wall”以及节奏都是清晰可辨的。

一、哪些脑区记录了音乐?

为了确定哪些部位的电极记录了歌曲声学信息的编码,研究者利用人工智能对2379个无伪迹电极记

录的脑波信息进行了谱时感受野(STRF)拟合,评估不同位置电极记录的神经元高频活动到底能够多准确地被歌曲的听觉频谱图预测:预测程度越好,则该位置的电极便与记录音乐越相关。

图1A表示所有电极的覆盖范围。图1B显示,347个电极具有显著的STRF拟合结果,位于左半球的有199个,右半球的有148个。这347个响应电极绝大多数(87%)集中在三个区域:68%位于双侧颞上沟(STG),14.4%位于双侧感觉运动皮层(SMC,位于中央前回和中央后回),4.6%位于双侧额下回(IFG)。图1C、D中,较深的颜色表示电极位于右半球,较浅的表示位于左半球;双因素ANOVA分析显示,两侧半球的对比有统计学意义,电极响应更加集中的区域均是右半球。

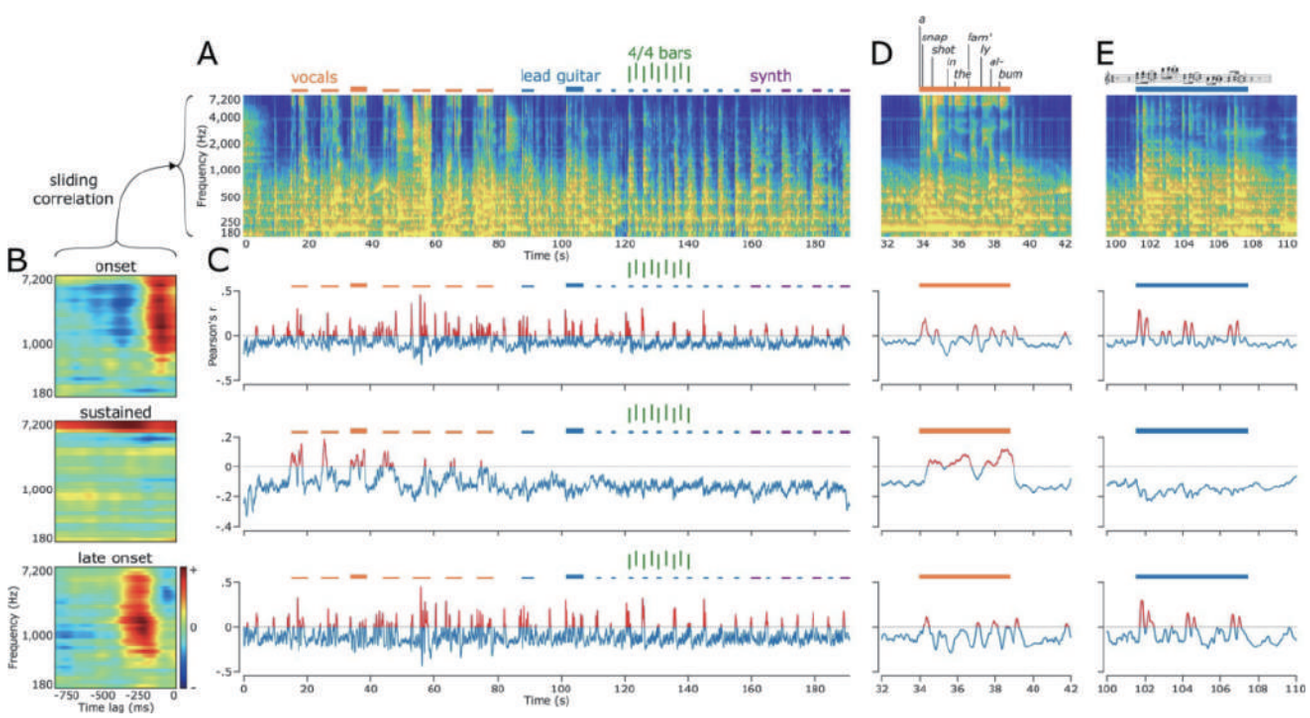


▷图1:响应电极的解剖位置

二、从脑海中重建歌曲需要多少信息?

科学家们从347个响应电极中随机抽取电极记录的数据,再利用人工智能解码其中信息、进行歌曲的重建。研究者发现,随机使用43个电极的数据即可达到最佳准确预测能力的80%;在单个患者上也类似,43个电极的信息已经可以进行解码,尽管解码的准确性较低;使用数据的持续时间与预测准确性之间也存在类似关系,例如,相比于使用完整的190.72秒的歌曲数据,使用69秒的数据即可以获得90%的重建准确性。

那么,放置电极的解剖位置对重建是否有影响呢?在移除不同解剖位对的电极信息后再解码,发现:(1)相比其他脑区,双侧STG具有独特的音乐信息;(2)相比左侧STG,右侧STG具有独特的信息;(3)左侧STG的部分音乐信息在右侧STG存在冗余编码。



▷图2:不同音素的解码

不同解剖位置的电极在解码音乐时是否具有不一样的功能?确实如此。在对所有响应电极的独立成分分析后,结果如图2所示:

(1)仅位于双侧后STG的“起始成分”,记录主音吉他或合成器的起始部分、及人声中音节核心的起始部分(图2B、C、D的第一行);(2)位于双侧中、前STG以及双侧SMC的“持续成分”,记录歌曲的人声部分(图2B、C、D的第二行);(3)位于双侧后、前STG,以及双侧SMC的“迟发型起始成分”,也与主音吉他或合成器的起始部分、及人声中音节核心相关,只是潜伏期更长(图2B、C、D的第三行);(4)位于双侧中STG的“节奏成分”,记录歌曲中速度为99bpm、贯穿整个歌曲的节奏吉他中的16分音符(图2E)。

了解了大脑是如何“接收”与“理解”音乐信号后,再来看响应电极的功能成分对音乐重建又有哪些影响?在移除不同相关功能的响应电极后再解码,发现:(1)右侧起始电极相比左侧具有独特的信息;左侧起始电极的部分信息在右侧起始电极中存在冗余编码。对于迟发性起始电极也观察到类似现象。(2)右侧节奏成分电极具有独特的信息,没有任何信息在左侧节奏电极中存在冗余编码。(3)尽管持续电极数量很多,

但移除它们的信息未发现任何影响。不过，由于电极的功能成分存在一定重叠性，所以对它们功能重要性的解读没有解剖位置那么准确。

三、为什么关注大脑中重建的音乐？

或许有人会对这项研究的目的产生疑问，为何我们要去听大脑对外界声音的映射？

其实，这类研究能够帮助我们更好地理解音乐和语言的处理方式。另外，在对疾病的认识上也能给我们一些实证性的启发，例如揭示为什么布罗卡失语症患者讲话费力，但通常却可以毫无困难地用唱歌的方式唱出相同的词。

这项研究也为将情感赋予合成语音奠定了基础。尽管研究重点放在音乐上，但研究人员认为，这项结果对于基于脑电波的语音合成将有很大帮助。无论哪种语言，人类的话语都包含着节奏、重音、抑扬顿挫等音乐性要素，这些要素构成了话语之中隐含的情感成分。“这些元素，我们称之为韵律，携带着无法仅仅用语言表达的意义。”研究者Bellier希望这个模型能够改进脑机接口技术，使语言辅助不仅能重构语音本身，还能重构话语中用韵律表达的意图。

过去，科学家们已成功重建脑内语音，能使中风或肌萎缩侧索硬化症等神经系统疾病患者通过植入式语音解码器来表达自己；但此类重建通常是机械、刻板的。研究者们希望这项成果最终能帮助失语患者恢复自然言语的音乐性。

“如果脑机接口能够用音乐中固有的韵律和情感来重新创造某人的言语，那么它所能重建的不仅仅是单词，不是机械地说，‘我，爱，你，’而是可以像真人一样大喊，‘我爱你！’”另外，这项研究重建的是研究对象听到的音乐，而研究者Robert Knight认为，未来的研究应向重建脑中想象的语音、音乐发展。“虽然他们没有记录受试者想象音乐时的大脑反应，但这可能是脑机接口未来的用途之一：将想象的音乐转化为真实的音乐。”不过，这项技术离实际运用还有很长的路要走，比如，目前获取的数据还基于有创的颅内植入电极；或许未来可以利用无创电极来实现数据收集，这将建立在脑机接口技术的发展之上。

总之，一旦这项技术成熟应用，是否我们就能依靠一个轻型头盔来创作心中的音乐呢？那时候会有怎样瑰丽的想象化为现实？让我们无限遐想。（编辑：韵珂）

参考文献：

Bellier L, Llorens A, Marciano D, Gunduz A, Schalk G, Brunner P, et al. (2023) Music can be reconstructed from human auditory cortex activity using nonlinear decoding models. PLoS Biol 21(8): e3002176. https://doi.org/10.1371/journal.pbio.3002176

音乐是一种语言吗?

编译:马文博

川北医学院精神卫生学院教师。参与国自然项目三项,发表SCI学术论文5篇。认知神经科学博士(毕业于北师大心理学部)。

扫码查看原文

语言和音乐贯穿人们的日常生活,人们用语言沟通,用音乐传递情绪,陶冶心境。从共性来看,语言和音乐的加工都涉及从基本单位生成结构化表征。虽然语言和音乐可以在不同类型的单位(例如,音素、语素、音符、和弦)上表现出来,但它们都是将离散单位组合成具有统一定义层次的结构。研究发现,与音乐有关的左脑区域通常与语言区域重叠,包括颞上回、额下回和顶叶下回。

然而,语言和音乐之间也存在一些实质性区别,如节奏结构、音高的使用、表征单位的"意义"、结构构建机制以及应用范围等的差异。有观点认为,"音乐和语言是可分离的资源"。语言的一些特性(例如,语音感知)已被证明在大脑中与音乐分离,但语言的许多其他特性却没有得到广泛的处理。特别是衡量语言和音乐复杂性的某些标准仍有待在高时空分辨率下的进一步探索。由于语言和音乐在结构构建机制上似乎有所不同,因此可能需要不同的结构敏感性测量方式来捕捉特定领域的效应。

为此,Nitin Tandon教授及其团队结合颅内记录的方法,在因果扰动的同时对大脑皮层活动进行高时空分辨率的记录。他们以一位癫痫患者为受试者,探究了音乐和语言结构处理的时空动态特征,通过利用特定领域的复杂性分析处理,再结合清醒开颅手术进行直接皮层刺激影像,对音乐和语言的神经机制展开了深入研究。结果发现,音乐和语言在大脑中虽然共享资源,但处理特定领域结构时具有不同的神经特征。该成果于2023年7月发表在iScience杂志上。

一、受试者与任务范式

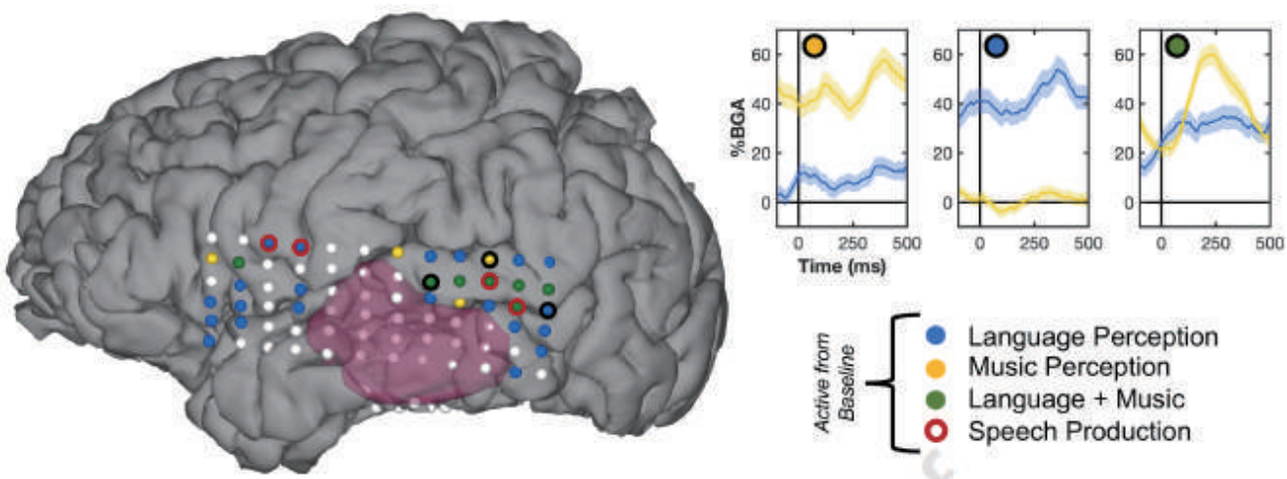
受试者为一名男性新发癫痫发作导致暂时失语的患者。他是一名专业音乐家,拥有丰富的钢琴经验,从童年时期就开始训练。在手术前,患者完成了在手术室开展的所有测试(基线期)。在基线期,患者在所有测试的语言任务和音乐任务中都达到了上限。他在蒙特利尔失音症评估电池(MBEA)测试中获得172/180分。患者在一系列琴键上进行了手指敲击测试,练习自由式演奏,以建立对键盘设置的熟悉程度。

Nitin Tandon教授及其团队对受试者进行了三个行为任务的数据采集，包括(1)听觉语句重复任务：重复口语语句；(2)旋律重复任务：单手使用MIDI键盘重复呈现的六音序列；(3)听觉命名任务：对常见物体的定义提供单字答案。

二、研究发现

(1)音乐和语言的ECoG影像特征

在句子和旋律重复任务期间，研究者同步记录了受试者的脑区ECoG活动(图2)。在每个单词呈现后100-400毫秒期间，发现28个侧电极明显高于基线。这些活跃的语言电极位点分布在颞叶后皮层和额叶下皮层。在旋律任务的音调呈现过程中，11个电极在同一时间窗口内对每个音调都显着活跃，大部分聚集在后颞上回(pSTG)周围。7个电极在音乐和语言感知方面均显著活跃，主要位于pSTG。在语言产生过程中，腹侧中央前回的两个电极明显活跃。



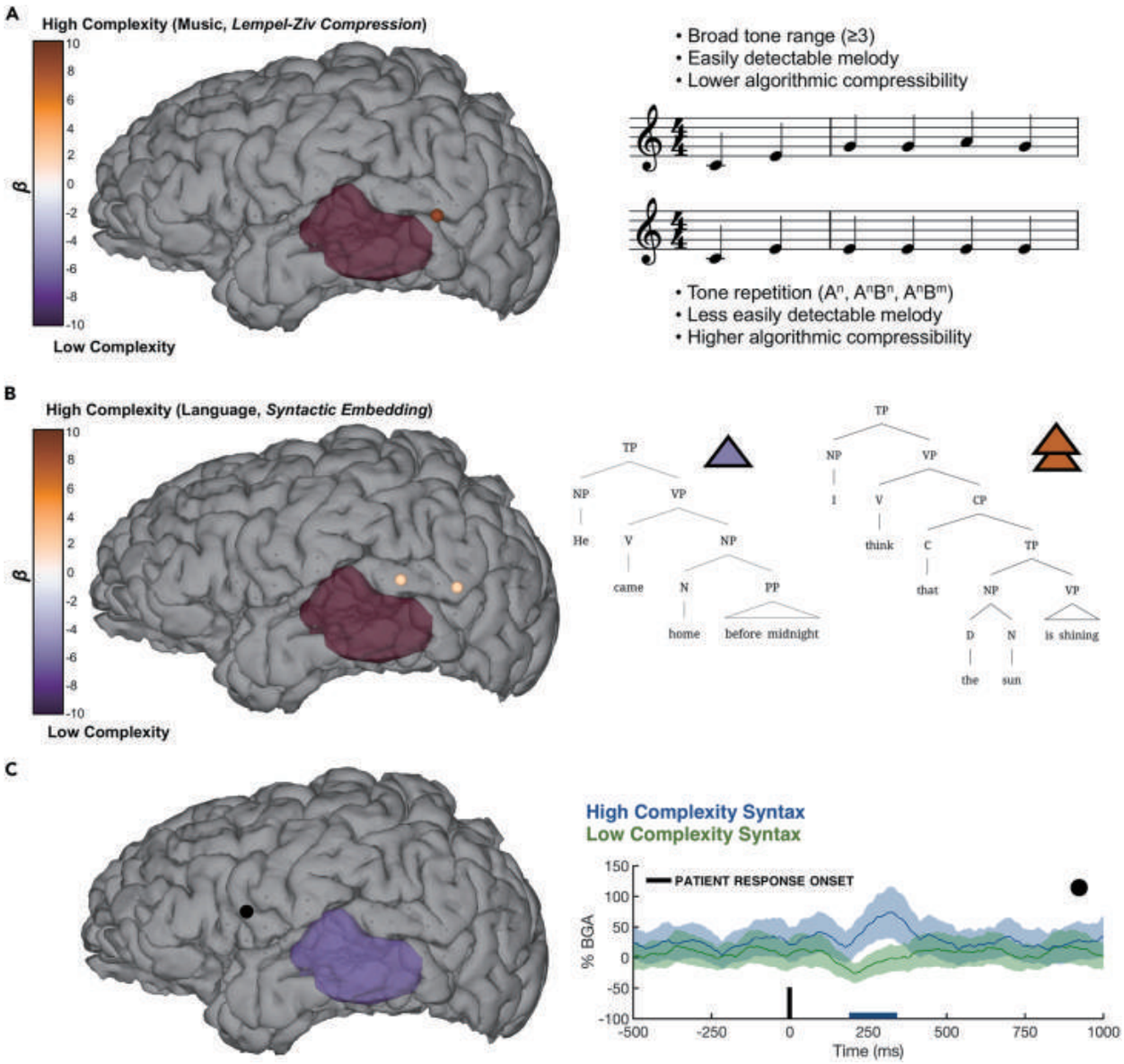
▷图2：音乐和语言的ECoG活动。图源：来自论文

(2)宽频带高伽马活动对音乐和语言刺激复杂性的调节

对于音乐任务，后颞中回(pMTG)中的一个电极(与pSTG的上边界)在复杂性和音符位置之间具有显著的相互作用，相对于低复杂度序列，高复杂度序列的宽频带高伽马活动(broadband high gamma activity，BGA)增加的幅度更大(图3A)。这说明该电极对语言和音乐都有显著的活动。

比较基于句法复杂性的语言刺激时，两个电极在pSTG上都表现出了BGA增加(图3B)。这说明这些电极对语言和音乐也有显著的活动。但是语言产生的BGA差异在百分比变化方面小于音乐产生的差异，这可能是由于音乐序列之间有更显著的差异。

研究还分析了这些相同语言和音乐对比的低频响应(2-15 Hz和15-30 Hz)，他们发现电极在位置和复杂性之间没有显著的交互作用。但是在句子生成过程中发现，在发音开始后大约200-300毫秒，额叶下皮层的一个电极对句法复杂性反应较强(图3C)。

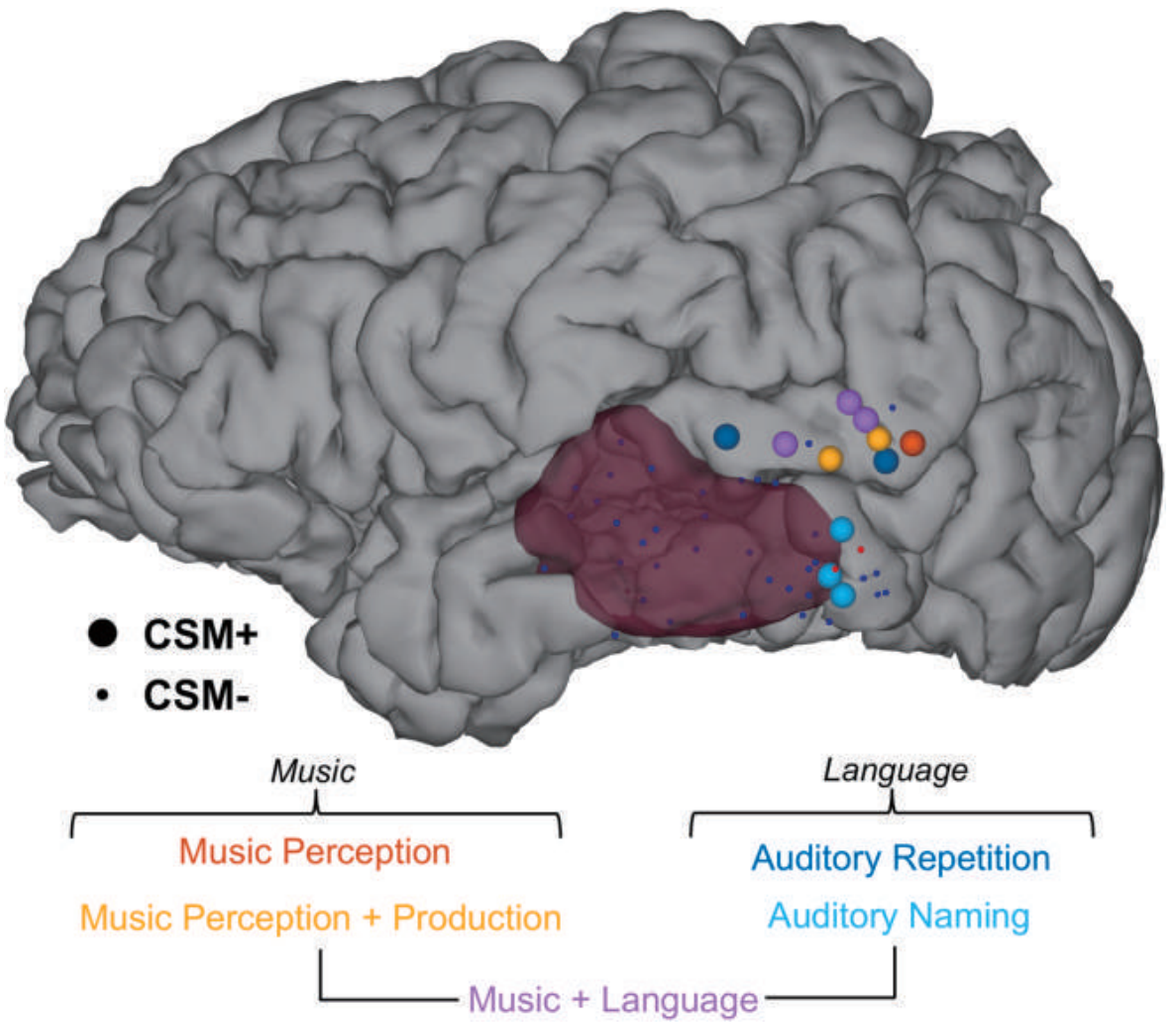


▷图3：术中音乐和句法复杂性的ECoG。图源：来自论文

(3)语言和音乐的皮层刺激影像

除了ECoG任务外，研究者还进行了直接皮层刺激影像(CSM)的分析，主要目的是比较切除后颞区和中颞区前后语言和音乐加工的变化(图4)。在CSM的分析结果中，有5个位点聚集在pSTG中，对句子重复呈阳性，导致理解中断。在pSTG测试中，只有两个位点在句子重复上呈CSM阴性。在听觉命名任务中，聚集在pMTG的三个位点的CSM呈阳性。在语言任务的产生和理解中，没有明显的影响。

在音乐任务中，CSM主要表征音乐感知或制作。其中聚集在pSTG中的6个位点为CSM阳性。他们根据CSM在试次中应用的时间，对六个CSM音乐阳性位点进行了分类。患者通常会在尝试演奏时引入“干扰”，或者只是无法选择旋律中的下一个音符。在腹侧颞叶皮层的两个位置测试音乐感知和制作过程中都表现出CSM阴性。如图4所示，在pSTG区域的语言和音乐相关位点之间存在明显的重叠。在CSM语言映射过程中，患者的主观报告表明语音工作记忆受到干扰，此外记忆访问也受到干扰。



▷图4：术中音乐和语言感知的刺激测试。图源：来自论文

三、结语

语言和音乐的加工都涉及将基本单位有效地组合成结构。目前人们尚不清楚大脑中对语言和音乐结构敏感的区域是否有相同的定位。这项研究发现，刺激pSTG干扰了音乐的感知和制作。随着语言的产生，pSTG和pMTG被语言和音乐激活。pMTG的活动受音乐复杂性的调节，pSTG的活动受句法复杂性的调节。这说明，音乐和语言在大脑中虽然共享资源，但处理特定领域结构时具有不同的神经特征。

这项研究的特点在于，它以癫痫疾病为模型，且受试者为从小接受训练的音乐家。研究者对受试者进行了多种行为任务以及颅内电生理活动的同步记录，这对于理解语言和音乐的神经机制具有重要的意义。（编辑：Lixia）

参考文献：

关联论文：McCarty, Meredith J., et al. "Intraoperative cortical localization of music and language reveals signatures of structural complexity in posterior temporal cortex." Iscience 26.7 (2023).

大脑如何感受美的存在?

讲者:Indre Viskontas

神经科学家,歌剧舞台导演,科学传播者。旧金山大学心理学副教授和创意大脑实验室(Creative Brain Lab)主任。

扫码查看原文

一、为什么大脑的物质结构相似,人的内心世界却天差地别?

近年来,神经科学家们开始认识到,个体的大脑并不相似——至少,每个人构建内心世界的方式天差地别。假设让你在脑海中想象一个球体,可以预见的是:每个人想象的都会不一样,大小、颜色、花纹、背景环境、材质都会有区别。

有一类人,他们无法想象出这个球体——他们甚至无法在脑海中构建任何视觉形象,这种想象被称为想像障碍(Aphantasia)。计算机科学家、皮克斯动画总裁Ed Catmull就有着这么一颗大脑:当他想象时,他是以一种概念化的、数学的形式来构建对象的;这份天赋让他成为了一位顶尖的计算机科学家、并在计算机图形学领域做出了许多革命性的贡献。

在皮克斯,诸如Catmull这样的科学家又和另一种极端的天才大脑密切地合作着——他们是世界顶尖的故事板艺术家,他们看完电影之后从不需要"二刷",因为他们可以从记忆中逐帧想象出整部电影。

人类大脑的最大优势便是其多样性和适应性。我们的大脑在出生时发育得如此不完善是有原因的:从技术上讲,在妊娠第 40 周,母亲已经无法为正在发育的大脑提供能量;在出生后,大脑需要通过学习来进一步发育。我们的大脑在生命的第一年体积增加了三倍,建立了各种各样的脑区和神经元之间的连结。这一切是如何发生的?答案是,通过经验。

二、大脑如何将信息碎片转化为内心体验?

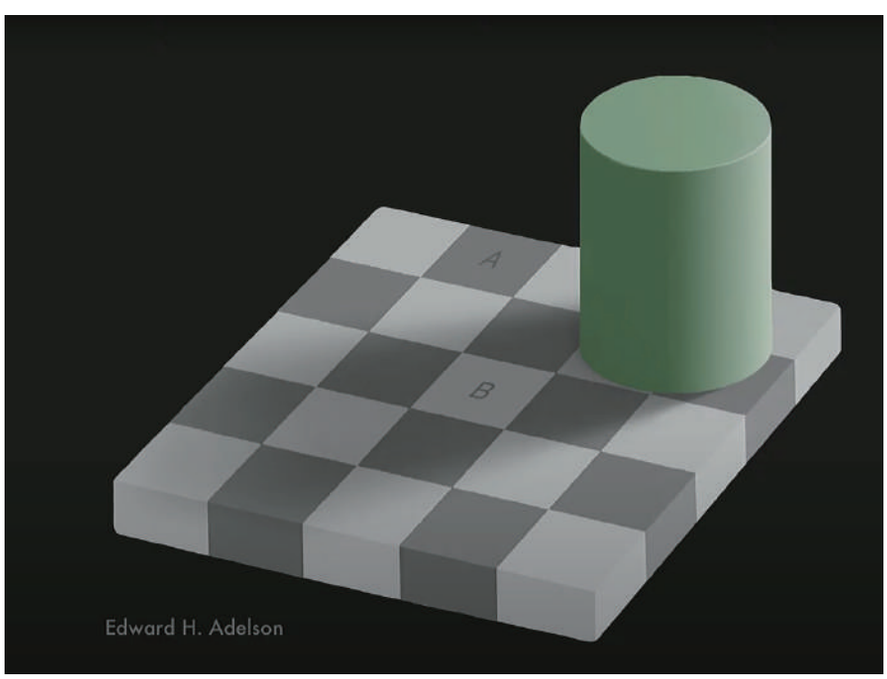
我们的大脑不是客观的"观察者" ,而是主观体验的"策展人"。感官从浩如烟海的外部世界中提取一丁点信息的尘埃,而大脑能将其加工为主观感受的浩瀚山峦。

我们常有一种错觉,认为我们看到的事物就是它们的本来面目。身处美丽的风景中,我们以为这壮美的景色是眼睛一下子全看到的,但实际上,我们一次只能只感受到一小部分:只有视网膜中央凹拥有可以

区分细节的光感受器，而其余部分只能看到黑色和白色；并且即使在这个最清楚的视野内，也有一个盲点，它是视神经离开视网膜、并将信息传入大脑的部位。所以，我们眼前看到的事物，其实是小片的信息在大脑中组装而成的。

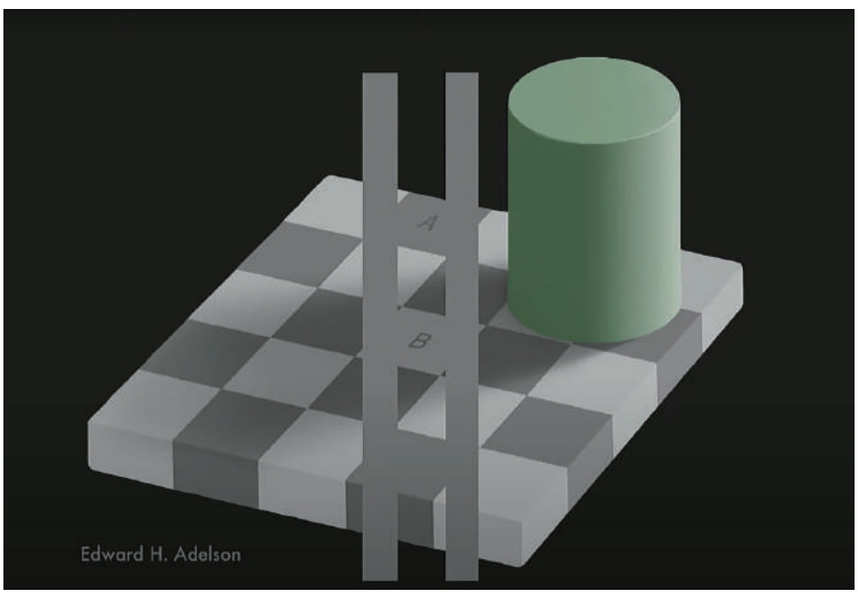
▷ 图注：美丽的风景在视网膜上实际收集的信息 图片来源：Chen Institute Brain and Mind Lecture

而事物的意义，更是大脑从组装的信息中提取出来的。例如，我们知道，我们所看到的不同颜色是因为它们对应着不同的可见光波长；但实际情况要复杂得多。大脑会自动考虑光照条件，因此当它接收到完全相同的波长、却假设它们们来自不同的光照条件时，我们会认为自己看到了不同的颜色。比如下图：



▷图片来源：Chen Institute Brain and Mind Lecture

方块A和方块B颜色一样吗?乍看上去,完全不同;A是深灰色的,而B则是白色的棋盘格。



▷图片来源:Chen Institute Brain and Mind Lecture

实际上它们的颜色是一样的!因为大脑认识到右边的圆柱体有一个投影,所以“认为”两者的颜色不一样。类似的视错觉还有很多。另外,大脑也会利用听觉来填补空白;我们会推断声音的前因后果、可能的来源,而不仅仅是听到声音本身。所以,一段客观、沉浸式体验,是由大脑中许多通路所追踪、不同脑区所组织起来的;但我们的感觉却是身临其境、连贯的。

三、大脑为什么要从感官体验中提取意义?

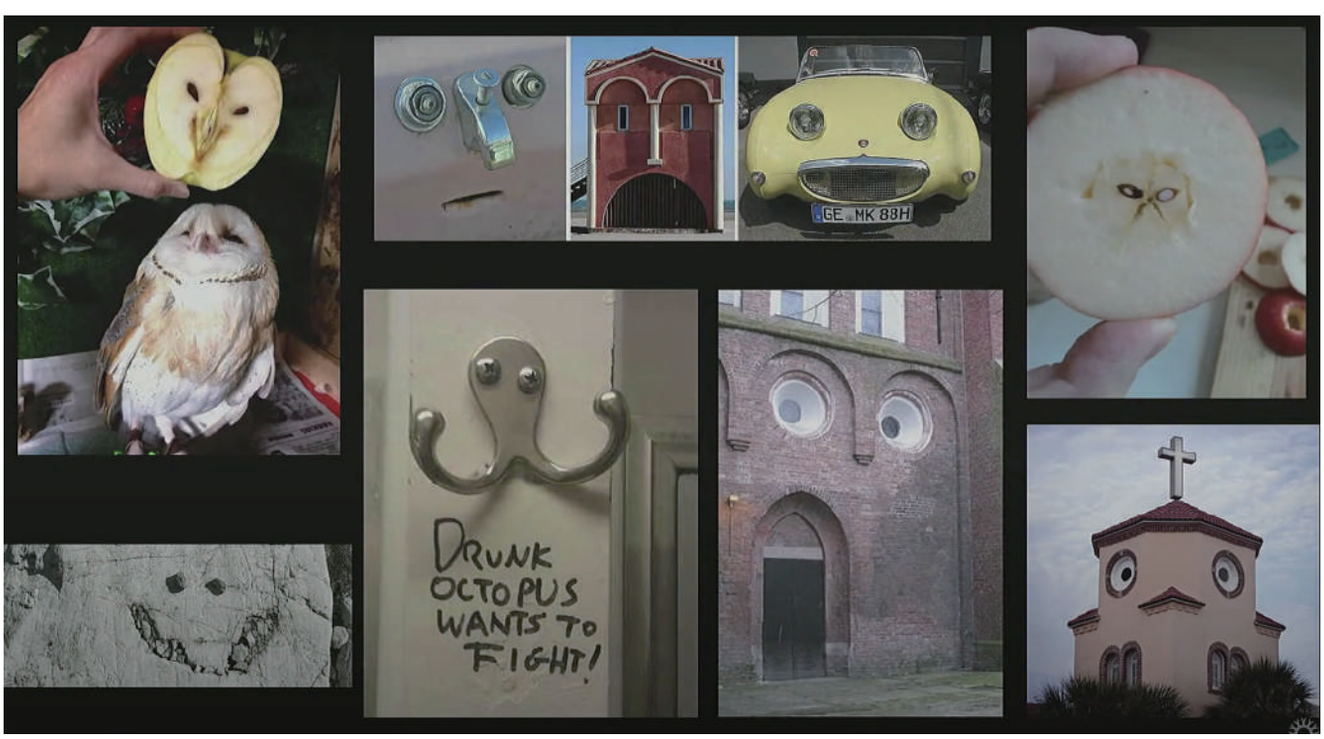
下面这张模棱两可的图片:如果说它与胡萝卜、长且毛茸茸的耳朵、复活节有关,你觉得看到了什么?

▷图片来源:Chen Institute Brain and Mind Lecture

实际上，这是一只鸭子。但看了之前的提示后，或许你脑海中会浮现一只兔子。你要么看到鸭子、要么看到兔子，你的大脑会在这两个意像中分配一个；现在，你可以有意识地在鸭子和兔子之间来回切换了；但是请注意，大脑为你做了决定，现在您需要推翻该决定才能看到另一个图像了。我们的很多主观体验都是在我们意识到之前发生的。

为什么大脑会自动为看到的事物标记意义、而不是将其抛给意识来决定？因为把无意义的事物看做有意义的、总比错过了重要的意义好，这可能来自于我们进化过程中的选择压力。大约 150 万年前，人类颅骨的容量呈指数级增长，同一时间也开始生活在更大的社会群体中。我们必须驾驭复杂的群体动态、将朋友与敌人区分开来、学会合作并共同建设社会，因此，了解他人的意图和感受成为一种强大驱动力——也许，那些更能胜任于此的人就被选中了。

这也是“人脸幻想性错觉”产生的原因：当看到一张脸时，识别脸上表达的情感对社会生物来说至关重要，因此具有进化上的优势；把别的事物认成“人脸”总比错过了这个关键信息要好。

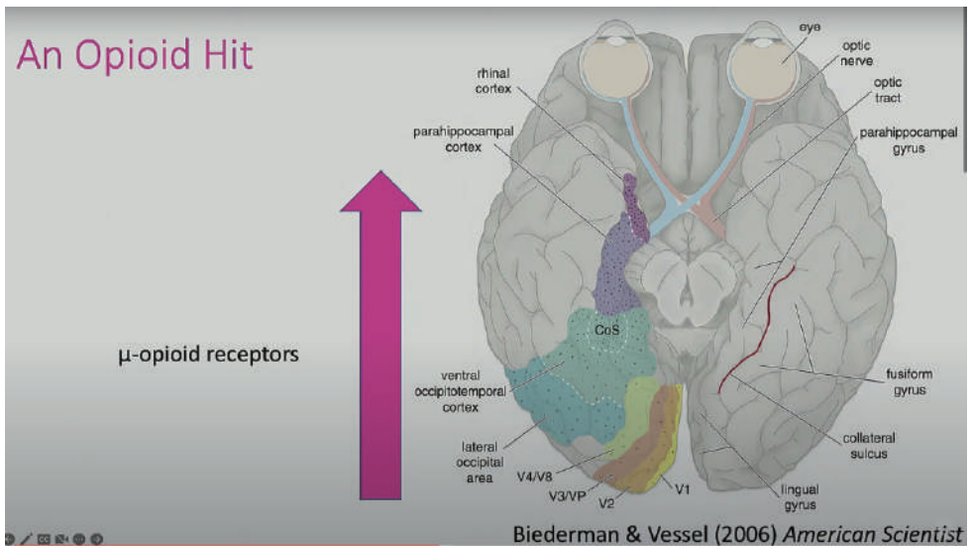


▷ 图注：人脸幻想性错觉。看这张图中是不是有很多表情各异的脸？图片来源：Chen Institute Brain and Mind Lecture

四、脑中的意义能用神经解剖学解释吗？

大脑中，在视觉通路的起始部分直至视皮质，阿片类受体的密度逐渐升高；这也就是为什么当我们体验到更深层次的意义，并将它们与我们过去的经验联系起来、投向未来时，我们总能感受到愉悦。

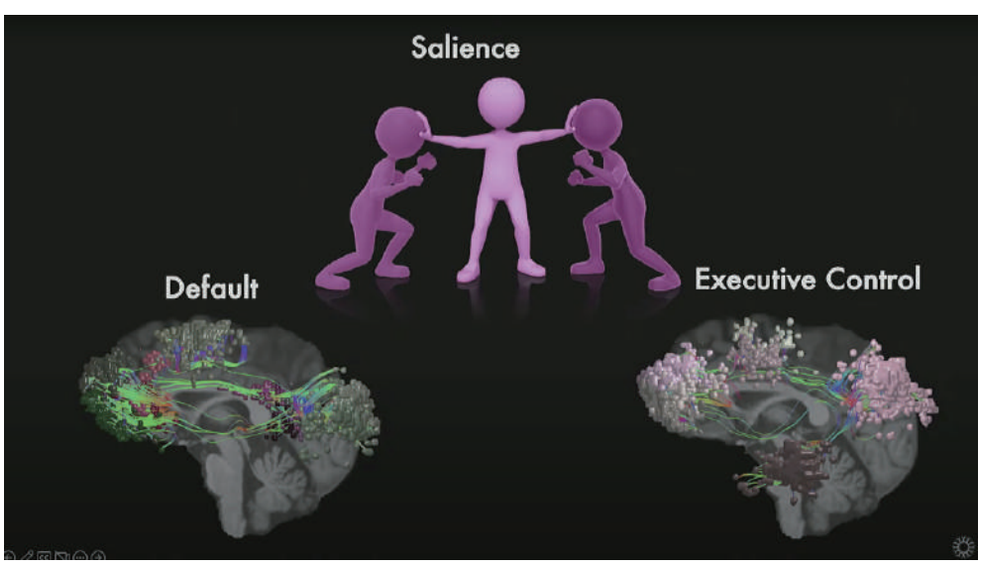
美丽的画面在我们眼前展示得越多，信号就越深入脑海中密集的阿片类受体通道。这就是参观博物馆、欣赏艺术展等带来的益处。一个好的展览会让我们得出自己的结论、而不仅仅是把结论一股脑儿倾倒给我们；因为大脑享受发现意义的过程，这给我们带来愉悦。



▷图片来源:Chen Institute Brain and Mind Lecture

大脑中有两种负责思考的网络。一种是默认模式网络,这个网络包括小部分的前额叶背外侧和大部分的大脑半球后侧神经元。神经科学家们发现,当他们让受试对象“什么也不做”时,受试者的一些神经元却开始活动,这些便是位于默认模式网络中的神经元。大脑不会无所事事,我们会思考过去、未来、做白日梦。诚然,太多的反思、担心与忧虑,会让我们痛苦不堪;但我们也会发现新的模式、得到新的想法,一直苦苦追寻的问题答案会突然跳进脑中。另一种是执行控制网络,主要由前额皮质驱动,负责目标、设定、计划和激励;这些很好,会帮助我们有条不紊地逐步完成任务,但它也可能导致固化思维,导致我们难以产生新的想法。

默认模式网络和执行控制网络经常相互矛盾,二者必须有一个占上风,才能更好执行当前的目标。而脑中还有一套网络,叫做凸显网络:在默认模式网络和执行控制网络之间的拔河比赛中,凸显网络可以介入并将两者结合在一起,能够对两者进行“开关”,让我们更能根据目标进行处理。



▷ 图注:默认模式网络、执行控制网络和调和的凸显网络 图片来源:Chen Institute Brain and Mind Lecture

五、同情心：人类情感外在的映射

虽然每个人的内心世界有许多不同，但当认识到他人时，我们会倾向于认为每个人的思想都像我们的一样。那么，同情心是如何建立的？

当我们将人类情感投射到外界，就会同情他人、动物，甚至于没有生命的事物。动物的听觉系统能听到人类听不到的频率的声音，如今，技术进步使我们能够将这些声音转化为我们可以体验到的东西，这样我们就能更有意义地观察这些动物的行为。如今我们知道了，雌性大象每 4 年才进入一次发情期，她会发出我们以前听不到的声音吸引雄性；雄象、雌象产下后代之后待在一起，但雄象最终要独自离开。所以，当我们看到偷猎大象的图片时，我们内心会受到更多震撼、感受到更多哀伤，因为我们了解到大象是群居动物，它们会哀悼死者、生活在家庭中，就像我们一样。

▷图片来源：Chen Institute Brain and Mind Lecture

当我们关心时，便会采取行动。当凸显网络将经验或有意义的刺激标记上意义时，它便可以切换我们大脑的状态，以产生新想法或激励我们行动。

当艺术与科学相交织时，体验让我们选择感动，让我们有动力采取行动；让我们有机会切换和调整大脑网络，产生新想法，建立新联系；摆脱日常习惯和陈规陋习、敞开心扉去想象无限可能。（记者：一碗萝莉面）

好的生活方式如何降低抑郁症风险?

作者:宋薇

中科院神经所博士在读,科普爱好者,计算神经生物学方向。对有趣的科学问题好奇,致力于传播神经科学的魅力。

扫码查看原文

在众多精神类疾病中,抑郁症(depressive disorder)是常见的一种,主要表现为情绪低落、兴趣减低、思维迟缓、食欲减退、睡眠质量差等。据世卫组织统计,全球有超过3.5亿人在遭受抑郁症困扰。然而,迄今为止,抑郁症的发病机制尚不清楚。

研究表明,生活方式与抑郁症之间存在多种共同的神经生物学机制,由遗传变异调节。

但健康的生活方式能多大程度降低患抑郁症的风险?它的神经生物学机制是什么?为了回答这些问题,近期,天桥脑科学研究院研究员、复旦大学附属华山医院神经内科教授郁金泰联合复旦类脑智能科学与技术研究院、英国剑桥大学等机构的研究人员在Nature Mental Health上发文,探究了饮酒、饮食、体育锻炼、睡眠、吸烟、久坐行为与社会关系这7种生活方式与抑郁症的关系,及其潜在的神经生物学机制[1]。

该研究基于大规模生物医学数据库与大数据统计建模方法,发现拥有健康的生活方式最高可降低57%的抑郁症患病风险,进一步机制研究表明健康生活方式能够通过影响大脑结构、免疫系统和新陈代谢,对预防抑郁症起到保护作用。

一、生活方式与遗传风险对抑郁症发病的影响

研究人员对来自英国生物样本库(UKB)的287282名参与者随访了9年,其中,参与者平均年龄为57.5岁,男女各占一半。为了探究7种生活方式对抑郁症发病的影响,研究者将多种生活方式因素整合到一个健康生活方式的综合评分中,根据健康生活的综合评分划分了三个生活方式类别:较差(得分0-1分)、中等(得分2-4分)、良好(得分5-7分)。

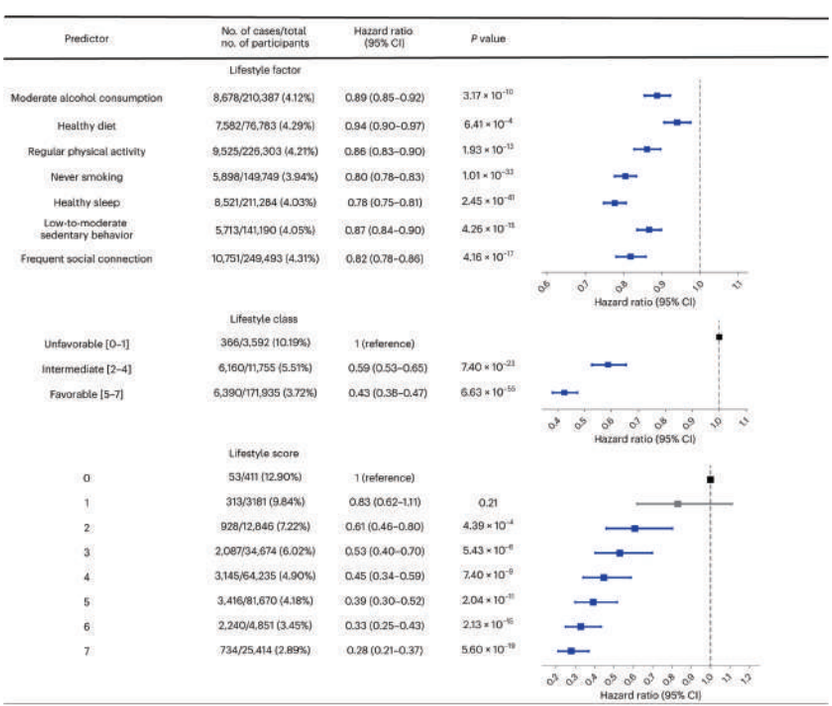
健康的生活方式如下:(1)健康睡眠:7-9小时;(2)健康饮食:7类食物中至少4类,包括水果、蔬菜、鱼、加工肉、未加工红肉、全谷物、精制谷物;(3)不吸烟;(4)适度饮酒:不喝酒或少量饮酒,女性每天低于14

克，男性每天低于28克；(5)频繁社交；(6)避免久坐：每天久坐时间低于4小时；(7)体育锻炼：150分钟中强度或75分钟高强度运动。

所有参与者的健康生活平均得分为4.75分，有1.25%的人属于较差生活方式，38.9%的人属于中等的生活方式，59.85%的人属于良好的生活方式。在平均9年的随访期间，共有12916名参与者出现了抑郁症。

研究人员采用多变量Cox回归模型检验了生活方式与抑郁症之间的关系。结果表明，7种生活方式均与抑郁症风险独立显著相关。其中，健康睡眠抑郁症风险降低22%；健康饮食抑郁症风险降低6%；不吸烟抑郁症风险降低20%；适度饮酒抑郁症风险降低11%；频繁社会抑郁症风险降低18%；避免久坐抑郁症风险降低13%；定期体育锻炼抑郁症风险降低14%。

研究发现健康的生活方式有助于降低抑郁症风险，与生活方式较差的人相比，拥有中等生活方式的人患抑郁症的风险降低了41%，而生活方式良好的人患抑郁症的风险降低了57%(图2)。

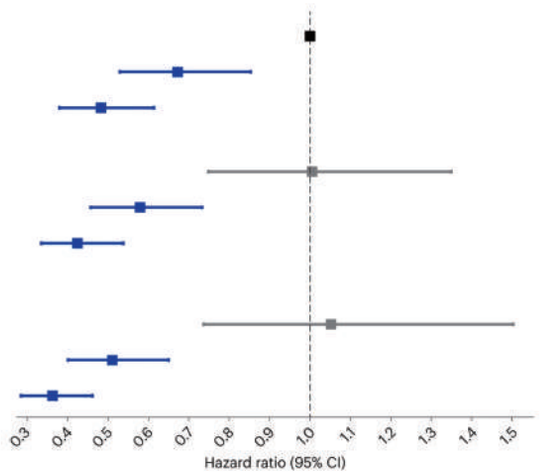


▷图2：生活方式因素与抑郁风险的关系。图源：论文

现有研究表明，精神风险的遗传结构是复杂的，由多种因素主导，因此，研究人员进一步探究了遗传风险与生活方式的交互影响。

以高遗传风险和不良生活方式的参与者为参照组，低遗传风险和良好生活方式的参与者的抑郁风险降低幅度最大(风险比为0.36，95%CI为0.28–0.4(6)，然而，多基因风险评分和生活方式评分与抑郁症风险之间没有显著的交互作用(Pinteraction=0.4(1)，表明生活方式可能不会被抑郁症的遗传风险显著改变，生活方式在不同水平的多遗传风险人群中具有很强的保护作用(图3)。孟德尔随机化分析显示，生活方式与抑郁症之间存在显著的因果效应。这一研究结果表明，无论个体的遗传风险高低，健康的生活方式都对预防抑郁具有重要的保护作用。

Predictor	No. of cases/total no. of participants	Hazard ratio (95% CI)	*P* value
High PRS			
Unfavorable [0–1]	72/690 (10.43%)	1 (reference)	
Intermediate [2–4]	1,247/19,638 (6.35%)	0.67 (0.53–0.85)	1.10×10^{-3}
Favorable [5–7]	1,221/28,882 (4.23%)	0.48 (0.38–0.61)	2.66×10^{-9}
Intermediate PRS			
Unfavorable [0–1]	115/1,129 (10.19%)	1.01 (0.75–1.35)	0.97
Intermediate [2–4]	2,033/37,854 (5.37%)	0.58 (0.46–0.73)	5.58×10^{-6}
Favorable [5–7]	2,205/59,628 (3.70%)	0.42 (0.34–0.54)	1.26×10^{-12}
Low PRS			
Unfavorable [0–1]	52/495 (10.51%)	1.05 (0.74–1.50)	0.78
Intermediate [2–4]	855/18,223 (4.69%)	0.51 (0.40–0.65)	4.56×10^{-8}
Favorable [5–7]	956/30,805 (3.10%)	0.36 (0.28–0.46)	1.56×10^{-16}



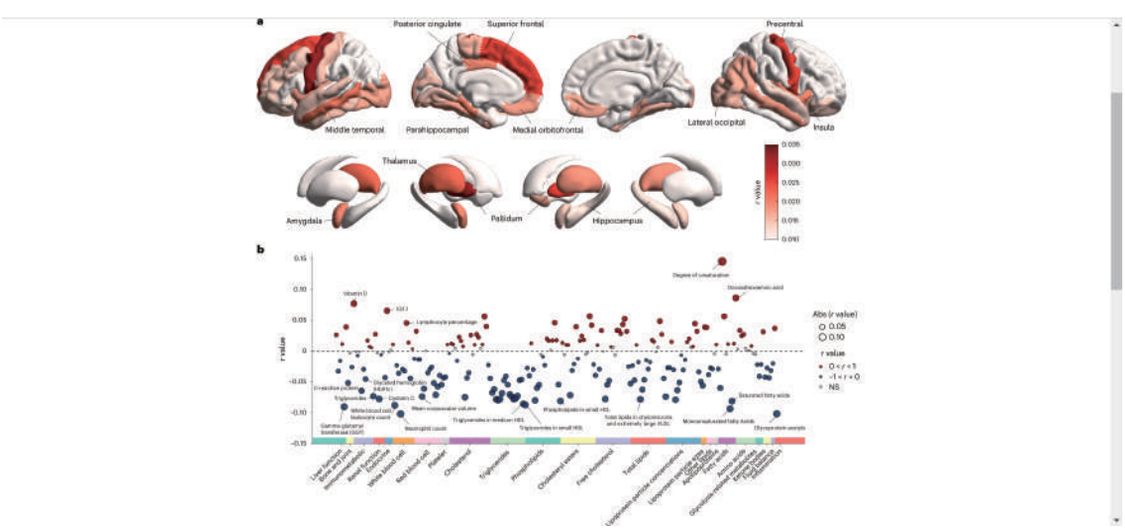
▷图3：抑郁症的风险取决于遗传风险和生活方式。图源：论文

二、生活方式是如何影响抑郁症发生的？

为了进一步阐明健康的生活方式能够降低患抑郁风险的原因，研究人员结合影像、生物化学等多维数据对其背后的神经生物学机制进行了探究。

首先，在神经影像方面，研究表明，生活方式得分越高，脑容量越大（图4a），涉及前额叶皮层、眶额皮层、中央皮层和岛叶等皮层结构，以及苍白球、丘脑、杏仁核和海马体等皮层下结构。这些脑区在认知控制和情绪调节方面具有重要功能。这些大脑结构与抑郁症状评分呈负相关，而结合神经影像学计算的结果也支持生活方式、大脑结构和抑郁症之间的关联。

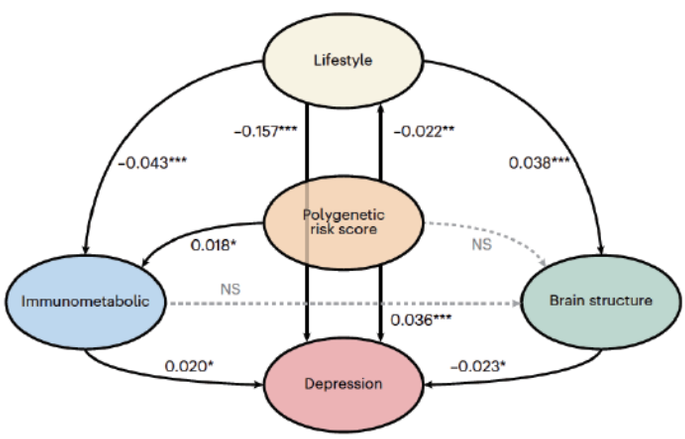
在各种生物标志物中，48种血液标志物和130种代谢标志物与生活方式具有显著相关性，血液生化标志物中关联最显著的为C反应蛋白（一种体内因应对压力而产生的分子）和甘油三酯（身体用来储存能量以备后用的主要脂肪形式之一），血细胞中则是与免疫相关的中性粒细胞和白细胞。在代谢标志物中，与生活方式具有正反相关性最显著的分别为不饱和程度和乙酰糖蛋白（图4b）。



▷图4：生活方式与大脑结构和外周标志物的关系。图源：论文

最后，研究人员进一步整合生活方式、抑郁、遗传、脑结构和免疫代谢等多方面数据，利用结构方程模型分析了18244名参与者这五个维度数据之间的相互作用关系，系统阐释了生活方式降低抑郁症发病风险的神经机制（图5）。

结果显示，生活方式与抑郁症风险、免疫代谢功能和大脑结构显著相关，多基因风险评分、大脑结构和免疫代谢功能也与抑郁风险显著相关。此外，多基因风险评分与生活方式和免疫代谢功能显著相关。除了多基因风险评分和大脑结构、免疫代谢功能和大脑结构之外，其余路径均具有显著相关性。该结论为我们理解健康生活方式如何通过影响大脑结构和免疫代谢功能进而降低抑郁风险提供了全面的视角。



▷图5：生活方式、多基因风险评分、免疫代谢功能、大脑结构和抑郁症之间的关系。图源：论文

总之，本研究证实了多种生活方式因素与抑郁症风险之间的因果保护关系。在具有不同遗传风险特征的人群中，坚持健康的生活方式有助于预防抑郁症。根据2019年的一项数据研究，中国抑郁症的终生患病率（在一生当中得过抑郁症的患者所占总人口比率）为6.8%，12个月患病率（12个月内得过抑郁症的患者所占总人口比率）为3.6%[2]。照此计算，超过9500万中国人一生当中得过抑郁症，而12个月内有5000万人患抑郁症。想要远离抑郁症，一方面要坚持健康的生活方式，健康多一点，抑郁少一点，另一方面，如果不幸患上抑郁症，应该积极接受规范的治疗。（编辑：Lixia）

参考文献：

关联论文：Zhao, Yujie, et al. "The brain structure, immunometabolic and genetic mechanisms underlying the association between lifestyle and depression." Nature Mental Health 1.10 (2023): 736-750.

[1]Zhao, Y., Yang, L., Sahakian, B.J. et al. The brain structure, immunometabolic and genetic mechanisms underlying the association between lifestyle and depression. Nat. Mental Health 1, 736–750 (2023). https://doi.org/10.1038/s44220-023-00120-1

[2]Huang Y, Wang Y, et al. Prevalence of mental disorders in China: a cross-sectional epidemiological study. Lancet Psychiatry. 2019 Mar;6(3):211-224. doi: 10.1016/S2215-0366(18)30511-X. Epub 2019 Feb 18. Erratum in: Lancet Psychiatry. 2019 Apr;6(4):e11. PMID: 30792114.

非侵入性神经调控，脑疾病精准治疗的新希望？

讲者：Alvaro Pascual-Leone

哈佛医学院神经病学教授，曾担任贝伦森-艾伦无创脑刺激中心主任，专注于控制大脑可塑性的机制研究。

扫码查看原文

随着人口老龄化进程的加快、生活竞争压力的增加及环境因素的变化，脑功能性疾病（如阿尔茨海默症、癫痫、抑郁症等）已成为全球主要的致残和致死原因之一，不仅给个人、家庭、社会带来了沉重的负担，其治疗和机制研究也是巨大的科学难题。借助于神经科学和生物医学工程技术的进步，非侵入性神经调控成为当前医学发展最快的领域之一，全球已有数十万脑功能性疾病患者从中获益。

2023年8月18日，天桥脑科学研究院（Tianqiao and Chrissy Chen Institute，TCCI）和哥伦比亚大学神经技术中心（NTC）、多诺斯蒂亚国际物理中心（DIPC）联合举办了在线学术会议NanoNeuro 2023。本次会议上，来自哈佛大学的Alvaro Pascual-Leone教授向大家介绍了非侵入性神经调控技术的最新进展。

我们知道，机体正常的神经环路是由电刺激和化学信号构成的一个固有平衡系统，但疾病（包括先天性和后天性的因素）打破了这个平衡，导致认知、感觉或运动受损。非侵入性神经调控又称无创脑刺激（Noninvasive Brain Stimulation，NIBS），它采用非侵入性的技术，通过物理（电、磁、光、超声等）或化学手段作用于大脑皮层，可逆性地调控大脑和神经元的活动，恢复和重建神经系统的平衡状态，从而达到治疗疾病的目的。

近年来，NIBS在临床上已被批准用于治疗难治性抑郁症、创伤后应激障碍、双相情感障碍、孤独症、阿尔茨海默症、帕金森病等多种神经系统疾病，其主要应用方式是经颅磁刺激与经颅电刺激。然而，由于没有一个标准化的方案，现有的NIBS研究表明，针对不同患者的NIBS在刺激部位、频率、强度、剂量和其他参数方面存在很大差异。由此，利用NIBS进行脑疾病精准治疗的诉求也应运而生。

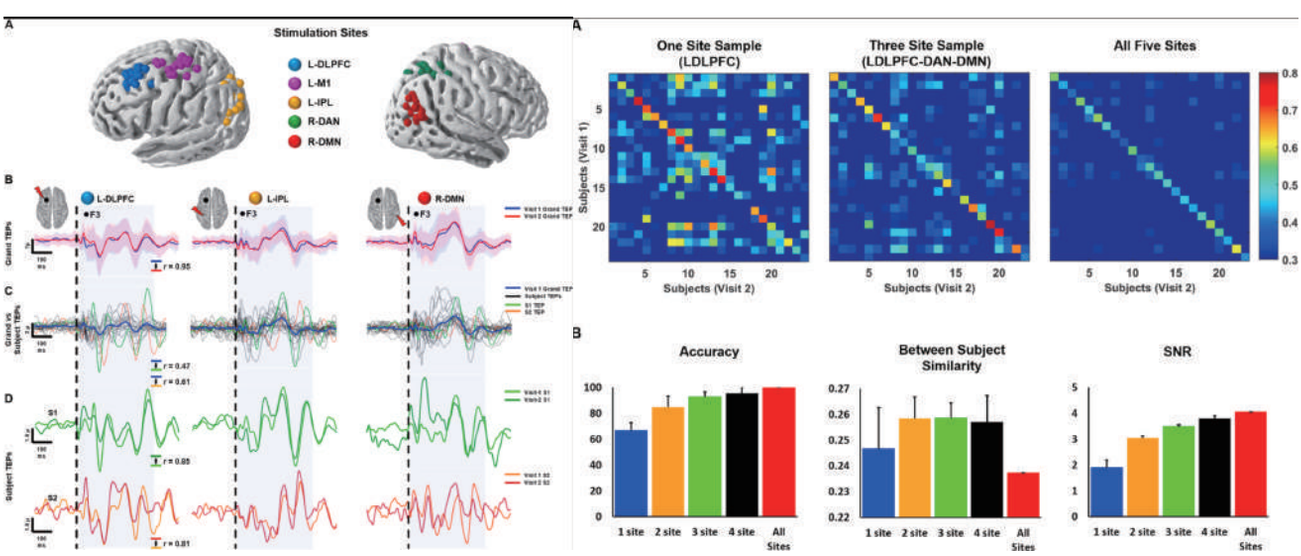
记者注：经颅磁刺激（Transcranial Magnetic Stimulation，TMS）是将一绝缘线圈放在大脑特定部位的头皮上，当围绕线圈的电流通过时，产生的磁信号可以无衰减透过头皮和颅骨，影响脑内代谢和神经活动。

经颅电刺激（Transcranial Current Stimulation，tCS）设备有阳极和阴极两个表面电极，使用时将电极放置在指定位置，刺激器输出恒定的低强度电流，电流穿过颅骨作用于大脑皮层，进而调节大脑皮层活动，影响相应的感知觉、运动和认知行为。可分为经颅直流电刺激（tDCS）和经颅交流电刺激（tACS）两类。

一、智能化神经调控，识别个体独特“脑指纹”

近年来，科研人员越来越明显地感受到，表征个体大脑结构、连通性和动力学对于理解健康和疾病中的大脑功能至关重要。非侵入性神经调控提供了调控大脑时空特征等的可能性。然而，大多数神经成像和脑刺激研究都是通过受试者群体的平均测量结果、人群水平的推断来表征人脑功能的。直接应用于特定大脑区域的外部刺激可以在个体水平上揭示关于人类大脑状态、连通性和动态的独特信息。

通过磁共振成像（MRI），可以绘制整个大脑的白质连接，即所谓的大脑连接组，提供关于全脑神经网络结构的丰富信息。连接组数据中个性化的结构连接模式，被称为指纹。在一项研究中，Alvaro Pascual-Leone教授和同事们对24名健康受试者的前额叶、顶叶、运动区等解剖学定义的区域与功能性定义的皮层节点进行单脉冲TMS，结果表明，TMS诱导的皮层传播模式在个体之间存在差异，而在个体内部高度一致，且与自发神经活动不同。这表明，扰动诱发的脑反应揭示了独特的“脑指纹”，反映了受刺激脑区域的因果联系动态，并可能作为个体脑功能的可靠生物标志物。



▷图一：使用fMRI引导的TMS-EEG获得个体化大脑指纹。图源：参考文献1和2

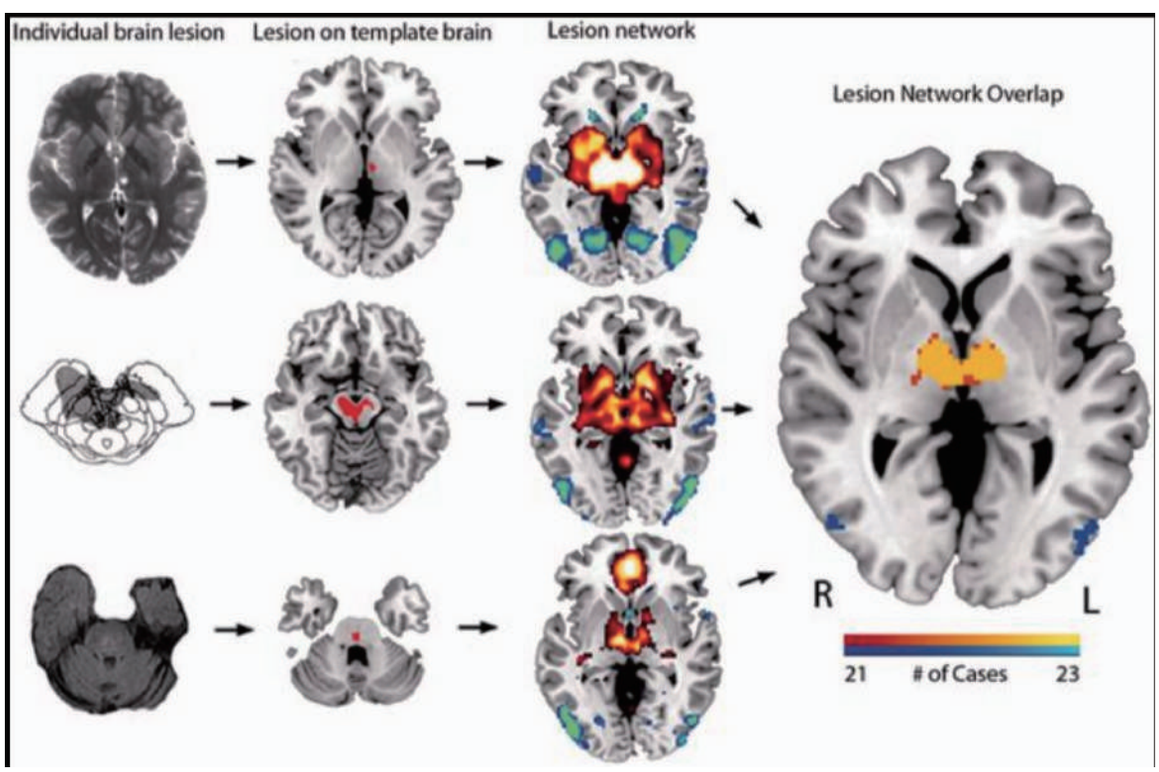
同时，Alvaro Pascual-Leone教授及其合作者发现，结合功能性磁共振成像（fMRI）等技术，可以实现精准医疗，即对患者进行个性化的神经调控。具体表现为：利用针对个体症状识别的生物标志物、个性化神经调控的靶标与刺激参数，结合fMRI和EEG等检测神经调控的生理效应，进一步使用生理反馈调整参数与刺激靶标，形成一个闭环治疗方法。

二、网络定位脑病变，优化神经调控疗效

Pascual-Leone教授指出，探索和选择最佳的刺激靶点以提升刺激效果，从而不断提高神经调控的疗效，这也是未来非侵入性神经调控的主要研究方向。

通常，将神经症状与特定大脑区域联系起来的方法涉及识别具有相似症状的患者之间病变位置的重叠，即病变映射（lesion mapping）。大多数精神类疾病的症状是由整个网络功能障碍而不是单一脑区病变引起的，当症状反映的是网络功能障碍时，传统方法的病变映射能力受到限制。为此，Alvaro Pascual-Leone教授等提出了一种新方法，利用规范的人类连接组数据将症状与病变相关网络联系起来。该方法包括三个步骤：(1)将脑病变的三维体积转移到参考脑上；(2)使用标准连接组数据评估病变体积与大脑其他部分的内在功能连接；(3)重叠病变相关网络进而识别临床综合征的共同区域。

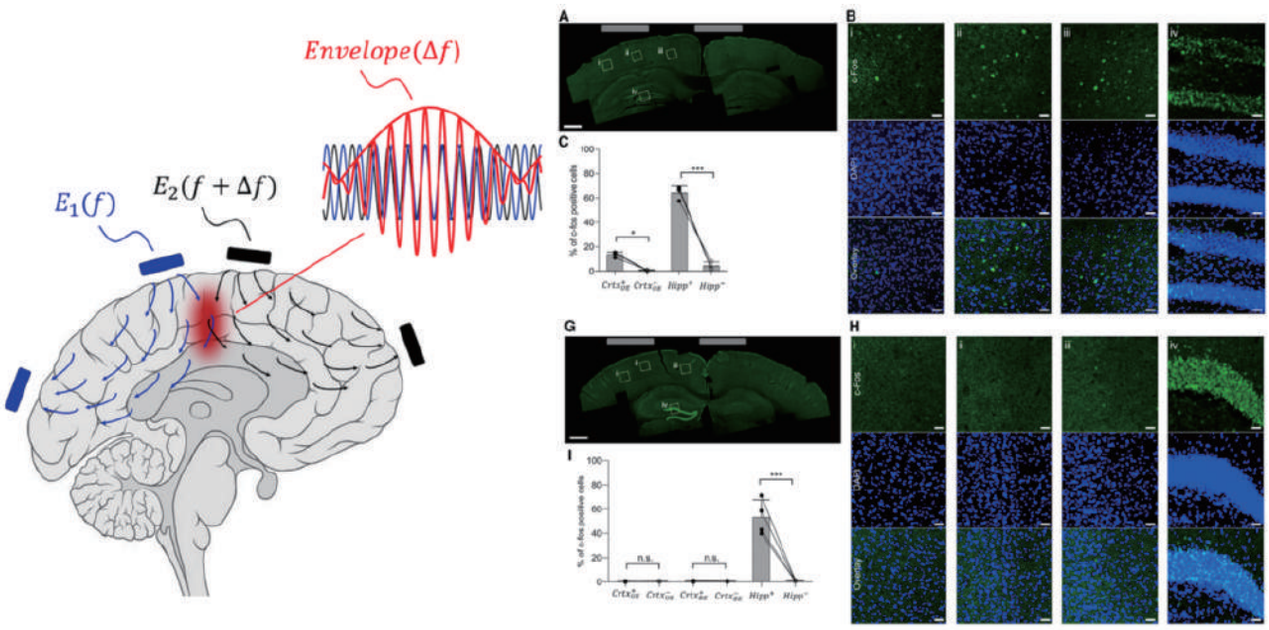
他们在四种病变综合征中验证了该方法，将皮层下病变与涉及症状表达的皮层区域联系起来，从而指导个性化调控靶标。此外，研究表明，多靶点刺激能够获得更为有效的治疗效果。



▷图二：局部脑病变神经症状的网络定位。图源：参考文献3

同时，由于神经元对高频率（>1　KHz）电场振荡无反应，他们还报道了一种非侵入式深部电刺激方式。该方法利用了以下物理原理：两个频率相近的电磁波相遇时能部分相干并形成一个包络的电场，且该包络电场的频率为两者之差。该刺激被称为时域相干（temporal interference，TI）刺激。

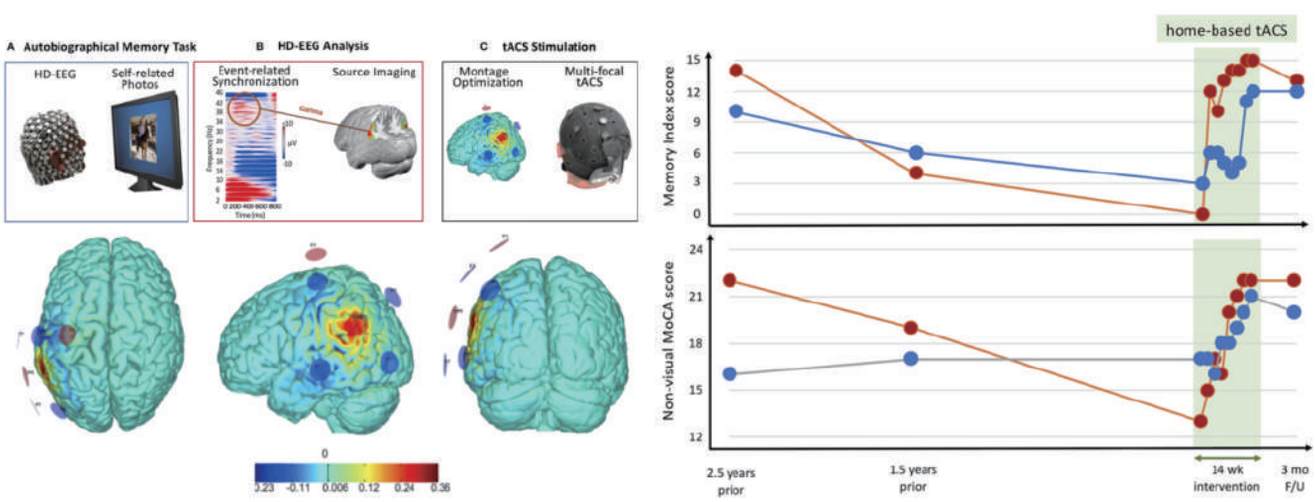
研究人员首先利用数学模型和计算机模型证实了TI刺激的可行性。接下来，他们在小鼠上进行了TI刺激海马的实验，以即早期基因c-fos作为神经元被激活的标志，TI刺激后小鼠海马处有c-fos表达而海马上方的皮层区域几乎无c-fos表达，证实其有一定程度的空间分辨率。之后，研究人员进行了进一步的行为学实验，在麻醉小鼠上给运动皮层施加刺激，可以观察到小鼠相应肢体的运动，而且改变电刺激强度可以刺激到不同的皮层区域，进而引起小鼠不同部位的运动。



▷图三:时域相干刺激在小鼠海马上的应用。图源:参考文献4

Alvaro Pascual-Leone教授还介绍道,个性化、以家庭为单位的干预治疗也是优化神经调控疗效的一种方式。tDCS作为一种非侵入性神经调节技术,在治疗抑郁症方面显示出令人鼓舞的疗效。他们研究了一种创新方案:在护理人员的帮助下,患有重度抑郁症的老年人在家中进行tDCS。同时,他们采用了新开发的多通道实时监控tDCS系统,以保证家庭tDCS的安全性和有效性。结果表明以家庭为基础、远程监督、护理人员管理的多通道tDCS方案,对老年重度抑郁症患者是安全可行的。

同时,基于tACS可以使高频神经元活动正常化,改善阿尔茨海默病患者的记忆,研究人员也制定了一项针对患者量身定制的以家庭为基础的tACS方案。这项初步研究表明,以家庭为基础、远程监督、护理人员管理、多通道tACS方案对老年阿尔茨海默病患者是安全可行的。



▷图四:基于家庭的tACS干预治疗有助于阿尔茨海默病患者的治疗恢复。图源:参考文献6和7

大多数精神类疾病具有不止一种症状，针对不同症状进行同步治疗同样有助于优化神经调控疗效。

三、总结

神经调控结合神经电生理及脑影像手段，为研究大脑调控的生理基础提供了丰富的可能性，对深入探索脑功能性疾病的功能网络机制有重要意义。

对于脑功能性疾病机制的揭示，一方面可依据采集的多种神经电生理、脑影像及临床指标，通过提取能够反映患者临床状态的疾病特异性生物标志物，以预测疗效。尤其是，基于刺激下神经活动的动态响应，利用机器学习及人工智能技术，建立刺激-响应关系的学习自适应神经调控模型，实时监测脑功能状态，为闭环刺激提供反馈信号，进而实现神经调控智能化。另一方面，对于脑功能性疾病机制的揭示有助于人们探索、选择最佳的刺激靶点，开发脑功能性疾病的新疗法，不断提高神经调控的疗效。

随着各项技术的发展进步，神经调控技术将稳步向前，迈进革命性的智能化时代，实现对患者个体化、精准化、动态自适应性的治疗，从而造福更多的脑功能性疾病患者。（记者：宋薇）

参考文献

[1] Ozdemir RA, Tadayon E, Boucher P, Momi D, Karakhanyan KA, Fox MD, Halko MA, Pascual-Leone A, Shafi MM, Santarnecchi E. Individualized perturbation of the human connectome reveals reproducible biomarkers of network dynamics relevant to cognition. Proc Natl Acad Sci U S A. 2020 Apr 7;117(14):8115-8125. doi: 10.1073/pnas.1911240117. Epub 2020 Mar 19. PMID: 32193345; PMCID: PMC7149310.

[2] Ozdemir RA, Tadayon E, Boucher P, Sun H, Momi D, Ganglberger W, Westover MB, Pascual-Leone A, Santarnecchi E, Shafi MM. Cortical responses to noninvasive perturbations enable individual brain fingerprinting. Brain Stimul. 2021 Mar-Apr;14(2):391-403. doi: 10.1016/j.brs.2021.02.005. Epub 2021 Feb 12. PMID: 33588105; PMCID: PMC8108003.

[3] Boes AD, Prasad S, Liu H, Liu Q, Pascual-Leone A, Caviness VS Jr, Fox MD. Network localization of neurological symptoms from focal brain lesions. Brain. 2015 Oct;138(Pt 10):3061-75. doi: 10.1093/brain/awv228. Epub 2015 Aug 10. PMID: 26264514; PMCID: PMC4671478.

[4] Grossman N, Bono D, Dedic N, Kodandaramaiah SB, Rudenko A, Suk HJ, Cassara AM, Neufeld E, Kuster N, Tsai LH, Pascual-Leone A, Boyden ES. Noninvasive Deep Brain Stimulation via Temporally Interfering Electric Fields. Cell. 2017 Jun 1;169(6):1029-1041.e16. doi: 10.1016/j.cell.2017.05.024. PMID: 28575667; PMCID: PMC5520675.

[5] Cappon D, den Boer T, Jordan C, Yu W, Lo A, LaGanke N, Biagi MC, Skorupinski P, Ruffini G, Morales O, Metzger E, Manor B, Pascual-Leone A. Safety and Feasibility of Tele-Supervised Home-Based Transcranial Direct Current Stimulation for Major Depressive Disorder. Front Aging Neurosci. 2022 Feb 2;13:765370. doi: 10.3389/fnagi.2021.765370. PMID: 35185515; PMCID: PMC8849231.

[6] Bréchet L, Yu W, Biagi MC, Ruffini G, Gagnon M, Manor B, Pascual-Leone A. Patient-Tailored, Home-Based Non-invasive Brain Stimulation for Memory Deficits in Dementia Due to Alzheimer's Disease. Front Neurol. 2021 May 20;12:598135. doi: 10.3389/fneur.2021.598135. PMID: 34093384; PMCID: PMC8173168.

[7] Cappon D, Fox R, den Boer T, Yu W, LaGanke N, Cattaneo G, Perellón-Alfonso R, Bartrés-Faz D, Manor B, Pascual-Leone A. Tele-supervised home-based transcranial alternating current stimulation (tACS) for Alzheimer's disease: a pilot study. Front Hum Neurosci. 2023 Jun 2;17:1168673. doi: 10.3389/fnhum.2023.1168673. PMID: 37333833; PMCID: PMC10272342.

Nature连发10篇，揭示迄今最全小鼠完整大脑细胞图谱

作者：宋薇

中科院神经所博士在读，科普爱好者，计算神经生物学方向。对有趣的科学问题好奇，致力于传播神经科学的魅力。

扫码查看原文

大脑是神经系统最高级的部分。2013年，美国NIH启动“美国脑计划”。随着技术的不断更迭，2017年，NIH进一步启动BICCN（BRAIN Initiative Cell Census Network）项目，这是脑计划的重要组成部分，旨在对人类、非人灵长类动物和小鼠大脑中的不同细胞类型进行全面识别和归类。

2021年10月6日，BICCN在Nature重磅发布16篇文章介绍大脑运动皮层细胞图谱；2023年10月13日，又在Science系列期刊连发21篇论文，揭示了迄今最全面的人类脑细胞图谱。

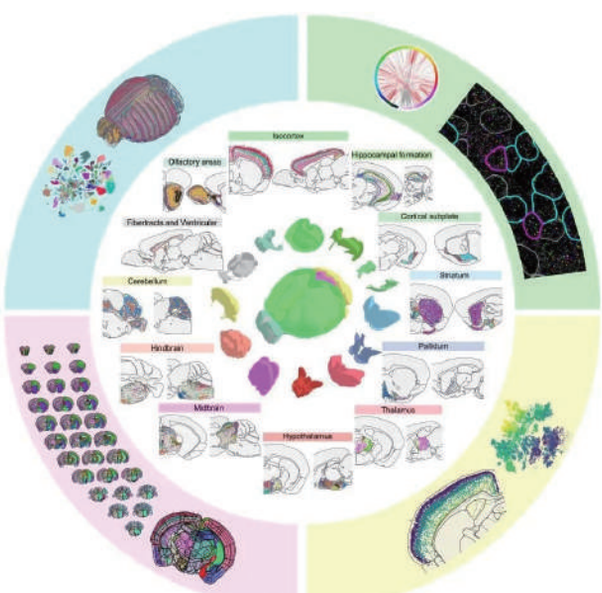
据Science发文仅两个月后，BICCN又有新动作，在Nature期刊同期发表了10篇研究论文，这10篇论文利用多种技术包括scRNA-seq、snRNA-seq、snm3C-seq、MERFISH（Multiplexed error-robust fluorescence in situ hybridization）、Slide-seq、STARmap、SMART-seq、snmC-seq、snATAC-seq分析了小鼠大脑中总计约3200万个细胞，鉴别出约5300个细胞类型，从而提供了迄今为止最全的小鼠完整大脑细胞类型的特性描述和分类。这些论文的研究团队来自艾伦脑科学研究所（Allen Institute）、哈佛大学、索尔克生物研究所（Salk Institute for Biological Studies）、博德研究所（Broad Institute）、加州大学圣地亚哥分校和加州大学伯克利分校等知名脑科学研究机构，论文通讯作者包括曾红葵、庄小威、何志刚、任兵、王潇、刘嘉、陈飞等多位华人学者。

▷图1：Nature特刊封面与完整小鼠大脑细胞类型图谱。图源：nature

过去对大脑的探测工作受限于选定区域，而此次的10篇论文突破了局限，提供了对整个小鼠大脑的详细调查。研究内容可分为三个主题：(1)构建小鼠全脑转录组学与空间图谱；(2)构建小鼠全脑脊髓投射神经元与脊椎动物视网膜神经元分类图谱；(3)剖析小鼠全脑表观遗传学特征与基因调控元件。

这些发现对完整的哺乳动物大脑结构和组织，以及单个脑细胞和神经回路的功能提供了详尽的信息。此外，它们也为进一步研究哺乳动物大脑的发育和演化提供了工具，包括不同类型的细胞组织引起神经系统疾病的微观机制。接下来，追问将带领读者一起了解此次的重要研究成果。

一、构建小鼠全脑转录组学与空间图谱

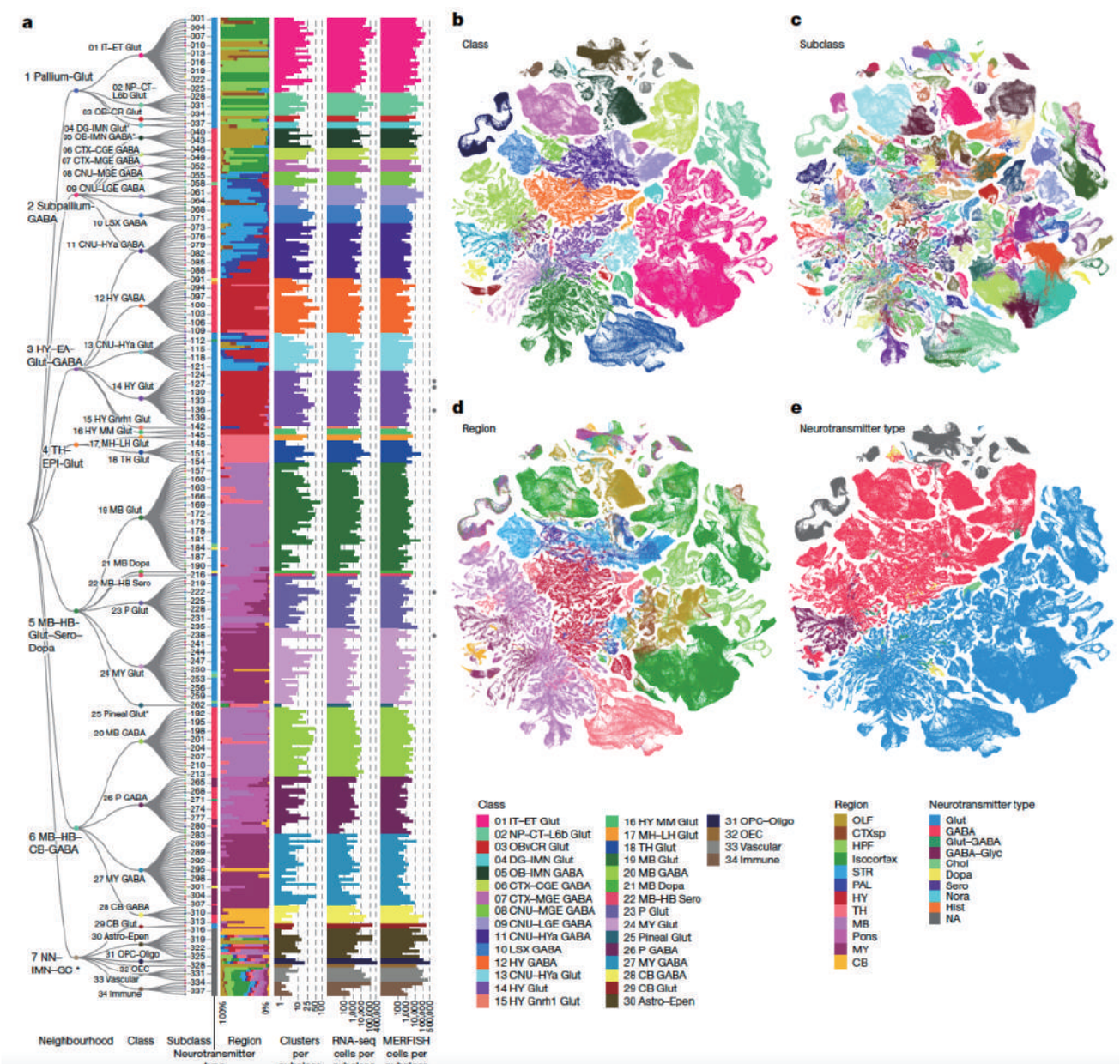


▷图2：利用MERFISH对约1000万个细胞中的1100个基因成像，生成分子定义和具有空间分辨率的全小鼠大脑细胞图谱。图源：nature

(1)小鼠全脑高分辨率的转录组及空间细胞图谱

在该专题的旗舰论文中,美国艾伦研究所的曾红葵等研究人员描述了他们是如何从结合了约400万细胞的单细胞RNA测序和约430万细胞的空间转录组学数据中,创建出首个完整小鼠大脑高分辨率转录组细胞类型图谱的。

研究团队结合了两个海量数据集——约700万个细胞的单细胞RNA测序(scRNA-seq)数据和约430万个细胞的空间转录组学(MERFISH)数据,将小鼠全脑细胞分为34个大类(class),再进一步细分为338个子类(subclass),1,201个超型(supertype),5,322个细胞集群(cluster)。其中,神经元的细胞类型显示出惊人的多样性,共有5,205个集群。



▷图3:小鼠全脑转录组细胞类型分类。图源:参考文献[1]

研究者指出，空间分辨率高是这份图谱的独特之处，由此该图谱可以将每种细胞的转录组特征与其空间位置高度对应，从而清楚地观察到大脑解剖结构与细胞类型之间的关系。

从图谱中可以观察到不同脑区细胞类型组织的独特特征，例如，在大脑的背侧和腹侧，细胞类型的组织存在明显差异：背侧区域含有的细胞类型较少，但高度分化；腹侧则包含了更多类型的神经元，但类型之间的关系较近。这种差异与不同结构的功能和演化有关。背侧部分包括皮质、丘脑和小脑等，主要执行生物体的适应性功能，如认知、感觉运动等，演化时间较晚，细胞类型在多样性上的拓展速度较快；腹侧包括下丘脑、中脑和后脑等较古老的结构，承担了进食、繁殖、新陈代谢平衡等更基本的生存功能，专用细胞类型和神经回路的变化就相对较小。

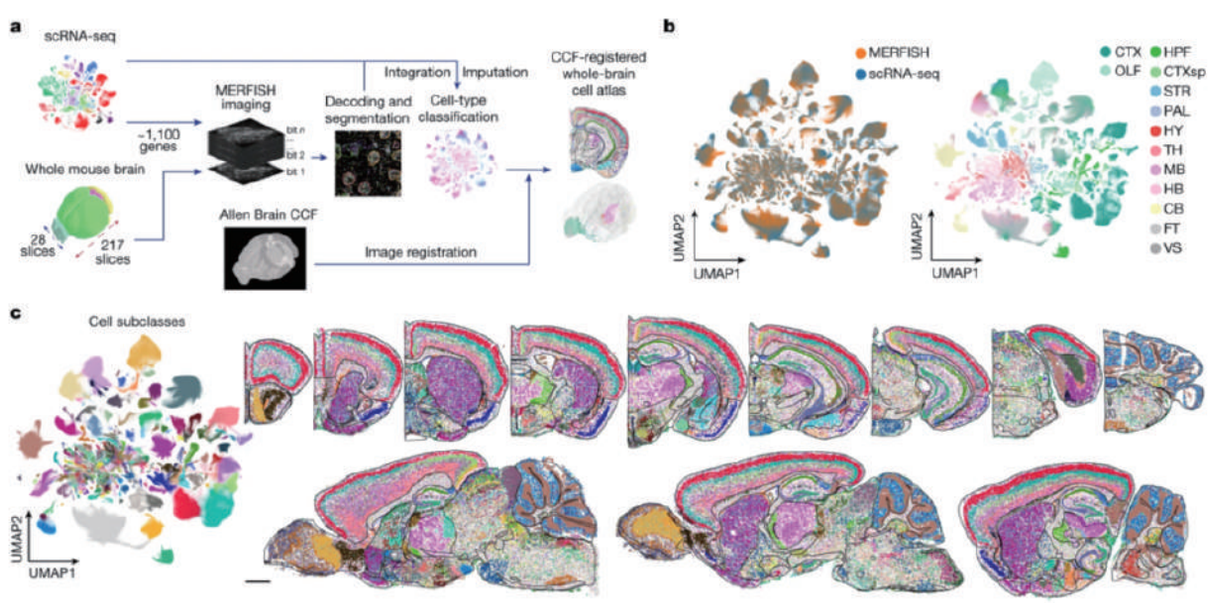
此外，这项研究还发现，不同细胞类型中神经递质和神经肽表达和共表达模式具有极为丰富的多样性和异质性。进一步研究表明，转录因子是细胞类型分类的主要决定因素。研究也确定了整个大脑各个部位决定每一种细胞类型的主要转录因子组合。

（2）小鼠全脑的分子与空间细胞图谱

哈佛大学庄小威实验室开发的多重容错性荧光原位杂交（multiplexed error-robust fluorescence in situ hybridization，简称MERFISH）技术在图谱创建中发挥了重要作用。这项技术可以让研究者在细胞或组织切片中同时检测多种RNA，从而绘制选定区域的基因表达空间图谱。

在这个由庄小威团队领衔的第二项研究中，研究者利用MERFISH对小鼠全脑约1000万个细胞中超过1100个基因进行了研究，绘制了一张超过5000个具有不同转录背景的细胞簇（归属于超过300个主要细胞类型）的高清分子与空间图谱（图4）。

过程中，研究团队将MERFISH与单细胞RNA测序数据结合进行成像，放入艾伦研究所的小鼠全脑通用坐标框架中，创建出单细胞空间图谱（每个细胞都有全转录组表达谱），可以系统地定量分析各个脑区的细胞类型和组成，推断细胞子类之间的特异性相互作用，预测配体-受体之间的分子关系，以及细胞-细胞相互作用的功能。



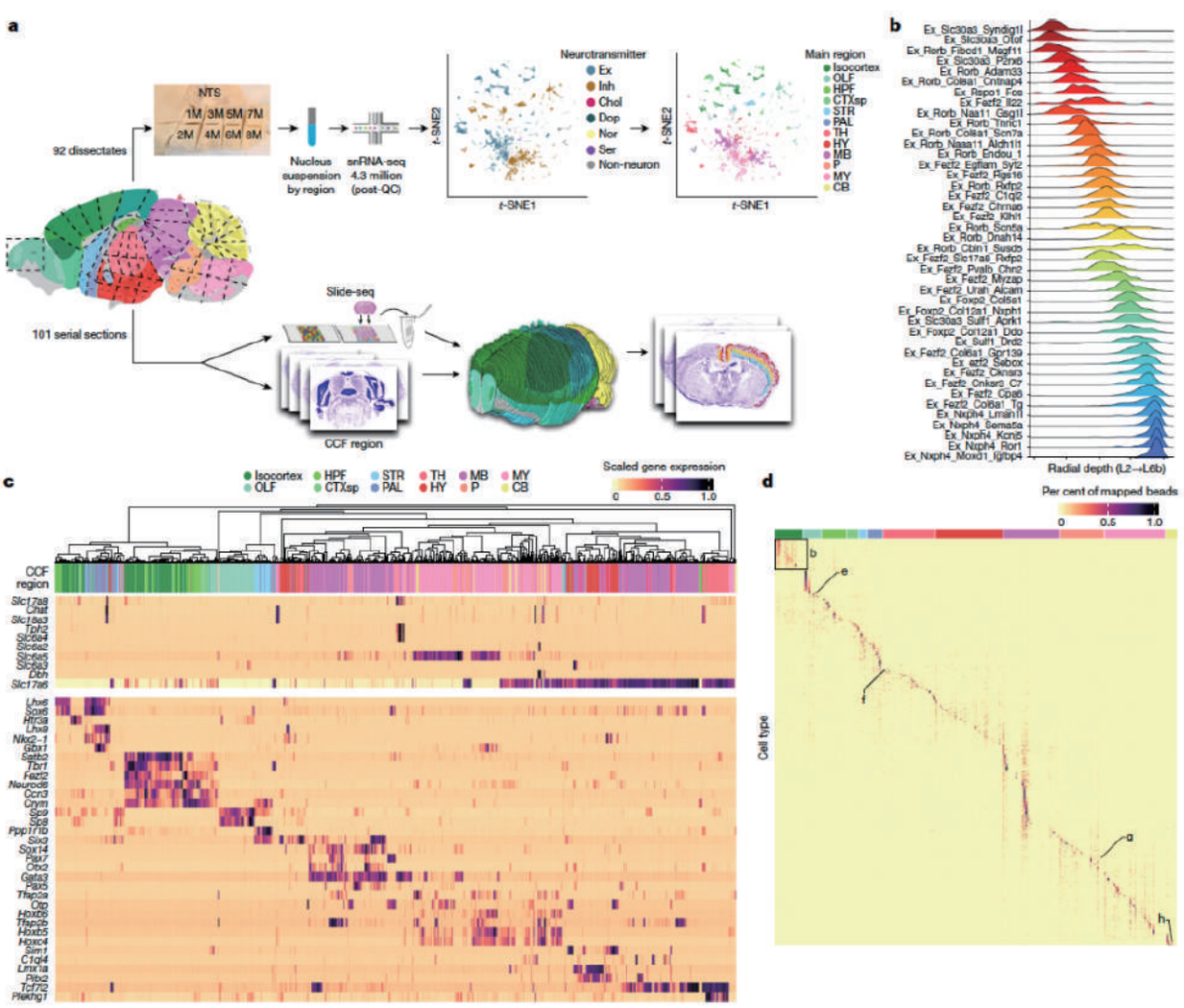
▷图4：构建全鼠脑细胞图谱的工作流程。图源：参考文献[2]

(3)成年小鼠大脑的分子细胞架构

论文集的第三篇来自麻省理工学院和哈佛大学博德研究所。研究采用了不同但互补的方法，将新开发的空间转录组学技术Slide-seq与高通量单核RNA测序相结合，以440万个10微米像素的分辨率量化全基因组表达，使细胞类型能够系统地定位到单个大脑区域（图5）。这揭示了小鼠全脑各个解剖结构中的细胞类型组成，包括过去关注较少的中脑、脑桥、延髓和下丘脑。研究人员估计，分析已涵盖小鼠大脑中约90%的细胞类型。

研究发现，在进化上更为古老的大脑区域，特别是中脑、后脑和下丘脑，它们存在的细胞类型多样性异常惊人。分析结果表明，在中脑和下丘脑内发现的细胞类型比整个端脑内发现的还多。此外，在全脑分布上，具有长程投射的神经元的分布比局部中间神经元的分布更具区域特异性；非神经元类型的细胞如星形胶质细胞、外胚层细胞、伸展细胞、血管软脑膜细胞等，也具有非常明显的空间特异性，这表明了神经元与胶质细胞、胶质细胞与血管之间的特异性相互作用。

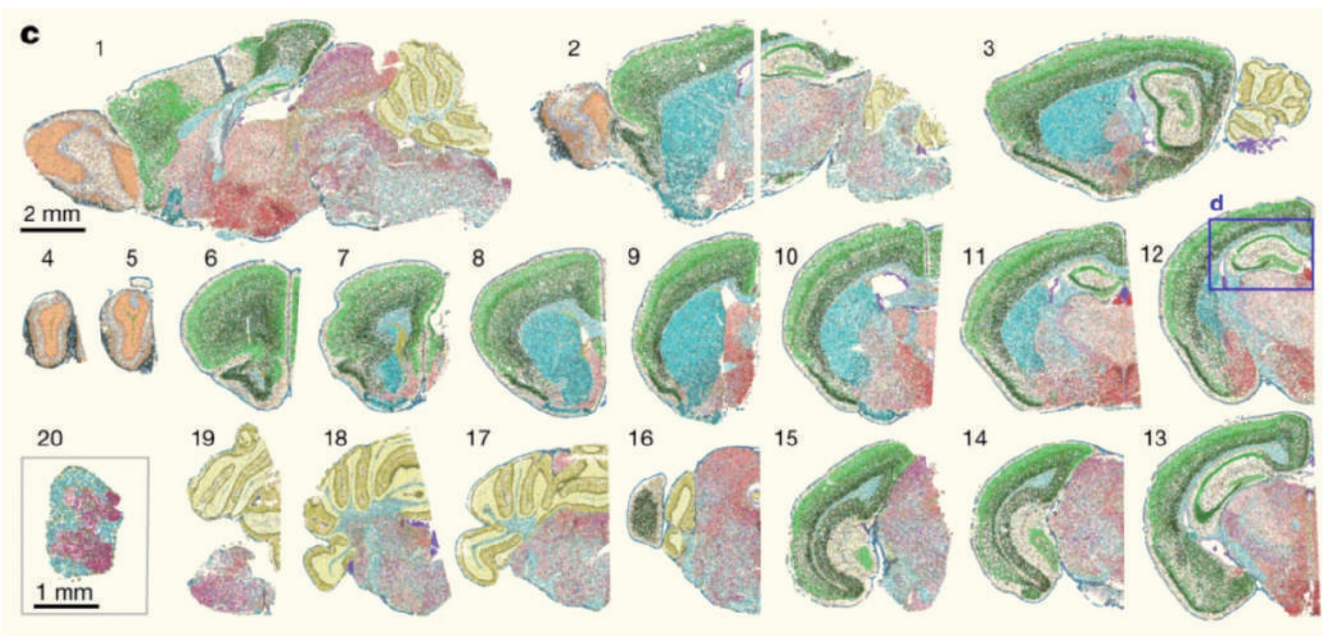
为了促进对相对未被充分研究的细胞类型功能的更多研究，研究团队通过在线开放资源的形式提供了这些数据：www.BrainCellData.org。



▷图5：利用全脑snRNA-seq和Slide-seq数据集绘制细胞类型的空间分布图谱。图源：参考文献[3]

(4)小鼠全脑高分辨率单细胞空间转录组图谱

第四篇论文由博德研究所王潇团队联合哈佛大学刘嘉团队共同发表(注:该论文已于2023年9月27日发表)。研究者采用原位空间转录组学技术STARmap PLUS,以194X194X345nm3体素尺寸在亚细胞分辨率水平上对成年小鼠全脑和脊髓中17个冠状切面和3个矢状切面组织切片中的1,022个基因进行了检测,并使用细胞分割算法ClusterMap得到了109万个高质量空间分辨单细胞基因表达(图6)。通过大规模的单细胞分析注释,该团队基于单细胞基因表达定义和注释了230个“分子细胞类型”并基于空间基因表达定义和注释了106种“分子组织区域”。在这一过程中,研究人员发现了部分端脑抑制性中间神经元亚型的脑区分布特异性。例如,纹状体中特有的中间神经元、嗅球的外网状层中表达多巴胺的中间神经元。

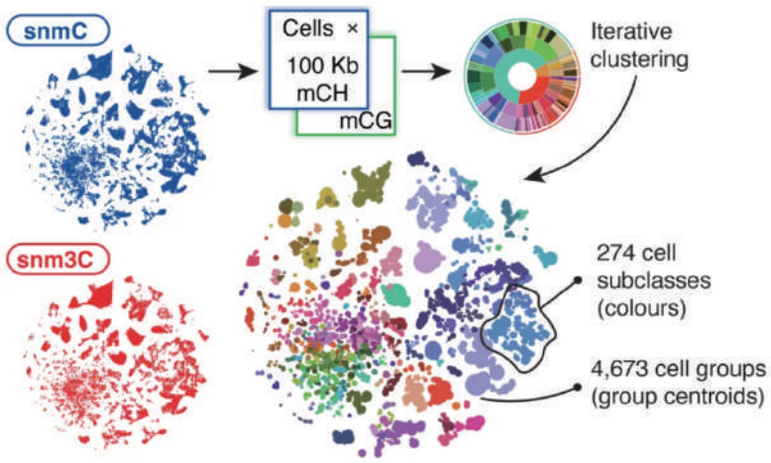


▷图6:小鼠中枢神经系统单细胞分辨率细胞类型空间图谱,230种细胞类型由彩色点标注。图源:参考文献[4]

基于空间上的分子表达,研究人员还补充和完善了小鼠大脑解剖学结构。例如,研究人员从分子表达和细胞分布的角度出发,划分小鼠大脑皮层的分区,并与传统解剖学的定义比较。一个有趣的发现是,解剖学上的压后皮层在小鼠大脑的前端和后端具有截然不同的“分子组织区域”组成;后端压后皮层在“分子组织区域”组成上与相邻的视觉皮层有着更高的相似性,这为理解压后皮层在视觉相关的行为和记忆中的功能提供了新的思路。

二、剖析小鼠全脑表观遗传学特征与基因调控元件

在论文集的第二个专题中,多个研究团队应用小鼠全脑转录组细胞类型图谱,分析了不同细胞类型的基因调控程序,并用于识别造成神经系统疾病和特性的遗传变异(图7)。

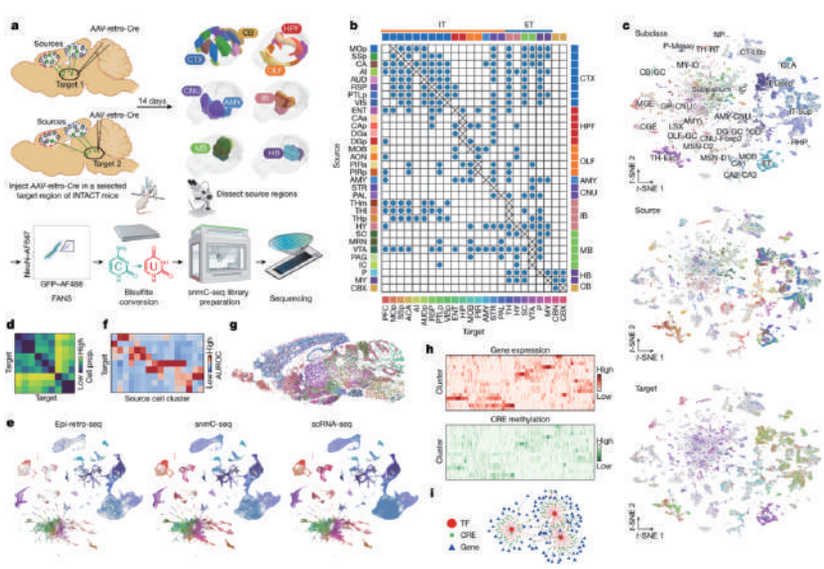


Cell and genome-based atlases across the whole mouse brain and the integration of molecular features such as DNA methylation, gene expression and chromatin accessibility. Credit: Hanqing Liu, Jingtian Zhou and Joseph Ecker.

▷图7：整个小鼠大脑的细胞和基因组图谱，以及DNA甲基化、基因表达和染色质可及性等分子特征的整合。图源：nature

（5）小鼠大脑神经元表观组及远端投射图谱

大脑中的许多细胞类型通过远距离投射来调节细胞功能，还有许多特定细胞类型之间具有相互作用。加州大学圣地亚哥分校的Edward M. Callaway团队和索尔克生物研究所Joseph R. Ecker合作，通过Epi-Retro-Seq将单细胞表观基因组和细胞类型与小鼠全脑中32个不同区域的33,034个神经元的长距离投射联系起来，这些神经元投射到24个不同的靶点（图8）。这个数据集涉及大脑32个不同区域超过3万个大脑神经元。借助该数据集，研究者可以量化投射到不同目标脑区神经元之间的遗传差异，对分子细胞类型和它们的投射靶标进行注释，并且在投射富集的细胞类型中构建基因调控网络。

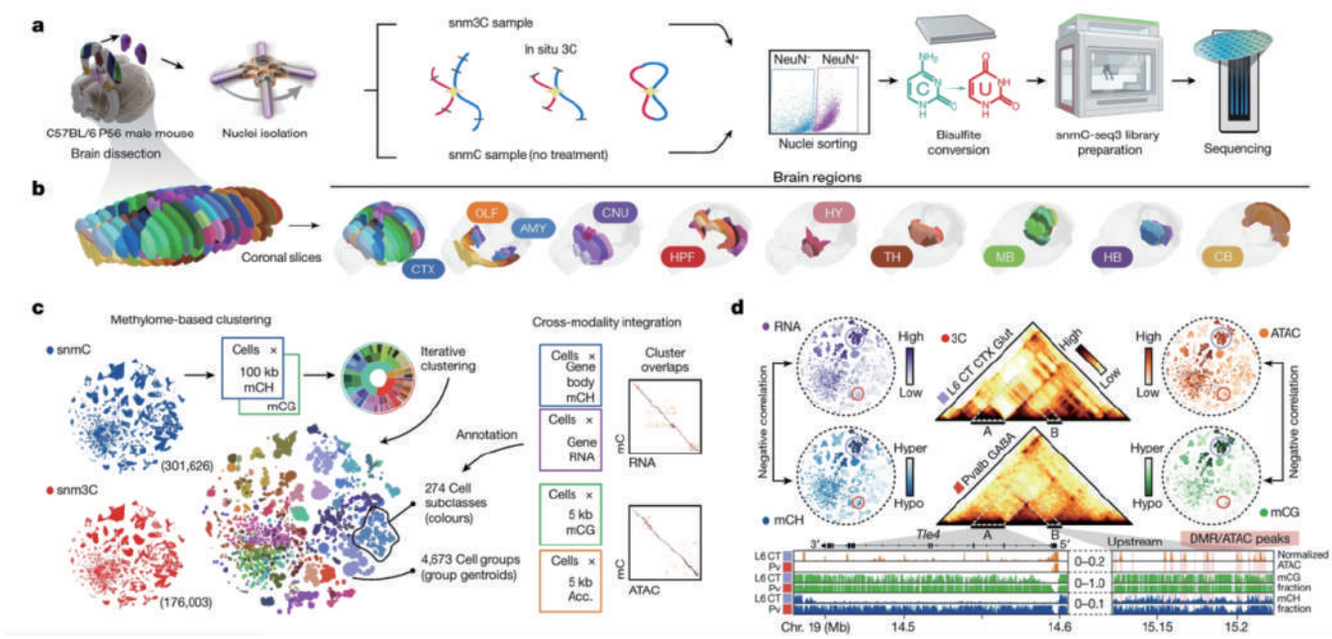


▷图8：全脑投射神经元的表观基因组图谱。图a.b表示用于逆行标记和表观遗传分析单个投射神经元的Epi-Retro-Seq工作流程。图源：参考文献[5]

(6)成年小鼠全脑单细胞DNA甲基化和三维多组学图谱

胞嘧啶DNA甲基化对大脑发育非常关键,并与各种神经系统疾病有关。索尔克生物研究所Joseph R.Ecker团队尝试在空间背景下理解大脑的DNA甲基多样化。借助单细胞DNA甲基化测序(snmC-seq(3)和甲基化联合三维基因组测序(snm3C-seq)技术,首次构建了完整的大脑细胞类型及其基因调控的分子图谱(图9)。

研究人员一共从成年小鼠大脑的117个解剖区域解析了超30万个甲基组细胞和17万个甲基组联合三维基因组细胞。通过迭代聚类和整合分析,联合了脑计划中的转录组数据集(10XRNA)以及基因组可及性数据集(snATAC-seq),构建了一个基于甲基化的细胞分类图谱,包括4,673个细胞群组和274个多组学注释的亚类。

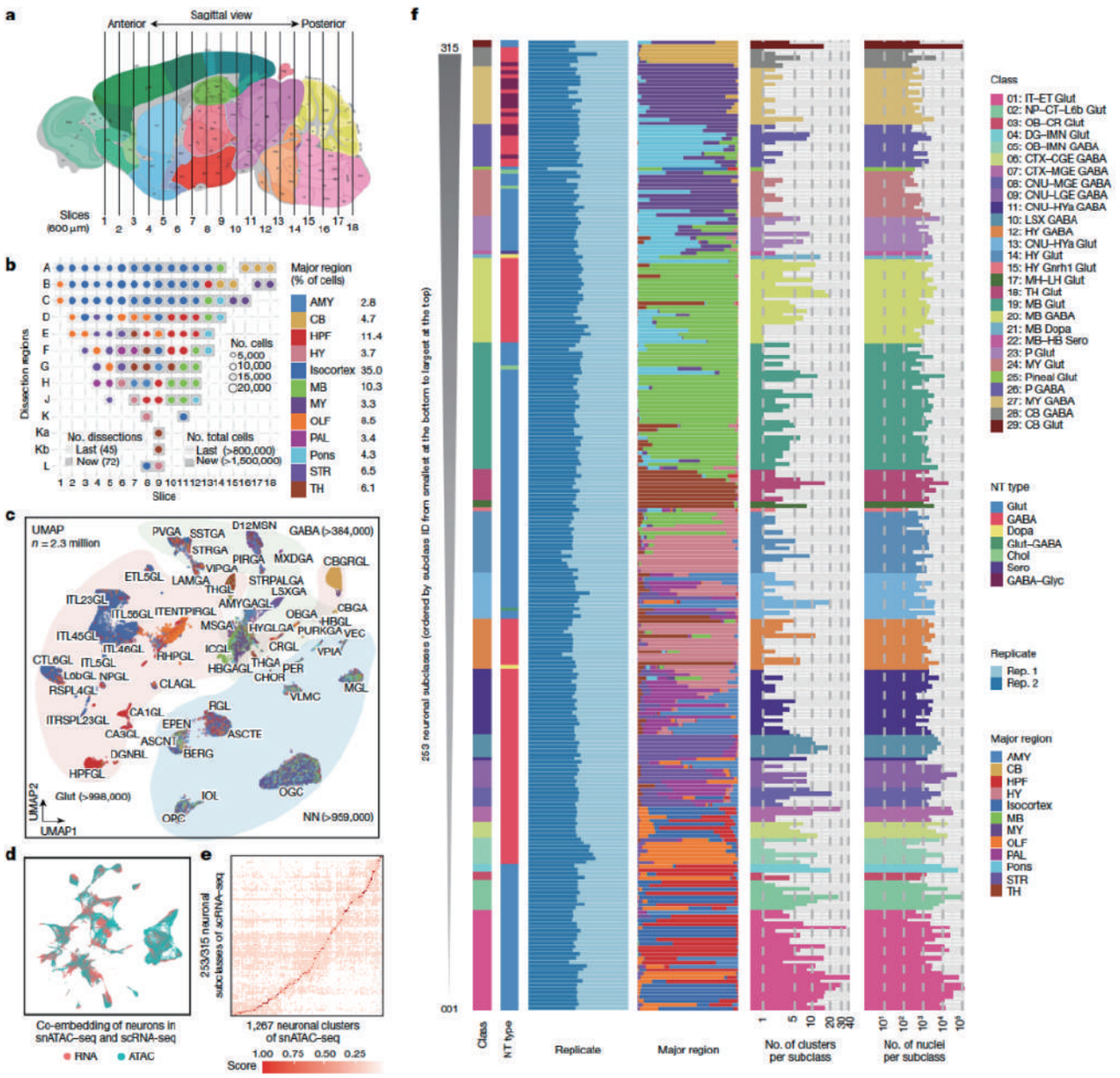


▷图9:snmC-seq3 + snm3C-seq技术工作流程示意图。图源:参考文献[6]

研究人员系统比较了全脑单细胞甲基化组学数据集与空间转录组学(MERFISH)数据集的关联性,结果表明,不同细胞的基因组甲基化模式与其在大脑中的空间位置高度相关。通过比较数百种不同的神经元亚类,研究人员还发现三维基因组结构也与DNA甲基化高度相关,尤其是在众多与神经细胞功能相关的长基因周围。这些全脑单细胞DNA甲基化组图谱可以为了解小鼠大脑细胞空间、基因组调控多样性提供大量资源。

(7)成年小鼠大脑单细胞染色体可及性图谱

由加州大学圣地亚哥分校任兵团队绘制的成年小鼠大脑中顺式作用DNA元件(candidate cis-regulatory elements, cCREs)综合图谱,通过分析230万个单细胞ATAC-seq数据而生成。整个图谱包括约100万个cCREs,以及它们在1,482个不同脑细胞群体中的染色质可及性。与早期该类研究相比较,新研究为小鼠基因组新增加了44.6万多个cCREs(图10)。



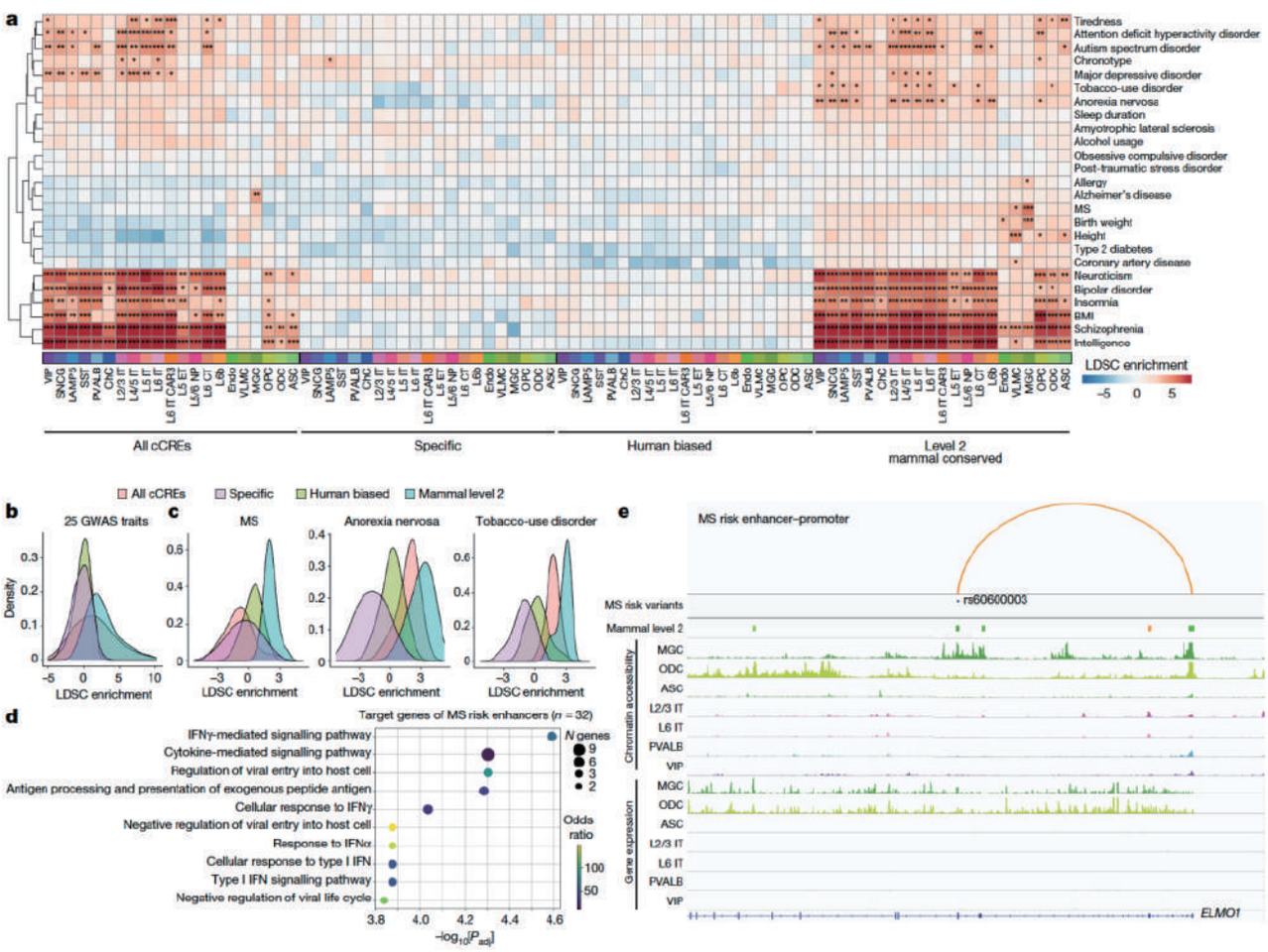
▷图10：成年小鼠全脑染色体可及性的单细胞分析。图源：参考文献[7]

根据该染色体可及性图谱，研究发现转座子（transposable elements，TEs）在新的调控程序和神经元多样性中具有潜在作用。例如，较其他小鼠脑细胞，在部分皮层兴奋性神经元中，研究人员观察到小鼠特异的cCREs更为显著地在TEs中富集。通过整合单细胞ATAC-seq数据和单细胞RNA-seq数据，研究人员刻画了260多个小鼠脑细胞亚型的基因调控网络，并利用深度学习工具，训练了一个可以仅根据DNA序列预测不同脑细胞类型中cCREs的染色质可及性的模型。这项工作为分析小鼠和人类全脑的细胞类型特异的基因调控程序提供了坚实的数据支持。

（8）多组学揭示哺乳动物初级运动皮层的基因调控

在该专题的最后一项研究中，任兵团队和Joseph R.Ecker团队强强联合，利用单细胞多组学方法研究了人类、猕猴、小鼠初级运动皮层的基因调控程序，从超过20万个细胞中获得了基因表达、染色体可及

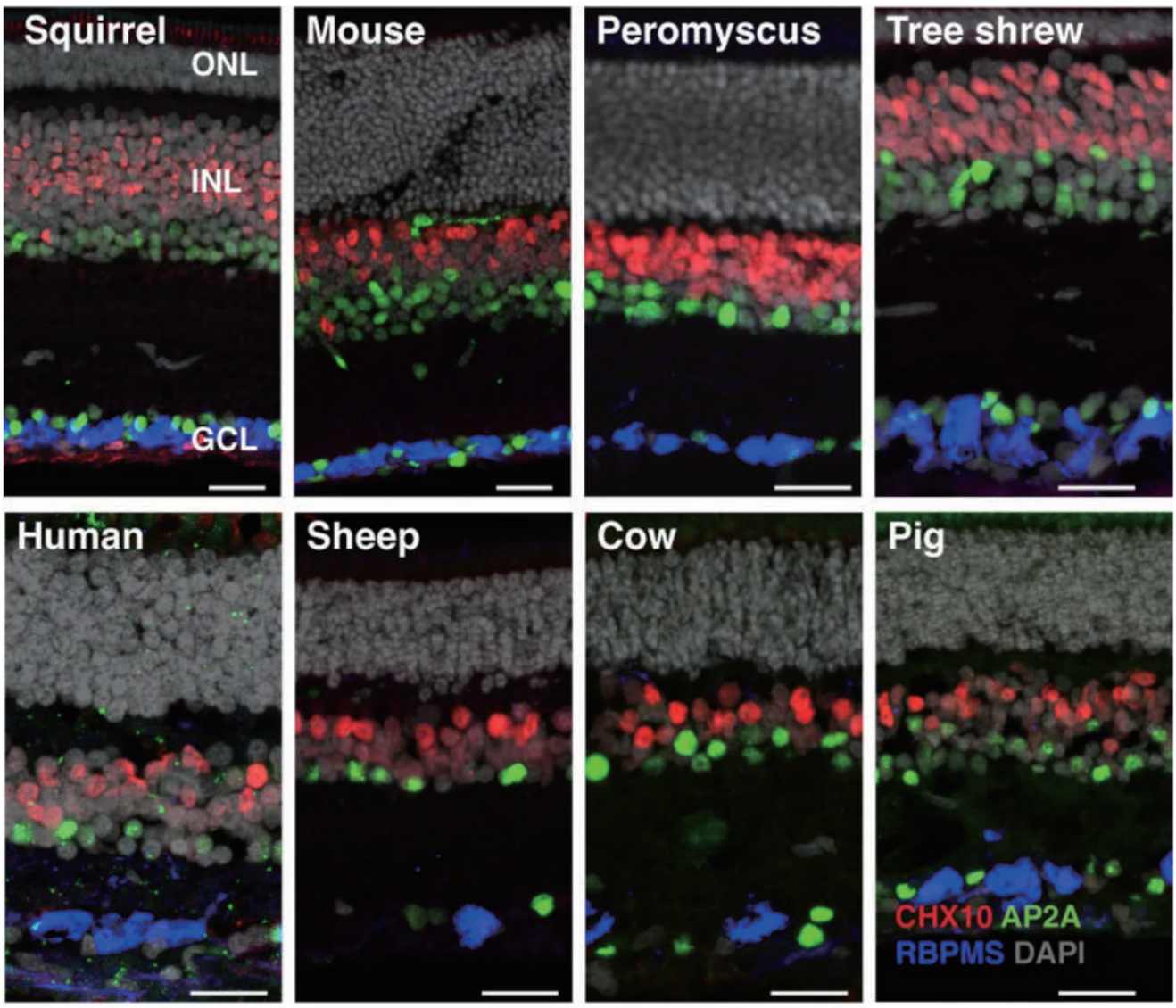
性、DNA甲基组和染色体构象图谱(图1(1)。这些数据显示,转录因子表达差异与物种特异性表观基因组相对应。研究人员发现,保守和分化的基因调控特征可以在三维基因组的演化中得以反映。在皮层细胞中,TEs占据了人类约80%的cCREs。利用机器学习手段,研究人员在不同物种中开发了可以预测顺式调控元件的预测因子,该模型揭示了从啮齿类到灵长类基因组调控规则的高度保守性。表观遗传的保守性结合序列相似性有助于挖掘功能性cCREs,帮助我们理解神经系统疾病的遗传变异机制。例如,该研究还发现了与多发性硬化、厌食和烟瘾相关的遗传变异中的保守特征。



▷图11:利用表观遗传学保守特征解释与神经系统疾病风险相关的非编码变异。图源:参考文献[8]

三、小鼠全脑脊髓投射神经元与脊椎动物视网膜神经元分类图谱

基于对全脑细胞转录组的研究,论文集的第三个专题针对特定神经元类型,即小鼠脊髓投射神经元与脊椎动物视网膜神经元图谱进行了探索。

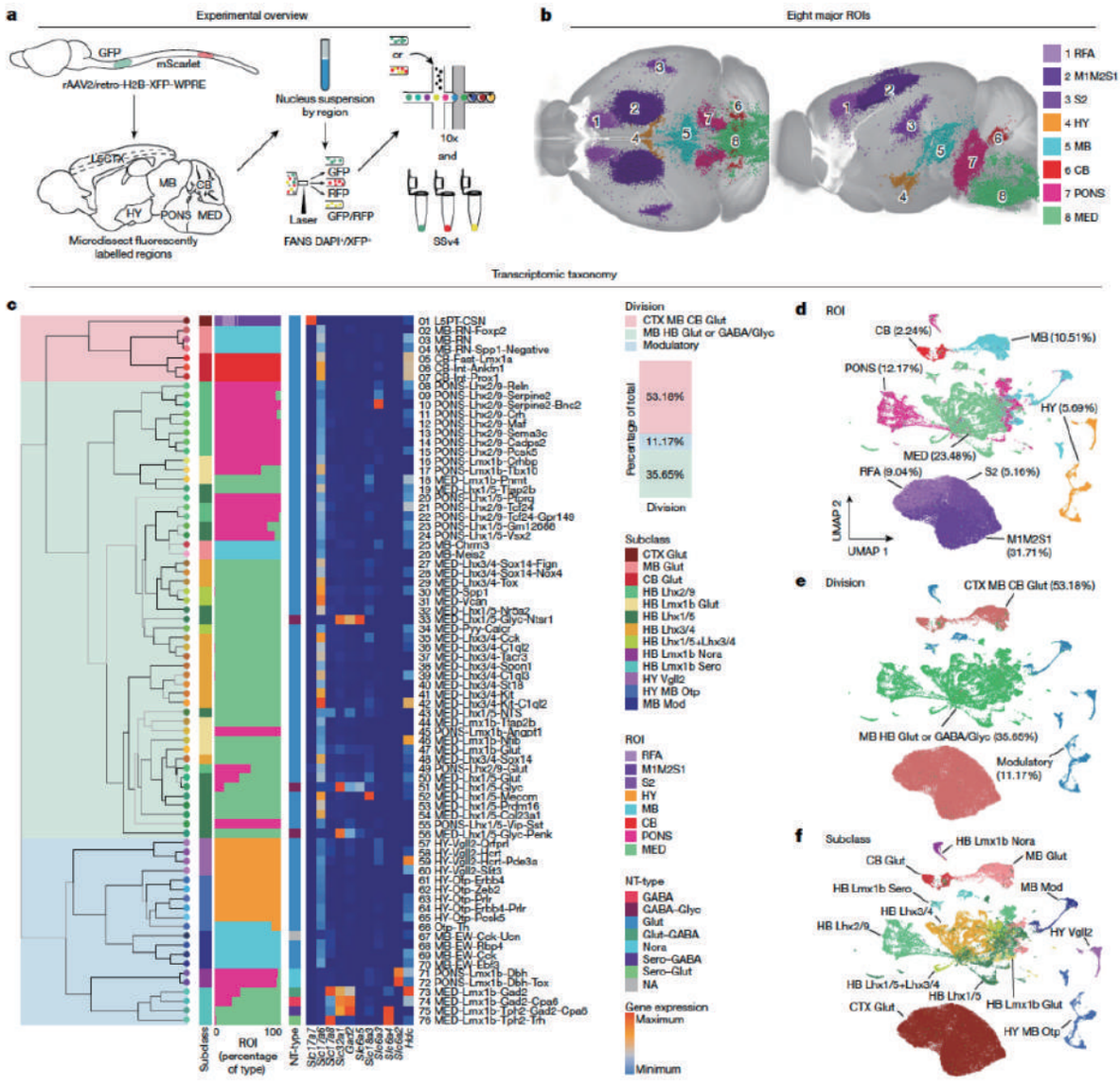


▷图12：对17种脊椎动物的比较转录组学分析揭示了灵长类动物中主导视觉处理的视网膜细胞类型的深度进化保守性，并确定了其他物种中的相应细胞。图源：nature

(9)小鼠全脑脊柱投射神经元的转录组分类

下行通路由直接从大脑各个区域投射到脊髓的脊柱投射神经元(spinal projecting neurons, SPNs)组成，SPNs将信号由大脑传递至脊髓，将大脑的指令(“思想”)转化为身体行为的指令(“行动”)。这些SPNs不仅对自主和非自主运动至关重要，而且在感觉调节和自主功能(如血压、心率和恐惧反应)中也起到关键作用。然而，对全脑SPNs的全面分子表征仍然不足。

哈佛医学院何志刚和艾伦研究所曾红葵合作，对65,002个SPNs进行了转录组分析，确定了3个群体、13个亚类和76种类型，遍及大脑皮层、下丘脑、中脑、桥脑背侧被盖体、桥脑网状结构和小脑深核，并将这些类型映射到整个小鼠大脑图谱中[1]，以获得每种类型的精确空间位置(图1(3)。



▷图13：全脑脊柱投射神经元的解剖学转录组图谱。图源：参考文献[9]

具体来说，该研究将全脑SPNs分为三大群体。(1)来自皮质、红核和小脑的具有分子同质性的兴奋性SPNs，这群神经元可能适用于脑于脊髓的“点对点”信息传递；(2)核分子异质性SPNs种群，将信息传递至脊髓多个区域，适合传递与整个脊髓活动相关的指令；(3)在下丘脑、中脑和网状核中表达缓慢作用的神经递质和/或神经肽的调节神经元，用于脑脊髓信号的“增益”。

这项研究揭示了全脑SPNs的全面分类，并深入探讨了SPN在影响大脑对身体功能控制中的功能组织。

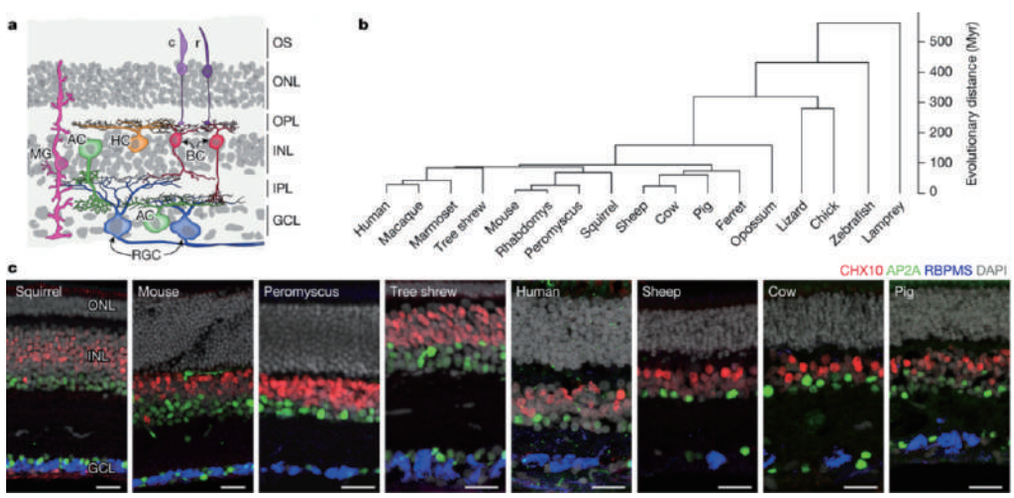
(10)脊椎动物视网膜中神经元细胞类型及演化

在该主题的另一篇论文中，加州大学伯克利分校Karthik Shekhar和哈佛大学Joshua R. Sanes研究团队合作，共同解开了利用细胞类型图谱窥视动物眼睛的演化之谜。

视网膜是眼睛直接与脑部相连的结构，也是视觉感知的起点。不同动物的视觉差异巨大，但新研究却指出，不同脊椎动物视网膜的细胞类型有着惊人的一致性。研究人员首先生成并比较了包括人类在内17

种脊椎动物的视网膜单细胞转录组图谱，这些物种跨越了广泛的动物类型，但6种视网膜细胞类型（光感受器、水平细胞、双极细胞、无长突细胞、视网膜神经节细胞和穆勒胶质细胞）有着高度分子保守性，但随着物种进化距离也存在转录差异（图1（4）。基于保守的基因表达程序，众多细胞类型在不同物种间共享。

这些发现提供了一种视网膜进化的可能性：早在2亿年前，所有哺乳动物最后共同祖先的视网膜，就已经具有与现代哺乳动物相媲美的复杂性；其中一些细胞类型的历史，甚至可以追溯到4亿多年前所有脊椎动物的共同祖先。



▷图14：脊椎动物视网膜的切片示意图，显示其六个主要细胞类别：光感受器（包括视杆细胞（r）和视锥细胞（c））、水平细胞（HC）、双极细胞（BC）、无长突细胞（AC）、视网膜神经节细胞（RGC）和穆勒胶质细胞（MG）。图源：参考文献[10]

这项研究还为一类视网膜神经节细胞（RGC）的起源提供了全新的认知。侏儒视网膜神经节细胞（midget　RGC）占据了人类RGC细胞的80%以上，此前被认为源于灵长类动物，它们为人类高度灵敏的视力提供了基础。但基于对小鼠脑图谱的对比研究发现，小鼠视网膜中也有与侏儒RGC同源的细胞alpha　RGC。与人类的侏儒RGC相比，alpha　RGC的细胞体更大，但数量更少，仅占小鼠RGC的2%～4%。追踪它们的进化表明，它们可能随着视觉皮层的扩大而变得更小和更多，而视觉皮层是人类视觉处理的主要中心。这种对应关系表明，视网膜和皮层同时进化，为灵长类动物提供了高灵敏度的视觉。

四、BICCN2.0 的意义何在？

在过去的十年里，单细胞RNA测序的技术建立推动了细胞类型图谱的产生。到目前为止，多模式、高分辨率的图谱主要局限于大脑的特定部位，比如运动皮层。BICCN2.0图谱将单细胞RNA测序与高分辨率单细胞转录组学技术结合，极致详尽地展示出全脑而不仅仅是某个脑区有多少种脑细胞类型，以及它们在全脑中的比例和空间排列。小鼠是神经科学研究中最常用的脊椎动物实验模型，这张图谱为更好地理解人类大脑铺平了道路。细胞图谱还为开发新一代精确治疗大脑精神和神经疾病的疗法奠定了基础。

细胞图谱描述了小鼠大脑每个区域的细胞类型及其在这些区域内的组织结构。研究表明，一个区域中神经元类型的数量并不与该区域的大小或神经元总数成比例。不同的规则支配着整个大脑的神经元多样性，这种现象可能是因为大脑的每个区域都在不同的约束下进化。转录因子决定了细胞类型的分类。在这

些观察的基础上，研究人员进一步探究了脊柱投射神经元与视网膜神经元的转录特征。

此外，通过剖析小鼠全脑表观遗传学特征与基因调控元件，图谱揭示了在不同细胞类型和物种中，遗传密码（细胞DNA和染色体的化学修饰）如何被读取和解释的规则。研究表明，基因序列和细胞类型特异性表观基因组之间的联系在物种中是保守的。高度保守的调控元件往往位于它们所调控基因的起始位置附近，而物种特异性调控元件则位于更远的位置。小鼠或人类细胞特有的调控元件经常与转座元件重叠，这表明转座子可能被用于微调特定细胞类型和不同物种的基因激活。通过将表观基因组图谱与大脑不同部位细胞类型之间的投射联系起来，研究表明表观基因组可以塑造不同大脑区域细胞类型之间的相互作用，以及单个细胞的遗传活动。

2021年，BICCN的研究人员在大约1%的小鼠大脑中绘制了这些细胞类型。这些基因组技术的进步使研究小组在两年内就能对成年小鼠大脑中90%的细胞类型进行分类。目前，美国国立卫生研究院已经启动了另一个为期五年的项目，名为“大脑倡议细胞图谱网络（BICAN）”，其目的是扩大技术规模，绘制人类和非人类灵长类动物以及小鼠大脑在器官发育过程中的脑细胞类型。BICAN代表了一个创新性的倡议，它与另外两个重要项目并行发展：大脑跨尺度连接计划（BRAIN Initiative Connectivity Across Scales）和精确获取脑细胞的装备计划（Armamentarium for Precision Brain Cell Access）。这些项目共同致力于解析控制行为神经回路的基本原理，并为治疗人类脑部疾病开辟新途径，以期彻底革新神经科学的研究领域。（编辑：韵珂）

参考文献

[1]Yao,Z., van Velthoven, C.T.J., Kunst, M. et al. A high-resolution transcriptomic and spatialatlas of cell types in the whole mouse brain. Nature 624, 317–332 (2023). https://doi.org/10.1038/s41586-023-06812-z

[2]Zhang, M., Pan, X., Jung, W. et al. Molecularly defined and spatially resolvedcell atlas of the whole mouse brain. Nature 624, 343–354 (2023).https://doi.org/10.1038/s41586-023-06808-9

[3]Langlieb, J., Sachdev, N.S., Balderrama, K.S. et al. The molecular cytoarchitecture of the adultmouse brain. Nature 624, 333–342 (2023).https://doi.org/10.1038/s41586-023-06818-7

[4]Shi, H., He, Y., Zhou, Y. et al. Spatial atlas of the mouse central nervoussystem at molecular resolution. Nature 622, 552–561 (2023).https://doi.org/10.1038/s41586-023-06569-5

[5]Zhou, J., Zhang, Z., Wu, M. et al. Brain-wide correspondence of neuronalepigenomics and distant projections. Nature 624, 355–365 (2023).https://doi.org/10.1038/s41586-023-06823-w

[6]Liu, H., Zeng, Q., Zhou, J. et al. Single-cell DNA methylome and 3D multi-omicatlas of the adult mouse brain. Nature 624, 366–377 (2023).https://doi.org/10.1038/s41586-023-06805-y

[7]Zu, S., Li, Y.E., Wang, K. et al. Single-cell analysis of chromatinaccessibility in the adult mouse brain. Nature 624, 378–389 (2023).https://doi.org/10.1038/s41586-023-06824-9

[8]Zemke, N.R., Armand, E.J., Wang, W. et al. Conserved and divergent gene regulatoryprograms of the mammalian neocortex. Nature 624, 390–402 (2023).https://doi.org/10.1038/s41586-023-06819-6

[9]Winter,C.C., Jacobi, A., Su, J. et al. A transcriptomic taxonomy of mousebrain-wide spinal projecting neurons. Nature 624, 403–414 (2023).https://doi.org/10.1038/s41586-023-06817-8

[10]Hahn, J., Monavarfeshani, A., Qiao, M. et al. Evolution of neuronal cell classes and typesin the vertebrate retina. Nature 624, 415–424 (2023).https://doi.org/10.1038/s41586-023-06638-9

神经科学领域里程碑，全面构建人类大脑单细胞图谱

作者：宋薇

中科院神经所博士在读，科普爱好者，计算神经生物学方向。对有趣的科学问题好奇，致力于传播神经科学的魅力。

扫码查看原文

大脑是人类神经系统最为高级的部分，负责接收并发出我们日常生产活动中的各种指令。有关大脑的研究极具挑战，人类距离理解大脑还有多远的距离？多年来，不同国家的科学家都在试图破译大脑这一复杂而又精密的器官。

2013年，美国国立卫生研究院（NIH）启动了“美国脑计划”，全称为“使用创新神经技术的脑研究计划”（Brain Research Through Advancing Innovative Neurotechnologies Initiative，BRAIN Initiative）。2017年起，随着单细胞测序技术和单细胞成像技术的高速发展，BICCN（BRAIN Initiative Cell Census Network）项目启动，这是一个由分布在美国和欧洲的研究团队组成的联盟，旨在用前沿技术对小鼠、非人灵长类动物和人类大脑的细胞类型进行识别和分类，以及针对特定细胞类型开发新的遗传工具。

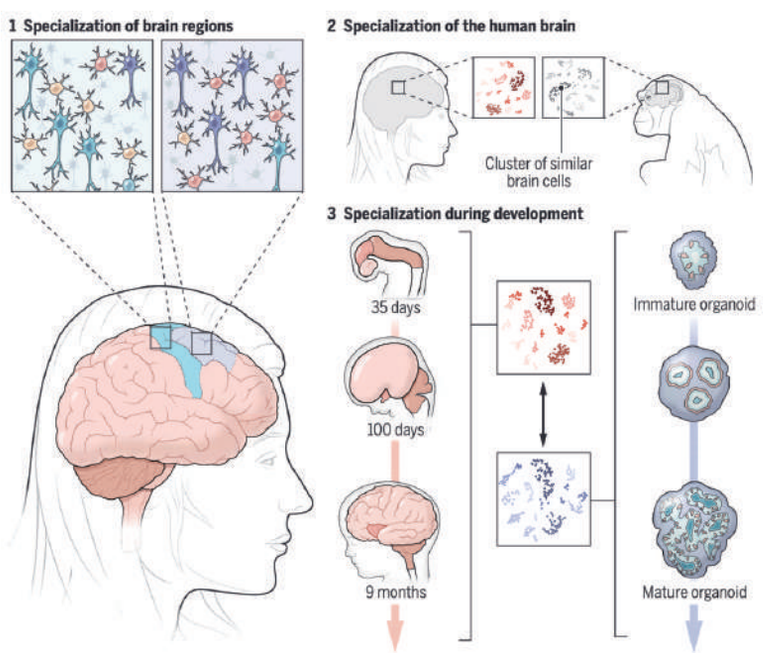
2023年10月13日，BICCN计划的科学家在Science及其子刊Science Advances和Science Translational Medicine上共发布了21篇相关论文，发布了迄今为止最大、最全面的人类、非人灵长类脑细胞图谱，揭示了超过3000种脑细胞类型，其中许多发现都具有“首创性”。

▷图 1:Science与Science Advances杂志封面合并为一个完整人类大脑图像。图源:Science官网

早在2021年10月,Nature杂志曾一次性上线16篇研究论文,报告了BICCN项目的阶段性研究成果,其主要强调在分子水平上全面绘制哺乳动物初级运动皮层细胞类型的特征。与两年前的研究不同,此次Science新发布的专题主要聚焦于人脑,尤其是从单细胞层面以前所未有的颗粒度综合转录组学、蛋白质组学、表观遗传组学解析了人脑的组织结构[1]。

这些研究主要围绕四个主题进行:(1)成体人类及非人灵长类动物单细胞图谱;(2)成体人类与非人灵长类单细胞图谱比较研究;(3)人类特殊细胞类型的建模分析;(4)人类及非人灵长类单细胞发育图谱。

这些研究识别和描绘出了人脑细胞类型的惊人多样性,为认识人类精神和神经疾病的机制提供线索,也让我们对人类这一物种的身份有新的认识。接下来,追问将带领读者一起了解相关的研究成果。

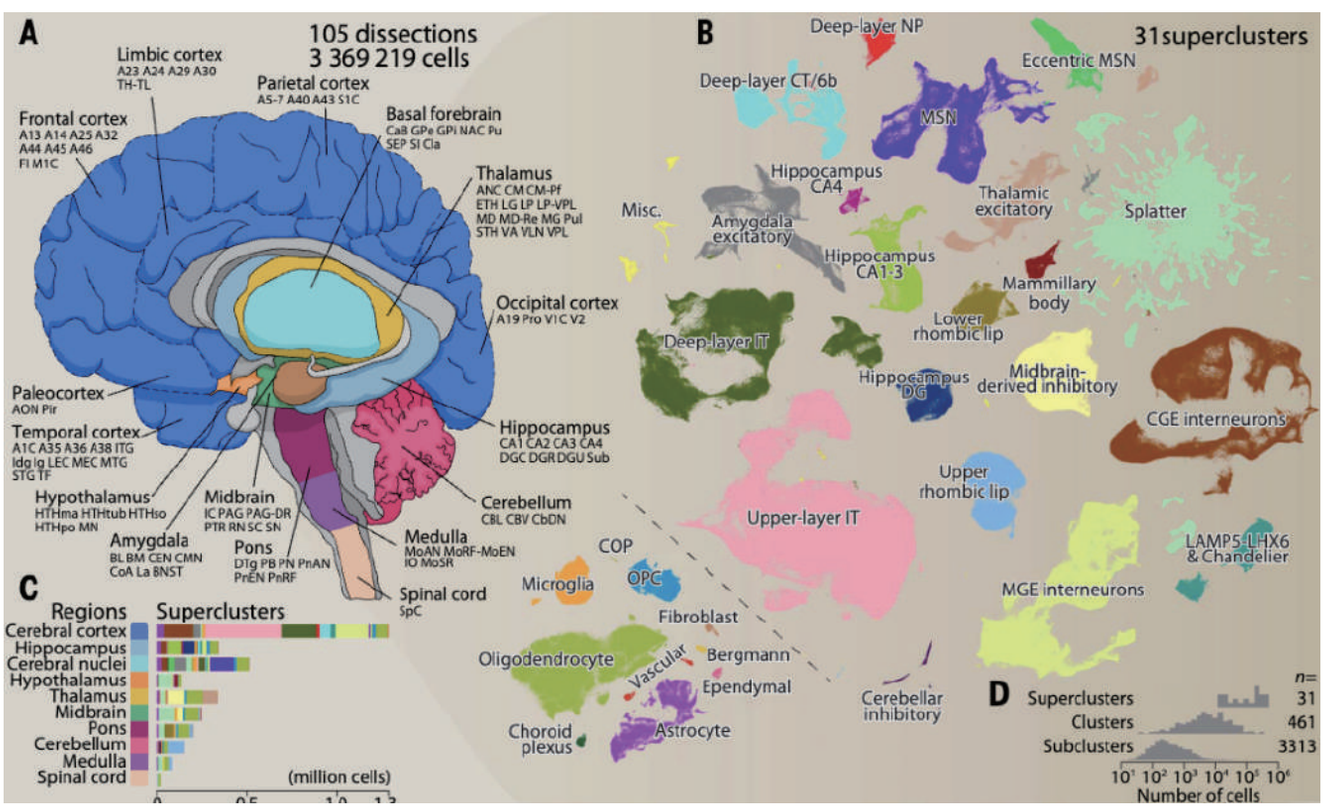


▷图 2:综合使用转录组学、蛋白质组学、表观遗传组学对人脑组织进行研究。图源:参考文献[1]

一、成体人类及非人灵长类动物单细胞图谱

第一个研究由瑞典卡罗林斯卡医学院Sten Linnarsson和荷兰乌得勒支大学医学中心的神经科学家Kimberly Siletti团队领衔[2]。他们从三位死者的前脑、中脑和后脑106个部位采集了300多万个单一细胞进行RNA测序(snRNA-seq),深入研究了脑细胞的多样性和亚型,为整个图谱的建立打下坚实的基础。

通过对测序数据展开分析,他们确认了461个脑细胞大类,包括3313个细胞亚群,其中约80%是神经元,其余是不同种类的神经胶质细胞(图3)。

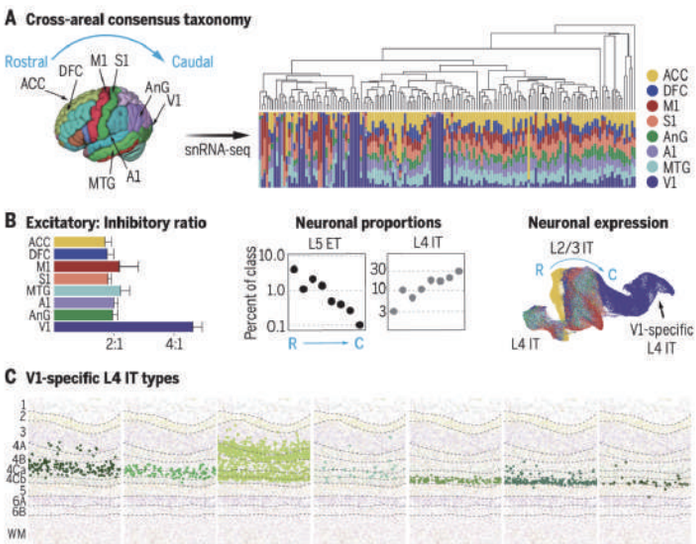


▷图 3:成人大脑细胞类型的区域特异性。图源:参考文献[2]

研究人员发现,除了大脑皮层外,神经元的种类非常多样化。尤其是在下丘脑、中脑和后脑,神经元的多样性非常明显,并且与它们不同的功能密切相关。值得注意的是,研究人员发现脑干是神经元多样性最丰富的脑区,并且其中一些细胞控制着先天行为,如疼痛反射、恐惧、攻击、性行为等。这些神经元的结构不像大脑皮层的神经元那样有明显的分层结构。很多神经元同属一个超级细胞簇,这个超级细胞簇包含不同类型的神经元,包括抑制性和兴奋性神经元。

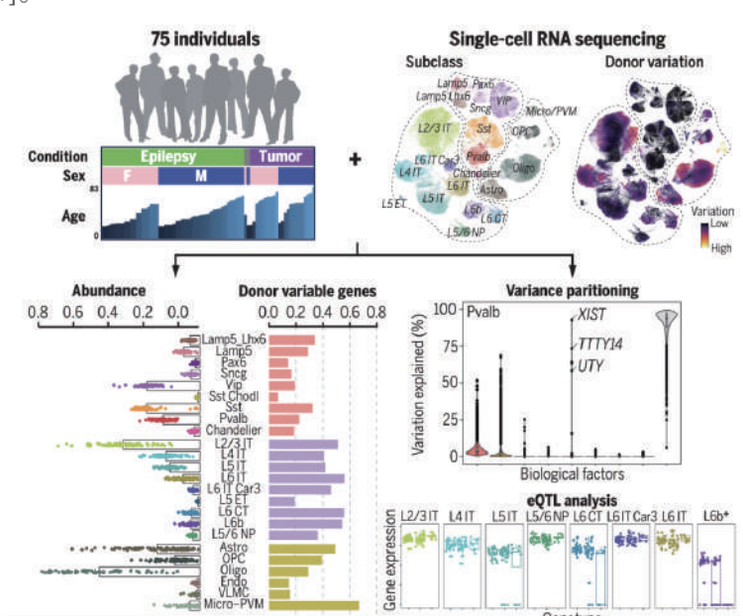
此外,研究人员还发现,虽然某些大脑区域具有不同的细胞类型,但许多区域的主要差异在于共享细胞类型组的相对比例。其中,连接大脑和脊髓的脑干有很多不同类型的神经元。来自美国西雅图艾伦脑科学研究所的Nikolas Jorstad等人的研究结果也得出了这一结论。他们对成人大脑皮层的八个区域进行了分析,结果表明大多数区域都包含相同的主要细胞类别,不同之处主要在于每种类型的比例以及非神经元细胞所在的皮层[3]。

有趣的是，初级视觉皮层（V（1）似乎是个例外：V1区拥有几种特异性的抑制性神经元类型，是人类和其他灵长类动物在这一区域特有的[3]（图4）。总体而言，这些数据支持区域功能多样化的进化策略，该策略在很大程度上不依赖于新细胞类型的产生，而是利用细胞类型内的微小变化及其相对分布的变化来构建不同的环路。



▷图 4：除V1外，大脑皮层所有区域的兴奋性神经元和抑制性神经元的比例相似。图源：参考文献[3]

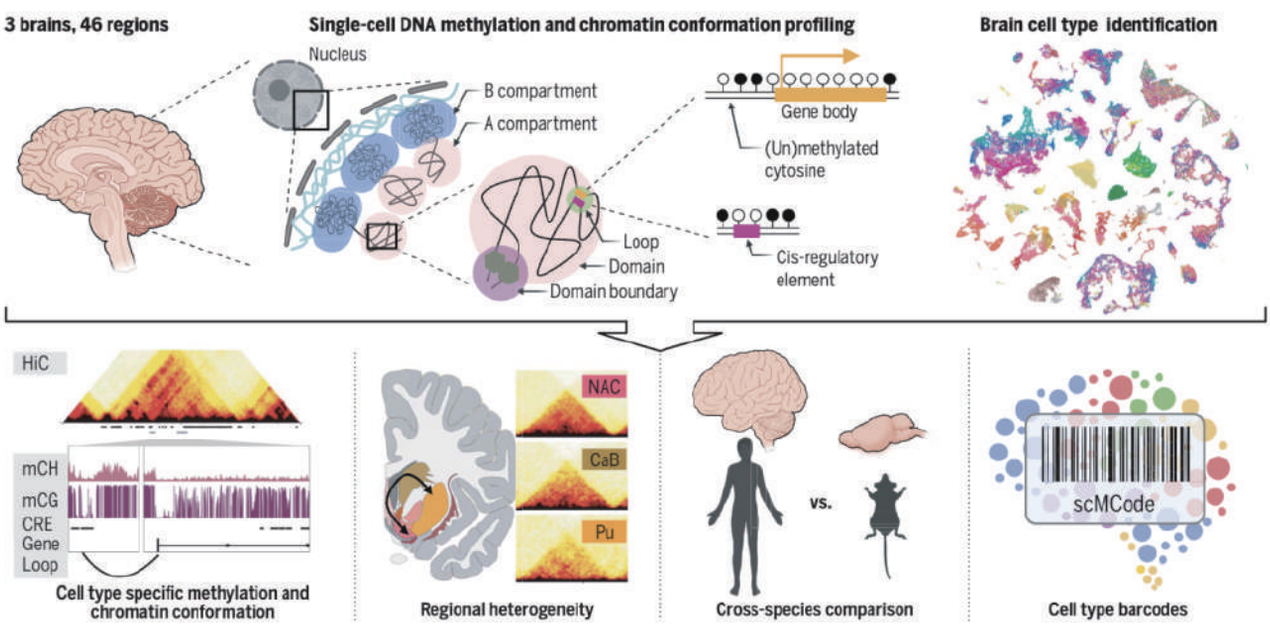
除了区域差异性外，脑细胞的基因表达还存在个体差异。其中一项研究首次使用单细胞技术比较了大量个体的脑组织后得出了这一结论。在这项工作中，75名因为难治性癫痫或脑部肿瘤而接受外科手术的成年志愿者向科学研究提供了珍贵的大脑皮层样本，研究人员对颞中回的脑细胞类型组成和细胞基因表达的变化展开了详细分析（图5）。结果显示，尽管不同个体之间存在高度一致的细胞构成，但也存在可以反映出个体特征、疾病状况和遗传调控变异的大量变异。这些结果也为未来研究健康和疾病状态下的细胞分型提供了参考[4]。



▷图 5：转录组学（snRNA-seq）和基因组学（WGS）分析揭示人类皮层细胞类型的个体差异。图源：参考文献[4]

来自美国索尔克生物研究所的分子生物学家Joseph　Ecker的团队，则从表观遗传学角度来剖析大脑。2020年，Ecker教授领导的研究团队根据DNA上的甲基化标记，分析了小鼠大脑中的160多种细胞，这种方法可以解析基因的“开关状态”。此次，他们使用同样的工具分析了三个成年男性捐献的大脑，对46个脑区的50多万个细胞展开DNA甲基化模式的分析，确定了近200种脑细胞类型。

与此同时，研究人员还对每个细胞基因组的三维结构进行分析，以便了解DNA序列中哪些片段可能正处于被积极读取的状态（图6）。研究发现，神经元和非神经元之间的染色质接触距离存在明显差异，这对于理解基因调控有很大的帮助。最终，研究团队构建了一个人类大脑的单细胞DNA甲基化和三维基因组结构图谱，阐明了大脑各部分细胞的特异性以及其多样化的表观遗传学结构[5]。



▷图 6：成体人脑表观遗传学图谱。图源：参考文献[5]

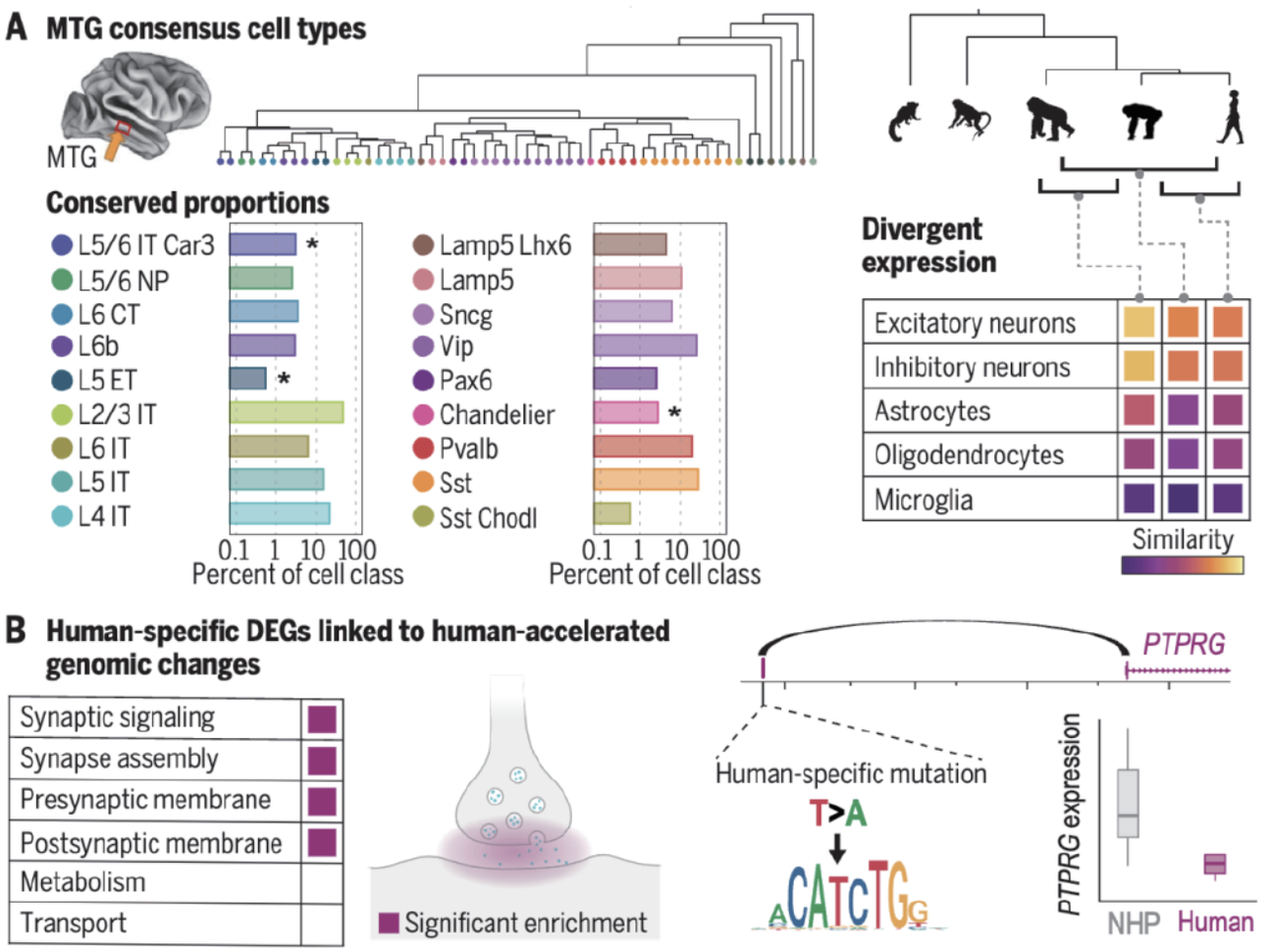
以上研究使用单细胞转录组和表观基因组分析表征了人脑与非人灵长类动物全脑的特征，揭示了大脑神经元类型惊人的多样性，为细胞类型多样性在人类大脑区域和物种特异性差异中的作用提供了新的视角。Ecker表示，“这些数据丰富的研究为在大脑中寻找特定疾病的位置提供了非常好的线索”。

二、成体人类与非人灵长类单细胞图谱比较研究

大脑细胞的哪些特征是人类与非人灵长类动物所特有的，也是BICCN项目想要回答的一个关键问题。

为此，Jorstad等人对成年人类、黑猩猩、大猩猩、猕猴和普通狨猴的颞中回（对语言理解至关重要的区域）进行了单细胞转录分析。他们发现，尽管灵长类动物在很大程度上具有相同的保守细胞类型，但它们在细胞比例方面表现出巨大差异。值得注意的是，尽管一般来说神经胶质细胞的多样性低于物种内的神经元，但与神经元相比，小胶质细胞、星形胶质细胞和少突胶质细胞在物种之间表现出更大的转录差异，并且在转录组中具有更快的进化趋异。只有几百个基因显示出人类特有的表达模式，并且这些基因不成比例地靠近具有人类进化选择迹象的基因组区域。这些结果表明，成年人类皮层的特定特性可能源自相对较少的细胞和分子变化[6]（图7）。

研究组还有一个有趣发现，尽管黑猩猩和人类有着更近的共同祖先，但黑猩猩的神经元特征与大猩猩更接近，而不是与人类更接近[6]。

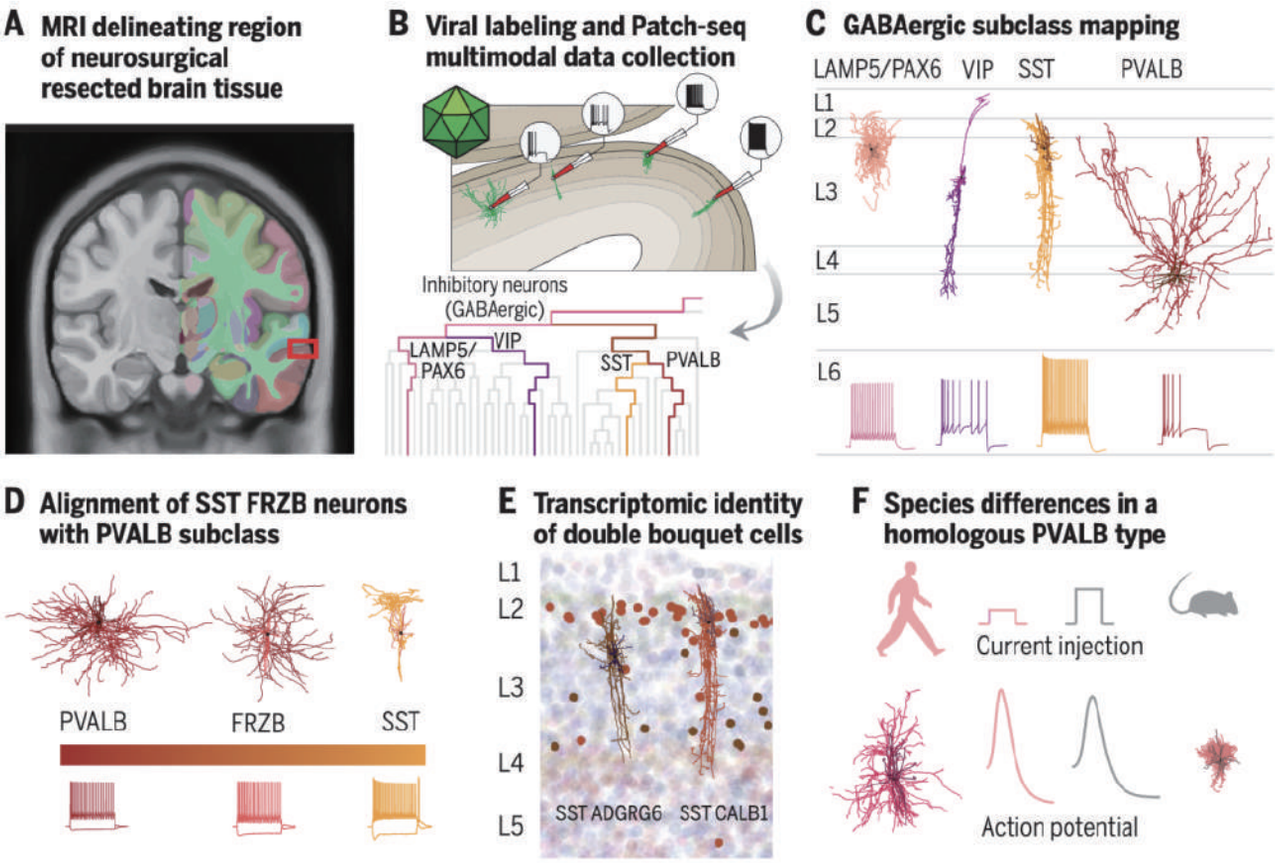


▷图 7：人类、黑猩猩的大脑皮层颞中回（MTG）显示出高度相似的细胞类型组成和分层结构。图源：参考文献[6]

另外，还有两项研究重点分析了在小鼠中没有、但在人类大脑中发现的细胞类型，例如根据外形命名的“双花束细胞”（double-bouquet cell）[7]、“玫瑰果神经元”（rosehip neuron）[8]。

艾伦脑科学研究所的Brian Lee等利用快速病毒基因标记和膜片钳电生理学结合RNA测序（Patch-seq），靶向神经外科手术切除的活组织，从中分析了这些神经元，包括它们的放电特性、复杂的形态学，以及开启的基因，证明了多模式Patch-seq数据对于转录组细胞类型分类学的细化至关重要。这项工作可以作为未来人类脑细胞类型功能研究的路线图，以解决新兴的转录组细胞类型分类学的问题，并提供丰富的开放数据集，用于探索各种人类新皮层GABA能神经元类型的基因功能关系[7]。

同样来自艾伦脑科学研究所的Thomas Chartrand等则利用单细胞转录组学来定义人类和小鼠大脑皮层L1中的神经元细胞类型，并定量识别了跨物种的同源亚种。他们观察到的跨物种差异表明，人类和小鼠对新皮层回路高阶输入的调节存在差异[8]。



▷图 8:利用新技术分析人脑皮层中多样化的GABA能神经元。图源:参考文献[7]

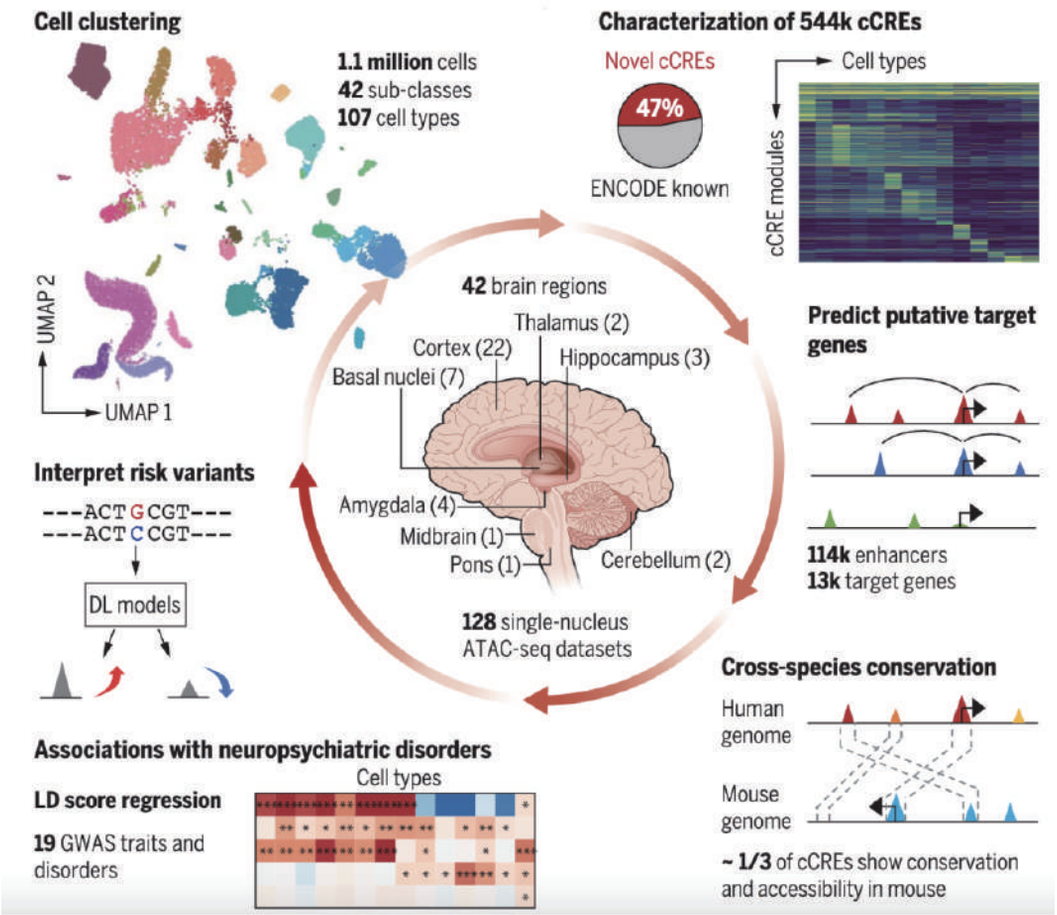
这些研究着重于对人体及非人灵长类动物大脑细胞组成进行单细胞比较分析。由此我们得知,一些将人类与灵长类动物区分开来的变化,导致了更高的认知能力,但也使我们面临更大的疾病风险,通过了解这些差异可以解开疾病风险的原因。

三、人类特殊细胞类型的建模分析

人类基因组中有数以万计的序列变异与神经精神疾病的病因相关,但它们主要位于编码区之外,因此解释这些已知风险变异的作用始终是一大挑战。这些非编码疾病风险变异可能通过扰乱转录调控元件来发挥作用,调控与神经精神疾病相关的细胞的基因表达,而转录调控与染色质的可及性密切相关。

为厘清其中的关系,美国加州大学圣地亚哥分校任兵课题组在单细胞水平上进行了人类大脑染色质可及性的综合分析,涉及42个不同脑区的110万个细胞。他们使用这个染色质图谱定义了107种不同的脑细胞类型,并揭示了这些细胞类型中54万个推定的转录调控元件的染色质可及性状态,以及特定脑细胞类型与精神分裂症、双向情感障碍、阿尔茨海默病、重度抑郁症等19种神经和精神疾病之间的关联。其中近三分之一的转录调控元件在小鼠脑细胞中表现出保守性和染色质可及性,强调了其功能重要性[9]。

此外,他们还开发了机器学习模型来预测疾病风险变异的调节功能。该图谱结合其他分子和解剖学数据,有望促进人们对脑功能和神经病理学的理解,最终为解决神经精神疾病提供更有效的方法。



▷图 9:对染色质可及性的单细胞测序分析。图源:参考文献[9]

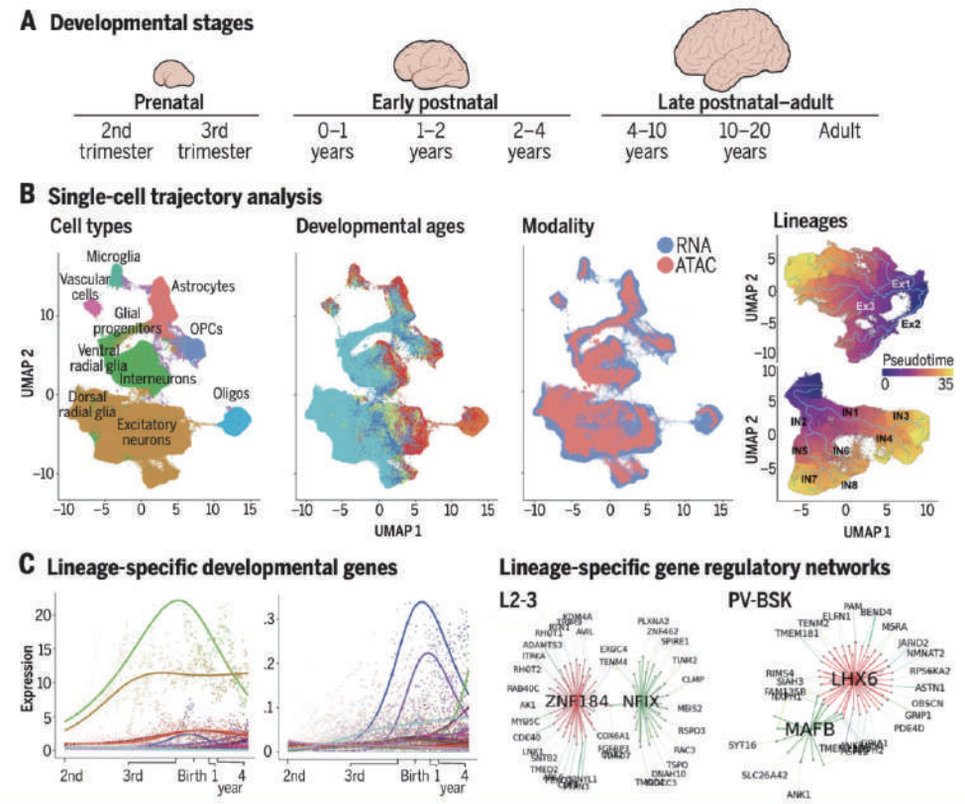
以上这组研究对人类神经元细胞的类型、功能进行了解剖分析,并解析了活体组织的细胞特性。大脑由许多类型的细胞组成复杂的环路,可以根据其分子、电生理和形态特征分为多种细胞类型,许多细胞类型在脑环路的计算中起着不同的作用。此外,脑部疾病并不平等地影响所有细胞类型,鉴定受每种疾病影响的主要细胞类型可以为有针对性地干预和治疗提供依据和启发。

四、人类及非人灵长类单细胞发育图谱

在这一系列论文中,有4篇重点关注了大脑的发育过程。

人的大脑皮层在胎儿出生前数月开始发育,并延续到出生后数月。该过程会受到精细调控,不同细胞的生长、分化和成熟,对大脑的复杂结构和功能至关重要。为了跟踪大脑皮层的发育轨迹,美国加利福尼亚大学旧金山分校Arnold R. Kriegstein团队,从不同阶段的人类皮层样本中收集了大量小核RNA测序(snRNA-seq)数据,对产前和产后发育阶段人类皮层样本中生成的单细胞数据进行了单细胞轨迹分析[10]。

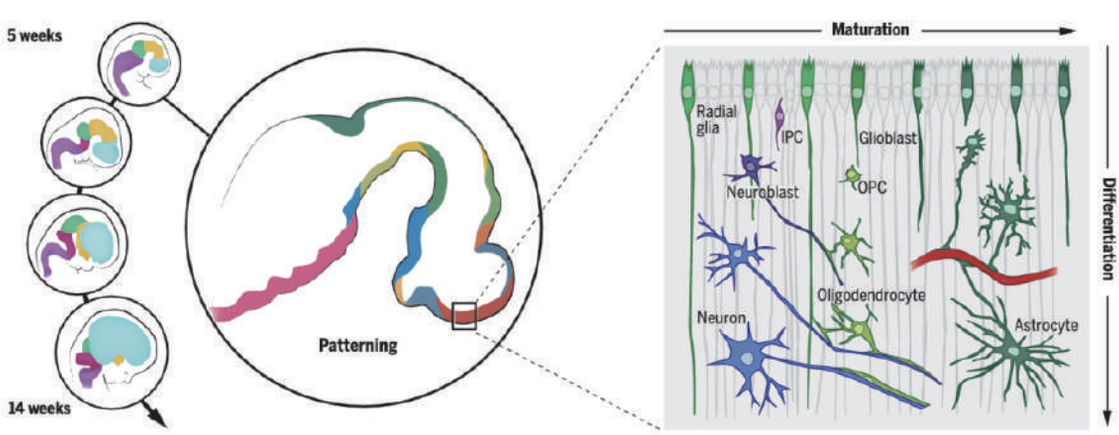
该研究阐明了人类皮层谱系发育背后的分子变化。通过整合单细胞转录组表达和染色质可及性图谱,他们绘制了一个全面的皮层谱系图谱,涵盖了产前和产后发育,确定了关键的转录网络,突出了性别特异性的发育变化。总之,这项研究成功揭示了正常皮层发育的谱系特异性机制以及性别二态基因表达在自闭症发病机制中的作用。



▷图 10:不同发育阶段的细胞类型和分子特征变化。图源:参考文献[10]

在此基础上,Emelie Braun等人对至关重要的人类发育前三个月的大脑进行了全面研究,绘制出了全面的皮层谱系转录组图谱[11]。

他们发现,尽管神经元是最多样化的,但前星形胶质细胞和少突胶质细胞前体细胞在各区域也都是不同的,它们的基因表达差异表明其具有区域特异性和细胞类型特异性的支持功能(图1(1)。这些发现强调了早期模式事件的重要性,并为确定影响特定脑细胞群体的人类疾病的治疗靶点提供了丰富的数据资源。此外,研究还确定了皮层发育时的关键转录网络,以及性别特异性的发育变化,对探索发育性大脑疾病、自闭症有重要意义。这也是科学家首次对大脑发育前三个月进行全面研究。

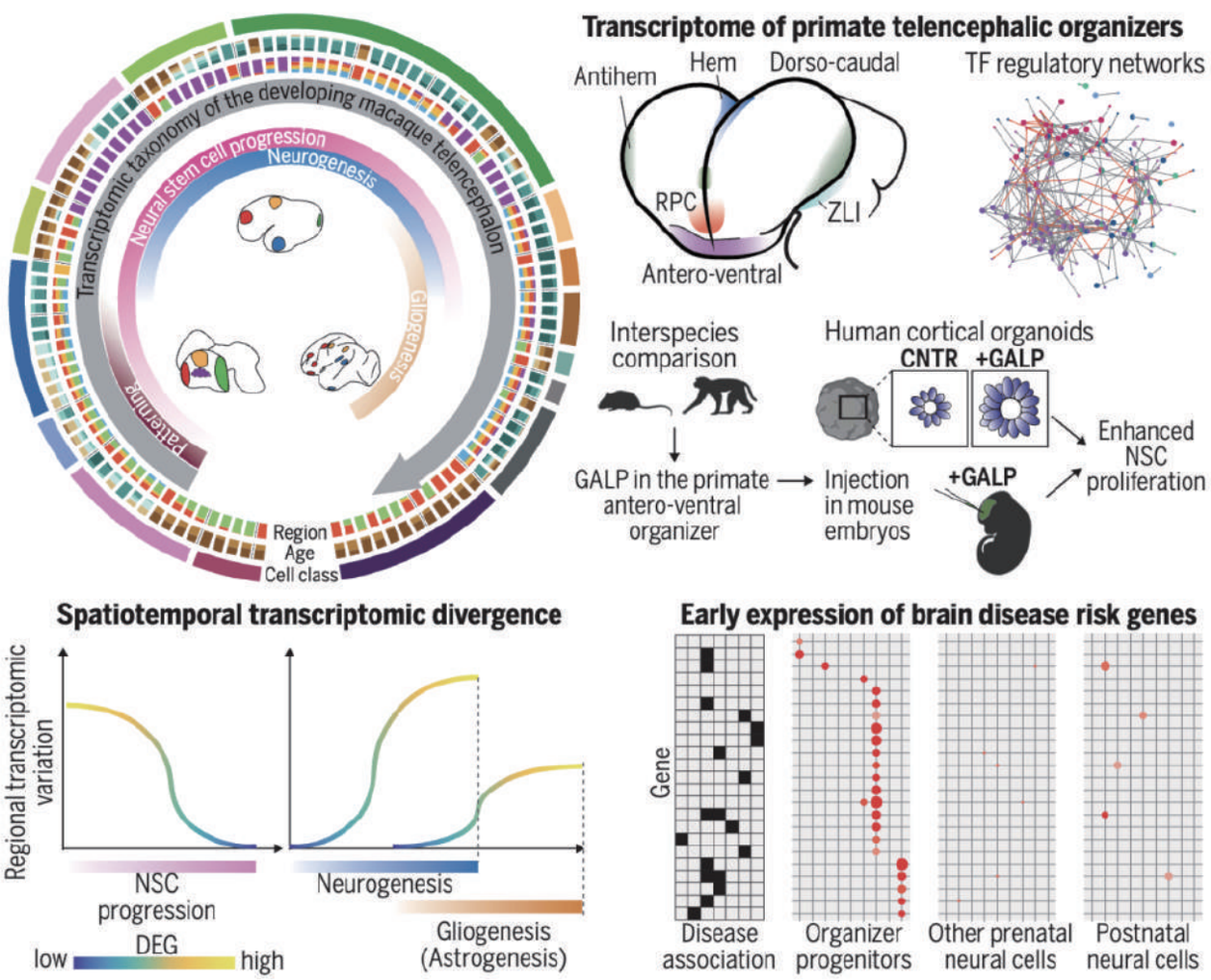


▷图 11:胚胎前三个月的人类大脑细胞图谱。图源:参考文献[11]

美国耶鲁大学医学院Nenad Sestan和Pasko Rakic的团队则借助恒河猴的样本，关注了整个端脑的发育过程，以及在此过程中调节神经细胞空间分布的分子过程[12]。

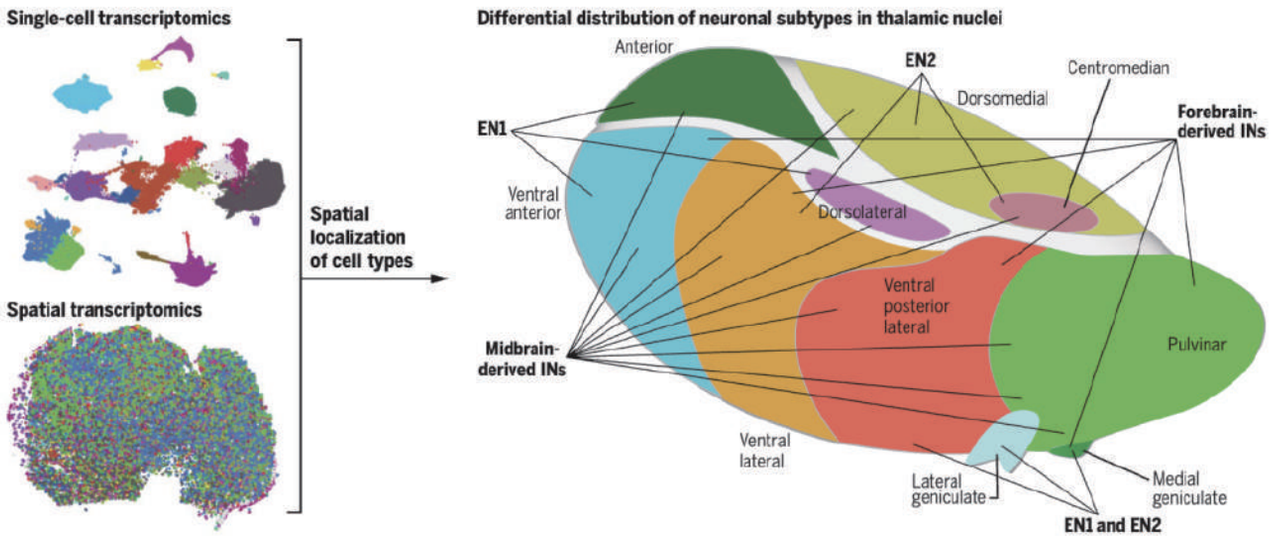
研究团队解剖了恒河猴产前端脑的多个区域，并对76.1万个细胞进行了单细胞转录组测序（scRNA-seq），他们区分了转录组定义的细胞亚型，包括背侧和腹侧神经干细胞、兴奋性和抑制性神经元、胶质细胞和非神经细胞，并描述了它们在谱系发育过程中跨区域的分子动力学特征。不仅确定了端脑组织中心的早期祖细胞，还预测了协调其模式功能的基因调控网络，包括ZIC转录因子。

这些数据资源为未来进一步探索灵长类动物和人类大脑的发育、进化和疾病铺平了道路。



▷图 12：恒河猴端脑发育图谱。图源：参考文献[12]

最后一篇论文关注的是丘脑的发育过程。丘脑是大脑与外界沟通的关键节点，并且哺乳动物的丘脑核团数量与投射模式相对保守。来自加州大学旧金山分校的Tomasz Nowakowski团队重点分析了人类丘脑发育时的细胞类型和空间组织等特征[13]。



▷图 13: 人类丘脑发育时的细胞类型和空间组织等特征。图源: 参考文献[13]

根据分析结果, 在妊娠早期, 丘脑会出现神经发生, 产生谷氨酸能和GABA能神经元; 到妊娠中期, 谷氨酸能神经元会分化成两种亚型, 而丘脑中GABA能神经元数量会显著增加, 并开始广泛分布。研究结果提示, 该GABA能神经元群体可能有助于人类进化。除此之外, 他们观察到神经胶质细胞的空间分辨模式, 以及随时间的发生与成熟过程, 比如星形胶质细胞的两种亚型, 一种会在丘脑富集, 另一种则会在丘脑附近的大脑区域富集。这些结论有助于帮助理解人类大脑进化过程, 尤其是提升对丘脑发育的认知。

除了上述发表在Science上的12篇文章, 还有8篇研究发表在Science Advances杂志, 由René Wilbers领导的工作探索了人类的快速脉冲中间神经元是如何在神经元到神经元的距离比老鼠大的情况下保持快速同步频率的。发表在Science Translational Medicine杂志的1篇研究中, Seth Ament及其同事聚焦于生命早期的炎症, 这是决定临床上几种神经系统疾病的危险因素。他们专注于小脑区域研究, 发现炎症主要与浦肯野神经元和高尔基神经元这两种抑制性神经元亚型的变化有关。

在对这一系列工作的总结中, NIH脑计划主任John Ngai博士评论道: “目前的一系列研究代表了一个里程碑式的成就, 为进一步阐明人类大脑在细胞水平上的复杂性架起重要桥梁。”

一百多年前, Santiago Ramón y Cajal绘制的神经元图像为我们理解大脑中细胞类型的多样性和复杂性开启了一扇大门, 由此神经科学开启了对大脑细胞类型和大脑功能的研究。如今, BICCN通过全球性的大规模合作, 对人类大脑的基本知识进行了全面的、高分辨率的描述。

BICCN这些研究的广度和深度证明了这种大型合作倡议在产生人类大脑基础知识方面的潜力。这一巨大进展将为了解人类大脑如何形成以及人类神经系统疾病的研究提供宝贵的资源, 同时也为理解人类大脑的构成和功能奠定了基础, 并为人类神经系统疾病病因的探究开辟了一个新时代。(编辑: Lixia)

参考文献

[1] Alyssa Weninger, Paola Arlotta, A family portrait of human brain cells. Science (2023). Doi: https://doi.org/10.1126/-science.adk4857

[2] Kimberly Siletti et al., Transcriptomic diversity of cell types across the adult human brain. Science (2023) Doi: https://doi.org/10.1126/science.add7046

[3] Nikolas L. Jorstad et al., Transcriptomic cytoarchitecture reveals principles of human neocortex organization. Science (2023) Doi: https://doi.org/10.1126/science.adf6812

[4] Nelson Johansen et al., Interindividual variation in human cortical cell type abundance and expression. Science (2023) Doi: https://doi.org/10.1126/science.adf2359

[5] Wei Tian et al., Single-cell DNA methylation and 3D genome architecture in the human brain. Science (2023) Doi: https://doi.org/10.1126/science.adf5357

[6] Nikolas L. Jorstad et al., Comparative transcriptomics reveals human-specific cortical features. Science (2023) Doi: https://doi.org/10.1126/science.ade9516

[7] Brian R. Lee et al., Signature morphoelectric properties of diverse GABAergic interneurons in the human neocortex. Science (2023) Doi: https://doi.org/10.1126/science.adf6484

[8] Thomas Chartrand et al., Morphoelectric and transcriptomic divergence of the layer 1 interneuron repertoire in human versus mouse neocortex. Science (2023) Doi: https://doi.org/10.1126/science.adf0805

[9] Yang Eric Li, et al., A comparative atlas of single-cell chromatin accessibility in the human brain. Science (2023) Doi: https://doi.org/10.1126/science.adf7044

[10] Dmitry Velmeshev et al., Single-cell analysis of prenatal and postnatal human cortical development. Science (2023) Doi: https://doi.org/10.1126/science.adf0834

[11] Emelie Braun et al., Comprehensive cell atlas of the first-trimester developing human brain. Science (2023) Doi: https://doi.org/10.1126/science.adf1226

[12] Nicola Micali et al., Molecular programs of regional specification and neural stem cell fate progression in macaque telencephalon. Science (2023) Doi: https://doi.org/10.1126/science.adf3786

[13] Chang N. Kim et al., Spatiotemporal molecular dynamics of the developing human thalamus. Science (2023) Doi: https://doi.org/10.1126/science.adf9941

“超声读心”:用意念控制方向,实现“未动先知”

作者:赵诗彤

中科院神经所硕博连读。在我看来,从数据中理解大脑正在做什么,并用模型对其进行解读是一件很有趣的事情。目前开展的课题集中于想象的神经机制。

扫码查看原文

脑机接口,作为当下神经科学领域最为关注的科技前沿话题,在23年二月被《自然-电子》杂志评为2023年年度技术。该杂志还强调,2023年是我们对脑机接口技术做出反思,并考虑其发展方向的关键时刻。23年七月,联合国教育、科学及文化组织(UNESCO)发布了一份关于神经技术的报告,而脑机接口正是这个领域的核心[1]。

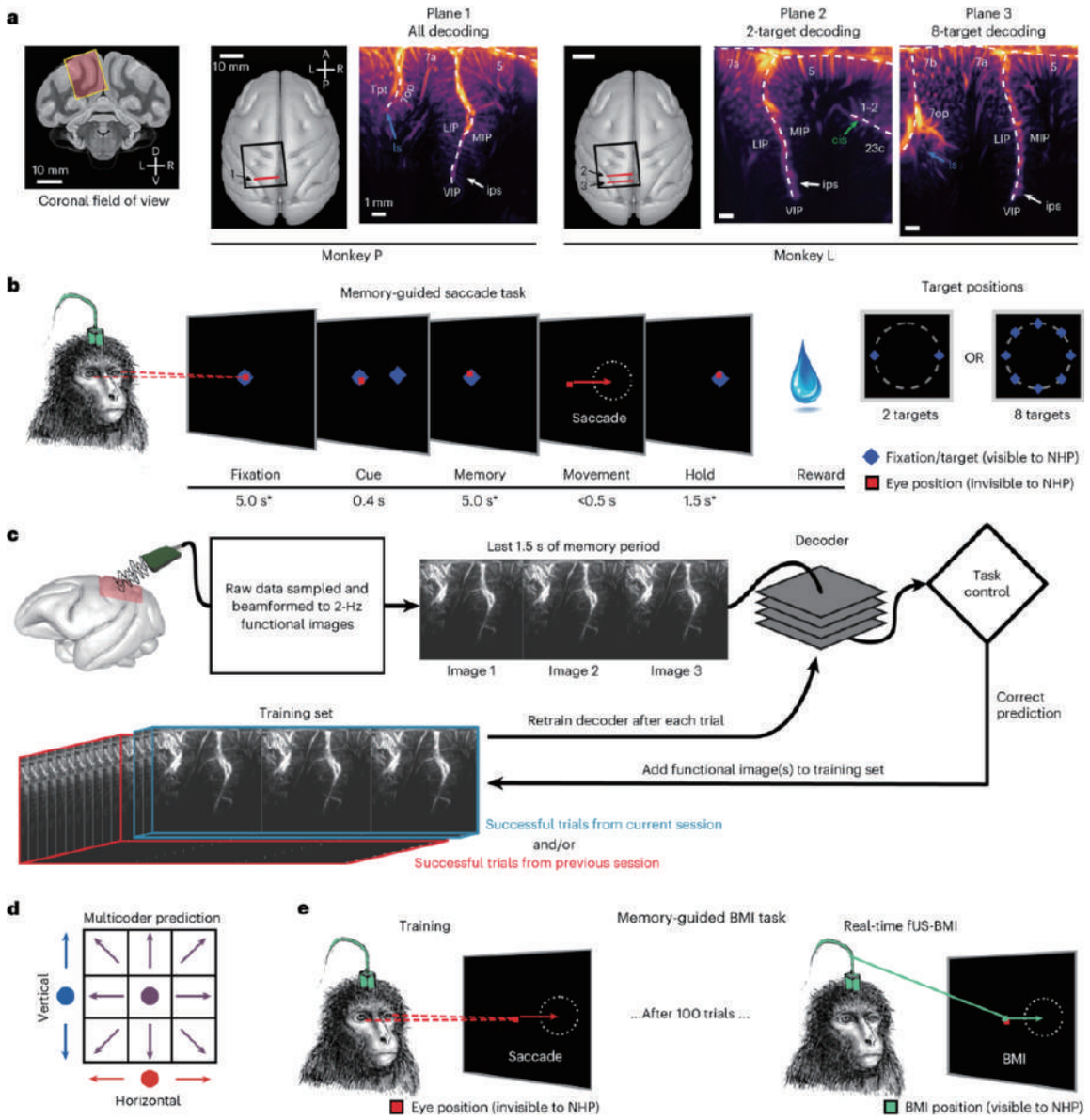
在获得高关注的同时,脑机接口技术本身也不断刷新着自身记录:23年10月在《自然-神经科学》上发布的一项研究表明,借助功能性超声(fUS)神经影像技术,可以做到“未动先知”,在不植入电极的前提下,在猴子没有做出任何行动时就可以解读出它们的运动计划[2]。研究者来自美国加州理工学院神经科学教授、天桥脑科学研究院(TCCI)脑机接口中心主任Richard Andersen团队。

简单来说,研究者通过fUS从两只恒河猴的后顶叶皮层获取它们执行眼睛和手部运动时的数据。经过训练,这些猴子能够使用脑机接口(BMI)实时控制最多八个运动方向。他们还开发了一种使用先前会话的数据进行预训练的脑机接口(BMI)方法。这样就能在随后的几天,甚至相隔数月的实验期间实现即时控制,而无需进行大量的重新校准。这项研究将人们的视野再次聚焦在这个带有浓浓科幻色彩的脑机技术上,从它身上,我们可以窥见“超声读心”技术背后的特点与突破。

为了测试fUS-BMI的可行性,研究者首先进行了在线闭环的两个运动方向上的解码。在猴子根据记忆引导完成左右眼动任务时记录实时的fUS数据。在得到100个成功试次的数据后,实验切换到闭环解码阶段:猴子只能通过运动的意图来控制任务方向,并每次都会被告知BMI预测出来的是左边还是右边。在第二个阶段中,解码器在55次训练试验后准确性显著提高,并在第114次试验达到82%准确性。

为了测试跨时段的重新校准问题,研究者通过预训练的方法在不同会话之间保持了fUS-BMI的稳定解码性能。他们比较仅使用当前会话数据和使用前一次会话数据进行预训练的fUS-BMI性能,发现预

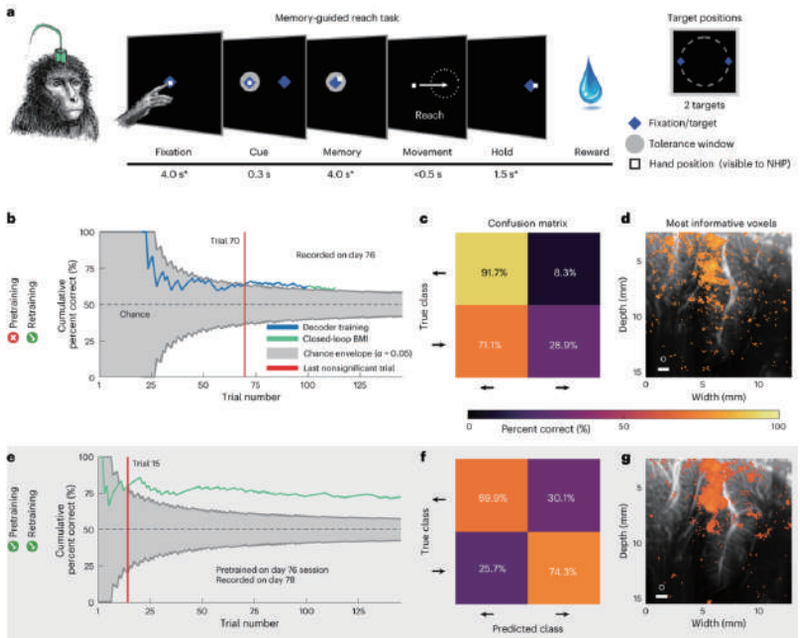
训练显著提高了性能且缩短了训练时间。



▷图:跨时段解码的预训练流程。图源:参考文献2

在解决了跨时区重新校准问题,保证了在线闭环解码的性能之后,为了进一步提升fUS-BMI的性能,研究者这一次尝试实时解码八个运动方向。首先分别预测运动意向的垂直(上、中、下)和水平(左、中、右)分量,然后将这些独立的预测组合成最终的预测(例如,向上和向右)。这样的设计与后顶叶皮层(PPC)神经元的响应特征是相符的:对相邻运动方向有相似的响应,但对具有更大角度分离的运动方向有不同的响应。换句话说,这种多编码器方法融入了相邻方向之间的神经表示相似性,而不是将八个方向视为八个独立的类别。解码器在86个训练试验后准确度达到显著(高于与12.5%的随机水平),并稳定在34-37%的准确度。这时,研究者再次测试了预训练的效果,再次发现预训练减少了达到显著解码所需的试验次数。

得益于fUS技术,研究者得以记录到多个功能脑区的活动,例如编码手和眼动的区域。为了验证这一点,他们将上述训练、解码过程在手动任务中重复了一遍,也获得了很好的效果。



▷图:眼动与手动实验流程。图源:参考文献2

一、“超声读心”的原理

人们研究大脑最直接的目的是理解、解读、最终自主控制大脑,脑机接口就为此提供了大脑与外界设备直接通讯的连接,可以记录、解码和刺激神经活动。我们都知道,神经信号在神经元上以电信号传播,最直接的测量方式是使用电极记录局部场电位(如ECoG)或者穿过头皮的电信号(如EEG),除此之外还有很多间接的方式来测量神经活动,如fMRI(功能性磁共振成像)、Ca2+(钙离子)等。此次研究所使用的超声波信号便是这样一种间接信号,它被用于大脑成像只有十余年的历史。

从成像原理上看,fUS与fMRI类似,依靠的是神经血管耦合效应。利用超声波在大脑小动脉和毛细血管内成像血流动力学参数(血容量),以约1赫兹的频率来间接推断神经活动。位置信息则是基于脉冲波多普勒效应,对超声波接收的反射声音,根据静止发射源和移动反射目标之间的距离推断反射目标的运动信息。同时,为了达到较深的成像距离,fUS必须使用极短的超声脉冲,这使得直接测量频率变化难以被实现,然而发送多个脉冲时,整合多个反射信号便可以间接计算出频率的短时间延迟,这便是脉冲波的多普勒效应。

在收集到成像数据后,所得到的神经元信息还需要经历一道关卡——滤波。血液细胞被移动的脑组织包围,这些脑组织也会产生多普勒信号,而分离血液和组织信号的过程至关重要。由于心脏、呼吸和行为活动引起的脑组织移动通常较慢,因此产生的多普勒信号的绝对频率较低于血液。使用高通滤波器可以消除这些组织信号,仅保留血液信号。更先进的滤波技术则可以优化这种分离,因此高效的滤波对清醒状态下的记录尤为重要[3]。

二、“超声读心”的前世今生

截至目前,大多数fUS研究都是在啮齿动物中进行的[4]。最初,研究者只能在固定头部的麻醉大鼠上做实验,得益于临床技术的进步,长期耐受性良好的颅窗使得在清醒的头部固定小鼠上采集fUS信号成为可能[5]。在啮齿类动物身上使用fUS信号主要是为了检测技术的可行性。直到最近,对固定头部清醒小鼠

的研究才开始着重于观测全脑成像与行为、环路操纵的关系[6]。与fUS固定成像同时发展起来的还有在不同物种动物身上应用的拓展，这证明了这项技术可以广泛应用于多个物种，只需要调整频率以适应不同的大脑体积。

fUS的一个明显优势是它能够跟踪自由移动的啮齿动物的全脑活动。从技术上讲，只需在动物头上植入颅窗和支架，在实验期间将超声波探头插入其中。当动物奔跑或执行行为任务时，就可以同时通过电缆连接到超声波扫描仪。自由移动成像技术的发展得益于小型线性超声探头的制造。这些微型探头有一个坚硬的电缆，一次只能成像一个成像平面。尽管如此，当啮齿类动物执行比头部固定的任务更自然的任务时，它们为观测大脑深处的活动提供了途径。在论证了该技术的可行性后[6]，大鼠自由移动的fUS主要用于系统层面的行为状态研究，如运动或睡眠，这些行为状态在无约束条件下才可以更自然地发生[7]。此外，将fUS成像与同时进行的电生理记录相结合，还可以使大规模血流动力学模式与特定区域的神经元活动相关联。

除了啮齿类动物外，fUS对于非人灵长类研究也十分具有吸引力，从技术上来说研究方法均与啮齿类动物应用类似，但为了适应非人灵长类动物（如猴子）头皮有褶皱的结构，研究者对深层脑区成像使用了更低的频率。例如近期有一项研究就利用fUS对清醒的非人灵长类动物深层视觉皮层进行精细研究[9]。此外，fUS还在猕猴身上被用于进行记忆引导的运动任务的成像，这被认为是用fUS解码灵长类动物大脑运动计划的首项工作。尽管由于血管信号的干扰，时间分辨率较低，但这为利用深层大脑信号控制外界机械设备作下了重要铺垫。自这篇文章发表之后，科学家便在临床与数据分析两方面持续突破。

23年，代表这项技术在信号处理与数据分析方面重大突破的便是此次来自Richard Andersen教授团队的研究[2]，而在十分相近的时间内，作者还发表了首例将fUS用于完全完整的成年人头骨中监测大脑活动的工作[10]。这一突破为研究者提供了在不侵入人体的情况下获取高灵敏度、大规模、高分辨率神经成像的新途径。通过这一技术，研究者首次实现了对手指运动的皮层响应的完全非侵入式映射和解码，生成了高分辨率（200微米）和大规模（50毫米x38毫米）的脑成像。这项研究将有助于深入理解脑活动，推动神经科学和脑机接口技术的发展，为神经疾病治疗和脑机接口应用提供创新性的方法。

三、“超声读心”与其他信号的比较

到此，我们已能够更好的理解fUS成像的特征，而与其他成像技术对比，就可以更直观的感受到不同技术收集信号性能的优劣之处。它们的差异主要体现在性能、侵入性、覆盖范围、时空分辨率、便携性和跨时段的解码稳定性上。

从表现上来说，目前最先进的使用膜下ECoG或侵入式电极的脑机接口可以以高准确度解码每分钟>15-60个单词、每分钟>29-90个字符以及个别手指的运动。

非侵入式的EEG是另一种常用于控制脑机接口的技术，而当代基于EEG的BMI有很大的个体差异性，一般来说使用EEG可以实现70-90%的准确率来对运动意图进行二分类。

再来看fUS，fUS记录到的信号既能实现广泛的覆盖范围，又能达到较高的空间分辨率（100微米）。15.6MHz的超声波探头可以提供12.8 mm × 20 mm这样大而深的视场，可以同时可靠地记录来自多个大脑皮层区域、广泛分布于多个脑区的运动信号。此外，许多脑机接口技术仅限于从大脑表面几毫米深的浅层皮层记录（例如ECoG、Utah阵列和钙成像）。但在本次研究中，研究人员采集到了解码眼动信息最有

信息量的位于中到深层的LIP脑区信号，大约在大脑皮层下5-9毫米，这是别的技术无法实现的。

另外，虽然目前fUS与EEG的性能指标相近，但相较EEG、fNIRS和fMRI，fUS更具发展潜力。先前的研究发现，用PPC脑区的fUS信号进行解码，准确率随着空间分辨率下降而迅速恶化。这意味着，对于其他宏观空间分辨率的脑机接口技术，将难以有效地解码在微观和中观神经群体中变化的信息。

此外，fUS还具有易于重定位和能够穿透软组织这些便携性的优点。侵入性电极阵列通常只能插入一次，并且常由于定位不佳或为了避免影响主血管而经常选择放置在次优位置。植入电极阵列还需要额外的手术从而难以重新定位。并且对这些技术来说，组织反应会降低硬脑膜下和颅内慢性电极的性能。而当使用fUS的时候，超声波探头可以在锁定位置之前进行多次定位、测试和重新定位，因此就更容易找到并记录特定感兴趣的区域。原则上，fUS可以无限期地通过硬脑膜操作，实现在很长时间内的慢性成像，并且信号质量几乎不受损耗。

可见，fUS虽然是一项新兴技术，却具有很大的潜力。不过我们同时也需要看到它的不足与受限之处。它受制于中观神经血管耦合（秒级），神经血管反应充当每个体素信号的低通滤波器，导致时间分辨率较低。在2赫兹的fUS下实时延迟约为800毫秒。这也是fUS无法实现实时成像的原因。在目前的技术条件下，只能通过先前获取的离线fUS数据进行滞后相关分析，近似得到100Hz fUS的成像结果。

但综合考虑时空分辨率，fUS依旧具有独特的优势，可以检测到脑内的神经血流动力学信号，例如监测神经精神疾病的生物标志物。在一些脑机接口的应用场景下，即使只有较慢的中观神经血流动力学反应，也可以提取出意图运动的轨迹有关的信号。

四、“超声读心”的未来发展方向

那么，这项研究的意义何在？简单来说，研究者们展示并验证了一种在线、闭环功能性超声脑机接口（fUS-BMI）的有效性，这代表了目前最前沿的结果。对此，他们主要提出了两大突破点。

（1）解码更多运动方向：基于以往解码两个眼动方向或者两个手动方向的结果，成功实现了实时解码八个运动方向。（2）实现跨阶段的解码稳定性：使用颅内电极的脑机接口（BMIs），例如犹他阵列，特别擅长在行为或刺激期间从空间定位的区域（<1厘米）感知快速变化（毫秒级）的神经活动，该活动与这些空间特定区域的活动相关，例如运动的M1区域和视觉的V1区域。然而，颅内电极在较长时间内（例如在后续记录会话之间）难以跟踪个别神经元，因此解码器通常每天重新训练。类似的神经群体识别问题也存在于超声设备中，包括在实验会话之间的视野偏移。在当前研究中，作者使用了一种对齐方法，该方法在一个月以上的时间内稳定了基于图像的脑机接口，即使需要重新训练，也可以在最小程度下从相同的神经血管群体解码。

而这是一项重要的发展，它使得研究者可以轻松地将前一天的模型与新一天的数据对齐，并允许在最小或没有新的训练数据的情况下开始解码。与传统基于流形的方法相比，作者采用的解码器对齐算法利用了fUS神经影像学提供的内在空间分辨率和视野，以直观、可重复和高性能的方式保证解码器稳定性。他们使用单个fUS帧（~500毫秒）生成了当前会话解剖的图像，并将前一会话的视野对齐到这个单一图像上。值得注意的是，这不需要任何额外的行为数据来进行对齐。

至于未来应如何提高超声成像在脑机接口领域的性能，作者也提出了几点。从采样区域上，作者建议更改放置头柱与探头的位置，使得可以对于更大范围的LIP脑区进行成像。其次，在成像技术上，过往主要

关注如何获得单层脑片的图像，但为了更全面的获得大脑活动信息，3D体积成像是一种可能的解决方案。然而目前，由于带宽、内存和计算能力的限制，高质量、低延迟的实时3D fUS成像尚不可能。然而，硬件和算法的不断进步很可能很快就会实现3D fUS-BMI。

最后，提高性能的另一途径可能是使用更先进的解码器模型来代替当前的线性解码器。卷积神经网络专门用于识别图像特征，并且对于fUS图像中常见的空间扰动（如与呼吸或心率相关的脑脉动）具有鲁棒性。此外，循环神经网络和Transformers模型使用“记忆”的过程，可能特别擅长表征fUS时间序列数据的时间结构。不过，这类人工神经网络（ANNs）的一个潜在缺点在于它们需要更多的训练数据。由此，作者也介绍了一种跨会话图像的对齐方法，允许将先前记录的数据汇总并组织成大型数据语料库。这样的数据语料库可能足以训练许多ANNs。除了训练ANNs所需的数据量外，最近的研究还强调了在为闭环运动BMI控制训练深度学习模型时所面临的额外挑战，特别是避免模型对先前记录数据中的时间结构过度拟合。尽管目前ANN还未被用于当前的脑机接口实验研究中，但它在未来研究更复杂的fUS-BMI有很大的潜力。

五、总结

当我们畅想大脑与机器的互动时，一定会好奇人类究竟能做到哪一步？通过回看fUS一步步的发展以及最近令人兴奋的研究成果，我们对这一新兴技术有了更全面的认识。也许在不久的将来，人类就能看到它得以更广泛地在人类身上得到应用，来解读人类更复杂的“想法”。那时候它不会是神奇的“天外来物”，而是凝聚了人类医学、工程学、数学等多领域智慧的结晶。（编辑：韵珂）

参考文献

[1]The year of brain–computer interfaces [J]. Nature Electronics, 2023, 6(9): 643-.

[2] GRIGGS W S, NORMAN S L, DEFFIEUX T, et al. Decoding motor plans using a closed-loop ultrasonic brain–machine interface [J]. Nature Neuroscience, 2023.

[3]MONTALDO G, URBAN A, MACÉ E. Functional Ultrasound Neuroimaging [J]. Annu Rev Neurosci, 2022, 45: 491-513.

[4]MACÉ E, MONTALDO G, COHEN I, et al. Functional ultrasound imaging of the brain [J]. Nature Methods, 2011, 8 (8): 662-4.

[5] MACÉ É, MONTALDO G, TRENHOLM S, et al. Whole-Brain Functional Ultrasound Imaging Reveals Brain Modules for Visuomotor Integration [J]. Neuron, 2018, 100(5): 1241-51.e7.

[6]FERRIER J, TIRAN E, DEFFIEUX T, et al. Functional imaging evidence for task-induced deactivation and disconnection of a major default mode network hub in the mouse brain [J]. Proc Natl Acad Sci U S A, 2020, 117(26): 15270-80.

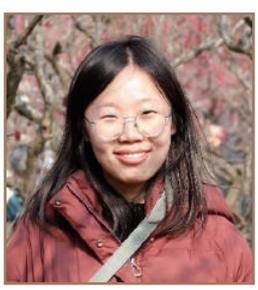
[7]BERGEL A, DEFFIEUX T, DEMENÉ C, et al. Local hippocampal fast gamma rhythms precede brain-wide hyperemic patterns during spontaneous rodent REM sleep [J]. Nature Communications, 2018, 9(1): 5364.

[8]SIEU L-A, BERGEL A, TIRAN E, et al. EEG and functional ultrasound imaging in mobile rats [J]. Nature Methods, 2015, 12(9): 831-4.

[9] BLAIZE K, ARCIZET F, GESNIK M, et al. Functional ultrasound imaging of deep visual cortex in awake nonhuman primates [J]. Proc Natl Acad Sci U S A, 2020, 117(25): 14453-63.

[10] CLAIRE R, SUMNER L N, WHITNEY S G, et al. A window to the brain: ultrasound imaging of human neural activity through a permanent acoustic window [J]. bioRxiv, 2023: 2023.06.14.544094.

光遗传——诺奖的种子选手



作者：轻盈

复旦大学博士生在读，计算&进化神经生物学方向。视科研和科普为人生的两大志业。想做有趣有意义的科学研究，也想把收获到的知识和乐趣分享给世人。

扫码查看原文

23年的诺奖已落下帷幕，mRNA疫苗技术的开创者们斩获2023年诺贝尔生理学或医学奖。对于很多人而言，有一项技术是诺奖“遗珠”。

光作为一种电磁波，影响着地球生物的方方面面，譬如光合作用、昼夜节律等生物学过程。这肉眼看来一团耀眼的、五彩缤纷的、却又似乎摸不着的光，神秘又迷人。光对于神经科学的发展，具有重要的影响。具体可以分为三个方面：其一，得益于绿色荧光蛋白GFP（获2008年诺贝尔化学奖）的应用，研究人员可以用荧光标记神经元。其二，将细胞膜上的电压敏感蛋白或细胞内钙敏感蛋白与荧光蛋白偶联，研究人员可以通过荧光的变化监测神经元的活动；而第三个方面，就是用光来操纵神经元的活动，即光遗传学（optogenetic）。

早在1979年，弗朗西斯·克里克（Francis　Crick，曾和詹姆斯·沃森共同解析DNA双螺旋结构，1962年诺贝尔生理学或医学奖获得者之一）就曾提出，神经科学目前面临的主要挑战是：在不影响其它细胞的情况下，控制大脑中的某一种细胞。而在光遗传技术出现之前，人们主要通过物理或者化学的方法控制细胞。可惜的是，现有电极无法精确定位特定的细胞群，而化学药物的作用速度又太慢。直到21世纪初，光遗传的火花照亮了整个神经科学领域。研究人员只需通过光照，控制一个小小的分子开关，“啪——”就可以快速开启或者关闭特定神经元的活动。

一、光遗传的英雄榜

（1）从细菌中发现的光敏元件

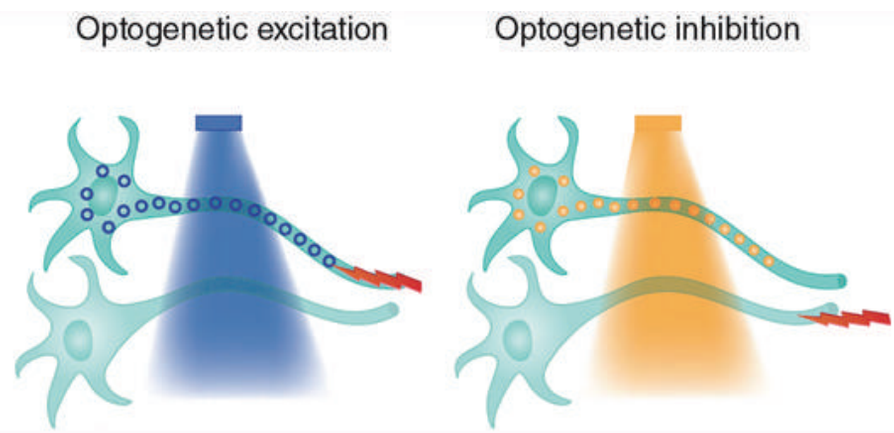
故事要从五十多年前说起。1969年，29岁的年轻人迪特尔·厄斯特黑尔特（Dieter Oesterhelt）博士毕业后来到美国加州大学旧金山分校进行博士后的研究。他要研究一种可在高盐浓度上生存的细菌——盐沼盐杆菌（Halobacterium salinarum）的紫色膜结构。而也正是这次经历，厄斯特黑尔特邂逅了其一生的

“学术缪斯”——细菌视紫红质(bacteriorhodopsin)。厄斯特黑尔特发现盐沼盐杆菌的细胞膜上的这种蛋白,可以作为一种光敏感离子泵,被光快速激活,驱动质子运输,产生化学能,从而使细菌在低氧的条件下仍得以靠光生存。

后来,厄斯特黑尔特的学生彼得·黑格曼(Peter Hegemann)继承了老师的衣钵。他不仅与厄斯特黑尔特一起分离了另一种氯离子特异性的光敏感离子泵,即盐视紫红质(halorhodopsin),还将对光感受器的探索延伸到了其它生物体系——莱茵衣藻(Chlamydomonas reinhardtii)中。基于以往的研究经验,黑格曼敏锐地猜测莱茵衣藻的趋光行为可能也和一种光敏蛋白有关。

功夫不负有心人,在长达数十年的探索后,黑格曼等人终于在21世纪初成功鉴定了通道视紫红质1(ChR(1)和通道视紫红质2(ChR(2)。两者感光之后,引起细胞去极化,产生光电流,最终触发衣藻鞭毛的趋光运动。

上文描述的光敏蛋白——细菌视紫红质、盐视紫红质和通道视紫红质,已在光遗传学中找到用途。天然存在的细菌视紫红质(该家族的第一个发现的成员,将质子泵出细胞)和盐视紫红质(将氯离子泵入细胞)通常在神经系统中具有抑制作用,因为这两种类型的超极化电流使神经元更难激发动作电位;相反,天然存在的通道视紫红质大部分允许带正电荷的离子自由流过视蛋白孔,因此倾向于使神经元去极化和兴奋。



▷图注:靶向激活(如蓝光激活的通道视紫红质)或抑制(如黄光激活的盐视紫红质),可以实现对神经元快速且细胞特异的甚至是投射特异的操纵。图源:参考文献Deisseroth, K. Optogenetics. Nat Methods 8, 26–29 (2011).

(2)三支力量的击鼓传花

光敏蛋白的发现也许为光遗传的诞生开了一个好头。但如何将光敏蛋白与神经元操纵结合,仍有待时代弄潮儿们的临门一脚。于是时间推进到20世纪末21世纪初,有三股力量对于光遗传的产生,起到了尤为重要的推动作用。

第一支是来自奥地利的神经科学家格罗·米森伯克(Gero Miesenböck)团队。米森伯克想到可以通过光敏蛋白来实现对神经元的操纵,遗憾的是一直没有找到比较高效的光敏蛋白的分子工具。直到2002年,米森伯克开创性的工作表明,在脊椎动物神经元中异源表达三个果蝇的光感受器基因,可以通过光刺激激活特定类型的神经元。

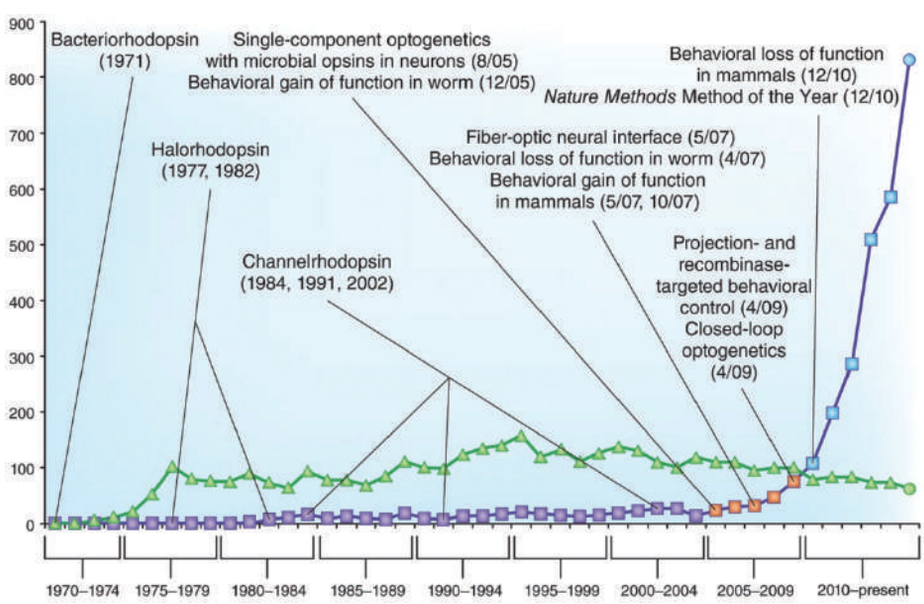
另一支便是彼得·黑格曼和他的合作者格奥尔格·纳格尔(Georg Nagel)团队。2002-2003年间,纳格尔和黑格曼先后鉴定了莱茵衣藻趋光行为中关键的通道视紫红质ChR1和ChR2蛋白,并在其它物种的

细胞中，创新性地异位表达这些蛋白，证明其异位表达仍保留光激活的离子通道的活性，并且仅需表达一个蛋白就可以实现。纳格尔和黑格曼的工作为光遗传技术提供了非常简单有效的分子工具。

最后一支则是爱德华·博伊登（Edward S. Boyden）和卡尔·迪赛罗斯（Karl Deisseroth）团队。他们则是将ChR与神经元操纵完美结合。2005年，博伊登和迪赛罗斯等报告了一项具有里程碑意义的研究工作：他们在海马神经元中仅表达单分子ChR2，就可以通过光以毫秒级的精度触发特定神经元动作电位，简洁又优雅地构建了光遗传的基本范式。（博伊登和迪赛罗斯团队后续还改良了这种工具，使研究人员可以快速且精确地沉默特定神经元。）

2006年，迪赛罗斯将这种工具正式命名为“光遗传学”。光遗传学很快被广泛应用于生物学各个领域。

近年来光遗传所获主要奖项：2018年，彼得·黑格曼、爱德华·博伊登、卡尔·迪赛罗斯获加拿大盖尔德纳奖国际奖（Canada Gairdner Award）。2019年，彼得·黑格曼、格罗·米森伯克，爱德华·博伊登和卡尔·迪赛罗斯获沃伦·阿尔珀特奖（Warren Alpert Foundation Prize）。2020年，彼得·黑格曼、奥尔格·纳格尔和格罗·米森伯克获邵逸夫生命科学与医学奖（The Shaw Prize）。2021年，迪特尔·厄斯特黑尔特、彼得·黑格曼和卡尔·迪赛罗斯获拉斯克基础医学研究奖（The Lasker Awards）。遗憾的是，迪特尔·厄斯特黑尔特先生已于2022年逝世。



▷图注：截止2015年统计，光敏蛋白和光遗传学的出版时间轴。近些年来，与光遗传有关的研究的数量急剧上升。图源：参考文献 Deisseroth, K. Nat Neurosci 18, 1213–1225 (2015).

二、光遗传方法详解

光遗传学能够以毫秒级的速度、细胞类型特异性的精确度，对诸如自由移动的哺乳动物这样复杂的生物系统进行光调控。这项神奇的技术，主要涉及三个核心环节：（a）光敏蛋白家族，这类蛋白可以在光刺激的条件下发生结构改变，触发跨细胞膜的离子流动；（b）将目标视紫红质蛋白靶向表达在特定脑区的特定细胞类型中；以及（c）时空精确性地在特定的脑区和细胞类型中施加光刺激。

而在一般的光遗传学实验中，主要有以下六个步骤：首先，研究人员需要构建一个同时包含ChR2基因

（或其它光敏蛋白）和控制其表达的遗传元件（如细胞类型特异性的启动子序列）的表达载体；随后，将表达载体包装到病毒中；接着，将病毒注射到动物的大脑中。虽然病毒会广泛地感染神经元，但视紫红质蛋白仅在具有激活其特定启动子所需机制的细胞亚群中表达。由此，实现了光敏感的视紫红质蛋白的细胞类型特异性的表达。表达的视紫红质蛋白定位到这些神经元的细胞膜表面；第四，在动物头骨埋入光纤；第五，沿着光纤，用特定波长的光触发这些视紫红质蛋白的活性；最后，对监测到的电生理和行为学的数据进行记录。

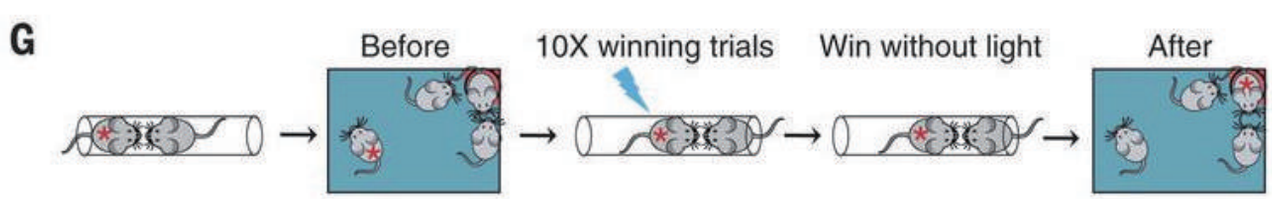
三、光遗传如何发挥作用？

2010年，光遗传技术获评为《自然-方法》（Nature Methods）杂志的“年度方法”，同年被《科学》（Science）杂志评为近十年来的突破之一。这项技术的应用范围可总结为两个方面：一个方面是，对于神经环路功能的基础研究；另一个方面则是对特定疾病的临床治疗。

2007年发表的第一篇在动物体内应用光遗传技术的文章，研究了分泌下视丘分泌素（hypocretin）的神经元的功能。研究人员通过光遗传的方法刺激侧丘脑区域下视丘分泌素分泌神经元的活性，明确该神经元参与小鼠睡眠-觉醒状态的转换。

光遗传技术还被应用于解释一些精神类疾病发病的生理机制。比如，光遗传学已经优雅地证明了伏隔核的胆碱能中间神经元在可卡因条件反射中的重要作用，中脑多巴胺神经元在慢性轻度应激诱导的抑郁样表型中的重要作用；或皮层-纹状体神经环路在强迫行为机制中的重要作用以及内侧前额叶皮层到基底内侧杏仁核的神经环路在焦虑中的作用。

值得一提的是，2017年，浙江大学胡海岚团队通过光遗传等技术，在小鼠中验证了“胜利者效应”。证明了从背内侧丘脑到前额叶皮层区域的神经环路介导了竞争中社会支配地位的长期改变，并且这些改变受到过往胜利经历的影响。



▷图注：光遗传让“败鼠”变强。光刺激丘脑-前额叶的神经环路，让原先竞争中占弱势地位的小鼠获得了战斗力。图源：参考文献doi: 10.1126/science.aak9726.

四、临床治疗——“让失明者复明”的光治疗

2021年，《自然-医学》上发表了一篇将光遗传学技术应用于治疗神经退行性眼疾——视网膜色素变性病的研究工作，首次实现了患者功能的部分恢复。实验中，研究人员在失明患者眼内注射携带光敏蛋白的病毒载体。随后对患者的眼睛施加光刺激，激活通过光遗传技术改造的视网膜神经节细胞。

经过上述治疗过程，患者的眼睛就能感知、定位、计数和触摸不同的物体。而在患者视觉感知过程中，多通道脑电图记录显示出与物体相关的活动位于视觉皮层之上，意味着患者视觉功能的部分恢复。

正如《自然-方法》中提到的那样，利用光来调节特定细胞中的特定活动，在生物学研究中取得了巨大成就，并照亮了尚未探索的科学道路。（编辑：Lixia）

用光束在脑中“绘制”电极，让纳米金颗粒标靶特定神经元

作者：孙子涵

中科院生物与化学交叉中心在读，喜欢遛弯。研究领域为神经退行性疾病。

扫码查看原文

多细胞生物体系，例如我们的大脑，拥有约860亿个神经元，超过100万亿个突触连接，结构高度复杂。要在这种生物体系中搭建具有细胞特异性的生物界面，想必极具挑战。

通过近几十年的努力，科学家们已然将生物电子设备尺寸缩小到了纳米级别，并不断提高制造工艺，但目前针对大脑的设备只能实现与几百个细胞的连接，且不具备细胞特异性。

此前，有研究者开发出了一种方法，在完整生物体内编辑活细胞，使其原位形成特定形式和功能的人造结构。也有研究表明，导电聚合物可以通过电化学反应在活体细胞上合成，或在具有天然氧化环境、氧化酶的生物组织中合成。但这些方法都没法实现细胞特异性这一关键目标。

发现这个问题后，鲍哲南教授、卡尔·迪赛罗斯（Karl Deisseroth）教授团队迈出了关键性的第一步。他们开发了一种基因靶向化学组装方法（GTCA），利用细胞特异的基因表达谱，在特定类型神经元细胞外膜表面表达抗坏血酸过氧化物酶（APEX（2）或辣根过氧化物酶（HRP）作为催化剂，引发过氧化氢（H2O（2）介导的具有导电特性或绝缘特性的聚合物在活神经元原位合成。

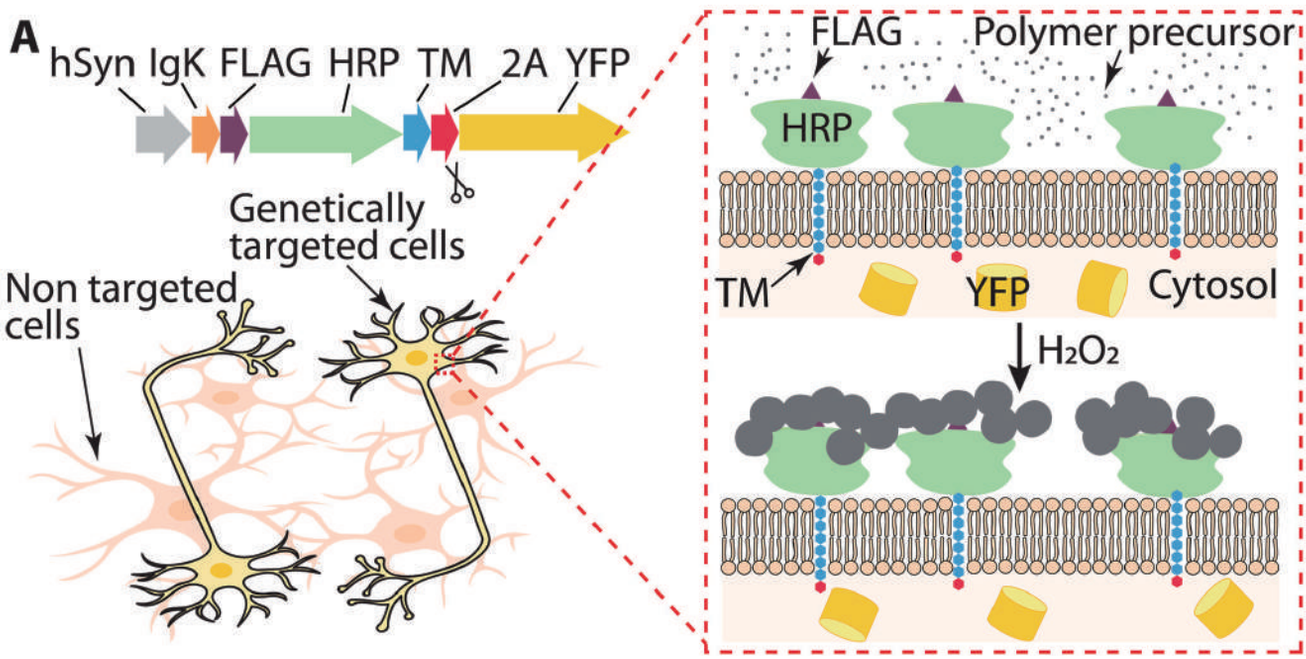


图1:在特定类型活神经元细胞外膜表面原位合成聚合物策略

尽管这种方法取得了初步成功,但主要问题在于效率不高。大部分催化酶游弋在细胞质中,而不是大量表达在神经元细胞外膜表面。研究团队面临的巨大挑战是如何将酶安置在正确的位置,即神经元细胞外膜表面。

*编者注:利用细胞外空间作为反应中心主要有以下几点原因:(1)较大分子的反应前体或材料难以透过活细胞的完整细胞膜,因此,定位于细胞外膜的酶不足,可能导致化学反应多聚物产量低;(2)在细胞外空间增加催化反应的酶,可以降低引发反应所需条件的阈值,例如过氧化氢(H2O(2)浓度,提高生物相容性;(3)在细胞内发生聚合反应可能具有细胞毒性,诱导细胞凋亡。

为了解决需要催化酶在细胞外膜最大量定位的难题,研究组成员张安琪想到,何不利用“膜蛋白从细胞质中产生,被有效分类和包装后,最终运送到细胞膜”这一大家熟知的天然细胞机制,就像是给归巢的信鸽腿上拴上信纸那样,寻找一种可以有效搭载催化酶定位于细胞外膜的蛋白质。

耗时一年半,筛选了超过50种组合方式后,她尝试了一种存在于免疫系统T细胞表面的蛋白CD2。结果带来了意想不到的惊喜:CD2既有效又耐受,不仅成功引导催化酶大量定位于细胞外膜,还能成功引导其他几种她正在编辑的蛋白。

基于此,研究组开发了新一代基因靶向化学组装方法(next-generation GTCA),利用CD2蛋白的细胞膜锚定功能,联合基因工程,在特定类型的神经元细胞外膜表达催化酶,介导神经元外膜表面多聚物的原位合成。该研究成果于23年8月发表在知名杂志Science Advances上。

针对新一代基因靶向化学组装方法,张安琪首先验证了这套方法中催化酶的膜定位和氧化反应活性:通过免疫荧光染色,使用去垢剂通透细胞,染色定位于细胞膜和滞留在细胞中的酶总和;或不使用去垢剂通透细胞,仅染色定位于细胞膜上的酶,比较两种情况下的荧光强度,证实CD2蛋白能够强有力地引导催化酶锚定到神经元膜表面,并且发现HRP的表达量整体大于APEX2。利用荧光和明场显微成像,证实了多聚反应发生后,仅HRP表达的神经元膜表面发生多聚物的原位合成。

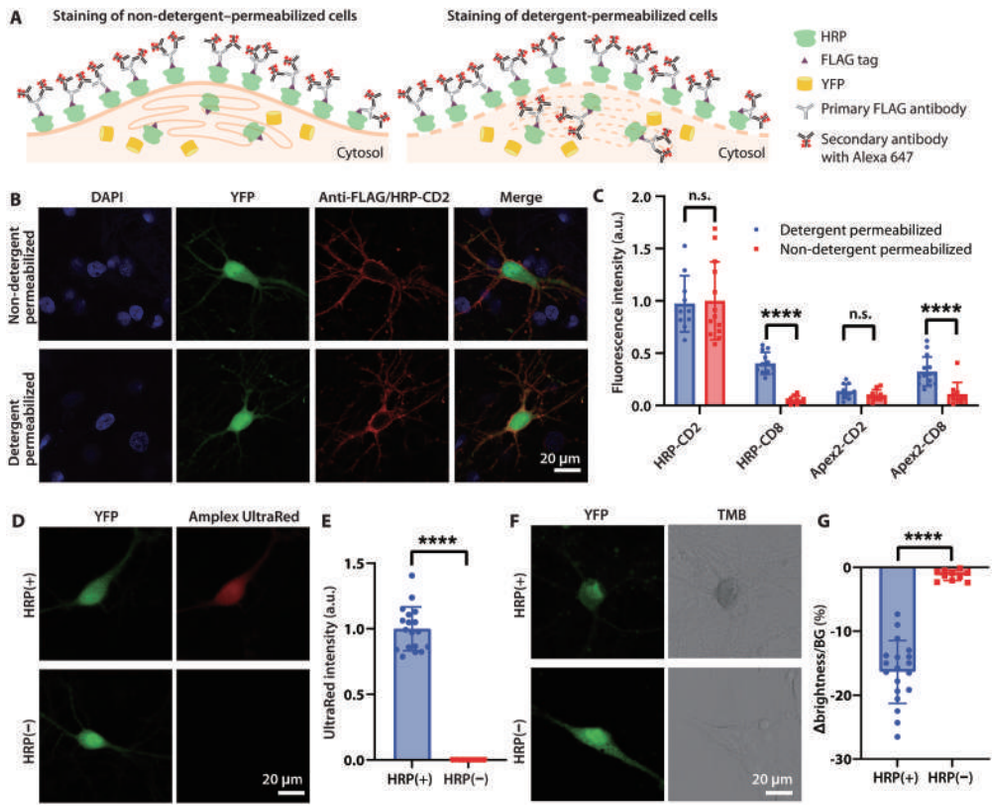


图3:验证催化酶的神经元细胞外膜定位和氧化反应活性 图源:论文

除了验证新方法的功能性和效率,其生物相容性也需要特别关注,即反应发生后,神经元活性不能受影响。于是,研究者利用溴乙啡锭二聚体(EthD-(1),一种只染色死亡细胞细胞核的红色荧光染料,指征聚合反应发生后的死亡细胞,发现HRP阳性的细胞没有和死亡细胞染料共定位,表明这种原位合成聚合物的方法没有细胞毒性,具有良好的生物相容性。

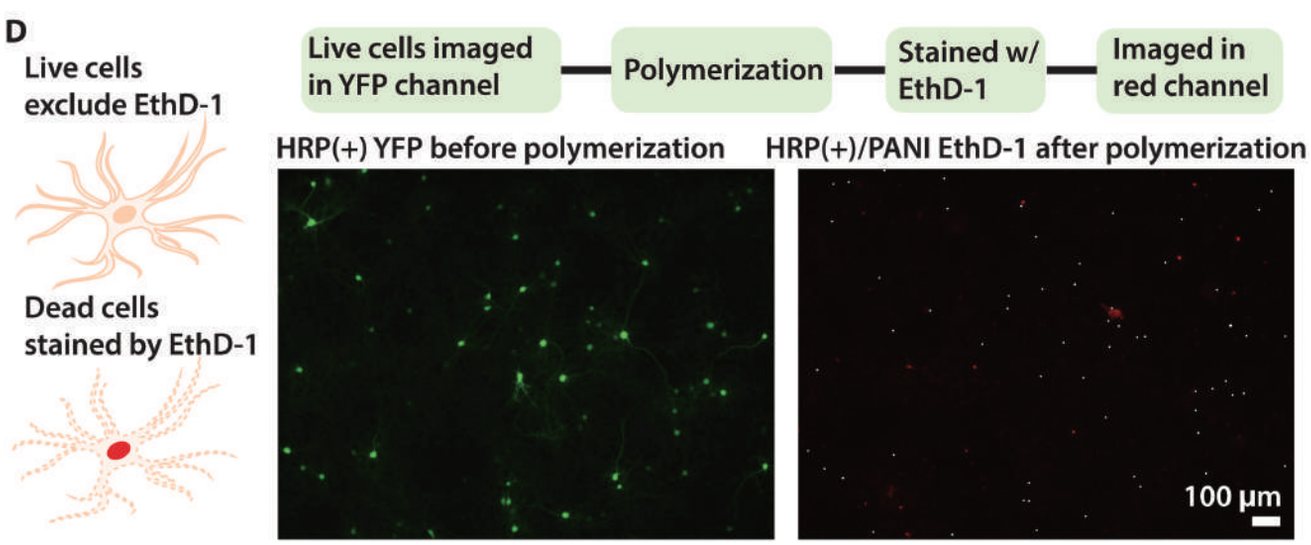


图4:聚合反应发生后,HRP阳性细胞不与死细胞共定位 图源:论文

有了CD2蛋白这个能够锚定细胞膜的主力军，基因靶向化学组装的方法就能扩展出许多新的玩法。比如，将过氧化物酶换成一种光敏感的催化酶，在基因工程范式下，可以实现一种独特的功能：当有光刺激时，表达这种光敏催化酶的特定类型神经元就可以在细胞外膜表面合成导电或绝缘聚合物，就像是用光束在大脑中“绘制”电极一样。

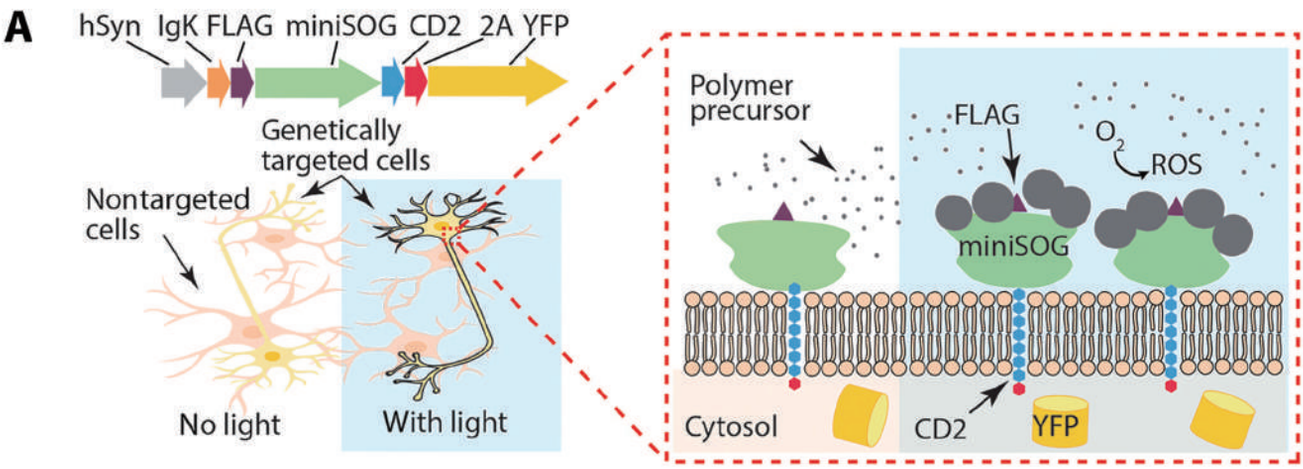


图5：感光催化酶介导的基因靶向化学组装方法 图源：论文

另外，研究者还尝试用这套系统在神经元表面富集事先合成的材料，例如纳米金颗粒。将纳米金颗粒与生物素连接，在特定神经元表达链霉亲和素，CD2蛋白引导链霉亲和素在神经元细胞外膜表面定位，与细胞外空间的生物素结合，使得纳米金颗粒富集在特定神经元。

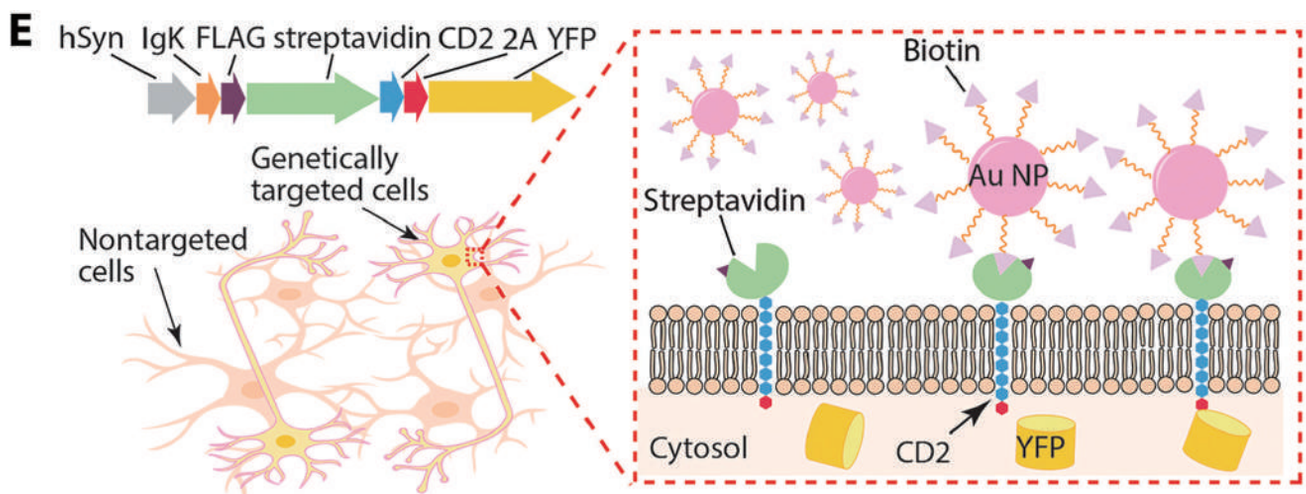


图6：链霉亲和素与生物素结合介导预制金颗粒在特定神经元表面富集

新一代基因靶向化学组装方法无疑让研究人员非常兴奋。未来，CD2蛋白可能用于构建微创型脑机接口，同时，如果CD2蛋白能够把各种编辑后的蛋白质引导到细胞膜上，那么对其他科学家的研究也存在巨大的科研价值。这套方法实质上是在细胞膜上搭建生物平台，为各类材料的创建提供基础。

值得一提的是，本研究的主要负责人是张安琪博士，她在10年前被誉为“门萨女神”“哈佛全奖学者”。如今，她加入斯坦福大学化学工程教授鲍哲南的实验室，并师从生物工程及精神病学和行为科学教授卡尔·迪赛罗斯（Karl Deisseroth），潜心投入脑科学科研工作，为脑机接口技术的推动做出贡献。针对这项研究的报道也获得由斯坦福大学可穿戴电子设备倡议（eWEAR）发起，由eWEAR行业联盟计划成员盛大集团及天桥脑科学研究院（TCCI®）资助落地的eWEAR-TCCI科学写作奖。（编辑：韵珂）

 # 关于研究者

张安琪是斯坦福大学鲍哲南实验室和迪赛罗斯实验室的博士后研究员以及本论文的第一作者。斯坦福大学的其他合著者包括：化学博士生 Kang Yong Loh；迪赛罗斯实验室的研究助理兼实验室经理 Chandan S. Kadur 和 Charu Ramakrishnan；化学工程博士后 Lukas Michalek；生物工程博士后 Jiayi Dou；鲍振南是 K.K. Lee 化学工程教授，同时也是斯坦福大学 Bio-X、斯坦福心血管研究所、妇幼保健研究所 (MCHRI)、普雷考特能源研究中心、Sarafan ChEM-H、斯坦福森林环境研究所、吴蔡人类表现联盟、吴蔡神经科学研究院的成员及扎克伯格生物中心的研究员；卡尔·迪赛罗斯是生物工程系和精神病学与行为科学系的D.H. Chen教授，霍华德·休斯医学研究所的研究员，也是斯坦福 Bio-X和吴蔡神经科学研究院的成员。

 # 关于eWEAR-TCCI科学写作奖：

eWEAR-TCCI科学写作奖是由斯坦福大学可穿戴电子设备倡议（eWEAR）发起，由eWEAR行业联盟计划成员盛大集团及天桥脑科学研究院（TCCI®）资助落地的项目。

参考文献：

关联论文：Anqi Zhang et al.,Genetically targeted chemical assembly of polymers specifically localized extracellularly to surface membranes of living neurons.Sci. Adv.9,eadi1870(2023).DOI:10.1126/sciadv.adi1870

纳米技术助力探索大脑中的“星辰大海”

讲者：段小洁

北京大学未来技术学院研究员，致力于新型植入式神经电极阵列的研制与脑机接口应用研究。

扫码查看原文

人们的大脑里有着数百亿个神经元，它们宛如浩瀚宇宙中的群星，以极其纷繁复杂的形式相互连接，构成了一个复杂而精妙的仪器，控制着人们每一瞬间的思绪与活动。受限于记录神经活动的技术工具的尺寸与精细程度，过往对于大脑不同层级的研究以及对于脑部疾病的精准干预都存在“瓶颈”，而纳米材料与技术的出现和发展成为了改变这种困境的有力工具。纳米材料的小尺度效应，使其在磁学、电学、力学等方面呈现出与传统材料迥然不同的性能，依此设计的纳米探针、纳米电极等也使大规模记录神经元的活动不再遥不可及。

2023年8月18日，天桥脑科学研究院（Tianqiao and Chrissy Chen Institute，TCCI）和哥伦比亚大学神经技术中心（NTC）、多诺斯蒂亚国际物理中心（DIPC）联合举办了在线学术会议NanoNeuro 2023。本次会议上，来自北京大学的段小洁研究员做了题为“纳米技术实现的全脑神经接口（Nano-enabled brain-wide neural interfacing）”的精彩报告，与大家分享了利用纳米材料与技术实现的微创大规模生物电位记录和磁共振兼容的深部脑刺激（DBS）两方面的最新研究成果。

一、高信号质量与微创，能否兼得？

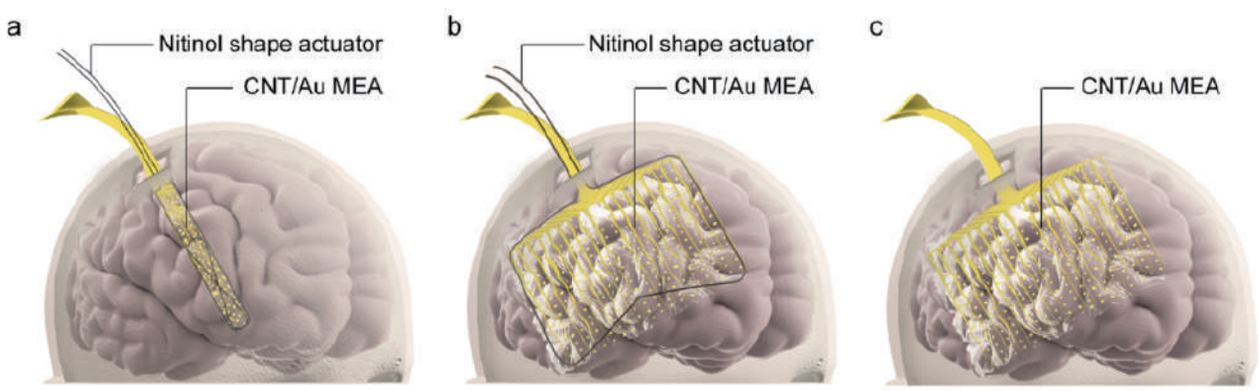
人类行为并非单一脑区活动的结果，而是基于全脑多个脑区和网络的协同参与。以视觉认知为例，位于枕叶的视觉皮层除了接收来自外部的视觉输入，还会接收负责物体识别、运动检测、视觉注意等功能的多个脑区的信号输入，这些脑区共同作用从而形成对外部视觉信息的最终感知。因此，想要全面理解人类行为，全脑神经活动的记录至关重要。

目前常用的记录全脑活动的神经成像技术主要包括脑电图（EEG）、脑磁图（MEG）、功能性磁共振成像（fMRI）和近红外脑功能成像（fNIR）等非侵入性技术，但这些技术记录的信号都有一定的局限性。例如，fMRI和fNIR并非记录直接的大脑活动，而是以血氧水平依赖信号（BOLD signal）作为神经活动的

间接测量，并且其时间分辨率只能达到秒级别；EEG和MEG虽然具有毫秒级别的时间分辨率，其信号的空间分辨率和带宽却十分有限。

与非侵入性技术不同，皮层脑电图（ECoG）是一种将电极阵列直接放置在大脑皮层表面来记录大脑信号的技术，需要在颅骨上开一个切口暴露大脑表面进而将电极植入。ECoG具有较大的皮层覆盖范围，记录到的信号具有毫秒级别的时间分辨率和毫米级别的空间分辨率，以及上至500 Hz的高带宽，在信号数据质量方面表现出了显著的优越性。然而，植入电极前的开颅手术和硬脑膜切开术可能给患者带来包括颅内感染、炎症反应、脑水肿或脑血肿等多种风险，而电极的长时间植入也可能损害植入区域附近的血管连接与大脑结构，甚至引起患者行为功能的变化。因此，ECoG的侵入性和手术风险限制了它在临床与科研中的应用范围。

为了解决这些问题，段小洁研究员团队开发了一种新型超薄、可形变的柔性电极阵列（shape-changing electrode array），简称SCEA，使得以最小创伤实现大面积的ECoG信号记录成为可能。SCEA使用可伸展的碳纳米管和金作为导电层，通过与一个形态调节装置相结合，在颅外实现将电极阵列从厘米级别大小的薄片压缩成只有几毫米宽的条状，这样具有开创性的巧妙设计使电极在颅骨仅开一个小孔的情况下便可插到皮层表面。随后，在形态调节装置的辅助下，SCEA条带完全展开，形状变回超薄片状，从而在皮层表面形成大面积共形界面，进一步实现硬膜外或硬膜下的大脑活动记录（图1）。



▷图1：SCEA的植入过程。图源：参考文献1

SCEA的微创性在小型动物大鼠和大型动物比格犬中都得到了验证。段小洁研究员团队成功通过颅骨中仅2mm×0.8mm的开口将SCEA植入到大鼠右侧大脑半球硬膜外，并通过仅有6mm长的硬膜裂缝将厘米级别大小的SCEA植入比格犬的大脑硬膜下。更重要的是，电极植入过程中均没有观察到出血或者对大脑组织和血管的破坏，表明手术过程没有给大脑带来急性损伤。同时，对比电极阵列形态改变前后，发现仅有很小一部分电极的电阻在形变后升高，表明形变后电极阵列的功能仍然可保持正常，信号质量维持稳定。在术后1-4周和第8周，对植入电极的大鼠大脑进行MRI扫描和组织学分析，仅观察到十分轻微、短暂的炎症反应，证明不仅SCEA的植入过程是微创的，SCEA还具有高度的慢性生物相容性，克服了现有ECoG在对大脑造成急性和慢性损伤方面的挑战。

因此，SCEA具有以微创和高度生物相容的方式获取高时空分辨率、高带宽和高信号质量的大规模生理或病理皮层活动的能力，使之在大脑的基础研究、脑机接口的研发等一系列应用中具有显著的优越性

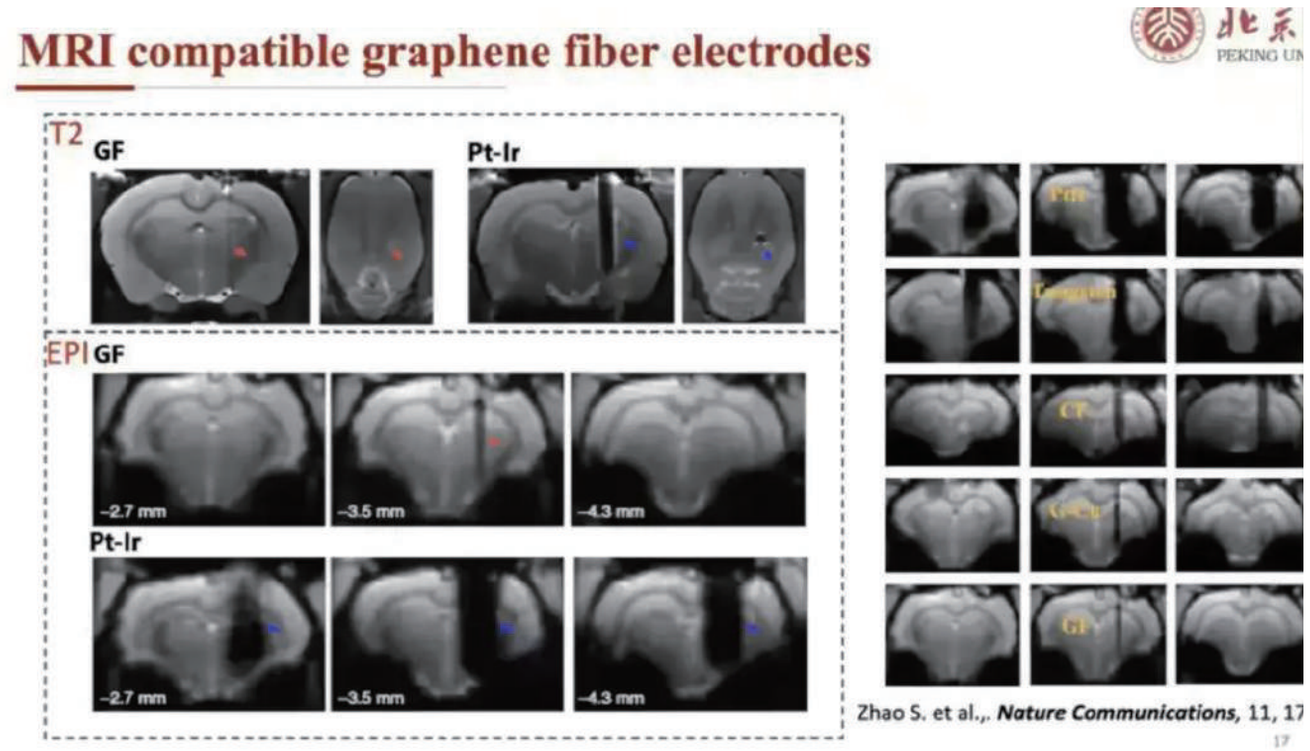
与广阔的前景。

二、DBS与fMRI能否成为“完美拍档”?

DBS常用于治疗运动障碍性疾病(如帕金森病),也被发现对难治性抑郁症具有一定的疗效。尽管DBS已经在临床中被广泛应用,对于电刺激的治疗机制和神经调节作用仍然了解甚少。对脑内特定靶点进行电刺激会在局部和全脑水平引起各种效应,仅依靠局部记录到的电生理信号无法对这些效应进行全面研究,DBS和fMRI联用(DBS-fMRI),则可以实现对全脑活动的捕捉,了解大脑局部和整体的功能状态和连接模式的变化,解释功能性电刺激对于神经疾病的治疗机制。

然而,fMRI和DBS的结合存在一大障碍:许多金属电极由于材料的磁敏感性与水/组织之间的磁敏感性不匹配,会引发强烈的磁场干扰,使电极周围区域的成像产生严重的伪影或扭曲变形;除此之外,刺激电极的尺寸也是伪影大小和严重程度的重要影响因素。

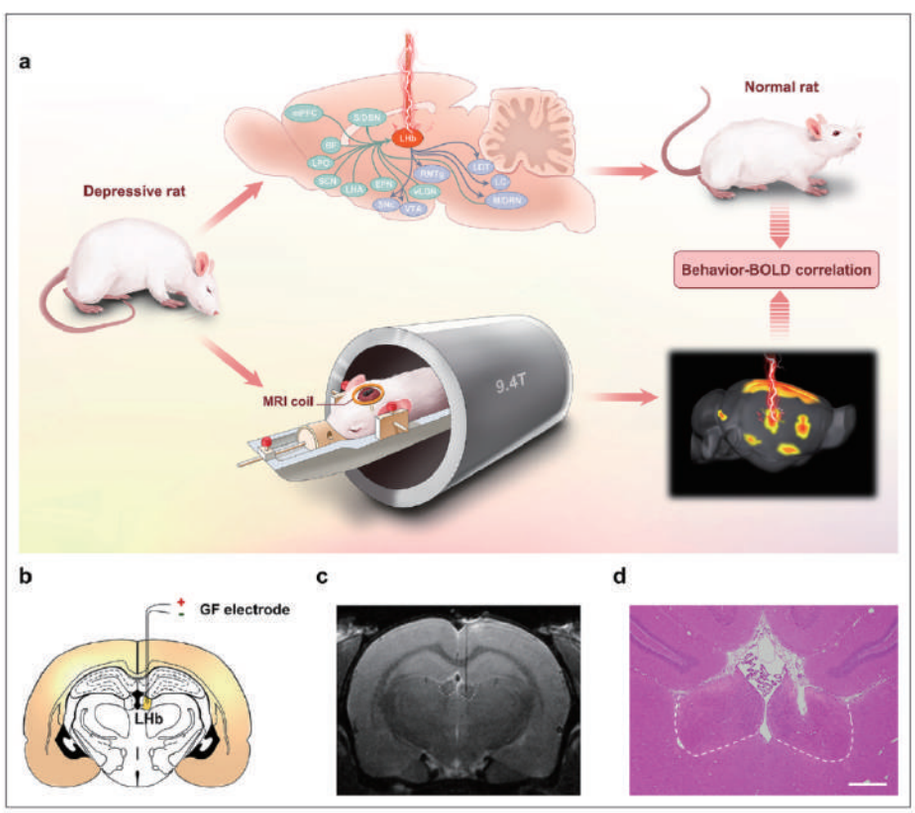
为了解决这个问题,段小洁研究员团队开发了以石墨烯纤维(graphene fiber, GF))为材料的微电极,用于DBS-fMRI的同步联用。与传统铂铱(Pt-Ir)电极不同,石墨烯电极的磁敏感性与水相近;同时,石墨烯纤维刺激电极还展现出了很高的电荷注入能力和很强的稳定性。更重要的是,在帕金森病大鼠模型中,利用该电极进行DBS-fMRI同步联用,不仅帕金森大鼠的运动障碍有明显改善,同时GF微电极也展现出了很强的MRI兼容性,在T2图像和EPI图像中伪影大小分别是铂铱电极的1/8和1/4;与其他电极材料如钨等相比,石墨烯纤维电极造成的伪影也更小(图2)。



▷图2:石墨烯纤维(GF)电极与其他材料电极的MRI/fMRI成像的伪影大小比较。图源:讲者PPT及参考文献2

接着，段小洁研究员介绍了实验室在抑郁症方面做的研究。外侧缰核（lateral habenular nucleus，LHb）是大脑内的一个“反奖赏中枢（anti-reward center）”。近年来，对于难治性抑郁症治疗的研究发现，以LHb为刺激靶点的DBS治疗（LHb-DBS）具有快速抗抑郁效果，但其治疗的相关机制仍然是一个未解之谜。同时，LHb在人脑内是一个大约仅有3mm×3mm×3mm大小的核团，而现有的用于病人身上的DBS电极直径约为1.28mm，使得现阶段在病人被试身上进行LHb-DBS研究依然充满挑战。

研究人员将新开发的石墨烯纤维电极植入两种抑郁症大鼠模型的LHb中，并进行MRI/fMRI扫描，与上一个实验相同，图像中GF电极带来的伪影很小，这种MRI兼容性使研究人员可以在MRI结构像（T（2）上轻松且准确地验证电极在LHb中的植入位置（图3），并且采集到高质量的EPI数据，即能够完整且无偏倚地反映全脑激活模式的功能数据。在疗效方面，当通过DBS电极在LHb中施加高频电刺激时，大鼠的抑郁样症状在几秒到几分钟内迅速得到缓解，具体表现为蔗糖偏好的显著增加、强迫游泳测试中静止时间的减少以及自主活动的增加。

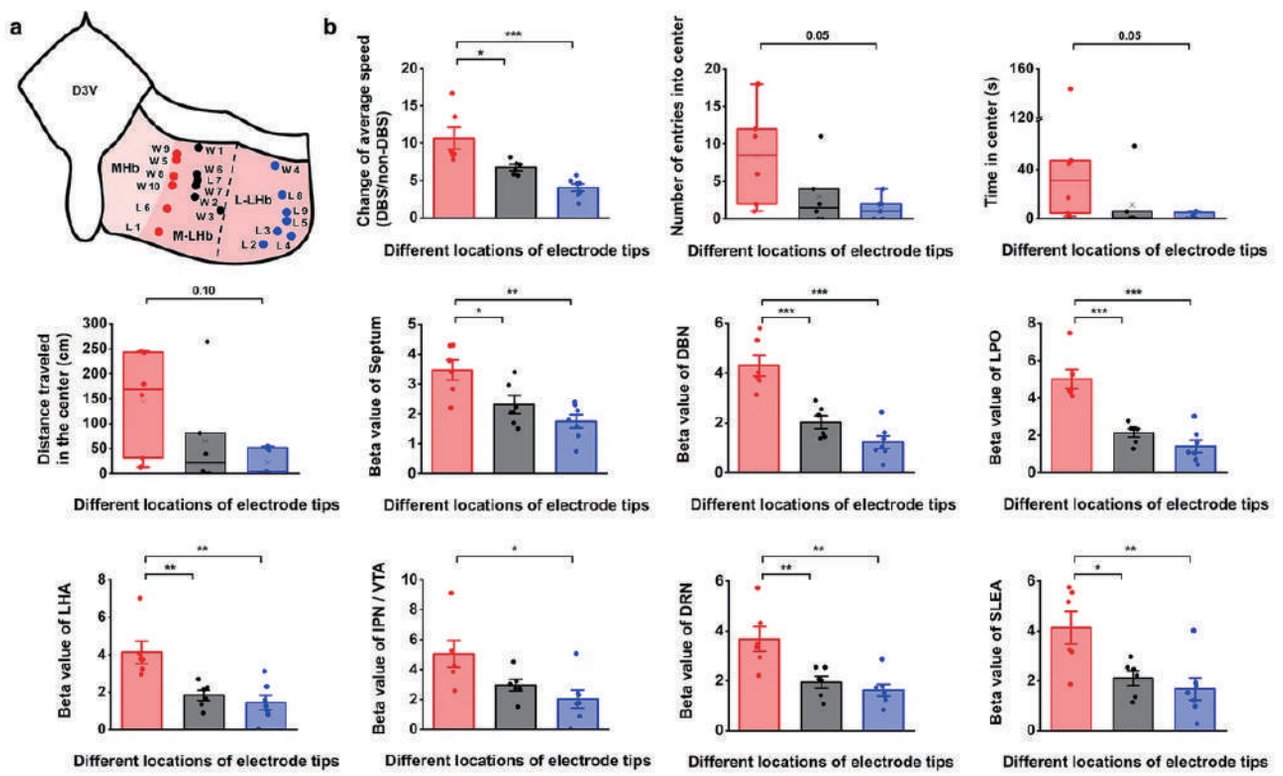


▷图3：利用fMRI研究LHb-DBS对抑郁症大鼠模型的影响。图源：参考文献3

为了进一步揭示LHb-DBS快速抗抑郁效果背后的机制，研究人员通过DBS-fMRI同步联用对植入DBS电极后行为改善的大鼠进行扫描，发现LHb-DBS激活了LHb传入和传出环路中的多个区域，包括位于边缘系统、5-羟色胺能系统和多巴胺能系统中的多个区域。此外，他们还观察到了三个与LHb没有直接连接的区域中的激活，包括扩展杏仁核（sublenticular extended amygdala，SLEA）、扣带皮层（cingulate cortex，Cg）和压后皮层（retrosplenial cortex，RS）。更重要的是，研究人员在fMRI神经信号数据与行为数据的关联性分析中发现，与内侧LHb相连的脑区的BOLD激活水平与大鼠抑郁行为的改善程度显著

相关，而在与外侧LHb相连的脑区中则没有发现这样的相关性。

基于T2图像对电极位置的定位，研究人员进一步根据电极在LHb中的植入位置将实验大鼠分为内侧、居中和外侧三组进行对比，发现了DBS刺激的位置越接近LHb的内侧，所产生的的抗抑郁效应则更强（图4）。这项研究成果揭示了内侧LHb可能是DBS实现快速抗抑郁治疗的重要靶点，为包括难治性抑郁症在内的多种神经系统疾病的DBS治疗策略提供了新见解，并进一步证明了使用石墨烯纤维作为电极的DBS-fMRI同步联用为DBS治疗机制和调节效应的转化研究提供了强大的工具平台与技术路径。



▷图4：DBS电极在LHb内的不同植入位置对抑郁症大鼠的行为改善和部分脑区的BOLD激活水平的影响。图源：参考文献3

最后，段小洁研究员强调，无论是微创且具有高时空分辨率的大规模电极阵列的开发，还是应用磁共振兼容的DBS电极进行疾病治疗机制的研究，都离不开纳米材料和精密加工技术的使用与发展，纳米材料的创新应用和加工技术的日新月异将不断拓宽脑科学与脑机接口研究的边界。（编辑：Lixia）

参考文献：

[1] Wei, S. et al. Shape-changing electrode array for minimally invasive large-scale intracranial brain activity mapping. bioRxiv, doi:10.1101/2023.06.29.547140 (2023).

[2] Zhao, S. et al. Full activation pattern mapping by simultaneous deep brain stimulation and fMRI with graphene fiber electrodes. Nat Commun 11, 1788, doi:10.1038/s41467-020-15570-9 (2020).

[3] Li, G. et al. Instantaneous antidepressant effect of lateral habenula deep brain stimulation in rats studied with functional MRI. Elife 12, doi:10.7554/eLife.84693 (2023).

脑与AI

汪小京：将神经元变为数学模型和算法，在人脑和 AI 间架起桥梁

讲者：汪小京

纽约大学神经科学教授，斯沃茨理论神经科学中心主任。研究重点为认知功能的脑机制，尤其以在工作记忆、决策的神经机制。他也是“计算精神医学”的创始人之一。

扫码查看原文

大脑是宇宙中最复杂的系统。人们一直试图揭秘大脑的内在机理，但囿于技术和方法的限制，进展始终有些缓慢。如今，人工智能技术的到来给脑科学研究带来了全新的机遇。

6月21日，天桥脑科学研究院（Tianqiao and Chrissy Chen Institute，TCCI）和neurochat神聊共同发起了第四届TCCI-neurochat神聊在线学术会议。会议首日，纽约大学汪小京教授作题为“Theoretical Neuroscience in the age of Artificial Intelligence”主旨报告，从理论神经科学的角度，分享了人工智能应如何与神经科学相结合来推动神经科学的研究。

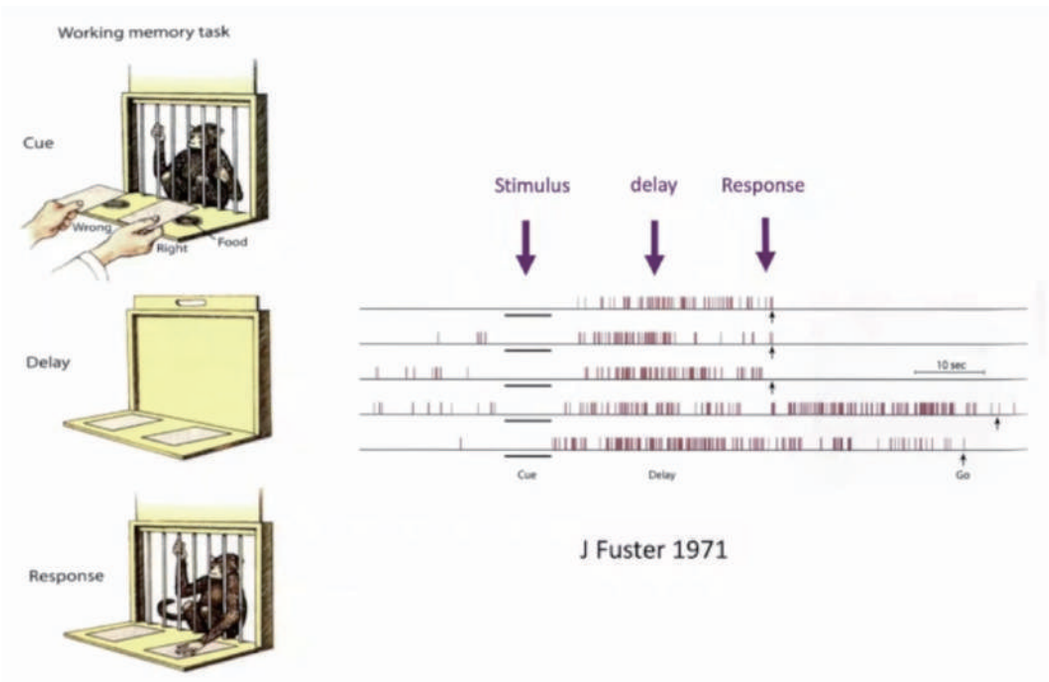
所谓理论神经科学或者计算神经科学（Theoretical/Computational Neuroscience），是利用数学理论来理解大脑是如何工作的。如同理论物理之于物理学，理论神经科学在大脑研究中也扮演着重要的角色，与实验研究相辅相成。仅仅通过实验，我们可能无法真正了解大脑如何在多个不同层面上运转。理论神经科学可以将神经元和神经网络转化为数学模型和算法，在脑科学和AI之间架起桥梁。随着实验工具的进步和大数据的出现，进行理论研究和建模显得尤为必要。

一、AI难以企及的脑功能

当下AI领域的技术革命约始于2012年。当时，杰弗里·辛顿（Geoffrey Hinton）的研究小组证明了深度神经网络可以执行视觉对象识别任务。众所周知，不同部位的皮层对应不同的功能。深度神经网络的不同层级可以映射到大脑的V1、V2、V4等区域。但眼睛上方的前额叶皮层（prefrontal cortex，PFC）直到现在也没有被研究透彻。人们也没有找到它的功能在当下的机器系统中的对应关系。

这个仍待阐明的PFC区，对各种认知功能都非常重要，例如工作记忆和决策。何为工作记忆？它是大脑内部维持和处理信息的能力，使大脑可以专注于内在，不被环境所役使。而决策则是指，人们如何在不确定性中做出选择。

通常情况下，在实验室中可以使用延迟依赖性任务研究工作记忆。比如，先给猴子看食物的位置，然后设置一个5或10秒的延迟阶段。在这段时间里，猴子必须牢牢记住食物是在左边还是在右边。延迟结束后，猴子会根据记忆来检索食物，而不是基于直接的感官刺激（食物本身）。上世纪70年代早期，人们使用单细胞记录和电刺激发现，在延迟阶段PFC单个细胞持续活跃。这个神经元在刺激之前并不活跃，它甚至没有对刺激作出反应，但却在延迟阶段持续活跃，并且随着延迟时间加倍而加倍。这就是所谓的刺激选择持续活动（stimulus-selective persistent activity）。



▷图注：延迟依赖性任务。图源：汪小京教授提供

对于决策的研究，有一个简单且巧妙的实验——双选项强迫选择（two-alternative forced choice）。猴子被训练注视屏幕上的一个点，并在点消失后做出决策：判断点的运动是朝左还是朝右。猴子需要通过眼动来表达其决策。诀窍在于，在任何给定的试验中，点移动方向的比例是可控的，这个比例被称为一致性（coherence）或运动强度（motion strength）。如果所有的点都向一个方向移动，那么就很容易做出决定。如果只有50%的点向同一方向运动，另外50%的点则是随机移动的，这就有点难以抉择了。如果只有5%的一致性，甚至是0%呢？这时候做出主观判断是非常困难的。单细胞生理学研究发现，神经元活动的持续递增（ramping activity）发生在几毫秒间。这种瞬时的剧烈活动是神经元整合信息的一种机制，通过积累关于不同选择方案的证据，做出主观判断。

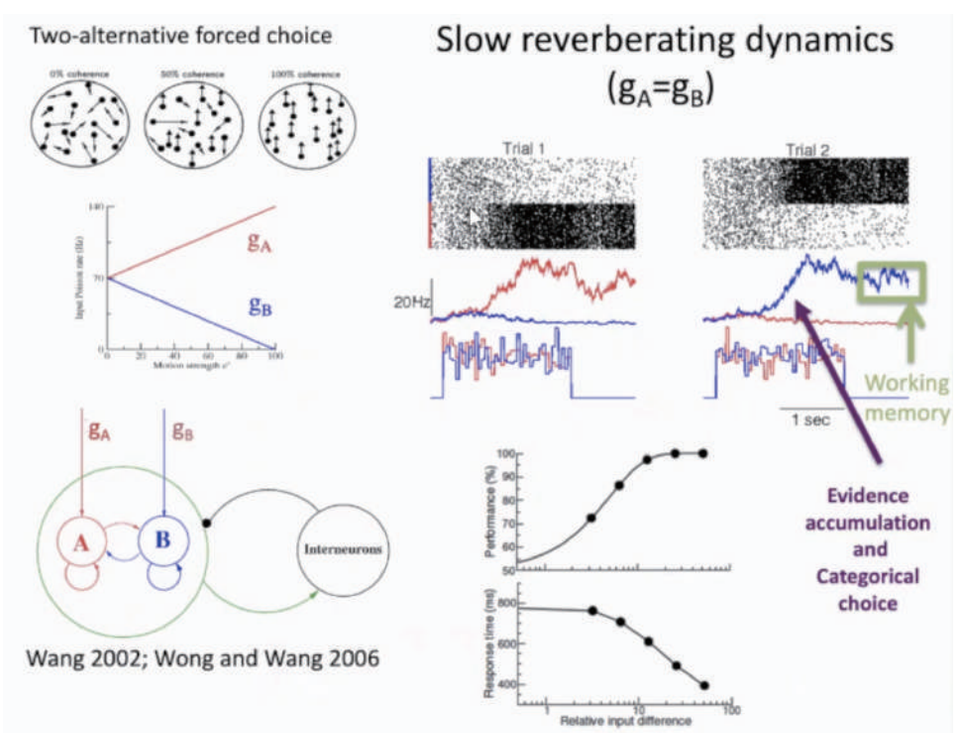
汪小京教授通过这两个实验例子，直击了最复杂问题的内核。正如著名心理学家唐纳德·赫布（Donald Hebb）所说的那样：“要用一种中枢神经机制来解释刺激和反应之间的延迟的话，它似乎与思维的特征不谋而合。”当我们思考时，不是由直接的外部刺激所驱动，而是大脑内部的一种行为。延迟活动，则是内部神经元群体活动的具体实例。

然而，人们尚不清楚决策或选择的生物基础，而AI系统也没有这种能力。诚如诺姆·乔姆斯基（Noam Chomsky）所言：“一旦涉及意志或决定或理由或行动选择的来源问题，人类科学就陷入了迷茫。”

二、基于生物学进行建模

即便真实的生物过程如此复杂，我们可以用相对简单的行为范式，尝试了解跨层次的机制。在物理学单摆实验中，如果空气中存在摩擦力，摆锤的幅度会越来越小，摆幅随时间呈指数式衰减。神经元的放电，亦是如此，会随着时间的推移而衰减（时间常数为10毫秒）。要想在工作记忆中获得持续的活动，就需要解决这一问题。为此，汪教授提出了一个大胆的假设：回响（reverberation）。也就是说，如果一组神经元相互激发，那么即使输入消失，它们仍然可以持续活动。在灵长类动物PFC的第二三层，有强大的水平连接，也为这一假设提供了结构基础。

为了验证这一假设，汪教授化繁为简，构建了一个只有两个动态变量的系统，展示了如何将随机点、感知决策和工作记忆结合起来进行判断和选择。这个模型有两个选择性发放神经元群体——a和b。神经元群体内部发生回响，而两个群体之间的兴奋选择性要弱得多。同时，它们通过抑制性神经元进行竞争，最终产生一个赢家。而这，就是系统的选择。系统的选择是由运动强度和对a组和b组的相对输入决定的。当输入消失时，选择信号在整个延迟期是自我维持的。这个模型可以解释工作记忆和决策的基本过程。

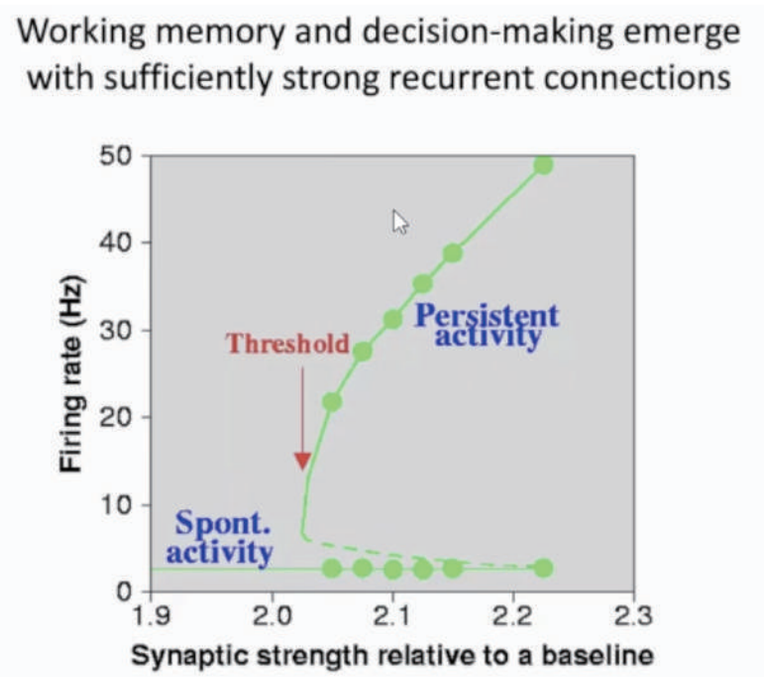


▷图注：（左）双选项强迫选择；（右）慢回响动力系统。图源：汪小京教授提供

并且，这个模型是跨层级的。以反应时间作为运动强度的函数，会发现，运动强度较低时，任务越困难，做出决定所需的时间就越长。就像在日常生活中，要做一个艰难的选择时，我们通常需要很长的时间来考虑不同的选择。而这一点在猴子的实验中也有体现。这也阐明了为什么循环神经回路的动力学可以解释行为。

同时，汪教授从细胞分子角度提供了一些洞见。比如，慢速时间积分主要由NMDA受体介导的兴奋性回响实现。NMDA受体对于决策和工作记忆非常重要，这在猴子实验中已经得到测试和确认。如果找到一种药物来减少或消除NMDA的信号，那么神经元活动就会结束。

如果以自我维持状态、内部状态为反复激活强度的函数，并将这些状态称为一个参数，则会发现：在这个模型中，当参数高于一个阈值时，会出现一系列刺激选择性的、持久性的状态。这些状态包括工作记忆中的不同记忆项目。这种状态的出现可以解释为动态系统中的分岔，即参数的分级变化导致了不同的功能和能力。这一观点对于循环神经回路非常重要，并且在多区域、大规模的大脑系统中也有应用。



▷图注：工作记忆和决策中足够强的循环连接。图源：汪小京教授提供

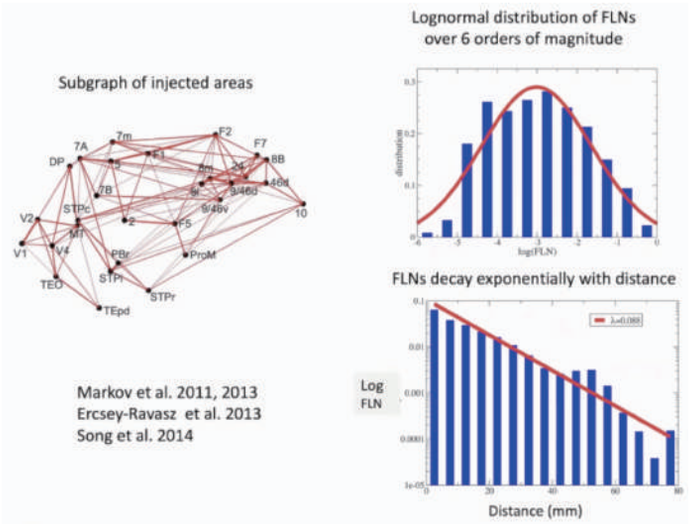
在此基础上，我们可以借助机器学习建立一个近似生物学上的神经回路，用于执行20个不同的任务，比如可以涉及工作记忆、决策、分类、多感官整合、延迟采样、延迟分类等等。一旦网络成功地执行了所有微小的任务，就可以打开“黑匣子”进行分析，提出自己的问题，分析新的回路、神经群体的动态。神经元的活动可通过任务差异来衡量；神经元之间的相互作用和组织方式可以利用聚类分析来厘清。

然而，即便我们设计出类似PFC的回路，可以基于规则灵活地指导行为，也不应止步于此。我们有必要超越目前为止所描述的局部回路建模。因为，事实上，任何一种功能都涉及多个脑区，不仅仅是PFC，还有后顶叶皮层，甚至还可能有一些超级皮层结构。

三、从局部网络到大脑系统

大脑不同区域的互动存在非常强的循环性，作为一位理论家，汪教授表示，仅将它们之间的反馈连接定性地勾勒出来远远不够，他需要数字这种清晰明确的描述。而近十年内，得益于科技的发展，利用收集数据绘制区域间有向加权连接矩阵（directed and weighted connection matrix）成为可能。

在猕猴的皮层中，如果用直方图表示两个区域有向连接的权重，会发现多个区域的直方图呈对数正态分布。而任意一对区域的权重与布线距离（wiring distance）呈指数分布。这意味着我们不能用传统的图形描述皮层网络，因为图形描述是顶层逻辑的，并没有实现这种特殊的嵌入，无法体现皮层间的特殊关系。汪教授的团队创造了一类新的数学模型，这些模型是专门为特殊嵌入而设计的，可以用来研究真实的皮层网络。

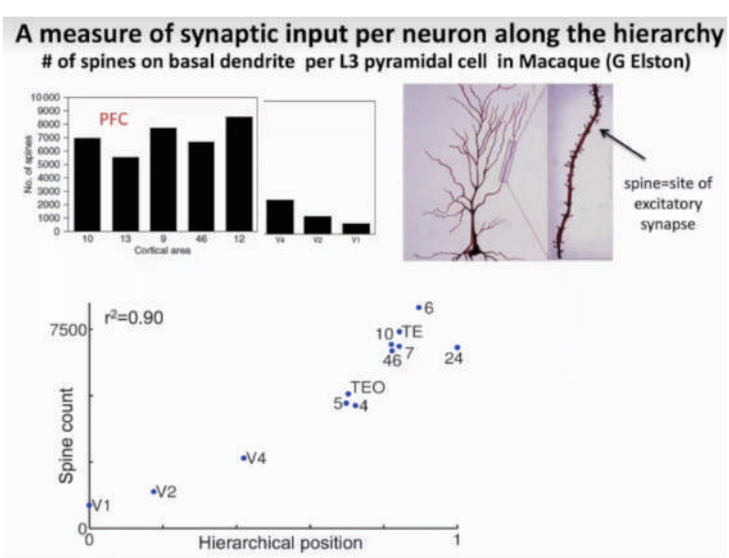


▷图注:(左)图形描述;(右)不同区域的有向连接权重直方图以及权重与布线距离直方图。图源:汪小京教授提供

构建大型皮层模型时,我们需要对每个区域进行数学建模。不同的区域可能需要使用不同的数学模型。因此,我们需要从制定科学问题开始,并根据问题建立相应的模型。

在神经科学中,有一个公认的原则是存在皮层柱(cortex column)。这是由一组垂直排列的神经元组成的柱状结构,这些神经元具有相似的功能和结构特征。因此,存在一个重复了很多次的局部回路单元,在啮齿类动物、猴子和人类中,这个局部回路都是差不多的。但是,即使在相同的局部回路中,也存在数量上的差异,这将导致质变,即功能和能力的不同。因此,在建立模型时,需要考虑数量上的差异。

汪教授展示了在大规模系统中,数量差异的影响。以树突棘数量作为层级位置的函数,可以观察到每个细胞树突棘的数量系统性地增加了。这种微观梯度显示了不同区域里神经元的兴奋强度之间的关系。而在前段时间发表于《自然·神经科学》的论文中,汪教授则研究了猕猴皮层神经递质受体的表达梯度,如多巴胺受体和5-羟色胺受体,发现了一个皮层层级,受体在高层级的皮层区域具有更高密度、更大树突和较少髓鞘。这是用数学模型量化生物特性变化的又一实例。



▷图注:一种沿锥体细胞基底树突棘层级测量神经元突触输入的方法。图源:汪小京教授提供

除了数量上有层级外，时间常数也具有层级。不同区域的活动发生的时间尺度存在差异，其中一些区域的时间常数较短，而其他区域的时间常数较长。比如，视觉网络运行得非常快，当看电影的时候，感官输入一直在变化，时间常数必须非常短，以快速反应。而在决策时，神经元活动持续增加，这意味着时间常数必须非常长。这对于工作记忆和决策等认知过程具有重要意义。

汪教授还表示，通过对孤立区域的分析，可以发现区域间的连接不仅仅是局部的，而是长距离的，这种连接形成了分叉空间。大脑皮层大规模模型的建立过程，既包括线性梯度的添加，也需要对时间常数的分析。层级结构和时间常数可以帮助我们理解大规模动力学和功能的影响。最后，汪教授强调，不是所有的神经回路都是一样的。像PFC这样的认知神经回路，它能够同时进行工作记忆和决策。从AI角度来看，这是最值得思考的问题。

四、写在最后

神经科学是一个复杂而迷人的领域，本次主旨报告以大量计算神经科学的例子为引，由表及里，发现隐藏在数据中的规律和趋势，层层剥开神经网络的运作方式。汪教授基于优雅的物理、严谨的数学，讲述了如何进行精细的生物学研究，内容深入浅出，发人深思。

从模仿到理解，计算模型会是大脑的最终归宿吗？

作者：吴宇轩

一个希望利用人工智能来推进自然科学和基础科学的学生。兴趣是数据科学以及与自然科学的交叉。尤其是天文物理学和生物信息。目前正在与苏黎世理工进行一个用于生物信息embedding的python库的开发。

扫码查看原文

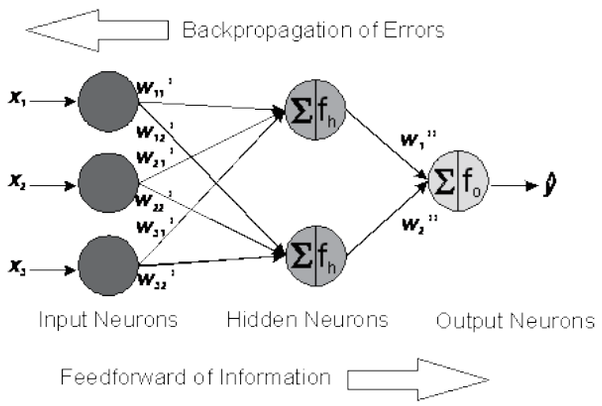
在人类探索的历史中，大脑仿佛是宇宙留给人类的最后一块版图。长期以来，神经科学家们一直致力于勾勒出这块版图上的线条，试图解答大脑如何执行那些看似不可能的复杂任务。尽管我们取得了一些进展，但大脑的高度复杂性和惊人的效率仍然让人望尘莫及。于是，受大脑结构及其信息处理方式的启发，我们设计出了神经网络，以帮助解决现实世界中的复杂问题。

但随着技术尤其是深度学习的飞速发展，这种模仿与理解的关系正在经历一场根本性的转变。神经网络，特别是深度学习模型，已不再局限于单纯模仿大脑的工具，它们正成为理解大脑之谜的关键钥匙。这些模型以其高度复杂和精细的处理能力，正在帮助我们揭开大脑是如何在多变和复杂的环境中学习和做出决策的秘密。这种从单向模仿到双向理解的转变，不仅在神经科学领域开辟了新的探索之路，更为我们提供了一个独特的窗口，透过它，就能更深入地洞察那个蕴藏在我们头颅中，重仅三磅，却包含着无限可能的宇宙。

一、大脑与计算模型的基本结构

神经元，是大脑核心构成单元。它们通过相互连接并发射电信号的方式，共同参与对事物的解释、推理和决策等复杂功能的执行，以帮助大脑处理不同信息，并灵活应对多变环境。神经元学习的关键在于突触的可塑性。当神经元之间频繁传递信号时，相关的突触连接会强化，形成记忆和学习。这种神经可塑性使得大脑能够根据经验调整神经网络的连接权重，从而适应不同的环境和任务。

1943年，McCulloch和Pitts就发现[1]，神经元的脉冲及其开关状态是一种逻辑门。他们认识到，大脑是一种由细胞组成的机器，类似于蜜蜂群体中涌现出复杂行为的现象。十年后，心理学家Frank Rosenblatt提出了感知器概念，这是一种单层简易神经网络，旨在通过监督学习模拟大脑的学习过程。感知器通过调整权重，使模型学会从输入到输出的映射关系。这类似于大脑中神经细胞之间的突触连接调整过程。



▷图1：计算神经网络示意图。图源：参考文献2。

Rosenblatt的感知器有三种不同类型的“细胞”（单元）组成，分别代表“投射”，“关联”和“响应”。它通过将权重与特征向量结合，使用线性预测函数进行预测，并从样本数据中学习权重，以应用于新数据。然而，这种方式很快在非线性问题上遇到了局限。

为了克服这一局限，研究人员引入了“隐藏层”和“激活函数”等概念。这些神经元可用于解决早期构建感知器时遇到的一些基本问题，特别是在接受大量神经元的馈送和训练数据时，它们解决了感知器在处理非线性问题上的局限性。由此，研究人员终于发现了一个能够有效解决非线性问题的“公式”——以深度学习（DL）为核心的神经网络。

虽然人工神经网络和生物神经网络在行为层面上具有相似之处，它们的学习方法却大不相同。人工神经网络使用梯度下降来最小化损失函数并达到全局最小值。其梯度下降需要反向传播，而反向传播只能在生物神经网络中的一个神经元的范围内进行。相比之下，生物神经网络采用的是赫布学习原则，通过尽可能多的学习实例，提高一个神经元激活另一个神经元的效率，进而增强连接，使其更容易传递信号。这种基于时间顺序的连接强化是生物神经网络学习和形成记忆的基础。反之，如果这种激活模式不再发生，连接可能会减弱，表现为我们所说的“遗忘”。尽管方式不一，但行为的相似，也足以帮助我们借用人工神经网络类比和理解生物神经网络。

二、自监督学习模型与大脑活动相似

近期，麻省理工学院的K.Lisa Yang与计算神经科学中心的研究人员发布的两项实验，为人脑可能使用类似于人工神经网络运作（自监督学习）的方式来理解世界的观点，提供了新证据。他们发现，当他们使用特定的自监督学习模型时，模型能够从未标记的数据中理解环境，表现出了强大的迁移学习能力和可重用性，从多种层面展现出了与哺乳动物大脑相似的活动模式。

更为引人注目的是，这些自监督模型能够学习到物理世界的表征，从而准确预测物理世界将要发生的事情。他们认为，哺乳动物的大脑可能具有相同的学习策略。例如，哺乳动物的大脑也会通过观察环境来学习和理解环境，而无需外部的指导或标签。这种学习方式使得哺乳动物能够适应各种各样的环境，并在面对新的挑战时，能够利用过去的经验来做出反应。

（1）视觉模型

在视觉处理领域，早期的神经网络模型主要依赖于监督学习，即在大量有标签的图像上进行训练以学习分类。这种方法虽然在特定任务上表现良好，但它的一个主要局限在于对大量人工标记数据的依赖。因此，自

监督模型逐渐成为更为有效的替代方案。

自监督模型，旨在从未标记的数据中学习有用的表示，摆脱了对外部注释或标签的依赖。其核心在于让模型自行从输入数据中生成目标，并优化生成目标与原始输入之间的关系，从而实现对数据潜在表示的学习。这种学习方式的独特优势在于，它能够有效地利用大量未标记的数据。由于不需要人工进行繁琐的标注工作，这使得自监督学习成为在数据稀缺或标注成本高昂的情况下的理想选择。

在麻省理工学院的一项新研究中[3]，研究人员通过使用数十万描述日常场景的视频，训练了一个自监督模型，该模型可以预测未来场景的状态。与传统难以适应不同任务的模型不同，他们发现通过对自然数据进行自监督学习，可以使模型成功推广到其他任务。

研究人员将训练完成的自监督模型应用于一个名为“Mental-Pong”的任务中，这是一种类似于用球拍击球的视频游戏。在这个任务中，球在即将被击中前会突然消失，玩家需要通过预测球的轨迹来成功击中它。

研究人员发现，他们的自监督模型能够准确地追踪隐藏球的轨迹。在他们的研究中，该模型能够成功模拟看不见的球的轨迹，表现出类似于人类进行“心理模拟”的认知现象。

在动物玩类似游戏时，其大脑的背内侧额叶皮层常会显示特定的神经激活模式。背内侧额叶皮层不仅会对空间位置和变化作出响应，而且在规划未来行动时表现出活跃性，包括对于如何达到目标、选择适当策略等方面的规划。自监督模型在执行任务时展现的神经激活模式，与动物在游戏中大脑的这一部分所表现的模式惊人地相似。研究人员表示，没有其他类型的计算模型能够像这个自监督模型那样与生物数据如此接近。

这一发现深化了对自监督学习模型与大脑相似性的理解：大脑在执行各种任务时展现出特定区域的神经激活模式，而自监督模型似乎能够在类似的任务中产生相似的模式。这不仅突显了自监督学习模型的潜在优势，同时为揭示大脑运作机制提供了更多线索。

（2）空间导航

无独有偶，由Khona、Schaeffer和Fiete领导的另一项研究[4]，通过自监督学习模拟了网格细胞的行为，暗示着大脑可能采用类似的自监督机制来训练神经元，以学习和理解其所处的世界。

网格细胞，位于内嗅皮层，与海马体中的位置细胞协同工作以帮助动物进行空间定位与导航。其独特之处在于，它们在空间中的多个点激活，形成一种对称且极其精确的六边形网格，就像精细的内部GPS系统。每个网格细胞都有其独特的坐标模式，但一个单独的网格单元无法准确指示动物的具体位置，因为它在多个点都会激活。然而，当多个网格单元的图案重叠时，就能非常精确地确定动物的位置。这些图案在大脑中形成了一种内部坐标图，有助于测量空间中不同点的距离。

在先前的研究[5]中，研究人员训练了一种自监督模型，来模拟网格细胞的功能，即根据动物的起点和速度自主预测下一位置，完成这一“路径整合”任务。然而，这类模型始终需要绝对空间的信息，而这这是动物所不具备的。

受这项研究的启发，Khona等人训练了一种对比自监督网络，执行相同的路径积分并以此表示空间。与之前的研究不同，该模型可以像网格细胞一样，通过位置的相似与不同来相对的区分位置。

“这类似于图像训练模型。如果两张图像都是猫，它们的编码应该相似，但如果一张是猫，一张是卡车，那么他们的编码应该互斥。而我们采用同样的想法，但将之用于空间轨迹。” Khona解释道。

在网格细胞与计算模型的早期研究中，麻省理工学院的团队也曾调整模型，使位置编码单元更贴近生物的

位置细胞。在这个过程中，虽然模型仍然能够执行路径整合任务，但却不再产生类似网格细胞的活动。当研究人员要求模型生成不同类型的位置输出，例如在网格上的X轴和Y轴位置，或相对于起始点的距离和角度的位置时，类似网格细胞的活动也消失了。

Fiete曾指出："如果你要求这个网络唯一要做的事情是路径整合，并且对单元施加了一套非常具体而非生理的要求，那么就有可能获得网格细胞。但如果你放松对读出单元的这些要求，网络产生网格细胞的能力就会大幅降低。"

最终，通过引入分离损失、路径不变性损失和容量损失三种损失函数，他们优化神经网络，使其能够形成多种不同的网格图案，与网格细胞的自然活动相似，并能在训练分布之外良好地泛化。此外，他们还通过一系列数学属性，如代数编码、高容量表示、快速去相关性等，将网格细胞的编码理论属性表征出来。这都代表着大脑的复杂空间表征不是通过外部监督学习获得的，而是通过一种内在的、自主的学习过程（自监督学习）形成的。

三、意义

除视觉、空间导航外，Edward Chang等人利用自监督模型研究了语音模型与人脑听觉通路的相似性[6]；而在认知功能和精神障碍的机制[7]上，相关模型也发挥着重要作用。它们都暗示着，大脑活动与自监督学习的相似性。

因此，神经网络不仅是一种强大的预测工具，更是我们解读和模拟生物神经网络的关键窗口。我们可以通过训练一个模拟生物神经网络的计算神经网络，并观察其活动来解释和类比生物神经网络的运作方式。同时，生物神经网络也能指导我们考虑更多已知的生物层面的限制，使我们的计算模型更加接近现实。

模仿大脑设计神经网络，使得计算模型具有生物特征；借由自监督学习探究大脑原理，以期发现大脑的计算特征。这一探索过程的终点，机械与生物之间的界限正变得越来越模糊。正如凯文·凯利在其著作《必然》中所指出的，"机械的终点是生物，而生物的终点是机械"。在这个交错的领域，究竟是否存在明确的分界线？随着我们不断的探索，这个问题的答案也将越发清晰。（编辑：存源）

参考文献

[1] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115–133. https://doi.org/10.1007/BF02478259

[2] https://www.frank-dieterle.de/phd/2_7_1.html

[3] Nayebi, A., Rajalingham, R., Jazayeri, M., & Yang, G. R. (2023, May 19). Neural Foundations of Mental Simulation: Future Prediction of Latent Representations on Dynamic Scenes. arXiv.Org. https://arxiv.org/abs/2305.11772v2

[4] [2311.02316] Self-Supervised Learning of Representations for Space Generates Multi-Modular Grid Cells. (n.d.). Retrieved 1 December 2023, from https://arxiv.org/abs/2311.02316

[5] Schaeffer, R., Khona, M., & Fiete, I. R. (2022). No Free Lunch from Deep Learning in Neuroscience: A Case Study through Models of the Entorhinal-Hippocampal Circuit (p. 2022.08.07.503109). bioRxiv. https://doi.org/10.1101/2022.08.07.503109

[6] Dissecting neural computations in the human auditory pathway using deep neural networks for speech | Nature Neuroscience. (n.d.). Retrieved 1 December 2023, from https://www.nature.com/articles/s41593-023-01468-4

[7] Frontiers | Editorial: Computational models of brain in cognitive function and mental disorder. (n.d.). Retrieved 1 December 2023, from https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1230587/full

人工智能如何向人类智能学习?

作者:史冬平

中国科学院大学在读博士生,关注认知神经科学。

作者:赵诗彤

中国科学院神经所硕博连读,关注视觉记忆与认知。

作者:吴婷婷

中国科学院神经所在读博士生,关注视觉记忆与认知。

作者:刘风临

陆军军医大学在读博士生,关注认知障碍性疾病相关研究。

扫码查看原文

大脑是宇宙中已知的,也可能是唯一的“智能机器”。在行为学层面,人与动物的大脑能执行精细的、高水平的认知任务,包括灵活学习、长时程记忆、开放式环境决策等。在结构层面,认知与计算神经科学揭示了大脑通过极其复杂而精细的网络实现其功能。北京大学心理与认知科学学院吴思教授与清华大学社会科学院心理学系的刘嘉教授等人在23年1月发表的“AI of Brain and Cognitive Sciences: From the Perspective of First Principles”一文中提到,随着人类对大脑的结构与功能的认识不断深化,大脑的基本原理为改进人工智能提供了最重要的参考。这些基本原理指大脑提取、表征、处理、检索信息的规则,它们指引着大脑的运行,是大脑执行其他更高级认知功能的基础。

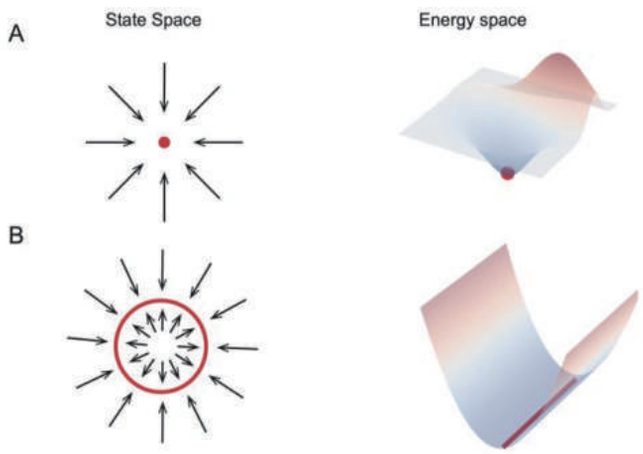
论文作者们将大脑的基本原理概括为:吸引子网络、临界性、随机网络、稀疏编码、关系记忆、感知学习。解读这些原理,并将其灵活用于人工智能,可能是使人工智能更“像”人类智能,并在性能上获得进一步提升的关键。

一、吸引子动力学:神经信息处理的规范模型

大脑由大量神经元组成,神经元之间通过突触形成各种网络。单个神经元的计算相当简单,而神经网络的动态变化才是完成大脑功能的关键。简单地说,神经网络接收来自外部世界和其他脑区的输入,其状态不断变化,从而进行信息处理。因此,动态系统理论是量化大脑如何通过网络进行计算的重要数学工具。

在动力学系统中,不同的状态演化规则和多样的外部输入可以产生各种动态学现象。在一个神经网络里,如果一个状态向量(state vector)的所有邻近状态都汇聚于它,那么这个状态向量就被称为稳定吸引子(stable attractor)。拥有稳定吸引子的网络被称为吸引子网络(attractor network),它们构成了大脑信息表征、处理和检索的基本构建模块。具有稳定状态的吸引子网络可分为两种类型,即离散吸引子网

络(discrete attractor network)和连续吸引子网络(continuous attractor network)。在神经网络中,吸引子对应网络的能量空间中的局部最小值,所有邻近状态的能量都高于它,所以会被“吸引”到这里(图1)。



▷图1:离散吸引子网络和连续吸引子网络示意图。图源:关联论文

在离散吸引子网络中,每个吸引子都有自己的吸引区域。如果以随机状态开始,网络的动态性会驱动随机状态向其邻近吸引子状态变化,这个过程也伴随着网络的能量降低。离散吸引子的这种特性使得网络能够纠正输入噪声并检索记忆表征。现在,离散吸引子网络经常用于模拟工作记忆、长期记忆和决策等。

与离散吸引子网络不同,在连续吸引子神经网络(continuous attractor neural networks,CANN)中,吸引子在状态空间中连续分布,形成一个平滑的流形。这个特性使得CANN中的吸引子状态能够迅速转变。它为CANN带来了许多潜在的利用价值,如路径积分、证据累积、预测性跟踪等。

已经有很多实验研究证明大脑中存在吸引子动力学,并且吸引子网络常被用于描述一些高级认知功能的潜在机制,这有赖于其在信息表征中的基本特性。

(1)特征一:吸引子网络稳健的信息表征

在离散吸引子网络中,记忆信息被存储为一个吸引子状态。在给定部分或有噪声线索的情况下,网络会动态演化到一个吸引子状态,并且相应的记忆会被检索出来。不同的吸引子对应于不同的局部能量最小值,并且有着各自的吸引域。如果噪声扰动不足以将网络状态推出吸引域,那么吸引子状态就是稳定的,所以记忆信息被稳健地编码。2

与离散吸引子网络不同,在连续吸引子网络中,吸引子在网络状态空间中形成一个平坦的流形,并且它们对噪声有部分的稳健性。如果噪声扰动与流形正交,网络状态在吸引子动力学作用下是稳定的。然而,如果噪声扰动沿着流形方向,网络状态将在吸引子流形上扩散,从一个吸引子移动到附近的吸引子。

(2)特征二:吸引子网络的记忆容量

记忆容量指的是能够可靠地存储在吸引子网络中的记忆数量。吸引子网络的记忆容量受到几个因素的影响。其一是噪声,当网络中存储了过多的记忆时,每个吸引子的吸引域会缩小,从而降低了吸引子对噪声的容忍度。另一个因素是记忆相关性,当记忆模式高度相关时,它们会相互干扰,破坏记忆的检索。为了增加吸引子网络的记忆容量,人们已经提出了许多方法,络(Hopfield network)4。

(3)特征三:吸引子网络的信息检索

除了具有大容量的记忆能力外,一个优秀的信息处理系统还需要高效的信息搜索能力。在吸引子网络中,记忆通常以内容寻址方式进行检索,即网络通过吸引子动态执行相似性计算,检索出与线索最相似的记忆。

在大容量网络中,从众多吸引子中找到正确的一个是具有挑战性的。例如,在一个自由记忆检索任务中,参与者需要尽可能多地搜索和回忆动物的名称。一个良好的回忆策略是在局部记忆搜索与记忆空间大跳跃之间合适地组合,表现出莱维飞行行为*。董行思等人证明,在一个带有噪声神经适应的CANN中,网络状态的动态显示出交替的局部布朗运动和长跳跃运动,呈现出莱维飞行的最佳信息搜索行为5。

*编辑注:莱维飞行行为(Levy flight behavior)是一种随机行为模式,其中个体在一定时间内以不规则的方式移动,具有长尾分布的步长。这种行为模式以法国数学家Paul Lévy的名字命名。莱维飞行行为与传统的随机行走模型(如布朗运动)不同,后者通常假设步长是固定的,而莱维飞行行为中的步长是从长尾分布中随机抽取的,因此具有更大的变化范围。

(4)特征四:吸引子网络间的信息整合

吸引子网络还可以相互交互,实现信息整合。张文昊等人研究了如何通过相互连接的CANN来实现多感官信息处理6, 7。在他们的模型中,每个模块包含两组神经元,每组神经元形成一个CANN,这些神经元的调谐函数相对于模态输入要么是一致的,要么是相反的。该研究证明,具有一致神经元的耦合CANN可以实现信息整合,而具有相反神经元的耦合CANN可以实现信息分离,它们之间的相互作用能够有效地实现多感官整合和分离。这项研究表明,相互连接的吸引子网络可以支持不同脑区之间的信息传递。

最近,在全球范围内的技术进步和大型脑项目的推动下,大量关于脑结构细节和神经活动的数据正在涌现,现在是建立大规模网络模型来模拟更高级认知功能的时机。作为神经信息处理的规范模型,吸引子网络成为人们开展这一任务的基本构建模块,人工智能也可以借助这一基本模块,在信息处理和表征方面受到一些启发。

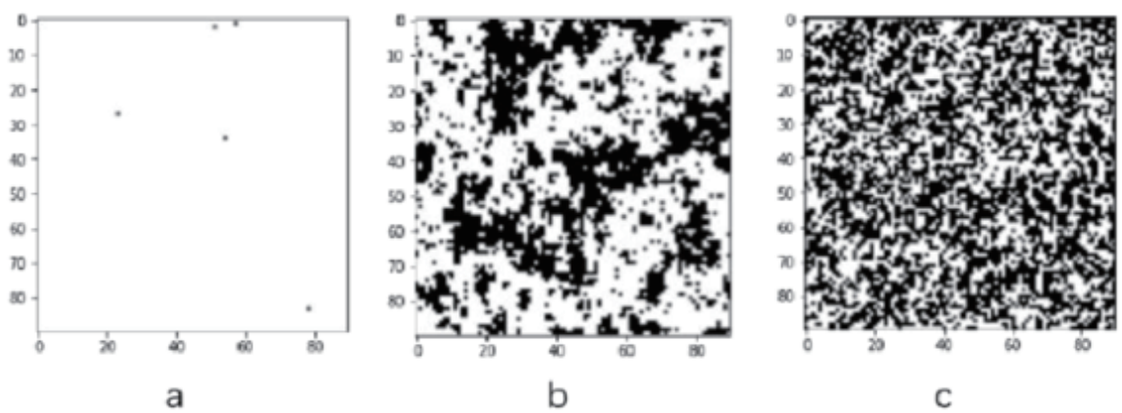
根据全局工作空间理论(global neuronal workspace theory)*,大脑分为一个共享的全局处理模块和许多分布式的专门处理模块8。每个独特的模块处理来自一个模态(如视觉、听觉、嗅觉或运动系统)的信息。相反,全局模块接收并整合来自所有专门模块的信息,并将整合后的信息“广播”回这些局部模态。为了实现这个目标,需要一个抽象的信息表示接口,允许不同模块之间进行通信。从这个意义上说,已经在实验和理论研究中被证明是规范模型的CANN,自然地成为了在模块之间表示、转换、整合和广播信息的统一框架。在未来的研究中,探索这个问题将是非常有趣的。

*编辑注:全局工作空间理论或全局工作空间模型是1988年心灵哲学家巴尔斯(B. J. Baars)首次提出的运用语境论解释意识运行的基本规则的模型。假设意识与一个全局“广播系统”相联系,该系统在整个大脑中发布信息,包括三个部分:专门处理器、全局工作空间和语境。专门处理器是无意识的,可能是一个单一的神经元,也可能是整个神经网络。

二、临界性:为大脑和人工智能带来新的视角

临界性(criticality)的框架是理解和分析复杂系统的强大工具,因为许多物理和自然系统处于临界状态。在过去的20年里,研究人员发现大脑中的生物神经网络的运行接近临界状态,这为研究脑部动态提供了新的视角。已知临界状态对脑部的活动/功能非常重要,因为它优化了信息传输、存储和处理的许多方面。在人工智能领域,临界状态的框架被用于分析和指导深度神经网络的结构设计和权重初始化,表明运行接近临界状态可被视为神经网络计算的基本原则之一。

在统计物理学中，一个具有相同物理和化学特性的材料系统中的均质状态被称为相（phase）。例如，水可以处于固相、液相或气相。当温度变化时，水可以从一个相变为另一个相，这被称为相变（phase transition）。临界状态表明系统正在从有序相向无序相转变。在有序和无序的边缘，或者称为“混沌边缘”，临界状态表现出许多特殊的属性。



▷图 2：蒙特卡洛模拟伊辛模型。图源：关联论文

图2通过伊辛模型的模拟展示了铁磁材料的相变过程和临界状态9。在伊辛模型中，自旋相互作用和热运动的竞争导致了有序相和无序相。图2a和2c分别显示了低温下的有序相和高温下的无序相。在相变边缘的温度下，如图2b所示，有序和无序处于平衡状态，两者都无法主导整个系统。在这个温度下，系统变得极其复杂，处于所谓的临界状态。

在有序相或无序相中，领域大小的分布集中在较大或较小的尺寸上，而在临界状态下，领域大小几乎分布在所有尺寸上。而且不同尺度上的分布是自相似的，这意味着这些分布是分形和无标度的。这种自相似分布在数学上可形式化为幂律（power law）10。如果使用对数-对数坐标系，分布将呈现为一条直线。幂律分布是临界状态的一个重要特征。

除了通过精确调节伊辛模型中的温度或其他系统中的控制参数来达到临界状态外，有些系统还能自发地达到临界状态，这被称为自组织临界性（self-organized criticality，SOC）。在过去的二十年中，通过记录体外培养的脑组织或体内完整大脑的神经元活动，许多实验表明大脑皮层网络也能够自组织成临界状态。

在大脑的皮层网络中，每个神经元通过突触连接接收来自大量周围神经元的输入。当输入达到其阈值时，将产生动作电位，该电位将传递给其他神经元，导致其他神经元发放动作电位。Beggs和Plenz通过多电极阵列记录大脑组织的活动首次确认了关于临界大脑的猜测11。他们发现神经元雪崩的大小和持续时间都服从幂律分布，这是临界状态的一个重要特征。随后，其他研究人员记录了不同物种的不同脑皮层在清醒和麻醉状态下的神经元活动，都再次证实了在网络的自发活动中神经元雪崩呈幂律分布12。这表明，大脑网络在临界状态附近运行是一个普遍特征，而兴奋性和抑制性的平衡在维持临界状态方面起着关键作用。

鉴于临界状态在包括神经系统在内的许多复杂系统分析中的成功应用，一些研究人员还尝试将临界状态应用于研究人工神经网络，例如，改进储层计算和增强深度神经网络的性能。

储层计算(reservoir computing, RC)通常指的是递归神经网络(recurrent neural network, RNN)的一种特殊计算框架, 其中可训练的参数仅存在于最终的输出层, 即非递归的输出层, 而所有其他参数则是随机初始化并在后续计算中保持固定状态。目前, RC模型已成功应用于许多计算问题, 例如时序模式分类、识别、预测和动作序列控制等任务。RC模型仅在网络处于临界状态时表现良好, 有时也被称为“回声状态(echo state)”。受生物神经网络中的短时突触可塑性(short-term synaptic plasticity, STP)的启发, 曾冠雄等人在RC模型中实现了一种SOC方案, 通过短时抑制(short-term depression, STD)来调整RNN的状态, 使其接近临界状态13。STD大大增强了神经网络的稳健性, 使其能够适应长期的突触变化, 同时保持由临界状态赋予的最佳性能。它还提示了大脑在不同时间尺度上组织可塑性的潜在机制, 在允许用于学习和记忆所需的内部结构变化的同时, 维持信息处理的最佳状态(临界状态)。

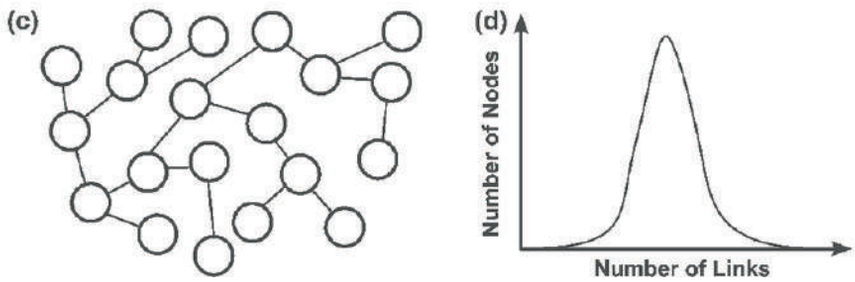
除此之外, 临界状态对于增强深度神经网络的性能也有启示作用。深度神经网络相较于浅层网络取得了巨大的成功。为了在理论上解释这一现象, Poole等人将黎曼几何与高维混沌的均场理论相结合, 揭示出深度随机网络中逐渐增加的深度与混沌状态(临界状态)的瞬时变化之间存在指数级表达能力的关系14。此外, 他们证明这个特性在浅层网络中是不存在的。这一发现对网络的结构设计具有重要意义, 为现有深度神经网络的卓越性能提供了理论基础。但是, 目前关于临界状态是否总是有益的问题仍然是一个未解之谜。

临界状态为人们研究生物和人工神经网络提供了一种全新的视角。目前, 临界性的框架不仅被用于理解神经动态和脑部疾病, 还被用于分析深度神经网络的运行并指导进一步的改进设计。在更好地理解应用于人工神经网络的约束条件, 以及设计更好的体系结构和动态规则来提高人工神经网络在复杂信息处理中的性能方面, 临界性的框架将发挥更加重要的作用。

三、随机编码:大脑编码信息的基础?

自从Hubel和Wiesel首次发现神经元对条状方向的调谐性之后[1], 神经生理学家便一直致力于寻找针对单一特定刺激具有清晰的调谐曲线的神经元。然而许多非线性混合选择性神经元的出现使得他们经常陷入混乱[2-7], 这些神经元同时、非线性地反映不同类型的特征。为了理解这一现象, 科学研究逐渐转向了对神经元的群体分析:群体编码(population coding)理论认为每个神经元在一个维度中活动, 而一群神经元的活动则组成一个高维状态空间, 引入这样一个高维空间后, 更多的信息便可以被更差别性的编码;增加了表征维度后, 原本线性不可分的表征变得线性可分了[8], 也让这些信息便于被大脑下游结构进一步处理。

另外, 为了解释混合选择性神经元可以同时参与不同信息编码的特性, 需要综合考虑多样性[3]与混合选择性, 同时还要保持尽可能的简洁性, 一些研究人员提出随机网络可能在支持多样混合性的大脑环路中起作用[9-15]。在随机网络中, 神经元突触连接的权重符合某些随机分布, 它们将信号混合后作为下游神经元的输入, 多样性便来自于这些连接的随机性;而在每个神经元内部, 输入会经历非线性映射(图3)。通过这种方式, 神经元便具有非线性混合选择性, 越来越多的生物学证据也开始支持这一观点。



▷图3:随机网络示意图。随机网络中各个节点的链接数量大致相似,在度分布图(degree distribution graph)上形成钟形曲线。图源:关联论文

在随机编码理论中,神经元群体构成的神经空间所能拥有的最大维度是神经元总数,为了增加维度,在完全随机连接的极端情况下,连接后的神经元群体应该比连接前的神经元群体更大。正因为此,网络应该具有发散性的架构。一方面,人们在不同物种的生物大脑中观察到这样的发散网络构架[16];另一方面,自上世纪九十年代早期以来,人工智能(AI)中就已经有了这样的观点[17]。

AI中的随机网络是指某些权重随机初始化并在训练过程中不被调整的一类人工神经网络。最初这些网络进入人们的视野,一是因为它们易于分析,二是因为它们的训练速度要快得多。然而,研究人员很快发现随机网络的表现出奇地好[18];在短期预测、图像识别和生物医学分类等应用中,它们的测试准确率接近完全训练的模型[15]。受到这些观察结果的启发,研究人员研究了各种随机网络的特性。

其中,前馈网络和类储层的递归网络这两类网络得到了广泛的研究。在前馈网络中,输入神经元通过随机权重连接到一个规模更大的隐藏层。在储层计算中,输入神经元连接到一个内部神经元组成的储层中,这些内部神经元之间随机连接。前馈网络的例子包括随机向量函数链接网络、径向基函数链接网络、带有随机权重的前馈网络、无反向传播算法、权重无关网络和随机卷积神经网络。储层计算中的例子包括回声状态网络、液体状态机和深度回声状态网络。

所有这些模型都有三个共同的特征:(1)隐藏层或储层创建了输入的高维表征[19],(2)连接到输出神经元的权重需要进行线性优化[17],(3)网络性能对不同随机权重的实现具有稳健性[20]。

从这些观察中可以得出的结论是,影响任务性能的是训练后的人工神经网络的架构,而不是精细调整的连接权重。更有趣的是,甚至有进一步的研究表明,架构本身甚至也可以是随机的:由随机图生成器创建的架构在ImageNet上显示出良好的分类准确率(随机连接网络为79%,ResNet-50为77%)[21]。这些观察表明,随机性并不是一个草率的操作,而可能是机器智能的基础。这一观点与上面总结的神经科学中的类似推测相呼应。随机网络的有效性和效率,以及它体现未知计算原则的潜力激发了许多人对其进行分析性研究。

生物学证据、工程实践和理论分析似乎都指向一个观点:分布式的随机网络足以实现认知功能。然而,这个结论过于简单化了,事实上,随机网络必须与其他网络特性相结合才能实现复杂的功能。这些特性包括收敛性读出[22]、可塑性[23,24]、兴奋性-抑制性平衡[10,25]和稀疏性[5,10,26]。所有这些附加特征,都是基于随机连接这一前提条件,它们对于神经环路而言是不可或缺的。

随机网络是产生神经生理学中常见的混合选择性的最简单的神经环路。尽管与“功能只能来自有组织的网络”的常识相抵触,但在过去几十年里,随机网络已在生物大脑的各种系统中被发现。与此同时,随机

性作为一种高效的计算方法，在人工智能中被用于构建人工神经网络。由于其独特性和有效性，随机网络已经吸引了许多理论研究者，来探究潜在的原则。

这些原则可以在三个概念层面上解释[27]：在计算层面上，随机网络与经过训练的神经网络一样，是通用的函数逼近器。通过发散性架构，随机网络创建了高维状态空间，在此空间中判别性解码更加灵活、可行。在算法层面上，随机网络就像计算机科学中的局部敏感哈希算法一样。这些算法可以大大节省训练深度网络所需的计算量。在实现层面上，随机网络是大脑中密集排列的神经毡中分布式网络最合理的物理实现方式。

但值得注意的是，随机网络的原则只有在与其他特性一起工作时才能完全发挥功能。过去十年间，人们对随机网络的重要性有了更多的认识，并澄清了一些关键概念，仍有更多的问题亟待解答：在计算层面上，尽管了解了维度和稀疏性的问题，但人们对随机网络中的表征仍然了解甚少。在状态空间中，内在状态流形是什么样子的？在算法层面上，用于随机采样权重的分布仍然是经验性的、任意的。那么应该如何指定这些分布？是否应该使用先验知识？生成的权重应该是固定的，还是经过缓慢的赫布型学习*？在实现层面上，大脑还具有模块化的特性，比如功能列。那么模块化应该如何与随机分布的网络结构相协调？当弄清楚这些问题时，人们对随机网络的认识将会进一步深入，届时，或许确实可以确认，随机网络代表了智能的基本原则。

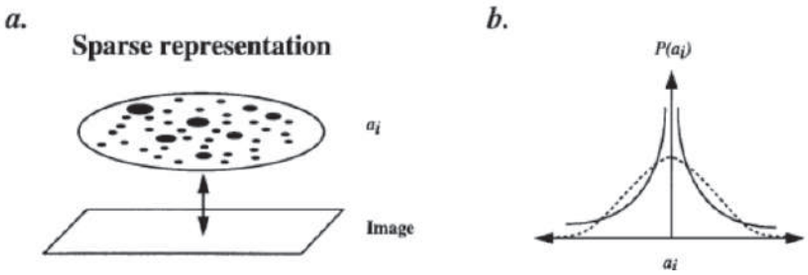
*编辑注：赫布型学习（Hebbian learning）指的是唐纳德·赫布（Donald Hebb）提出的一种神经网络学习机制，即两个邻接神经元若同时被激活，则联接权重增加的学习方法。

四、稀疏编码：大脑独特的特征

大脑是一个存储和处理信息的机器。为了实现这些功能，需要对外部信息进行准确的量化和合理的表征[28]。稀疏编码策略是实现这些目标的关键途径。大脑在多个层面上利用稀疏性机制，包括视觉、嗅觉、触觉等知觉层面[29]，讨论这些机制对于理解神经系统组织原则和智能形成至关重要。

稀疏编码的含义是：在任何给定时刻，发放神经元的数量仅占总神经元数量的一小部分（图4）[30]。“稀疏”本身只是一个相对的概念，没有明确的阈值，在与另外两种更极端的编码方案——局部编码和密集编码——比较时，其优势更加凸显[31]。局部编码，又被称为“one-hot”编码：每个神经元仅编码一个物体，任何两个物体的表征没有重叠，“祖母细胞”*便是最著名的例子。另一种极端编码情形是密集编码，又被称为完全分布编码：每个物体由神经元群体中所有神经元的联合活动来表征。而稀疏编码处于以上两种情形之间，同时拥有两者的优势[32]。

*编辑注：祖母细胞（grandmother cell）是20世纪60年代末一些科学家提出的视觉系统中一种假想的功能高度特异化的细胞。主要是对主体所熟悉的某些复杂对象的图像（如自己的祖母的不同照片）才有剧烈的反应，而对其他对象反应很小。这是对图像稀疏编码的一种极端形式的假设。



▷图4:稀疏编码示意图。在稀疏编码机制中,一幅图像由一小部分被激活的元件所表征,且被激活的原件分布随图像不同而有所差异(见图a)。另外,对于单一元件而言,由于稀疏编码过程中该元件大部分时间会处于静息态,这会导致其活动概率分布图上出现一个“峰”和两个长“尾”(见图b)。图源:关联论文

稀疏编码在编码能力、能源效率和解码难度之间起到了很好的权衡作用。局部编码不允许重叠,一个含有N个二进制神经元的群体最多可以表征N个不同的项目。编码更多物体时需要招募更多神经元,从而消耗更多能量,而大脑可用的能量是有限的,因此为局部编码设定了上限。相反,密集编码允许N个二进制神经元编码2N个物体,显著提高了表征能力。在稀疏编码中,即使只有少数(最多K个)神经元可以同时为一个物体所激活,可以编码的物体总数也可以达到,与局部编码相比,在编码相同信息片段时消耗的能量要少得多。分布式编码的困难之处在于读取,需要以生物学上合理的方式学习。但局部编码及其输出之间的关联可以利用简单的赫布学习机制来建立;因此,如果神经活动模式是稀疏编码的,学习会变得更有效[30,33]。

其次,稀疏编码还平衡了泛化和抗干扰之间的关系。在局部编码中,每个模式与其他模式之间是正交的,不同模式之间没有相似性,因此不可能从一个模式泛化到另一个模式。密集编码和稀疏编码允许部分重叠和不同级别的编码相似性,使得具有相似编码的项目之间可以进行泛化。然而,密集编码决定了许多物体(如果编码空间完全被占用,最多可以达到所有物体的50%)可能会激活同一个神经元,这种情况可能会导致不同发放模式之间的干扰[34]。稀疏编码则可以帮助应对这种灾难性遗忘[35],并减少模式之间的干扰[36]。在极端情况下,局部编码使得多个物体可以同时表征,相互之间完全不会受到干扰。

最后,稀疏编码显式地表征了刺激的自然结构,神经元响应有明显的调谐曲线式的反应。感受野复刻了环境中会遇到的频率结构,使得仅有一小部分神经元也可以表征一个自然刺激。结合过完备基*,稀疏编码可能会产生曲面流形的分段平坦表征,这个流形是自然刺激聚集的,从而简化了后续阶段的表征和分析[30]。这些优势支持生物体对信息进行更高效的编码、传输和存储。

*编辑注:过完备基(overcomplete basis)通常指的是一个向量空间的基,其中包含比必要的基向量更多的向量。这种情况可能导致冗余和过度表示,但在某些情况下也有其用处,例如在压缩感知和稀疏表示领域。过完备基的概念在信号重构、降维和特征选择等应用中起着重要作用。

在动物中,人们也广泛观察到了稀疏编码的存在,这些证据存在于各种感觉系统以及与运动、记忆有关的脑区中[29,37-39]。科学家们普遍认为,神经生物学中稀疏编码信号的实现可能是非常普遍的。

稀疏性和稀疏编码一直受到不同领域研究人员的关注。Horace Barlow于1961年提出了有效编码假设*[40],随后提出稀疏性可能是感知表征的基本原则[41]。其他研究表明,自然图像可以进行稀疏编码,而这种编码特性与V1区神经细胞的响应非常相似[42]。

*编辑注:有效编码假设(valid encoding assumptions)是Horace Barlow于1961年为脑的感知编码提出的一个理论模型。他认为,生物视觉系统初级阶段的一个重要功能就是尽可能地去除输入刺激的统计冗余。

许多人努力理解和解释稀疏性的潜在机制及其相关的生物学意义。而稀疏性在推动机器学习和智能算法发展方面的作用也受到关注，人们从多个方面探索了稀疏性和稀疏编码的优势，包括但不限于编码能力、稳健性和泛化性、压缩感知以及信息传输效率。这些研究促成了字典学习算法（dictionary learning algorithms）的发展，以及像分层时序记忆（Hierarchical Temporal Memory）这样利用稀疏性进行神经计算的新算法的出现。

正如Suryaz Ganguli等人所说的那样，“对任何神经系统来说，存储、传递和处理高维神经活动模式或外部刺激都是根本性的挑战。”处理和学习外部信息是神经系统的基本任务。此外，高维信息在本质上往往是稀疏的。稀疏编码策略可能是生物大脑处理外部信息的一种必要且可行的方法，可以提高处理效率和稳健性。

五、关系记忆：神经群体编码和流形

各种各样的信息充斥在人们的生活中，人们需要从中汲取生活经验并将其存储在大脑中。知识存储是重要的认知能力，也是神经科学和计算科学重要的研究问题。最近的研究发现，大脑的记忆系统以参考框架（reference frame）的形式在内侧颞叶（medial temporal lobe，MTL）精确地组织和存储不同信息之间的关系。目前，参考框架已经在空间记忆和非空间记忆的实验中被观察到，且驱动了一个新的研究方向——关系记忆的产生。

（1）长期以来关于MTL功能的争论

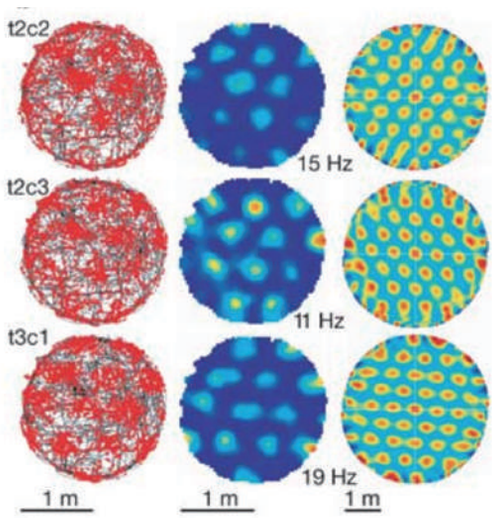
MTL由海马和嗅周皮层、内嗅皮层和旁海马皮层组成[1]。早期通过对病人的研究发现，双侧内侧颞叶的切除和海马CA1区双侧的损伤均会导致短期记忆的丧失，但是长期记忆完好。这说明MTL在记忆巩固中起重要作用，海马损伤会导致学习经验不能从短时记忆转化为长时记忆[2]。短时记忆和长时记忆都被归为陈述性记忆。空间记忆在当时一直被忽视。

在20世纪70年代，John O'Keefe及其同事在大鼠海马中发现了位置细胞，人们才对大脑如何编码空间信息有了一些理解[3]。位置细胞会根据大鼠处于空间中的特定位置选择性反应，即只有当大鼠处于空间中的特定位置时，相应的位置细胞才做出反应。这表明海马中的位置细胞能够对位置信息进行编码，而由海马中的位置细胞群构成的认知地图系统（cognitive map system）可以在脑内形成空间参考地图（spatial reference map）[4]。

（2）MTL系统通用的编码方式：参考框架

之后，May-Britt Moser和Edvard I. Moser和同事们在大鼠的背尾内侧内嗅皮层（dorsocaudal medial entorhinal cortex，dMEC）植入电极，找到了一些发放场（firing field）呈现显著的空间结构的神经元[5]。这些神经元被称为“网格细胞”，它们的发放场在整个环境中规律地分布于六边形的顶点处（图5）。这说明大脑中存在地图状结构，这些地图是由网格细胞组成的。接着，Alexandra O Constantinescu等人证实了网格细胞不仅仅编码物理空间，还编码抽象的概念空间（conceptual space），在他们的实验中，当被试进行一个二维的关于“bird space”（由连续变化的鸟的脖子和腿长构成的二维空间）的任务时，在其MEC也观察到与网格细胞类似的神经表征[6]。这些研究都支持Eichenbaum提出的关系处理理论（relational processing theory），即MTL展示了一种通用代码，能够将非空间知识组织到一个参考框架中，存储的知识可以通过全局关系进行存储或检索，类似于在物理

空间中组织空间关系。在2014年，诺贝尔生理学或医学奖授予了John　O’Keefe和Moser夫妇以表彰他们发现了大脑中编码空间坐标系统的位置细胞和网格细胞。这些细胞的发现为研究大脑如何组织各种日常经验提供了思路。



▷图5：大鼠在一个圆形场中奔跑时，三个网格细胞的放电情况。左图显示大鼠的轨迹（黑色），上面叠加了尖峰位置（superimposed spike location，红色）。中间图显示彩色编码的速率地图，用峰值速率表示。红色是最大值，深蓝色是零。右图显示每个速率地图的空间自相关性。图源：关联论文

James　Whittington及其同事于2020年首次成功实现了MTL的统一框架，建立了名为Tolman-Eichenbaum　Machine（TEM）的参考框架系统[7]。他们通过将空间记忆和关系记忆都视为结构的抽象和示例的概括，来解释海马既能表征物理空间又能表征抽象空间的特性[7]。空间推理可以被看做是对于结构的概括，因为不同的空间环境共享欧几里得空间的常规性质，这些性质定义了可以进行哪些推断，以及可能存在哪些途径[7]。例如，以当前位置向前或者向后移动可以使你返回起始位置。

六、基于参照框架的知识存储与检索的群体编码

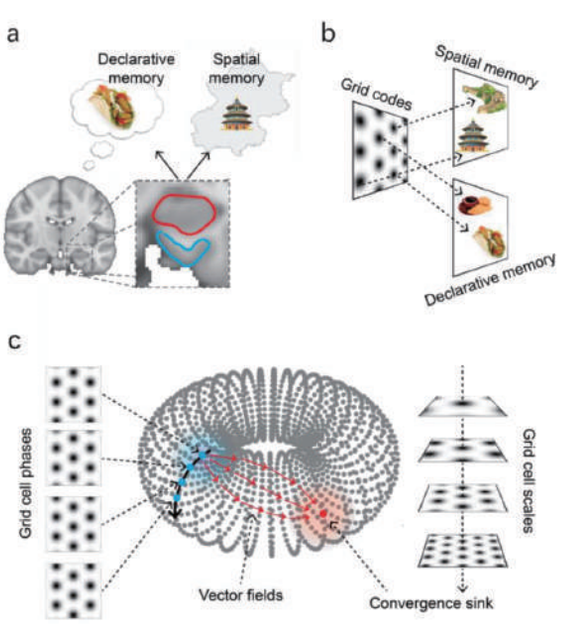
结构性的规律在非空间关系问题中也同样适用。例如，传递性推理问题可以将刺激物表示在一个抽象的有序轴上，由A>B和B>C可以推理出A>C[7]。类似地，对分层结构的抽象使得人们在遇到新的社交情境时可以进行快速推断，例如已知小A（男）和小B（女）是兄妹，小C是小A的女儿，由此可以推断出小B是小C的姑姑。这种结构性概括非常有利于学习新的知识和进行灵活推断，是人工智能领域的一个关键问题。TEM采用“分解和连接”的方法对知识进行结构泛化，从而可以在空间和非空间记忆任务中学习泛化的神经表征并用于预测，即TEM假设知识的各个方面是单独表征的，可以被灵活地重新组合以表征新的知识。同时，在查看TEM网络的学习表征时，人们也观察到了类似于MEC中的网格细胞，并证明了TEM用网格细胞表示结构，且在不同环境中通用，在不同环境中通过位置细胞的重新映射形成记忆[7]。

从理论上讲，参考框架能够以连续和定量的方式尽可能多地存储与特定特征对应的知识。单一维度的参照系不可能存储由多特征组成的知识。为了灵活记忆和应用知识，“传递性”和“对称性”是十分重要的。“传递性”表示判断具有共同特征的知识对的能力，“对称性”表示将以相反事件顺序呈现的知识对联系起来的能力。为了满足多样性的需求，大脑必须使用群体编码来连接来自多个维度的知识片段。

近年来，神经科学和计算科学的研究主要集中在神经元放电速率的拓扑结构上，结果表明，神经元群体能够协调嵌入参考框架的知识中（即知识存储），并在复杂环境中应用（即知识检索）。从群体编码的动态角度来看，这些研究为揭示神经元如何有效地相互作用以存储和检索关系记忆的原理提供了方向，这是一个几十年来仍未解决的经典问题。

对神经相互作用机制的研究最初来自计算建模领域，如Shun-ichi Amari提出的竞争-合作机制（competition-cooperation mechanism），该机制假设一个周期权重函数形成一系列吸引子[8]。在空间导航场景中，每个吸引子以二维欧几里得空间的某个位置为中心，然后通过周期函数的激励或抑制确定神经交互作用。吸引子网络中的“连续吸引子神经网络（CANN）”成功地模拟了位置细胞和网格细胞的动态模式，使空间知识能够稳定地被细胞编码。在CANN的基础上，Alexei Samsonovich和Bruce L. McNaughton[9]进一步提出，网格细胞群的放电速率依赖于被称为“环面”的拓扑结构（图6c）。与欧几里得空间的运动最终有界不同，环面结构没有边界来适应网格细胞的周期性模式，因此从环面上的任何位置开始的运动永远不会有界，而是最终返回原点。

在2022年，Moser夫妇及同事们提供了首个生理证据，证明了这种环状结构的存在[10]。在他们的实验中，当大鼠从事觅食行为或睡觉时，数百个网格细胞同时被高密度Neuropixels硅探针记录下来。在对单个尺度内149个网格细胞的放电率进行降维并将主要成分转换到3D空间可视化后，一个环状结构清晰地显示出来，并且在动物清醒或睡眠状态下保持稳定。在环面结构中，网格细胞群被组织为一个整体，环面的内圆和外圆分别对应于欧几里德空间的水平轴和垂直轴。因此，环面上的每个位置代表一个具有唯一相位对（分别对应于内圆和外圆）的网格单元，它无缝地编码给定空间模块下的物理空间。在给定拓扑形态的情况下，这直接证明了“数十万个参照系是由超越欧几里得空间的高维拓扑空间灵活组织的”这一假设。



▷图6. a）MTL中海马（红色轮廓）和内嗅皮层（蓝色轮廓）的解剖学位置，两者既编码陈述性记忆又编码空间记忆。b）内嗅皮层中六边形的网格细胞充当了存储知识的参考框架。c）网格细胞群体的几何结构（环面），其中每个环面位置表示一个具有独特相位的网格细胞（左列）。蓝色和红色阴影区域分别代表起始位置（蓝色圆圈）和目标位置（红色圆圈），这两个位置都由多尺度网格细胞群体决定的独特活动强度编码（右列）。目标位置处具有更强的活动，形成了通过向量场驱动路径积分的活动梯度（红色箭头）。图源：关联论文

从动力学角度看，环形结构表面的运动对应着一系列知识检索的过程。然而，每次移动决定下一个“位置”（例如，从A到B，然后从B到C）的原则是什么?受网格细胞的生理证据启发，研究人员发现dMEC背侧至腹侧的模块大小呈梯度分布，网格细胞相位呈随机分布，于是提出了大脑是模块化系统的假设[5, 11]。该系统说明了相位和位置之间的映射关系，动物的位置可以由一组相位唯一指定，

符号mod表示取模操作，x和λ分别表示当前位置和模块大小的梯度。理论上，网格细胞群在一个空间中最多编码个位置。在模块化系统的基础上，Andrej Bicanski和Neil Burgess提出的赢者通吃机制（winner-take-all mechanism）[12]，当多个神经元或神经元群体接收到输入信号时，这些神经元会相互竞争，以确定哪一个神经元或神经元群体会被激活[12]。竞争的结果是，拥有最强刺激或最高激活水平的神经元将被激活，而其他神经元则被抑制[12]。因此，赢者通吃机制通过激活最大的网格单元的相位以形成从当前位置到下一个位置的向量。

在2022年，O’Keefe和他的同事发现了支持这一算法的证据[13]。他们对5只大鼠的456个CA1区位置细胞进行记录，发现当大鼠在蜂巢迷宫中导航到一个有食物奖励的目标地点时，有142个细胞的放电模式显示出向量场（vector field）聚集在目标附近的位置上。将细胞间的向量场相加，可以从群体向量图中清楚地看到动物朝向目标时的最大发射率。此外，向量场还可灵活适应目标位置，即当原目标位置移动时，种群编码的最大发射率会朝着新目标重新组织。这些结果直接支持了赢家通吃机制，即海马-内嗅皮层（HPC-ERC）系统创建了一个基于向量的模型来支持灵活的导航。

近年来，随着对向量导航中的群体神经元编码的理解，机器学习领域的神经网络在记忆识别和空间导航方面取得了巨大成功。Bicanski和Burgess开发了一种视觉记忆模型[14]，利用网格细胞群驱动扫视，从而识别熟悉的面孔、物体和场景。在该模型的导航部分，图像上的每个凝视位置由9个空间尺度上的100个网格细胞的发射率的唯一向量表示。根据当前位置和给定的随机目标位置，使用赢者通吃机制可确定扫视的位移向量。最后，利用赫布关联*建立图像特征与位置细胞、位置细胞与网格细胞、网格细胞与位移向量之间的权重关系，成功模拟海马的导航功能，引导眼球运动。Andrea Banino等人在研究中训练一个递归网络在2.2 m × 2.2 m的正方形舞台上进行基于向量的导航，结果发现该网络的行为就像哺乳动物一样，具有精确的性能，这表明网格细胞群在编码非空间和空间知识方面都具有很大的能力[15]。

*编辑注：赫布关联（Hebbian association）用于描述神经元之间的突触权重如何随着时间的推移进行调整，以促进信息传递和学习。其基本思想是，当细胞A频繁和持续地参与到细胞B的兴奋性激活，就会增强或保持其连接强度。

七、参考框架系统

虽然参考框架的某些方面，如群体编码的作用已经被揭示，但要充分了解参考框架的本质，如参考框架的多样性、可塑性或脑区之间的相互依赖所起的功能，还需要做更多的工作。此外，参照框架的研究至少在以下两个方向，可以促进大脑机制研究向人工智能研究的转化[16]。一个是图论的发展，参考框架可以用来解决子图同构问题，而赢者通吃机制可以帮助解释路由问题*。更重要的是，群体编码将有助于从生物学角度理解节点和边缘的概念。另一个是认知推理和判断，通过这种方法，大脑不仅可以检索存储的知识（即判别模型），还可以产生超越经验的新知识（即生成模型）。例如，大脑可以利用“想象力”的能力，反复模拟未来的计划，直到找到最优的解决方案。综上所述，未来对参考框架的研究将有助于理解知识在脑内的灵活安排，这是揭示关系记忆机制和开发AI通用知识编码框架的关键[16]。

*编辑注：子图同构问题指两个图之间存在一种对应关系，使得一个图的一部分可以通过重排列和重新标记节点等方式，与另一个图完全相同。

路由问题是在计算机网络和通信领域中经常遇到的一个问题，涉及在网络中选择最佳路径以使数据从源节点传输到目标节点。在一个复杂的网络中，数据包需要通过多个中间节点进行传递，而路由问题就是确定这些中间节点的路径，以确保数据能够高效地到达目标。

八、神经可塑性：感知学习的经验

大脑是一个庞大有序的动态神经网络系统，几乎参与了所有重要的生命活动。在这个神经网络之中，经验可以诱导神经元和神经回路的功能发生特异性变化，这种变化被称为大脑可塑性，有助于个体适应环境。感知学习（perceptual learning）作为一种典型的感知系统适应外部环境变化的现象，是指通过反复的实践或经验，对物理刺激的感知发生持续而稳定的变化[17-19]。它表现为在几个月或几年的时间里，识别感知特征和物体的能力逐渐无意识地提高。这种感知能力的增加伴随着大脑多个结构和功能水平的神经变化，为研究大脑可塑性提供了一个很好的范例。

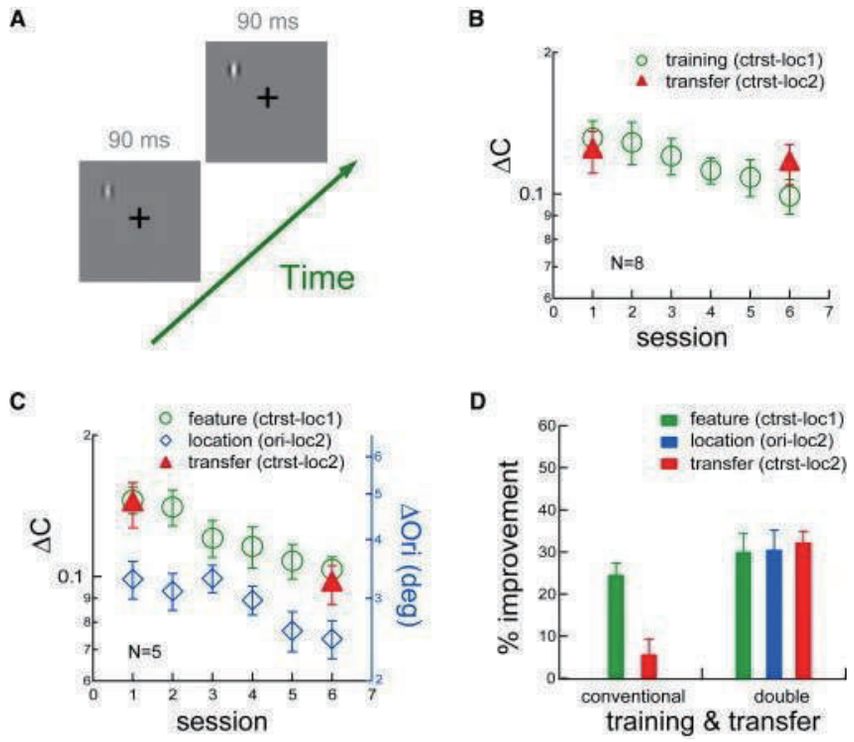
九、感知学习的特异性与迁移

传统观点认为，在个体发展的早期阶段，感知系统具有高度的可塑性。然而，这并不意味着成年人的大脑功能已经固化。感知学习是指通过实践而产生的长期的表现改善，它被广泛用作研究成人经验依赖的大脑可塑性的范式[17-19]。与一般意义上的学习不同，感知学习不获得显性知识，而是与内隐记忆相关，表现为区分或识别感知特征和对象的能力增加。感知学习可以以多种方式发生，如视觉、听觉、嗅觉、触觉和味觉。本文将聚焦于视觉感知学习。

*编辑注：显性知识是指能够有意识地表达、讨论和分享的知识，通常通过语言和符号进行传递和交流，例如数学公式和历史事件。

内隐记忆又称非陈述性记忆、程序记忆、反射记忆等，是指与直觉和意识无关的记忆。需要反复从事某种技能的操作，经过反复的经验积累才能缓慢地被保存下来的记忆。一旦建立，可保存较长时间，不再需要意识的参与。例如骑自行车、打字、游泳等运动技能。

感知学习具有特异性并且可迁移。Aniek Schoups等人在定向辨别任务中，精确评估视网膜定位特异性。被试首先在中央处进行了练习，然后在周围5°环状区域的一系列位置进行了练习[20]。结果表明，感知学习能力的提高对周边位置具有特异性。且在不同周边位置的学习训练的表现都优于在中央处的表现，可以说明感知学习得到的能力存在从中央到周边位置转移效应。另一方面，一系列研究表明，感知学习的特异性可以减少甚至完全消失。例如，余聪实验室使用双重训练范式，让被试在一个位置进行判别光栅对比度的训练，在第二个位置进行判别光栅方向的训练，实验结果显示被试在第二个位置进行对比度测试时，表现也较好（图7）[21]。这种感知学习的转移表明，感知学习过程可能还涉及高级皮层，以便对不同刺激进行更复杂的处理。



▷图 7. 用传统和双重训练范例研究视网膜位置特异性要。图源：参考文献21

研究人员使用单细胞和功能磁共振成像（fMRI），研究了感知学习机制，揭示在大脑的多个阶段发生的与学习相关的变化，协调感知学习的特异性和转移之间的冲突。

与感知学习的特殊性相对应，大量的电生理和fMRI研究表明，感知学习与早期视觉皮层区域的活动增强有关。2001年，Schoups对恒河猴的研究发现，经过训练的神经元表现的提高与定向识别行为的提高有关[22]。他们在经过训练的神经元中，观察到特定而高效的调制曲线斜率增加，表明这些神经元很可能编码已鉴定出的方向。同时，对于未经训练的方向，调制曲线没有发生变化。为了研究人类视觉皮层在感知学习过程中的激活变化，Yuko Yotsumoto等人在一项fMRI研究中采用了纹理判别任务（texture discrimination task）[23]。结果显示，V1区相应的BOLD激活随着被试的行为表现而增加，而在表现饱和时则减少。这表明，感知学习导致早期视觉区域响应增加是由于局部感受野的变化，而非来自高级视觉区域的反馈。此外，一项应用神经反馈技术的核磁共振成像（MRI）研究通过去除外部视觉输入，仅依靠来自视觉皮层神经信号进行训练，改善了行为表现，支持初级视觉皮层在感知学习中发挥决定性作用的观点[24]。之后，一些研究暗示，感知学习引起的神经可塑性可能早在丘脑层次就开始发生了[25]。

然而，目前的研究还发现，与感知学习相关的大脑区域与注意力和决策有关[18,26-28]。具体而言，感知学习被发现与前顶叶区域中的顶枕沟（intraparietal sulcus, IPS）、外侧顶叶区和前扣带回的区域选择性增强或神经响应减弱相关[29-31]。逆向层次理论（reverse hierarchy theory）主张，控制学习过程的是自上而下的信息流，而不是自下而上的信息流[32]。学习首先发生在特定任务的较高的神经层次，然后在必要时在较低神经层次实现。在一定程度上，这一理论解释了感知学习中转移的发生。同时，一些研究还指出，感知学习与视觉和决策相关区域之间的功能连接增强有关[33,34]。再权重理论（reweighting theory）表明，感知学习不会改变早期视觉皮层的功能特性；相反，它会改变表征视觉信息的神经元和决

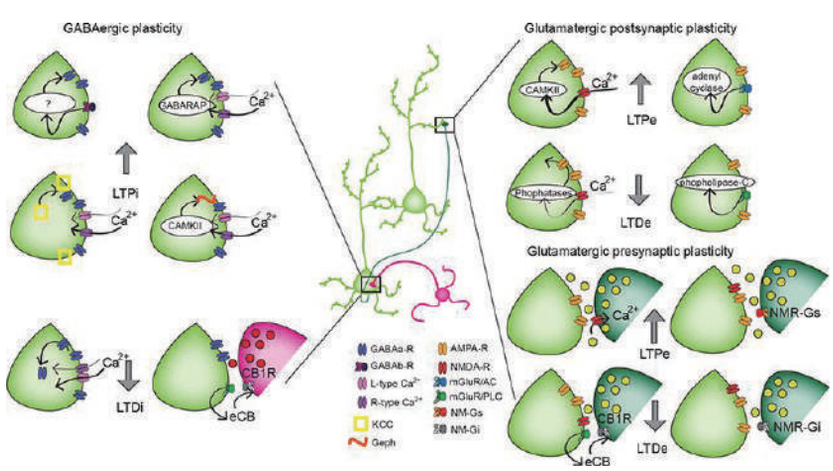
策相关的神经元之间连接(权重)的强度[35]。

基于该理论,Chi-Tat Law和Joshua Gold构建了一个模型,将学习视为一个高级决策神经元与感觉神经元的连接权重在特定训练方向上进行精细调整的过程[30-34]。一个关于运动辨别的感知学习研究显示,行为的改善可以通过V3A区中对于刺激的感觉表征和从V3A到IPS的连接性变化来解释[33]。同样,为了研究感知学习如何调节人脑中与决策相关区域的活动,研究人员使用运动方向辨别任务,采用了一种基于模型的方法进行探究[36],发现除了在前顶叶网络和决策网络中有增强的神经响应外,从V3A到腹侧前运动皮层和从IPS到额叶眼区的连接增强与训练呈正相关。总之,感知学习的机制并不是一个简单的过程,而是多个脑区之间的复杂相互作用,因此在不同条件下可以观察到感知学习的特异性和转移。

GABA是大脑中参与神经元抑制性调节的分子。以前的动物研究已经证明,GABA能在学习和突触可塑性中发挥重要作用[37,38]。人类的磁共振波谱成像(MRS)研究揭示出,视觉皮层中的GABA浓度与稳态可塑性相关[39],而运动皮层中的GABA浓度与个体能力和运动学习中表现的改善相关。

十、感知学习的赫布规则和计算模型

长时程增强(Long-term potentiation,LTP)和长时程抑制(long-term depression,LTD)分别由高频间断性刺激和低频间断性刺激诱发,都允许双向突触改变,被认为是学习和记忆的突触基础[40-42]。此外,脉冲时序依赖可塑性(spike timing–dependent plasticity,STDP)表明突触改变的方向取决于突触前和突触后脉冲的时间顺序[43]。具体来说,重复暴露于刺激会影响前后突触激活的时间,突触强度会发生变化。这些变化可能会减短突触潜伏期并减少神经元的激活时间。



▷图8. GABA能(左)和谷氨酸能(右)形式的突触可塑性的LTP和LTD机制概要。图源:https://doi.org/10.1093/acrefore/9780190264086.013.148

有人认为LTP可能是感知学习的基础。Sam F. Cooke和Mark F. Bear通过在切片中进行电生理记录,提供了直接的证据,表明感知学习影响了大鼠V1区突触反应的LTP[44]。众所周知,在体内研究突触可塑性是具有挑战性的。在一项人类行为研究中,类似LTP的被动高频刺激会改善被试在亮度变化检测任务中的表现,而类似LTD的被动低频刺激则损害了被试的表现[45]。

基于感知学习的现象和机制,再权重理论被提出来解释和预测实验结果。该理论认为,感知学习是处于与任务相关的决策集成的层次上[35]。感知学习在视觉系统的不同层次上改变了表征的权重。其原则是增强视觉区域的神经元协调曲线,加强学习分类的结构连接,并减少与任务无关的通道的输入。已经证

明，基于这一理论的相应增强型赫布再权重模型（augmented Hebbian reweighting model）能够生成与实验数据相媲美的泛化模式[46]。

建模研究还表明了GABA在感知学习中的重要性。在2011年，Osamu Hoshino等人使用皮层神经网络模型（cortical neural network model），研究GABA在形成感知学习记忆痕迹方面的作用[47]。他们通过一种新颖的神经元-星形胶质细胞网络模型，研究了星形胶质细胞的缝隙连接通信对感知学习的影响[48]。结果显示，局部的GABA的同步增强了STDP水平，通过影响学习过程中留下的记忆痕迹，促进感知学习。

十一、新技术和新见解

大脑的学习是一个复杂而灵活的过程，既涉及基于赫布规则的突触强度变化，也涉及多个脑区之间高层次的交互协调。理想情况下，人们希望以高时空分辨率全面观察特定行为（例如学习）背后的神经活动。但是目前的技术难以同时满足高时空分辨率的需求，借助单个神经元记录、EEG/MEG记录、fMRI采集和计算模型，人们可初步看到学习和神经可塑性之间的内在关联[16]。赫布规则已在从昆虫到人类的各种神经回路中得到证实，并为构建具有学习功能的神经网络模型奠定了基础。众所周知，当前人工神经网络经过训练后，节点之间的连接权重通常是固定的，这使其在转移和功能补偿方面与人类学习有所不同。在人类感知学习中，转移和特异性之间通过一种不明确的机制达到平衡。转移和特异性之间的平衡，使大脑成为一个高效而节能的智能中心。未来的研究可能可以从人脑中收集直接证据，揭示学习的突触水平机制[16]。人工神经网络可能会从中获益，进化出更强大的学习策略。

这次调研从吸引子网络、临界性、随机网络、稀疏编码、关系记忆、感知学习六个角度对大脑基本原理和作用机制进行了阐述，进一步分析了如何运用这些发现，使人工智能更“像”人类智能。研究大脑基本原理、开发下一代人工智能是一项规模巨大的工作，需要认知科学、计算机科学、心理学等多学科通力合作，希望这次调研能够起到抛砖引玉的作用，启发研究者们描绘出人工智能与人类智能的渐近线，为创造下一代人工智能提供思路。（编辑：韵珂）

参考文献

关联论文：Chen, Luyao, et al. "AI of brain and cognitive sciences: from the perspective of first principles." arXiv preprint arXiv:2301.08382 (2023).

大语言模型是如何发展而来的?

编译:吴婷婷

中国科学院神经所在读博士生。研究兴趣:大脑是如何实现视觉想象的呢?大脑是如何更新或维持工作记忆中的内容以满足实际需要?

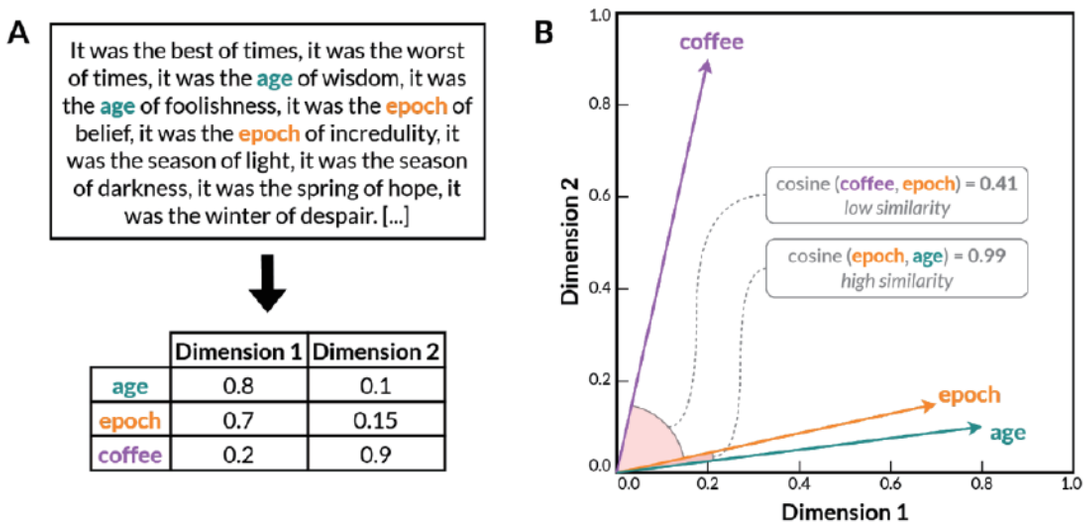
扫码查看原文

LLMs的起源可以追溯到人工智能研究的开始。早期的自然语言处理(natural language processing, NLP)主要有两大流派:符号派和随机学派。Noam Chomsky的转换生成语法对符号派影响重大。该理论认为自然语言的结构可以被一组形式化规则概括,利用这些规则可以产生形式正确的句子。

与此同时,受香农信息论的影响,数学家Warren Weaver首创了随机学派。1949年,Weaver提出了使用统计技术在计算机上进行机器翻译的构想。这一思路为统计语言模型的发展铺平了道路,例如n-gram模型,该模型根据语料库中单词组合的频率估计单词序列的可能性。

现代语言模型的另一个重要基石是分布假设(distributional hypothesis)。该假设最早由语言学家Zellig Harris在1950年代提出。这一假设认为,语言单元通过与系统中其他单元的共现模式来获得特定意义。Harris提出,通过了解一个词在不同语境中的分布特性,可以推断出这个词的含义。

随着分布假设研究的不断深入,人们开发出了在高维向量空间中表示文档和词汇的自动化技术。之后的词嵌入模型(word embedding model)通过训练神经网络来预测给定词的上下文(或者根据上下文填词)学习单词的分布属性。与先前的统计方法(如n-gram模型)不同,词嵌入模型将单词编码为密集的、低维的向量表示(图1)。由此产生的向量空间在保留有关词义的语言关系的同时,大幅降低了语言数据的维度。同时,词嵌入模型的向量空间中存在许多语义和句法关系。



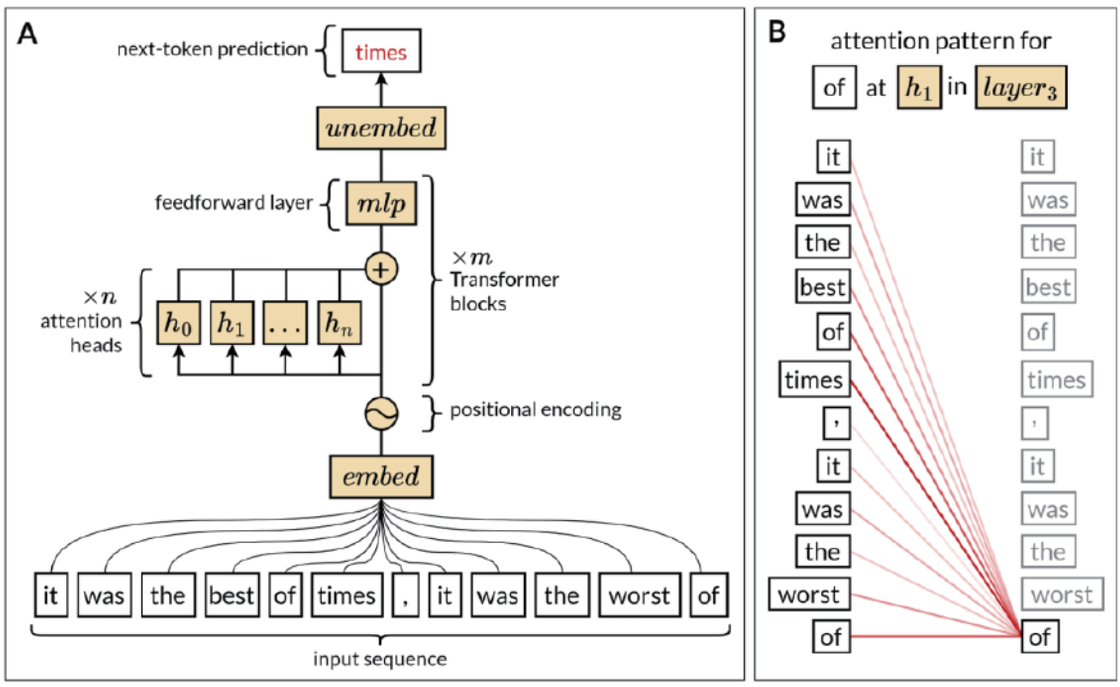
▷图1. 多维向量空间中词嵌入的一个例子。A. 一个在自然语言语料库上训练的词嵌入模型学会将单词编码成多维空间中的数值向量，为了视觉上的清晰性而简化为两维。在训练过程中，上下文相关的单词（例如“age”和“epoch”）的向量变得更加相似，而上下文无关的单词（例如“age”和“coffee”）的向量变得不那么相似。B. 在经过训练的模型的二维向量空间中的词嵌入。具有相似含义的单词（如“age”和“epoch”）被放置在更靠近的位置，这由它们的余弦相似度得分高度表示；而具有不同含义的单词（如“coffee”和“epoch”）则相对较远，反映在余弦相似度得分较低上。余弦相似度是一种用于确定两个非零向量夹角余弦的度量，反映它们之间的相似程度。余弦相似度得分越接近1，表示夹角越小，向量之间的相似度越高。（图片引自[3]）

词嵌入模型的发展是NLP历史上的一个转折点，为基于在大型语料库中的统计分布在连续向量空间中表示语言单元提供了强大而高效的手段。然而，这些模型也存在一些显著的局限性。首先，它们无法捕捉一词多义和同音异义，因为它们为每个单词类型分配了单一的嵌入，无法考虑基于上下文的意义变化。

随后的“深度”语言模型引入了类似记忆的机制，使其能够记住并处理随时间变化的输入序列，而不是个别的孤立单词。这些模型虽然在某些方面优于词嵌入模型，但它们的训练速度较慢，处理长文本序列时表现也欠佳。这些问题在Vaswani等人于2017年引入的Transformer架构中得到解决，Transformer架构为现代LLMs奠定了基础。

Transformer架构的一个关键优势在于，输入序列中的所有单词都是并行处理，而不是像RNN、LSTM和GRU那样顺序处理。这种架构不仅极大地提高了训练效率，还提高了模型处理长文本序列的能力，从而增加了可以执行的语言任务的规模和复杂性。

Transformer模型的核心是一种被称为自注意力（self-attention）的机制（图2）。简而言之，自注意力允许模型在处理序列中的每个单词时，衡量该序列不同部分的重要性。这一机制帮助LLMs通过考虑序列中所有单词之间的相互关系，构建对长文本序列的复杂表示。在句子层面之上，它使LLMs能够结合段落或整个文档的主题来进行表达。



▷图2. A. LLM的自回归模型变体的结构体系。来自输入序列的tokens首先被嵌入为向量，这涉及将每个token转换为一个高维空间，其中在语义上相似的token具有相应相似的向量。位置编码将关于每个tokens在输入序列中位置的信息添加到向量中。然后，这些丰富的向量通过连续的Transformer块进行处理。每个块包含多个attention heads，可以并行处理所有向量，以及一个全连接的前馈层，也称为多层感知机（multilayer perceptron，MLP）层。最后，在取消嵌入阶段，向量经历线性变换，将它们投影到一个与词汇大小相同的空间中，生成一组Logits。这些Logits表示词汇中每个潜在下一个token的未归一化分数。然后应用柔性最大值传输函数层，将这些逻辑转换为对词汇的概率分布，指示每个token成为序列中下一个token的相对可能性。在训练过程中，已知正确的下一个token并用于反向传播，而在推理过程中，模型在没有此信息的情况下预测下一个token。可以迭代地以自回归方式重复此过程，以生成多个token的预测。B. 自注意机制的可视化。每个attention head为序列中的每个标记ti分配权重或注意力分数，该分数适用于包括ti在内的序列中的每个标记t0-i。在这里，每条红线表示‘of’与输入序列中的每个其他标记之间的注意力分数，包括自身。在此示例中，注意力分数量化了每个标记相对于标记‘of’的相关性或重要性，较粗的线表示较高的分数。这个模式说明了注意力机制允许模型动态关注输入序列的不同部分，以得出每个标记的具有上下文细微差别的表征。每个attention head的注意力模式都不同，因为在训练过程中，每个头专门关注于选择性地关注标记之间的特定依赖关系。（图片引自参考文献[3]）

值得一提的是，Transformer模型并非直接操作单词，而是操作称为tokens的语言单位。tokens可以映射到整个单词，也可以映射到更小的单词片段。在将每个单词序列提供给模型之前，首先进行标记化，将其分块成相应的tokens。标记化的目的是尽可能多地表示来自不同语言的单词，包括罕见和复杂的单词。

基于Transformer模型的最常见变体被称为“自回归”（autoregressive）模型（图2），包括GPT-3、GPT-4和ChatGPT。自回归模型以准确预测下一个token为学习目标。在每次训练时，模型的目标是根据先前的tokens预测语料库中抽样序列的下一个token。在第一次预测时，模型使用随机参数初始化，预测结果并不准确。随着每次预测的进行，模型的参数逐渐调整，直至预测出的token和训练集中实际的token的差异最小。这个过程重复数十亿次，直到模型能够准确预测从训练集中随机抽取的内容的下一个token。Transformer模型的训练集包括百科全书、学术文章、书籍、网站，甚至大量计算机代码等多样化来源的大型语料库，旨在概括自然语言和人工语言的广度和深度，使Transformer模型能够准确进行下一个tokens的预测。

尽管这种方式训练的LLMs在生成文本段落方面表现出色，但它们对真实的、有用的或无冒犯性的语言没有固定偏好。为了让生成的文本更符合人类语言使用规范，近期的LLMs，如ChatGPT，使用了“从人类反馈中进行强化学习（RLHF）”的微调技术来调整模型的输出[8]。RLHF允许开发人员更具体和可控地引导模型的输出。这一微调过程在调整这些模型以更好地满足人类语言使用规范方面发挥着至关重要的作用。

LLMs具有出色的能力，能够利用文本提示中的文本信息来引导它们的输出。已部署的语言模型经过预训练，其参数在训练后保持固定。尽管大部分架构缺乏可编辑的长期记忆资源，但它们能够根据所提供的内容灵活调整输出，包括它们未经明确训练的任务。这种能力可被视为一种即时学习或适应的形式，通常被称为“情境学习”（in-context　learning）[9]。情境学习可被解释为一种模式完成的形式，如果序列构造为一个熟悉的问题或任务，模型将尝试以与其训练一致的方式完成它。可向模型发出具体的指令。

在所谓的“少样本学习”中，提示的结构包括要执行的任务的几个示例，后面跟着需要响应的新实例。在“零样本学习”中，模型不会得到任何示例，任务直接在提示中进行概述或暗示。少样本学习长期以来被认为是人类智能的重要方面。而老式机器学习则在少样本学习任务中表现较差。然而，经过训练后的LLMs在少样本学习上表现出色。在较大的模型（如GPT-（3）中观察到，少样本学习能力似乎与模型大小高度相关[9]。通过RLHF精调后，LLMs的零样本学习能力得到增强。

LLMs已经在NLP领域内外的许多任务中得到应用，且有不错的表现。除了传统的自然语言处理任务，LLMs还具有执行包括生成代码、玩基于文本的游戏和提供数学问题答案等。由于LLMs出色的信息检索能力，它们甚至已被提议作为教育、研究、法律和医学的工具。（编辑：存源）

写给神经科学家的大语言模型基本原理

编译：郭瑞东

科普作家，关注复杂系统与神经科学。追问nextquestion、集智俱乐部长期撰稿人，曾为知识分子，果壳等多家媒体撰文，科普书《机器学习与复杂系统》合著者。

扫码查看原文

语言不仅仅是交流的工具，它还蕴含着丰富的人类智慧和信息，在一个比特中，其内涵的丰富性远超过我们通常所接触到的任何其他数据形式。自然语言处理（NLP）是一门致力于让计算机拥有理解、解读以及创造人类语言的能力的科学，它在分析和处理人类文本资料方面已经取得了显著的进展。

在早期，研究者借助如n-gram模型这样的简单语言模型（例如，2-gram模型将词-词组合视为独特实体）用来研究语言和语义，以达到各种目的。这些语言模型不时地被用于研究各种认知任务，包括阅读理解、语言翻译和问题解答等。通过比较NLP模型在这些任务上的表现与人类的表现，研究人员获得了关于人类认知的洞见，就如心理语言学领域所示。

大约2010年之后，深度学习的兴起点燃了NLP建模中的语义“嵌入”时代——单个词语、句子、段落或整个文档，都可以封装在一个紧凑的浮点向量格式中，以此向量来表示对应词句的意义。从直观上讲，这种嵌入方式类似于在高维坐标系中定位，使得不同的语义实体（如词序列）根据它们在上下文中的相似性被映射到相近的位置[1-4]。两个语义实体表示的上下文越相似，它们的语义嵌入就越相似。使用像Word2Vec[5]和GloVe[6]这样的最新一代模型，研究人员开始使用这些可互操作的语义嵌入表示来量化意义之间的关系，如词语或句子之间的关系。

当前的大语言模型（LLMs）是在比一个人在数百或数千个生命周期中能阅读的文本还要多的数据上训练的。这种庞大的数据训练基础使它们涌现（emergent）出了诸多能力，如编写计算机编程代码、数学、规划、文献综述和总结，或玩基于文本的游戏等。这些能力并非在它们的各个组成部分中原本就有，而是随着系统复杂度的增加而涌现出来[7]。有时，这样的模型被用来研究大脑如何处理上下文信息以及人类心智如何产生语言（请参见Goldstein[8]、Caucheteux[9]和Schrimpf[10]的优秀示例）。随着当前研究范式的转变和大模型规模呈指数级增长，LLMs学习了迄今为止可能是最强大的意义内部表征。

人类语言反映了人类思维，这就是为什么最先进的NLP可能会为神经科学研究提供内生的优势。从这个角度来看，该文试图讨论大模型对神经科学和生物医学研究者带来的即将到来的影响。

一、数据科学的角度，大语言模型解决方案

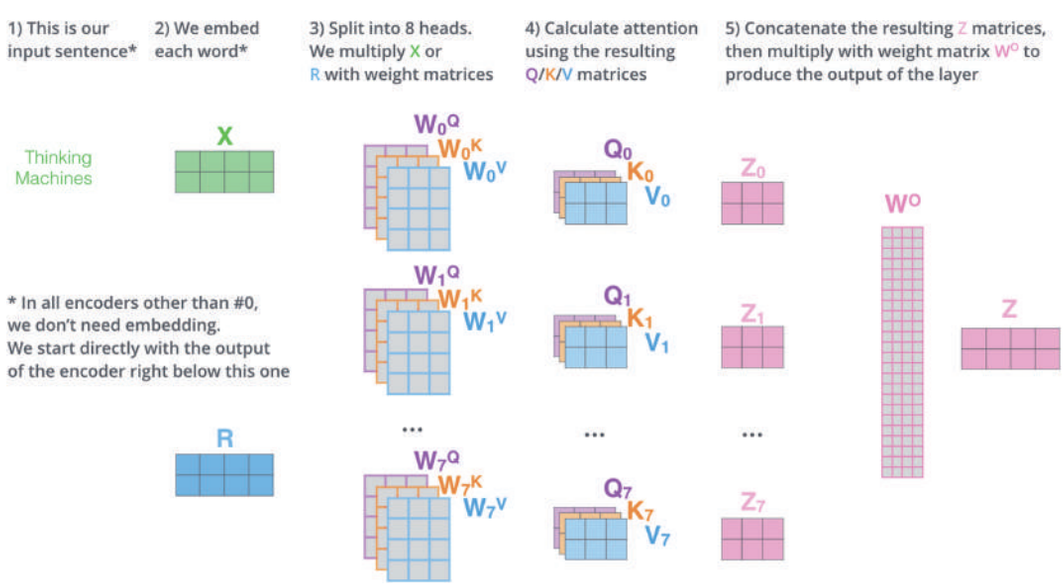
历史上，自2010至2012年以来，卷积深度神经网络（CNNs）因其在处理图像等网格结构数据上的优势，重燃了人工智能的热潮，而LLMs目前正在为AI生态系统注入另一波动力。特别是在引入了Transformer架构之后，语言模型取得了显著进展（例如，Vaswani等人[11]的研究在发表后的前5年内被引用超过9万次），推动了当前AI创新的推动力。

GPT-2在多个语言任务上表现出色，它由24个Transformer块组成，而最新的架构更是深入发展，一些细节仍待揭晓。作为所谓的“生成性AI”的一个实例，这些算法的输出不是类别（例如，患病与健康组的患者）、数字（例如，认知表现测量）或离散类别（例如，年收入的区间），而是一种结构化的内容，如语言（以及图像或音频信息），即从之前输入的内容中合成或幻想新内容。

相较于以往复杂的深度神经网络，Transformer架构以更简洁的特性成为NLP领域的新宠（见图1和图2）。这种简化的架构比之前的方案更具可扩展性，部分原因是这种架构非常适合并行计算工作流程。与之前的深度NLP解决方案不同，在Transformer架构中，无论是近还是远的词标记之间的相互依赖关系，都能同样好地被捕捉。

与某些之前的神经网络设计不同，Transformer模型是前馈深度学习架构，不包含显式的处理循环。相反，通过将已经生成的、之前的文本作为输入反馈到LLM中（“自回归”），创建了隐式循环。与围绕BERT（使用双向上下文来理解词义）的前一代LLM不同，生成性预训练Transformer（GPT）架构，如ChatGPT，在训练期间只关注当前词之前的词标记，这导致了其单向处理模式，即具有自回归性质。

Transformer中所谓的位置编码是该架构的一个特征，它帮助模型理解词序。在自然语言处理中，自回归模型会根据前面的词来预测下一个词。由于其单向性质，GPT式LLMs在预测下一个词时不会“看到”或“考虑”后续的词标记——它是在回顾给定句子的过去，而非展望未来，正如人类阅读书籍时的方式。

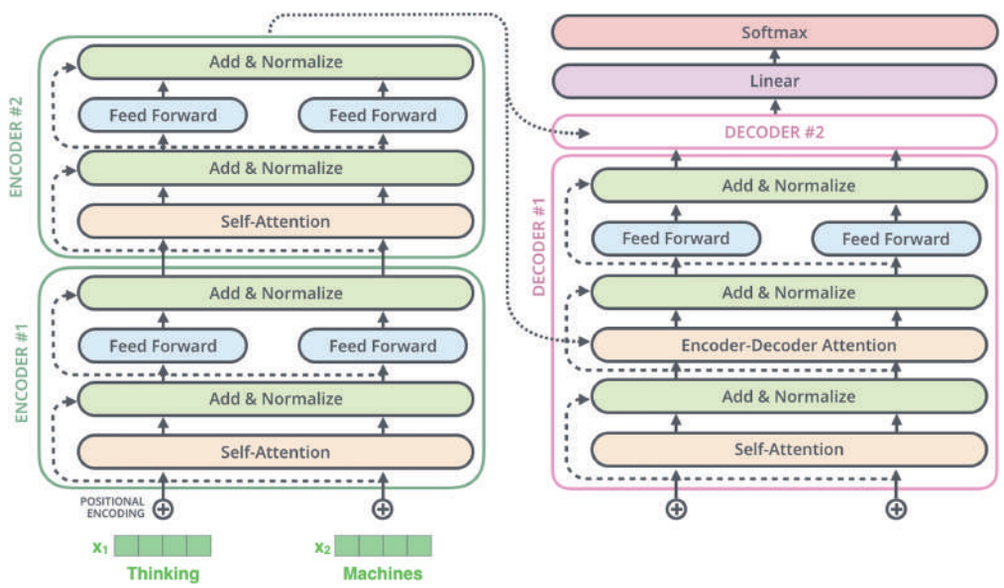


▷ 图一：transformer架构的大语言模型的核心-自注意力机制

正是自注意力机制构成了Transformer建模架构的核心。自注意力机制允许模型在处理序列数据时，对序列中的每个元素分配不同的注意力权重，从而更好地捕捉序列内的长距离依赖关系。

专注于更近或更远的词标记在算法上是相同的——不需要经过逐步迭代的过程来关联更远的信息片段，这与早期深度学习架构的要求不同。在Transformer中，关注句子中附近或远处的词的处理方式是相同的，该架构允许模型同时考虑句子或文本序列的所有部分。与早期的神经网络相比，这意味着不需要按顺序处理输入的远端部分。

自注意力机制的常见实现的计算复杂度与序列长度成二次方相关[12]。尽管在注意力机制上有所改进，但在处理特别长的序列时，大多数情况仍然遇到困难[13]。每个Transformer层可以一次性“看到”其范围内的所有标记。然而，递归信息处理的深度受到连续Transformer层数的限制，例如句子的嵌套意义或数字序列的连续乘法。



▷ 图二：自注意力（self-attention）层在transformer架构中的作用

此外，当前的LLM架构通常在每个连续的Transformer层中设置了几个并行的注意力机制。这种“多头注意力”（1）允许同时并行关注输入序列的几个不同方面，扩大了整体可以捕捉的复杂性范围；（2）因此允许同时识别和提取多个语义表示维度（有些类似于建模不同的潜在因子成分，如主成分分析或自编码器神经网络）[14]。

值得注意的是，温度参数（temperature，一个在0到正数范围内的标量值）是一个影响模型表现的超参数。这个超参数控制着模型输出的创造性程度，作为一种平衡探索与利用（exploration versus exploitation）的形式。设置高温度（例如，>（1）会在最后一层模型中产生更均匀的词概率分布。这导致输出更加模糊，因此可能不够准确，但也更具创造性。相反，低温度（例如，<（1）会导致输出词相关性的概率分布更加尖锐。在这种操作模式下，模型变得更加具有确定性，紧密遵循输出分布中最可能的候选词，从而减少了响应中的随机性。

尽管模型目标简单（例如，BERT调用词掩蔽，GPT3调用下一个词预测，而在GPT4/ChatGPT的情况

下涉及人类反馈），但由于其庞大的规模，Transformer赋予的架构已经引发了小样本学习（few-shot learning）和在多个情景中生成派生语义世界模型的能力[15]。这些能力是自监督建模制度的核心。这些次生能力甚至让这些模型的创造者在解释LLMs的成功时花了不少功夫[16]。

二、LLM解决方案的涌现标度定律

规模效应的极限是什么？作为影响模型性能的关键量，随着训练观察数量的增加，LLMs的模型生成实例的质量迅速提高。在拥有大约2到20倍于模型参数的训练词标记数量时，LLMs已经在多个场合取得了令人印象深刻的表现。

从数据的角度来看，很难感知到模型所需的可用文本、转换文本（ext-transformed）和可转换文本数据（text-transformable data）的上限。具体来说，根据简单的规范假设（根据ChatGPT查询，截至2023年9月，大约有120亿个网站，每个网站平均1500个词），互联网上所有文本的总量可能达到约2万亿词标记。从模型的角度来看，从2018年到2022年，LLMs的规模（参数）从大约10^8（例如，ELMo，BERT-L）增加到大约10^11（例如，PaLM）。作为许多应用场景中的一个基本原则，扩大模型的深度和宽度（增加参数数量）会带来明显的性能提升。了解模型性能如何随规模变化具有战略价值，因为此类见解可以指导资源分配决策：如何确定计算预算、数据资源和模型大小的优先级。

更具体地说，深度学习文献中一项全面的、广泛认可的经验研究，探索并仔细地基准测试了跨越七个数量级的模型规模所带来的影响[17]。这些研究者设计了计算实验，成功地确定了决定模型性能变化的三个关键因素：（1）模型参数的数量（N）；（2）可用数据量（D）；以及（3）用于模型估计的计算能力（C）。在这些实验中，模型性能仅轻微依赖于模型架构的实际形状。通过同时增加N和D，似乎在很大程度上防止了过拟合（即对训练数据中的特殊性的过度适配）。相反，如果只增加N或D（但保持另一个因素固定），性能会下降[18]。最后，N、D和C的持续扩大显示出回报递减的模式，遵循幂律法则。

然而，最新的研究进展指出，与最初增加模型规模的趋势相反，LLMs在所需参数数量上会随着训练越来越小[18,19]。对许多研究者来说，这似乎是反直觉的，再次减少模型规模，可能更好地与实际可用数据量对齐，提高了模型性能，放宽了内存要求，并减轻了计算成本。这些改进可能对LLM解决方案在现实世界问题中的应用至关重要，并增加了未来几年智能手机作为广泛使用的移动设备携带专用LLMs的潜力。简而言之，一个新兴的研究表明，相对于模型参数规模，更多的数据在某种程度上更为重要，尽管两者都是推动模型性能提升和发展的关键因素。

值得注意的是，衡量模型性能在很大程度上取决于研究者选择的评估指标[20]。这些作者认为，只有通过选择特定的评估指标，才能显现出LLMs的“涌现能力”（即在模型规模增大时，模型展现出的新能力或行为[15]）。与上述观点相反，Schaeffer等人[20]还展示了评估指标的选择可以在不同的架构和任务中诱导出看似涌现的能力。因此，最近的实证研究[20]表明，改变评估指标可以削弱或增强LLM架构中涌现能力的信号，这对AI安全和AI对齐有直接影响。

总的来说，较大的LLMs在微调和小样本学习场景中比较小的LLMs更具样本效率。也就是说，矛盾的是，需要估计的模型参数越多，实现相近性能所需的输入数据点就越少。正如在数据科学中一般，提高数据质量总是可以带来进一步的性能提升。重要的是要承认，神经网络的幂律定律在目前上几乎完全是经验性的，但这些幂律特征显示出稳健的趋势[21]。LLM架构的扩展和爆炸是由（1）transformer的发明，

这些transformer在最近的LLMs中变化不大；(2)大量数据源的可用性；以及(3)大规模计算能力的可用性推动的。与下一部分相关，模型的具体架构(如层数、层维度等)相对不那么重要，尤其是随着模型规模的增加。

三、LLMs展现出前所未有的迁移学习能力

为了使深度学习工具蓬勃发展，通常需要丰富的数据。然而，神经科学领域的许多领域并没有现成的大量数据可用，更不用说像AI社区中用于文本和图像分析的互联网规模数据集。这种差异引发了一个问题：我们可以利用哪些丰富的非神经科学数据来建模解决方案，之后迁移到神经科学问题上？

“迁移学习”(transfer learning)是一种数据分析模式，其核心在于解决一个问题时积累的结构化知识，可以之后被应用到一个不同但相关的问题上。迁移学习旨在提高在类似但数据资源可能严重受限的任务上的模型性能。在深度学习的背景下，这通常意味着首先在大规模数据集上预训练模型作为起点，然后，通过轻微调整(微调)模型参数，将这个模型应用于与特定任务相关的较小数据集上*。这个过程利用了这样一个假设：预训练模型学到的特征可以作为通用表示，对目标任务有益。历史上，迁移学习的成功通常依赖于预训练和微调任务之间的高度相似性。

*请参见https://www.ruder.io/transfer-learning/，了解LLMs微调技术的全面资源

LLMs和其他基于Transformer架构的模型在迁移学习方面展现出超出预期的能力，从而通过扩大可执行任务的范围，彻底改变了自然语言处理(NLP)。作为一个关键的转折点，直到最近，主导范式仍然是在大规模语料库上进行有监督的模型预训练。这需要大量的高质量标注数据，而这些高质量的标注不易获得，严重限制了互联网和其他来源可用数据的实际预训练和迁移学习。直到现在，通过无监督预训练，无需为每个数据点提供精确的注释就成为可能，这标志着性能的一个巨大飞跃。这一分水岭事件极大地扩展了LLMs预训练可用数据的范围。

更正式地说，LLM中需要估计的参数越多，模型开发过程就越慢。LLMs开启了全新的微调领域，超越了以往模式中学习算法能够实现的任务范围。研究人员已经提出的几种方法，可以在只更新或添加相对较少的参数的情况下，使模型适应新任务。其中一种策略是“冻结”(保持不变)预训练LLM的多个层的参数。这种方法接下来只调整下游任务的一小部分可调参数，从而避免在神经网络学习新任务时，遗忘之前习得知识的现象。

在微调过程中，通过向LLM中添加新的可学习层，可以进一步扩展这种策略。这样新增的“适配层”可以显著减少目标任务的训练时间和计算成本[22-24]。研究已经证明，选择特别高质量的数据用于微调，即使在目标任务的样本量较少的情况下，也能给迁移学习后的模型带来有竞争力的性能。LLMs在小样本学习方面表现出色。在极端情况下，即便没有为新任务提供示例，仅利用预训练的LLM进行零样本学习，也已证明LLM即使没有调整预训练模型，其零样本学习能力也使其在各种下游任务中表现出色[25,26]。

简而言之，LLMs包含着数十亿个可调模型参数，通过其庞大的规模，解锁了从大规模文本语料库中提取本质表征的能力，而不再迫切需要监督标签注释。无监督深度学习在实践中被证明更具可扩展性。因此，对于那些没有能力从零开始训练LLM的神经科学家来说，通过微调已经预训练好的模型来适应感兴趣的特定任务，可以充分利用这些模型的先进性能，同时减少对数据和计算资源的需求。LLMs可以更好地识别文本中的深层隐藏模式、关系和上下文，这使它们能够回答人类的查询、创造性地生成新内容，以

及形成准确的结果预测。

四、作为计算乐高积木的基础模型

基础模型最初是在大规模文本语料库上进行训练的，例如互联网内容和其他公共或私人来源的数据。这让它们能够发展并构建一个通用的内部语义表征，该表征包括语法和句法，尽管LLMs在多大程度上包含了对语义的理解目前还存在争论[20,27,28]。更进一步，这些模型学习了大量的通用知识、展现了一定的推理能力，以及对可能的语义世界的表征。基础模型的演变可以追溯到transformer时代之前的上一代NLP模型（2017年之前），如Word2Vec[5]和GloVe[6]，它们在连续向量空间中表达词语（参见1背景介绍），这暗示了语义空间的普遍性。

通过从不同的多样来源提炼和吸收精华，基础模型形成了一个通用表征，它包含了庞大、紧凑和密集的人类知识，作为下游建模的先验知识。这不仅仅包含记忆，且包含信息提取和结构化。从哲学上讲，这种对信息的成功压缩可以视为预测能力的一大飞跃，因为成功的预测本身就是一种信息压缩的体现。类似于共享基础设施或平台，这样的AI引擎可以作为多种任务构建的基础，使许多定量建模工作流程变得可行、高效且易于扩展。这些基础模型就像是乐高积木，因为许多下游应用可以在它们之上构建，就像堆叠积木一样。这种对定量建模的新态度与为狭窄任务部署训练专门模型相反。

利用数千个GPU处理数万亿个词元，几周时间内就能完成LLM的训练，其成果能被存储并部署至智能手机中。未来的基础建模框架将提供通用的计算单元，这将有可能使广泛的研究者能够民主化地访问高质量的AI解决方案。这对于神经科学尤为重要，因为神经科学家往往需要在比核心机器学习社区更小的数据集上进行操作。同样，在生物研究中，即使在人类细胞图谱项目中也是如此。截至本文撰写时，该项目也只产生了来自约6,000名捐赠者的约4千万个人类细胞的基因表达数据。

如何创新性地利用这些基础操作系统，以全新视角审视并解决经典研究问题，将是一场大胆的创新之举——在transformer类模型出现之前，这些应用是完全不可想象且和不可行的。使不同领域的研究人员能够启动共同的计算模型模块，也可能有助于提高研究之间的可比较性，并促进不同机构和地理位置间的团队合作。随着资源日益紧缩，深度学习的突破性成果将变得更加容易获取。基础模型在不久的将来，极有可能彻底改变神经科学和生物医学领域的生物信息学面貌。

五、延展：当前LLM的缺陷

尽管LLMs可能是有史以来发展最快的技术，但今天的这些模型版本仍面临许多挑战。

幻觉：幻觉是模型生成与现实或提供上下文无关的文本或信息的常见问题。模型可能会生成听起来合理但错误或捏造的信息，尽管表达得很自信。LLM的设计是生成文本，而不管模型对其输出是否确定。因此，当前的LLM变体可能在准确和可靠的信息查询（例如，给出确切的论文引用）方面处于不利地位[70]。

大数据依赖：LLMs需要大量的输入数据。现在，互联网的大部分内容已经被用于LLM的开发。因此，我们可能会想知道，我们是否已经耗尽了可用的训练数据。未来训练更强大的LLMs的数据生成模式是什么？一种可能性是，上一代LLMs将越来越多地在互联网或其他场所生成输出数据，这些数据将被反馈到下一代LLMs中。目前很难预测这种递归场景的后果。可能的一个后果是，针对评估方案的解决方案可能会越来越多，从而污染训练数据。

资源饥渴：部署LLMs需要大量的计算能力、信息存储容量和能源消耗；可能还包括持续的环境影响。对于那些打算从头开始训练LLM的目标，所需的计算存储资源的丰富程度可能使地球上的大多数工业、学术和政府机构无法参与。

推理：这类模型通常缺乏常识，在应对训练数据中未出现的新情况时，其反应能力有限。我们如何确保LLMs的行为符合人类价值观（所谓的对齐问题）？此外，这些模型有时可能会生成与提供给它的上下文不相关或不完全对齐的文本。作为解释的一部分，LLMs在单步推理任务中表现相当好，但在连续推理步骤的整合上面临挑战。

偏见和伦理考虑：LLMs继承了训练过程中可能存在于摄入数据集的偏见。模型可能会无意中生成有害、冒犯性或有偏差的输出。从人类反馈中进行强化学习，校准LLMs以产生人类期望的答案，可能是解决方案的一部分。此外，当前的LLMs在跨语言和文化方面可能表现不佳。

判定：判断文本是否由LLM生成可能极为困难，甚至不可能。

缺乏可解释性：对于用户和开发者来说，理解给定模型为何生成特定响应仍然很困难，这对于需要可解释性和透明度的应用来说是一个重大限制，尤其是在政治压力下要求机器学习解决方案必须是“白盒”的情况下（参见欧盟的GDPR法律）。闭源LLMs进一步使这个问题复杂化。

规模扩大的递减回报：随着数据量和计算/存储资源的持续增加，我们已经开始遇到递减回报的问题。未来可能需要采取替代策略，以将LLM的能力提升到新的水平。（编辑：存源）

参考文献

关联论文：Bzdok, Danilo, et al. "Data science opportunities of large language models for neuroscience and biomedicine." Neuron (2024).

模型与大脑以不同的“眼光”看待世界

作者：顾海成

施普林格自然计算机编辑，关注人工智能、大数据、信号处理、安全与密码学方向。

扫码查看原文

大脑利用感觉系统来感知和理解周围的环境，比如通过视觉识别物体，通过听觉辨别声音。人类感觉系统的奇特之处在于，对物体的识别具有不变性（invariance），不会受物体外观变化的影响，比如不论光线明暗，我们都能认出月台上父亲的身影。同样地，一段话不论是面对面说出，还是通过电话传递，语调不论是平铺直叙还是抑扬顿挫，我们都可以准确地听懂交流的内容。也就是说，大脑会忽略与核心特征无关的差异，通过一系列复杂转换，精准稳定地识别物体和声音。

神经科学工作者一直致力于构建一系列能重现大脑反应和行为的模型。而在众多计算机模型中，深度神经网络（deep neural network，DNN）模型具有与大脑感觉系统类似的层级结构，因此，人们尝试使用这类分层神经网络模型来模拟大脑的感知，将感官输入转化为与任务相关的表征。到目前为止，基于DNN的模型已成为性能最好的机器感知系统，同时也是大脑视觉和听觉系统领域的主要研究模型。如今，DNN模型在识别物体或声音方面，表现得如大脑一样出色。

但近期，麻省理工学院的研究者在Nature Neuroscience上发表的论文发现，这些模型似乎会“搭错神经”，对与目标无关的刺激作出同样的反应。进一步的研究表明，虽然在目标识别判断方面，DNN模型与人类感觉系统的表现类似，但它们的识别策略截然不同。DNN模型拥有自己独特的不变性，亦即它们会对在人类看来千差万别的刺激物作出相同的反应。

一、一种行为检测方法

人工神经网络模型之所以能够复制生物感觉系统中的运算，是基于这样一种假设——这些模型的不变性反映的正是生物感觉系统中的不变性。但有研究发现，人类与模型之间存在着差异，它们的不变性似乎并不完全匹配。为了确定DNN习得的不变性与人类感知的不变性是否相似，Jenelle Feather等人以执行分类任务的DNN模型为例，进行了深入探究。

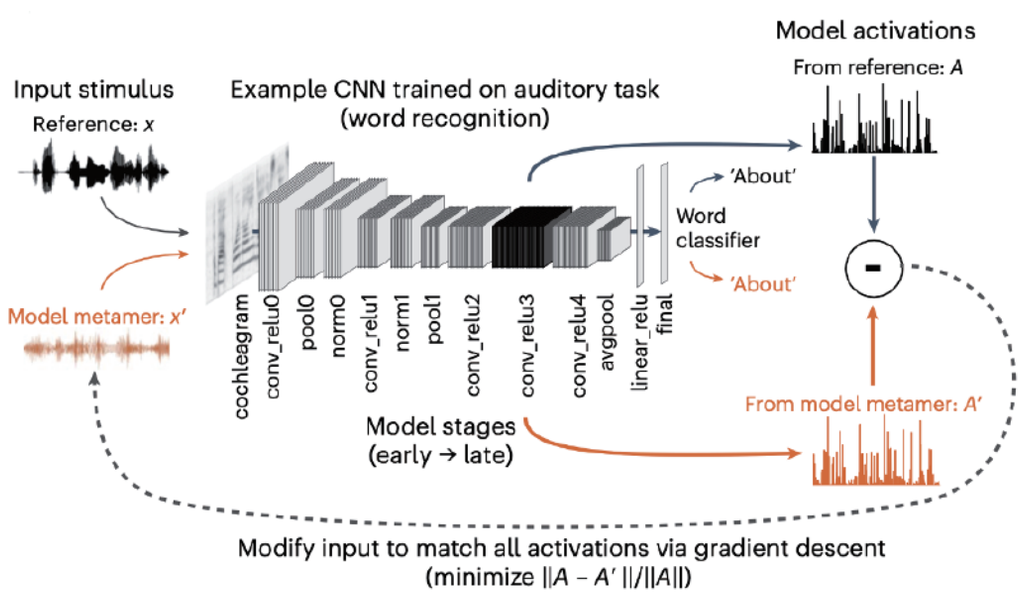
目前，大多数感知系统的神经网络模型，都是为了完成单一的行为任务而训练的。如果该模型成功地再现了人类在某项任务中的不变性，那么它的同色异谱（metamer）*在该任务中应该也会产生与人类相同的行为判断。因为它们与介导判断的人类表征应该是无差别的。

*注：同色异谱的概念源于对人类感知的研究，描述的是有些颜色由不同波长的光组成，但看起来却完全相同。根据格拉斯曼定律，人的颜色视觉系统只能分辨颜色的明度、色调和彩度三个颜色属性，只要在视觉上对这三个颜色属性的感觉相同，就认为是相同的颜色，便可以相互替代，不必考虑它们的光谱组成究竟是否相同。在听觉方面也有类似的现象。比如，尽管两群昆虫发出的声音具有不同的声学细节，但人们无法将它们区分开，因为这些声音具有相似的总体统计特性。借用此概念，以DNN模型替代人类感知，由模型生成的、与自然刺激相配对的刺激，被称作模型同色异谱。

基于此，该研究使用识别判断（recognition judgement）作为行为测定标准，来检测模型同色异谱是否反映了相关人类感知系统中实例化相应的不变性。如果人类无法识别一个模型同色异谱，那么他们也就无法将其与参照刺激划为一类。

所以，在研究中他们首先构建模型同色异谱，之后将获得的同色异谱呈现给人类测试者，让他们进行分类，看是否与最初的刺激物属于同一类。

具体步骤如下：（1）测量自然图像或声音在特定模型阶段引起的激活；（2）将自然图像或声音的同色异谱初始设置为白噪声信号。图像或声音波形均可，选择白噪声是为了在模型约束条件下尽可能广泛地对同色异谱进行采样，而不会使初始化偏向于特定的对象类别；（3）对噪声信号进行修改，使其在相关模型阶段的激活同与之匹配的自然信号的激活之间的差异最小。优化过程是对输入信号进行梯度下降，在模型参数保持不变的情况下反复更新输入信号。



▷图注：模型同色异谱的构建过程。图源：论文

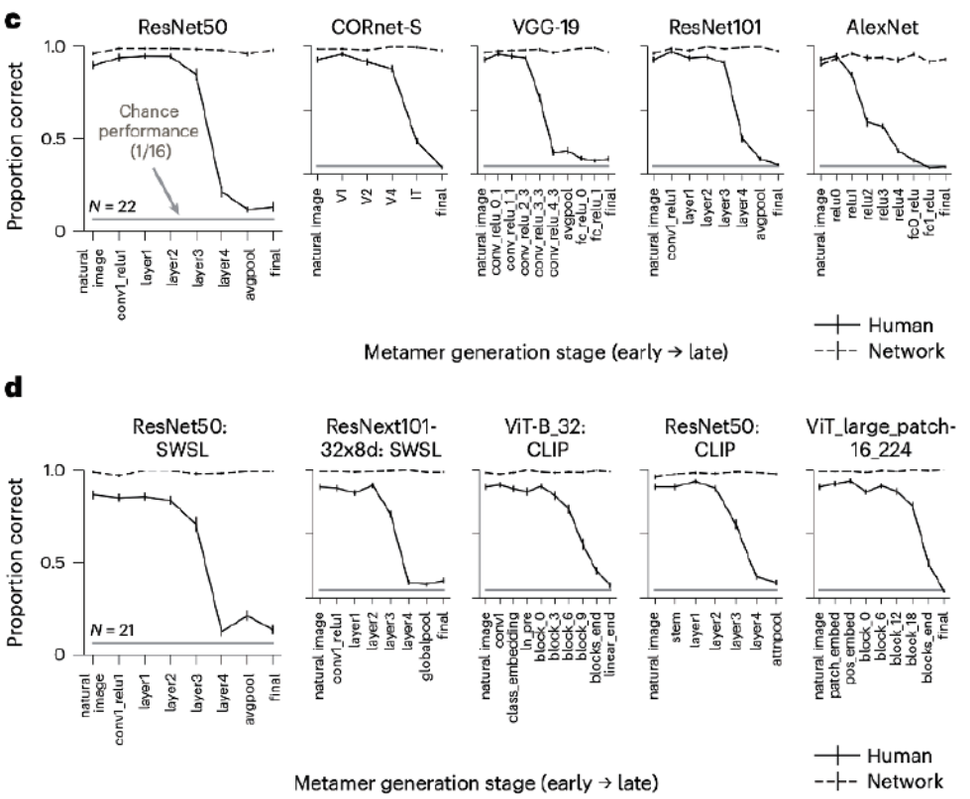
任何由可微分运算构建的模型阶段，都可以用这种方法生成模型同色异谱。由于本研究所考虑的模型是分层的，如果图像或声音在某一特定阶段得到了高保真匹配，那么随后的所有阶段也都会得到匹配，包括在监督模型中的最终分类阶段，它们会产生相同的决定。

（1）标准视觉DNN的同色异谱

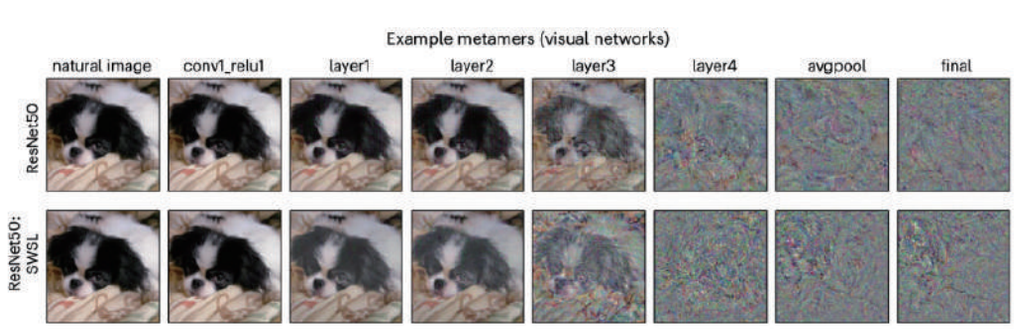
研究者为五个跨越不同结构和深度的标准视觉神经网络的多个阶段生成了同色异谱。这五个标准视

觉神经网络在ImageNet1K数据集上进行训练，可以捕捉到与灵长类动物视觉表征相似的特征。随后，又对另外五个模型进行了第二次实验，这五个模型是在项目后期获得的更大数据集上预先训练的。

为了评估人类对模型同色异谱的识别能力，人类测试者对自然刺激和模型同色异谱进行了一个包含16个类别的分类任务（16-way categorization task）。结果发现，与“训练有素的神经网络学会了类似于人类的不变性”这一想法相反，人类对模型同色异谱的识别能力在不同模型阶段都有所下降。



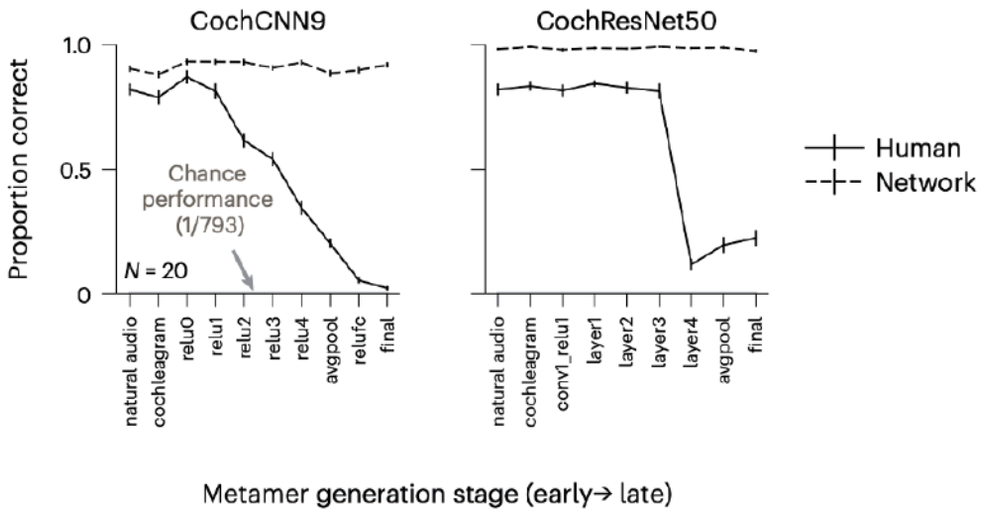
▷图注：标准训练的视觉深度神经网络的同色异谱，通常无法被人类测试者识别。图源：论文



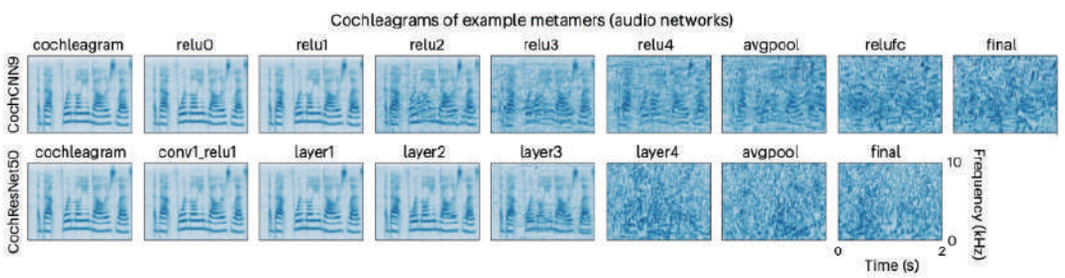
▷图注：来自标准训练和半弱监督学习训练的ResNet50视觉模型的同色异谱示例。图源：论文

（2）标准听觉DNN的同色异谱

研究者还用两个经过训练的听觉神经网络——CochResNet50和CochCNN9，进行了类似的实验。他们在Word-Speaker-Noise数据集中进行单词识别任务，模型同色异谱是根据验证集中的纯语音示例生成的。人类测试者执行了一项包含793个类别的分类任务（793-way categorization task），识别刺激中的单词。



▷图注:人类对听觉模型同色异谱的识别率。图源:论文



▷图注:来自两个听觉模型同色异谱示例的耳蜗图。颜色强度表示频率通道中的瞬时声音振幅(任意单位)。图源:论文

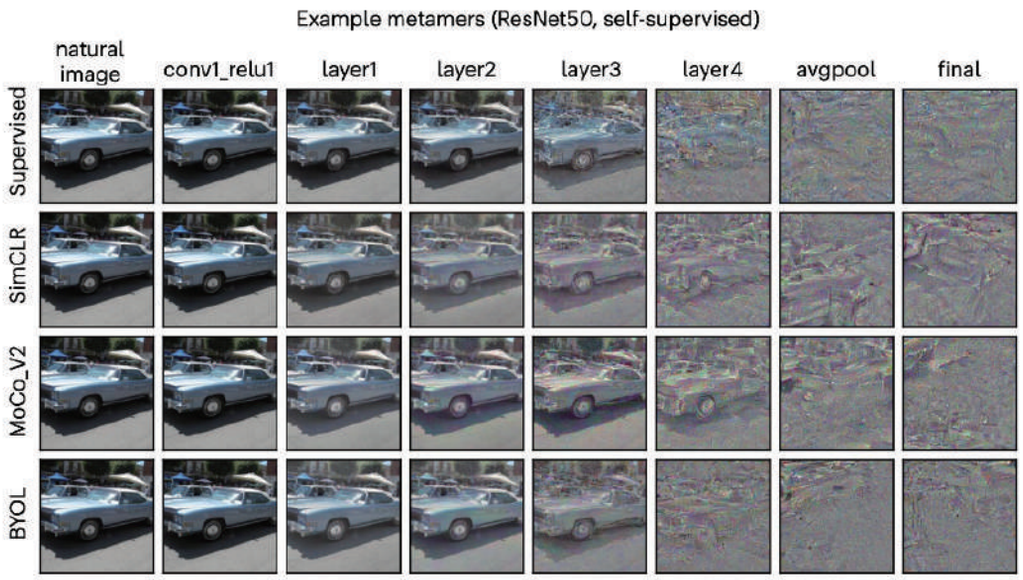
研究人员惊奇地发现,通过同色异谱方式生成的大多数图像和声音,看起来和听起来都与模型最初得到的示例完全不同。大多数图像只是杂乱无章的像素的堆叠,而声音则听起来更像噪音。将它们展示给人类测试者时,他们大多数都无法将其归到与原始刺激相同的类别中。这表明,尽管这些视觉和听觉神经网络模型目前是每种模式下大脑反应的最佳预测模型,但它们的不变性与人类感知的不变性严重不符。也就是说,模型形成了自己的不变性。在模型看来相同的刺激物,对人类来说有着天壤之别。

二、这不是个例

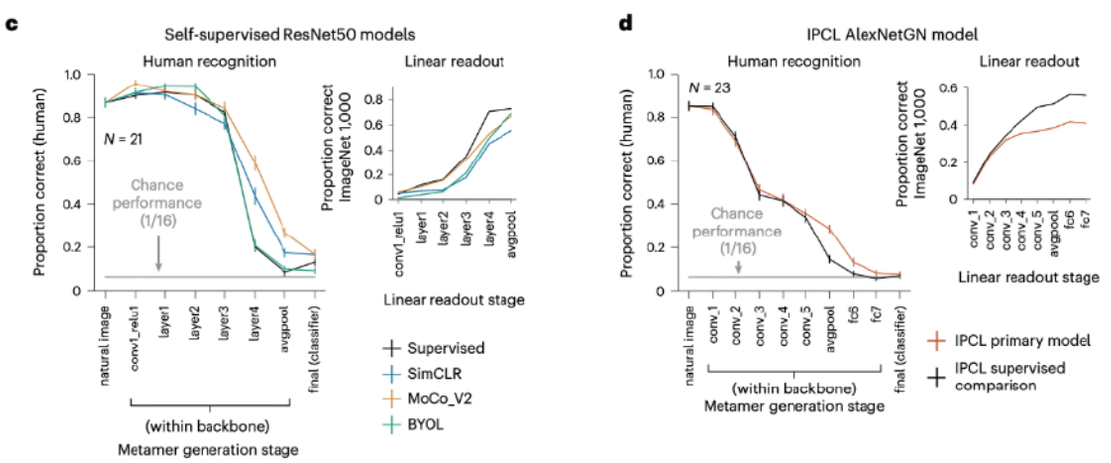
研究人员不仅试图回答,常用神经网络模型习得的不变性是否与人类感知系统相同,他们还好奇无监督学习模型(unsupervised model)中是否也存在这种不变性差异?

生物系统通常无法获得监督学习所需的大规模标签(label),在很大程度上依赖无监督学习。所以,有理由怀疑,神经网络模型中明显存在的不变性差异,在某种程度上可能源于带有明确类别标签的监督训练。同色异谱非常适合用于回答这个问题,因为其生成不依赖于分类器,任何感官模型都可以生成同色异谱。

目前,主要的无监督模型都是“自监督”模型,它们在训练时使用损失函数,该函数偏向于将单个训练示例的变体(例如图像的不同裁剪)表征为相似,而将不同训练示例的变体表征为不相似。研究者为四种此类模型(SimCLR、MoCo_V2、BYOL和IPCL)以及具有相同架构的监督比较模型生成了模型同色异谱。



▷图注：ResNet50监督和自监督模型中部分阶段的同色异谱示例。在所有模型中，后期同色异谱大多无法识别。图源：论文



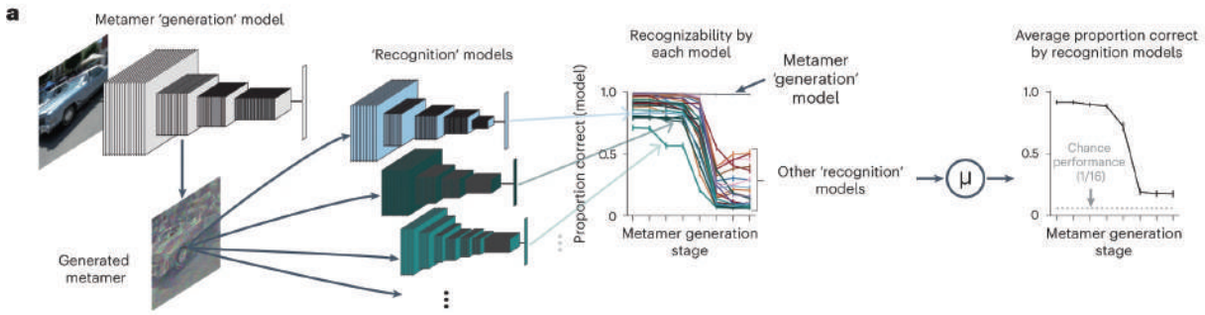
▷图注：人类测试者从监督和自监督模型中识别同色异谱的概率，以及在模型的每个阶段根据ImageNet1K任务训练的线性读出器（linear readout）的分类性能。图源：论文

对比发现，自监督模型的结果与监督模型相似，人类测试者对模型同色异谱的识别率在模型后期都有所下降，在最后阶段接近偶然水平。这表明，标准神经网络模型无法通过上述同色异谱测试，这并不是监督训练程序所特有的。

三、人类各模型之间是否共享同色异谱？

虽然模型同色异谱无法被人类识别，但它是否能被其他模型识别呢？也就说，各个模型是否拥有相同的不变性？

为了解决这个问题，研究者将所有针对一个模态训练的模型纳入研究，将其中一个模型作为“生成”模型，并将其同色异谱呈现给每个其他模型（“识别”模型），测量它们对类别预测的准确性。每个模型依次作为生成模型，重复上述过程。



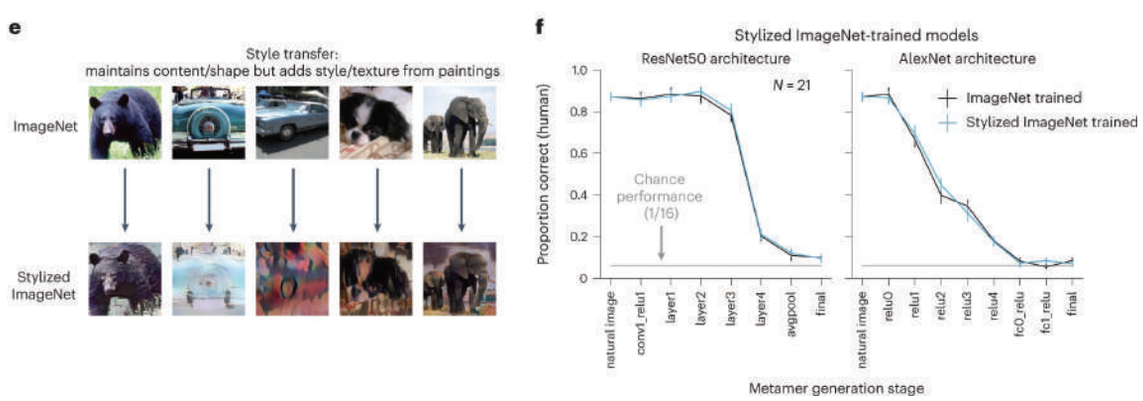
▷图注:为“生成”模型的每个阶段生成同色异谱。这些同色异谱被呈现给“识别”模型。图源:论文

结果发现,模型往往包含独特的不变性,即它们的同色异谱无法被其他模型识别。之前的研究发现,自然图像的表征不相似矩阵(representational dissimilarity matrix)在不同神经网络模型之间可能存在差异,本研究的发现与之基本一致。简言之,研究人员在不同的听觉模型和视觉模型中都得到了相同的效果——每个模型都形成了自己独有的不变性。当一个模型的同色异谱展示给另一个模型时,第二个模型和人类测试者一样,也无法识别。

四、如何使模型的同色异谱更易被人类识别?

目前模型与人类之间另一个常见的差异是,模型倾向于根据纹理而非形状来进行判断。这种“纹理偏差”(texture bias)可以通过“风格化”图像的训练数据集来减少,从而增加模型对形状线索的依赖,使模型在这方面更像人类。鉴于此,研究人员探究了纹理偏差是否也有助于减少模型同色异谱的差异。

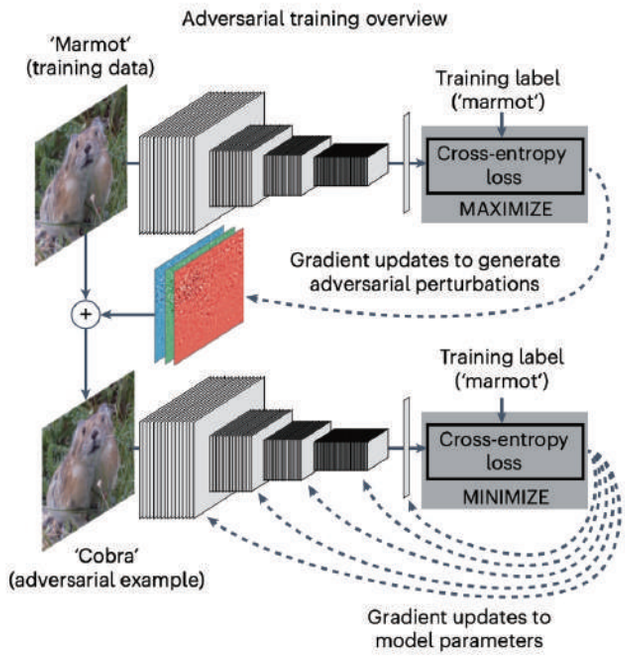
研究者选取了在Stylized ImageNet上训练的两个模型,为它们生成了同色异谱。结果发现,这些模型的同色异谱与在标准ImageNet1K训练集上训练的模型的同色异谱一样,同样无法被人类识别。这表明,纹理偏差无法解释同色异谱差异!这些差异并不只是由模型的纹理偏差造成的。



▷图注:(e)使用Stylized ImageNet增强的自然图像和风格化图像示例。(f)人类对使用Stylized ImageNet训练的ResNet50架构和AlexNet架构的模型同色异谱的识别。图源:论文

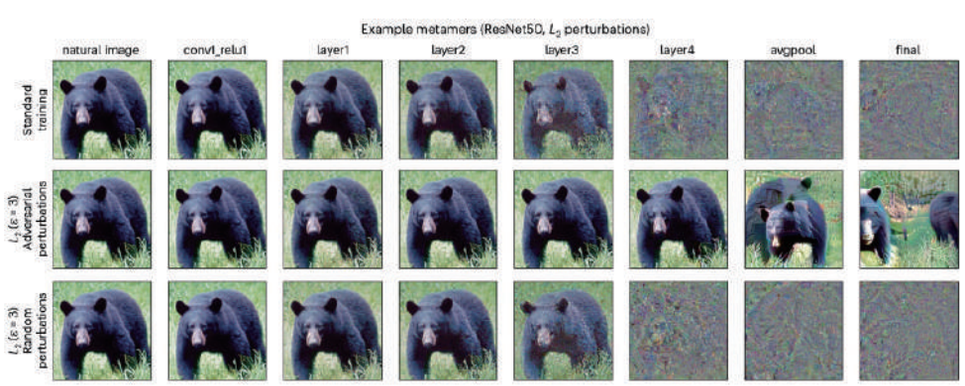
众所周知,目前人工神经网络的一个特点是,容易受到对抗扰动(adversarial perturbation)的影响。刺激物的微小变化就会改变模型的判断,但人类通常无法察觉这些变化。

降低这种脆弱性的一种方法是进行对抗训练(adversarial training),即在训练过程中产生对抗扰动,迫使模型学会将扰动图像识别为“正确的”人类可解释的类别。那么,这种对抗训练是否会有助于人类识别模型的同色异谱?



▷图注:对抗训练。图源:论文

研究人员为五个经过对抗训练的视觉模型生成了模型同色异谱,这些模型具有不同的结构和扰动大小。作为对照,他们还对模型进行了随机方向而非对抗方向的等量扰动训练,这种训练通常无法有效防止对抗性攻击。



▷图注:有对抗性扰动、无对抗性扰动或随机扰动的模型同色异谱示例。图源:论文

研究人员发现,与标准训练的模型或随机扰动训练的模型相比,在所有情况下,通过对抗训练诱导出的同色异谱都更容易被人类识别。不过,这些同色异谱的效果还是不如原始刺激物。

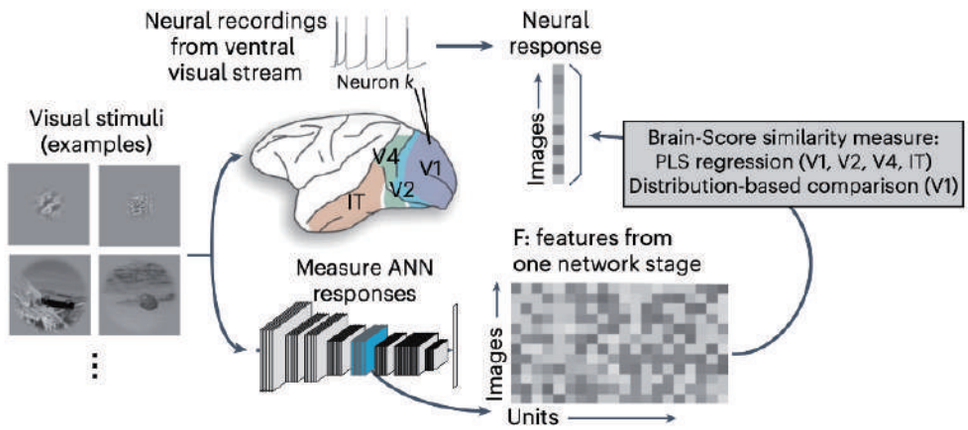
在听觉模型上的对抗训练研究,也得到了相同的效果。总之,对抗训练可以使模型的不变性在视觉和听觉领域变得更像人类。进一步的研究还发现,识别率的提升与对抗训练对模型鲁棒性的影响无关。

五、同色异谱vs模型-大脑相似性

既然不同模型之间同色异谱存在差异,那么同色异谱测试与传统的模型评估方法(如大脑预测或对抗脆弱性等)相比是否具有优势呢?

为了弄清楚这个问题,研究者使用标准模型-大脑比较基准,对上述提到的所有视觉和听觉模型进行了评估。对于视觉模型,他们使用Brain-Score平台,用以测量模型表征与视觉区域V1、V2和V4以及下颞

皮层(IT)的神经基准的相似性。



▷图注:神经基准流程。图源:论文

对于每个模型,他们都选取了在各视觉区域的保留数据中相似度最高的模型阶段,以此计算每个视觉区域的得分。然后,将这一神经基准得分与用于获得神经预测的同一阶段的模型同色异谱的可识别性进行比较。这项分析表明,V4和IT的两个测量值之间存在适度的相关性,但经过Bonferroni校正后并不显著,而且远低于预设的噪声上限。 此外,不同模型的神经基准得分总体上极其相似。因此,标准的模型-大脑比较基准并不能捕捉在同色异谱识别方面的差异。

同时,他们使用一个大型人类听觉皮层功能磁共振成像(fMRI)数据集,对听觉模型进行了类似的分析,最终得出相似的结论。也就是说,同色异谱测试在区分模型方面超过了这些传统指标,可作为传统的模型-大脑拟合度量的一种补充工具,与其相辅相成。

六、总结

在不同的模态(视觉和听觉)和训练方法(监督训练和自监督训练)下,由于DNN存在不同于人类感知系统中的不变性,其同色异谱通常无法被人类识别。这种效应是由模型特有的不变性驱动的。同时,研究者还找到了使模型同色异谱更易被人类识别的方法,比如在模型的中间阶段对模型进行对抗训练。

同样,人类是否也有个体特有的不变性?鉴于目前还无法对人类同色异谱进行采样,因此很难测试这种可能性。如果人类也存在特异的不变性,那么本文描述的现象可能就不是人类与模型之间的差异,而更可能是识别系统的一种共同特性。(编辑:Lixia)

参考文献

关联论文:Feather, Jenelle, et al. "Model metamers reveal divergent invariances between biological and artificial neural networks." Nature Neuroscience 26.11 (2023): 2017-2034.

Feather, J., Leclerc, G., Mądry, A. et al. Model metamers reveal divergent invariances between biological and artificial neural networks. Nat Neurosci 26, 2017–2034 (2023). https://doi.org/10.1038/s41593-023-01442-0

https://news.mit.edu/2019/differences-between-deep-neural-networks-and-human-perception-1212

https://neurosciencenews.com/neural-networks-sensory-perrception-24953/

AI语音模型与人的听觉有多相似?

编辑:Lixia

生物学硕士,资深科学编辑,致力于科学传播。

扫码查看原文

对于人类来说,我们依赖大脑的听觉通路实现高效精准的语音信号处理,能够轻松实现每分钟300个汉字或者150个英文单词的自然语音识别。那么,如何建模大脑的听觉和语言环路并解析自然语音感知的神经机制?这是长久以来认知神经科学关注的重要问题。

如今,计算机科学家花费了数十年才终于实现了较为接近人类水平的自动语音识别AI模型。这类纯工程的AI模型彻底抛弃了早期基于语言学理论的模型框架,完全采用数据驱动的端到端大规模预训练深度神经网络。那么,这样的模型与人脑听觉通路有多少相似性呢?

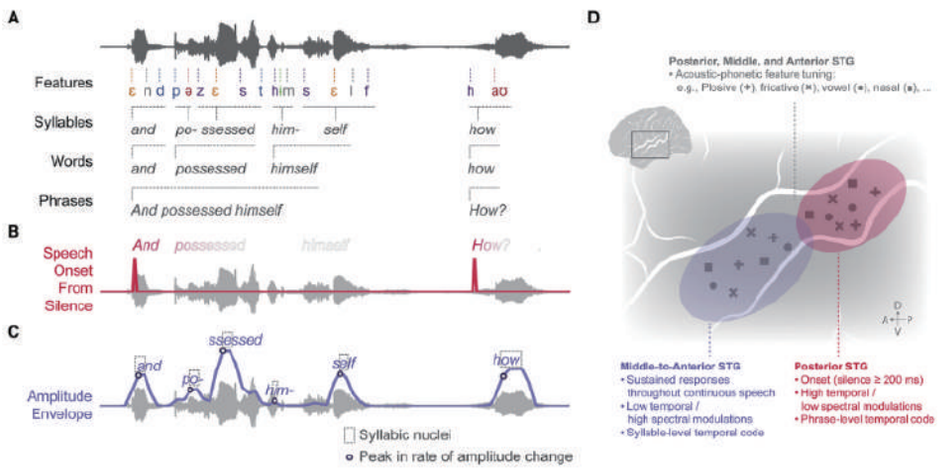
针对这一问题,上海科技大学生物医学工程学院李远宁教授团队与加州大学旧金山分校Edward Chang教授及复旦大学吴劲松/路俊锋教授团队合作,融合自监督预训练深度语音模型、高密度颅内脑电、单神经元仿真模型等多种技术方法,在中英文跨语言对照实验范式下,深入研究了AI语音模型与人脑听觉通路在计算与表征上的相似性。

2023年10月30日,该研究成果以“Dissecting neural computations in the human auditory pathway using deep neural networks for speech”《运用深度神经网络语音模型解析人脑听觉通路的神经计算》为题在线发表于Nature子刊Nature Neuroscience[1]。

传统上,神经科学家会利用线性编码模型来研究神经环路的信息处理机制,所谓线性编码模型,主要是利用从语音中提取的特征来预测神经响应[2]。这些特征是基于语言学和音韵学的假设或理论来定义的。研究者可以在不同层次上提取特征——从纯声学的声谱图特征,到语音学的元辅音、构音方式,再到包含上下文信息的相对音高等,然后使用滑动时间窗口来预测神经响应。如果某类特征可以准确预测某个区域的神经活动,通常认为这个区域的神经活动编码表达了此类特征。

在过去的十多年中,运用颅内电生理记录实验的方式以及神经编码模型,研究者们已经发现了很多

重要的神经编码的特征，例如，颞上回次级听觉皮层的不同神经群体的活动编码了从语音的包络、开头到具体的元辅音音素的特征等等[3]（图2）。

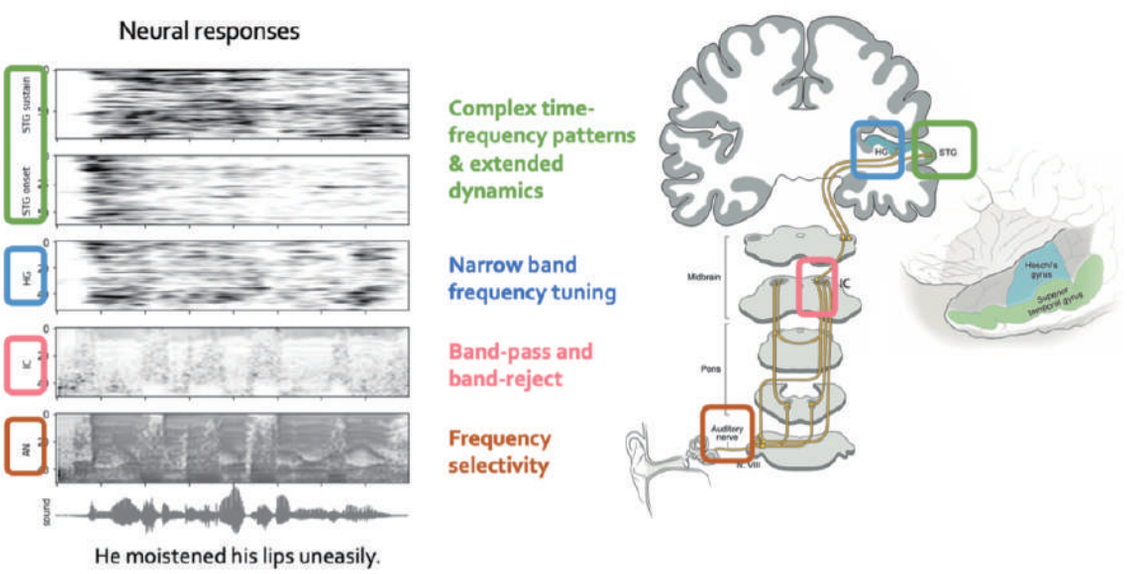


▷图2：基于语言和语音学理论的语音特征提取以及神经编码模型。图源：参考文献3

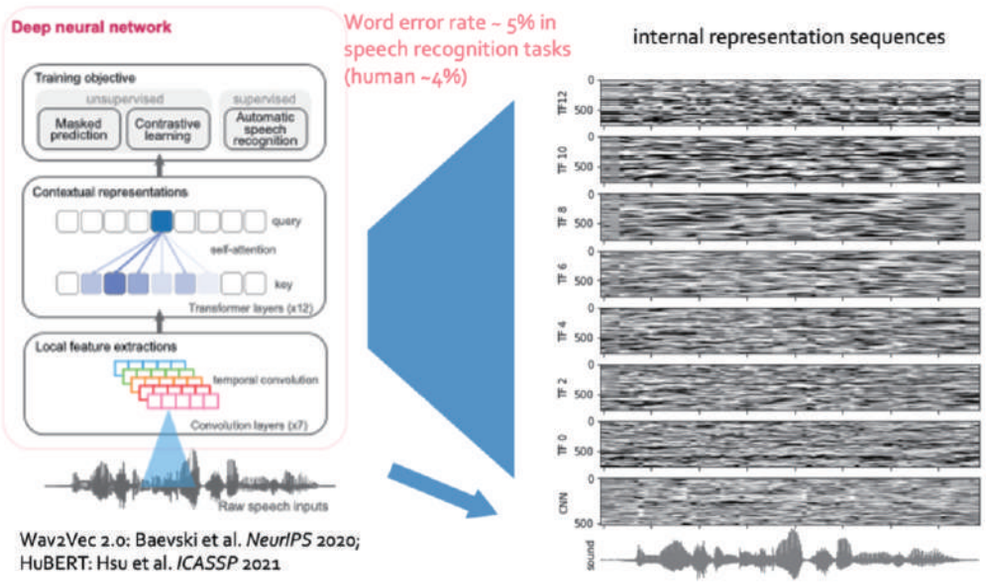
在本项研究中，除了与语言密切相关的次级听觉皮层之外，运用颅内高密度脑电记录技术以及高精度的单神经元级别生物物理仿真模型，研究者获得了覆盖整个听觉通路的、从听神经到脑干再到听觉皮层的神经响应（图3）。虽然传统的基于理论驱动的神经编码研究可以分析通路中各个环节的编码信息，却难以融合成一整个能够实现高效精确语音识别的计算模型。

在人工智能领域，情况则恰恰相反。基于大规模自然语音训练的语音识别模型在很多自动语音识别（ASR）任务上已经达到接近人类的识别水平[4][5]，但这类模型的内部特征表达呈现出复杂的动态模式（图4），其内在的表征与计算难以直接被理解与解释。

既然这些人工智能模型与大脑听觉回路能够接收相同的语音输入，并执行相似的认知功能，那么这两者之间是否存在计算和表征上的相似性呢？这便是这项研究聚焦的关键问题所在。

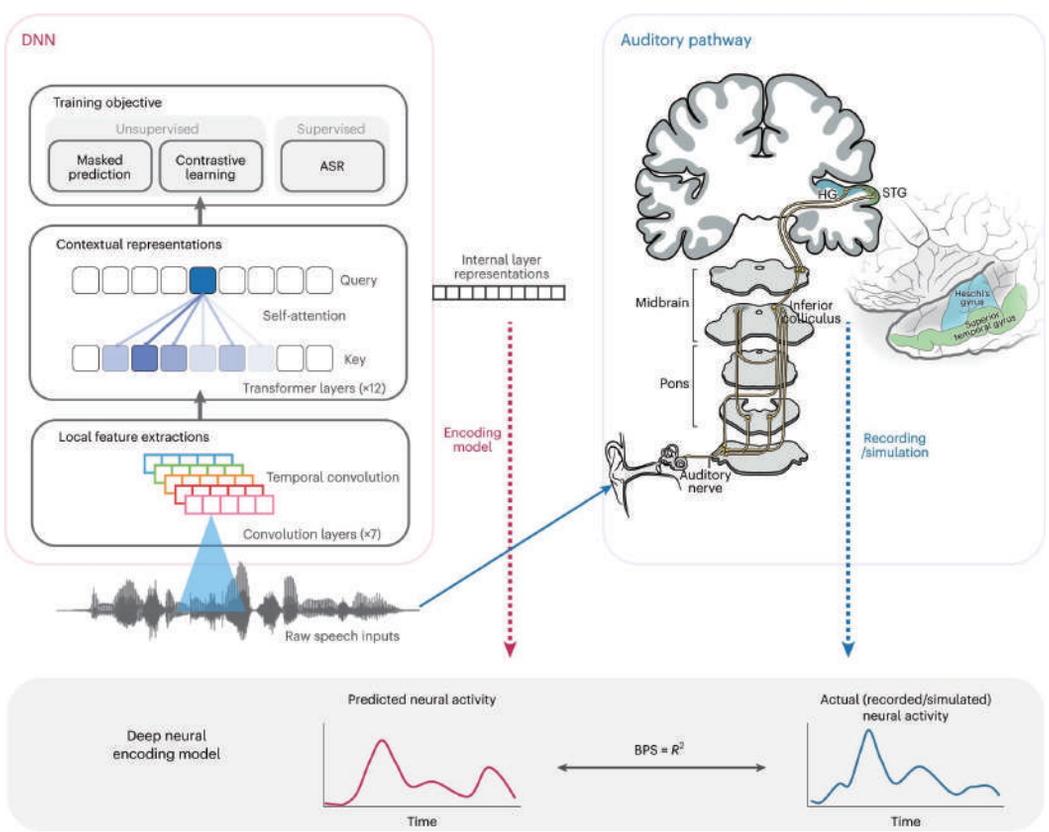


▷图3：人脑听觉通路的自然语音神经响应活动记录，图中包含听神经（AN）-下丘（IC）-初级听觉皮层（HG）以及颞上回次级听觉皮层（STG）。图源：来自论文作者（李远宁）



▷图4:基于Transformer的深度语音模型(HuBERT)及其内部动态特征表达。图源:来自论文作者(李远宁)

为此,研究者通过构建一种新的深度神经编码模型来研究这个问题。这是一种纯数据驱动的模型,从语音预训练的深度神经网络中提取特征表达,运用这些数据驱动的特征构建新的线性编码模型,并与真实的大脑听觉响应信号进行相关性分析,从而研究深度神经网络内在特征表征与大脑听觉通路内不同神经群体活动之间的相似性。



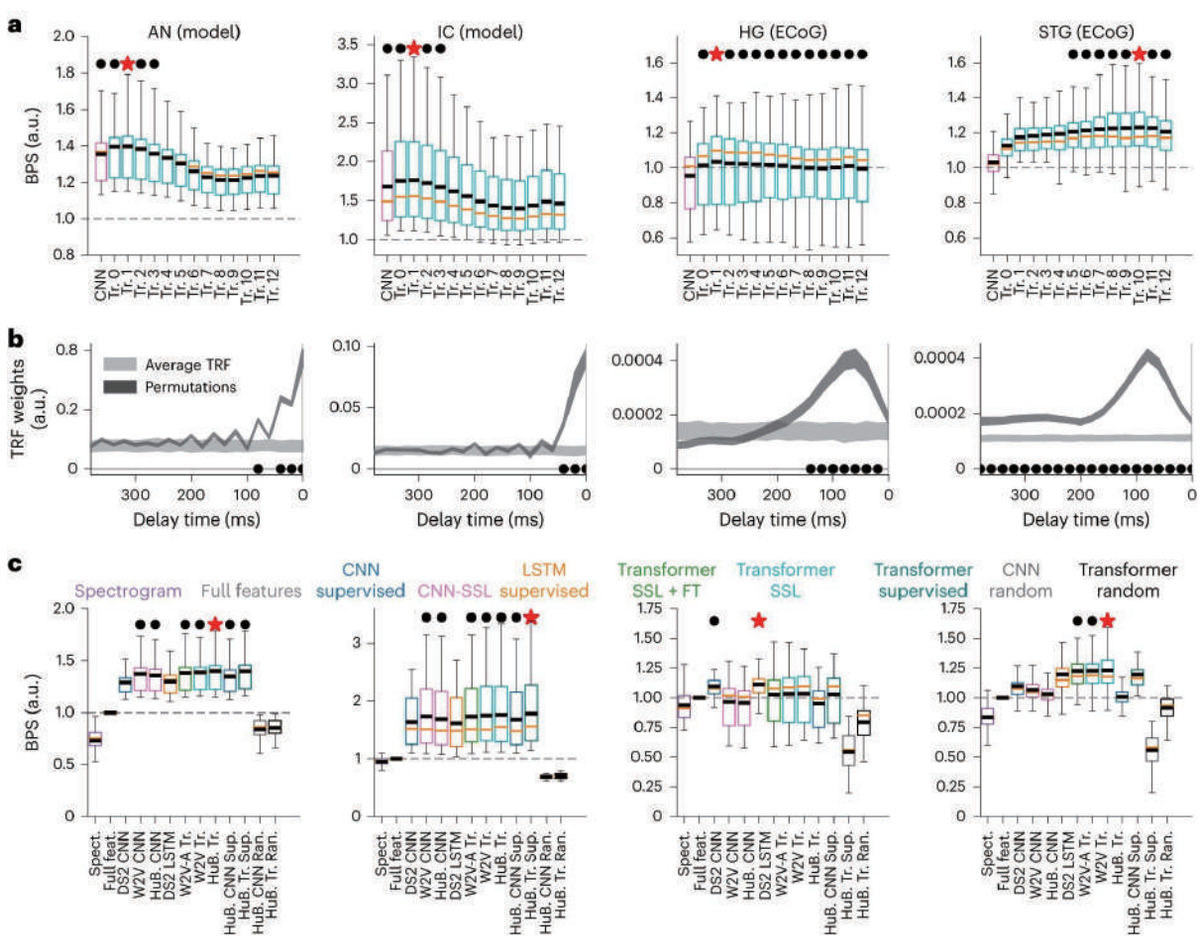
▷图5:基本研究模型。基于预训练深度语音网络特征,构建神经编码模型,预测听觉通路的语音神经响应。图源:本论文

研究者在1000小时英文自然语音上训练了多种不同的人工智能模型，包括基于卷积（CNN）、LSTM以及Transformer等不同架构，运用对比学习、掩码预测等自监督训练和ASR有监督训练等不同训练方式。通过比较基于这些模型建立的神经编码模型在听觉通路不同节点的神经活动预测表现，研究发现，端到端的语音预训练网络的层级结构，与听觉回路的层级结构之间确实存在着很大的相似性（图6）。

首先，对于整个听觉通路，基于深度神经网络特征的编码预测模型要全面优于传统的基于语言学理论的线性特征模型。这说明整个听觉通路具有很强的非线性特征，即便是在传统认为高度线性化处理时频特征的听神经上，额外的非线性特征也可以极大地提升神经编码模型的预测准确性。

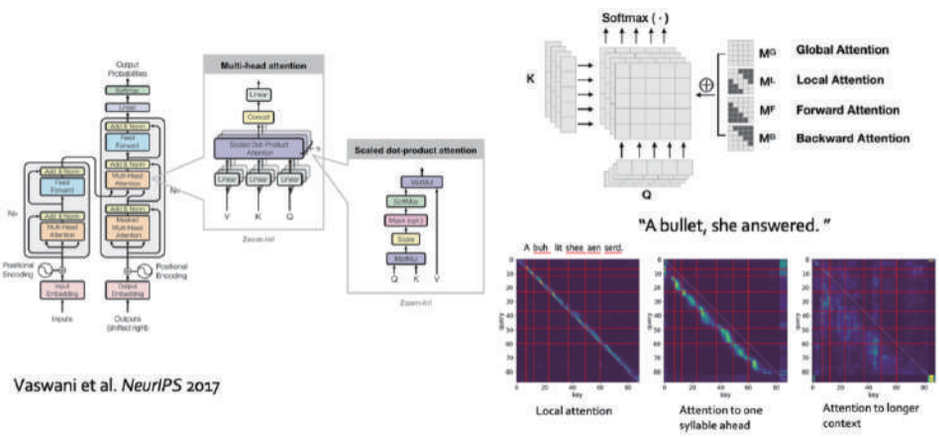
其次，不同复杂程度的模型对应于听觉通路中的不同区域：对于较为底层的听神经（AN）和丘脑（IC）神经活动来说，较为简单的卷积层即可较好地预测神经活动，而额外的Transformer结构并无法进一步提升预测准确度；而对于负责较为复杂语言信息加工处理的颞上回（STG），能够动态提取上下文信息的Transformer结构将显著地提升神经活动预测的准确性，而仅有静态有限感受域的卷积网络则无法与之相媲美。

此外，研究发现，对于同一个自监督语音模型，它的整体层级结构与听觉通路AN-IC-STG层级结构相对应，其中较为前端的卷积层更好地对应于听神经，而卷积输出层与前部Transformer层更好地对应于丘脑听觉神经元，颞上回次级听觉皮层则与中后部的Transformer层相对应。



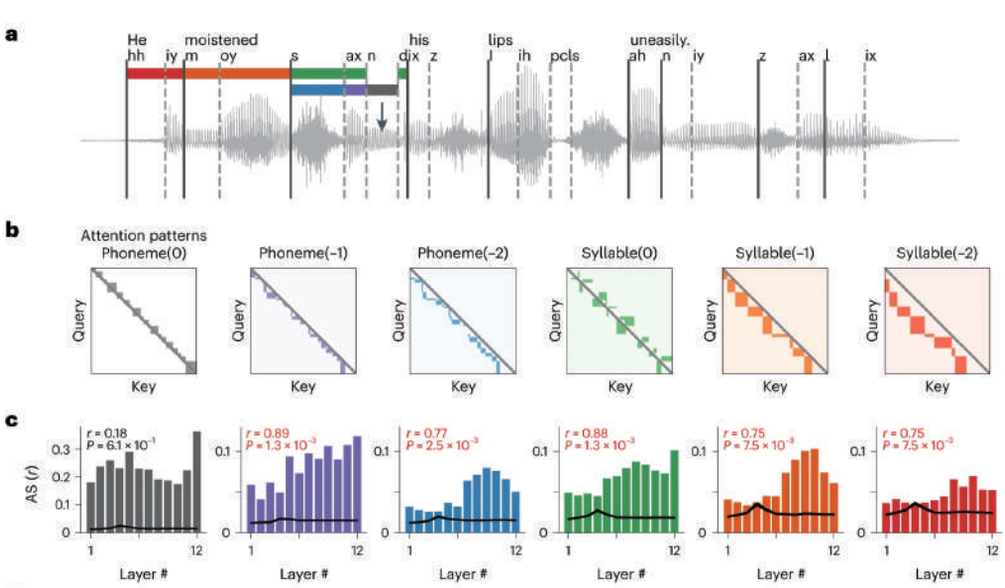
▷图6：不同深度语音神经网络模型与听觉通路呈现不同对应关系。图源：本论文

在建立了深度语音模型与听觉通路的表征相似性之后，研究者进一步探究了驱动这些表征相似性的计算机制，并聚焦在表现性能最好的HuBERT模型上。这是一种类似BERT结构的Transformer模型，其中最重要的计算单元是自注意力机制[6]（图7）。它的内部隐藏层的特征由skip connection和multi-head attention两部分叠加而成，skip connection反应的是当前时刻的序列状态，而attention则是上下文信息的加权组合。研究者通过分析注意力矩阵的权重信息来分析神经网络如何提取语音序列中的上下文特征。



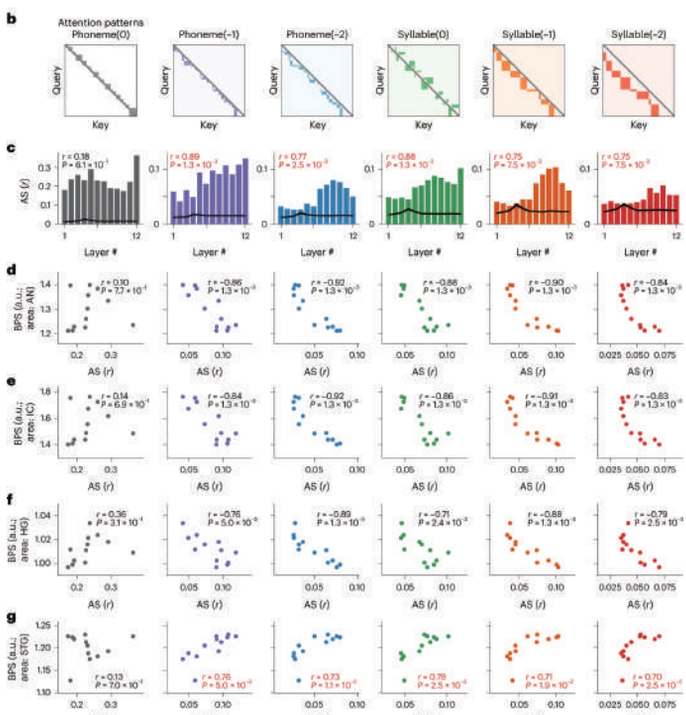
▷图7：Transformer与自注意力机制。图源：参考文献6

依据音素（phonemic）和音节（syllabic）级别的语音上下文结构，研究者定义了随输入动态变化的自注意力模板。随后，使用实际的语音数据，研究者计算了Transformer预训练的网络的自注意力分布，究竟在多大程度上对齐到了这些上下文结构上。结果表明，随着网络的加深，对齐到长距离上下文结构的注意力权重也逐渐变大。值得强调的是，此处使用的HuBERT模型是完全自监督模型，训练过程不包含任何显式的上下文结构信息以及语音内容信息。这一结果表明，自监督训练的语音模型可以学习到自然语音中与语言和语义相关的关键上下文结构信息。



▷图8：自监督学习模型从语料中学到语言相关的语音上下文结构信息。图源：本论文

这种通过自注意力计算获得的关键语音结构的准确性，是否与深度语音模型对大脑语音听觉皮层的相似性有关呢?研究者将这两者进行了相关性分析，结果表明(图9)：在与语音处理密切相关的颞上回次级听觉皮层，这两者呈现显著的正相关，也就是说，自注意力权重与语音中的上下文结构对齐程度越高，神经网络对于大脑活动的预测能力就越强；而反之，在初级听觉皮层以及听神经、脑干这些区域，这两者则是负相关，说明对上下文注意的越少，即对时域上局部瞬态信息的表达越多，神经网络与大脑信号的相似度也就越高。因此，通过自注意力机制对语音上下文信息的动态提取的过程，是解释自监督深度语音模型与大脑听觉通路表征相似性的关键计算机制。

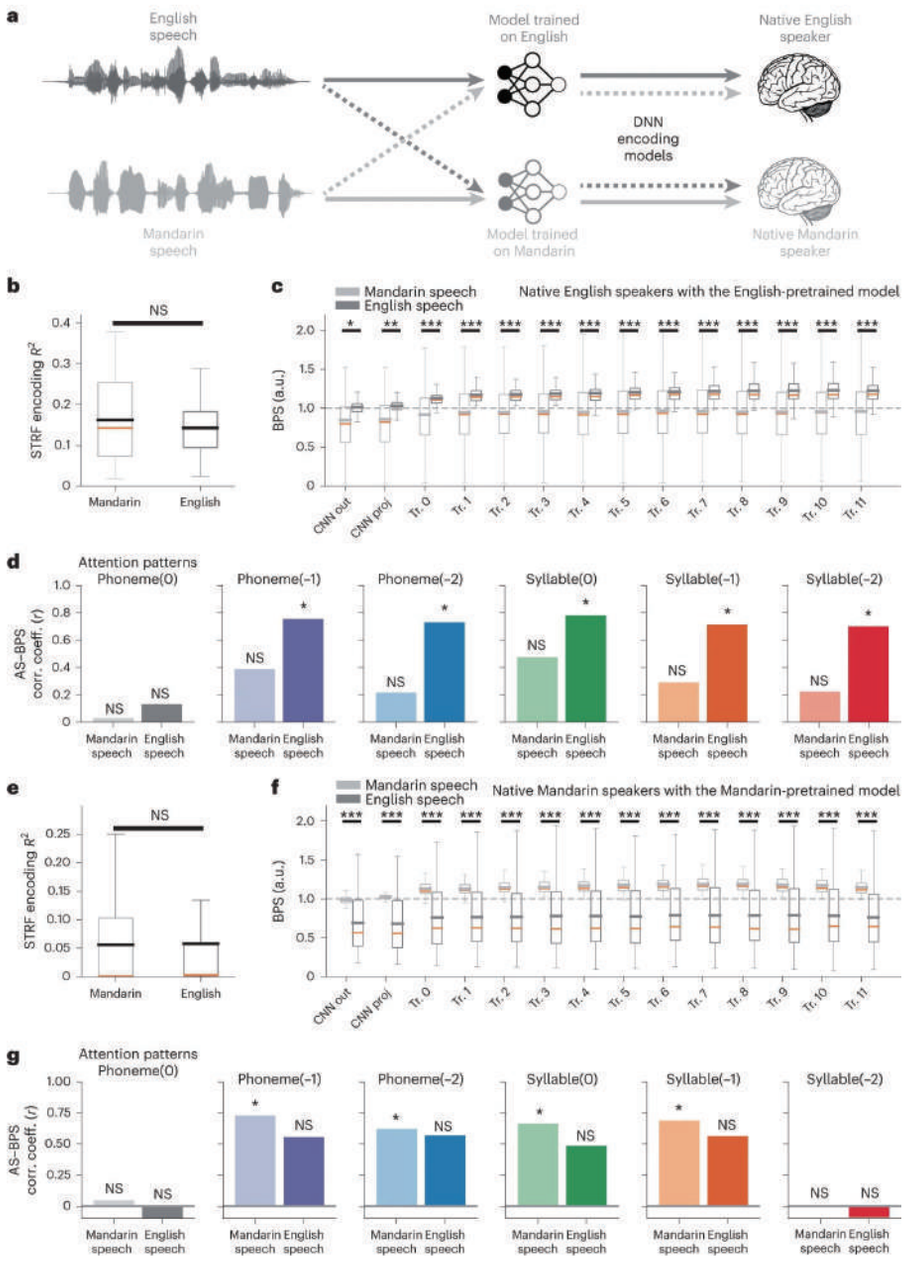


▷图 9：自注意力机制与语音上下文结构的对齐程度解释了深度神经模型对听觉通路语音响应的预测能力。图源：本论文

最后，研究进一步分析了自监督模型是否能够学习到更高层级的上下文信息，通过跨语言比较这一独特的范式[7]，分析了深度神经网络与大脑听觉皮层的语言特异性。为了模拟母语者的语言特异性，研究使用了在英文数据上预训练的英文语音模型，以及在中文数据上预训练的中文语音模型(图10)。

如果仅仅使用线性时频编码模型，也就是STRF model，是无法体现出英文母语者在听英文和听中文时候的语言特异性的，这一点在研究者的前期研究中也验证过，底层声学信息的处理是跨语言通用的。但是如果用英文预训练神经网络来预测神经活动，则可以体现英文母语者对不同语言语音的特异性响应，英文模型更好地预测听英文时的神经响应，并且模型注意力权重与英文的上下文结构信息的对齐程度，与模型性能显著正相关。类似地，如果使用中文预训练模型预测中文母语者的听觉皮层神经响应，则可以体现中文母语者对不同语言语音的特异性响应，中文模型更好地预测听中文时的神经响应，并且模型注意力权重与中文的上下文结构信息的对齐程度，与模型性能显著正相关。

这一双重分离的结果表明，自监督模型能够学习到更高层级的与语言特异性相关的上下文信息，并且这一特异性信息与大脑语音皮层的计算与表征是显著相关的。



▷图　10:不同语言预训练的语音神经网络体现出语言特异性的计算与表征,并与次级听觉皮层的语言特异性神经活动呈现显著对应关系。图源:本论文

从神经科学的角度来看,这项研究与近期发表的多项相关研究[8][9][10]共同提出了基于大规模自监督模型建立语言相关的认知功能计算模型的新思路,展现了自监督语音模型与大脑听觉通路的计算与表征的相似性。从人工智能的角度,这项研究也为打开深度神经网络,特别是自注意力模型Transformer的“黑箱”提供了新的生物学视角。

上海科技大学生物医学工程学院李远宁研究员为本文第一作者,加州大学旧金山分校神经外科Edward Chang教授为本文通讯作者,复旦大学附属华山医院吴劲松教授、路俊锋教授,上海科技大学研究生陈佩利参与了此项研究,该研究参与者还包括来自加州大学伯克利分校、Meta AI以及罗彻斯特大学的研究者。

参考文献

[1] Li, Y., Anumanchipalli, G., Mohamed, A., Chen, P., Carney, L. H., Lu, J., Wu, J., Chang, E.F. (2023) Dissecting neural computations of the human auditory pathway using deep neural networks for speech. Nature Neuroscience, 26, 1-30.

[2] Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network: Computation in Neural Systems, 12(3), 289.

[3] Bhaya-Grossman, I., & Chang, E. F. (2022). Speech computations of the human superior temporal gyrus. Annual review of psychology, 73, 79-102.

[4] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

[5] Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451-3460.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[7] Li, Y., Tang, C., Lu, J., Wu, J., & Chang, E. F. (2021). Human cortical encoding of pitch in tonal and non-tonal languages. Nature communications, 12(1), 1161.

[8] Millet, J., Caucheteux, C., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J. R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. Advances in Neural Information Processing Systems, 35, 33428-33443.

[9] Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118.

[10] Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. Nature neuroscience, 25(3), 369-380.

35年激辩尘埃落定！这项能力不再是人类独有

作者：赵诗彤

中科院神经所硕博连读。在我看来，从数据中理解大脑正在做什么，并用模型对其进行解读是一件很有趣的事情。目前开展的课题集中于想象的神经机制。

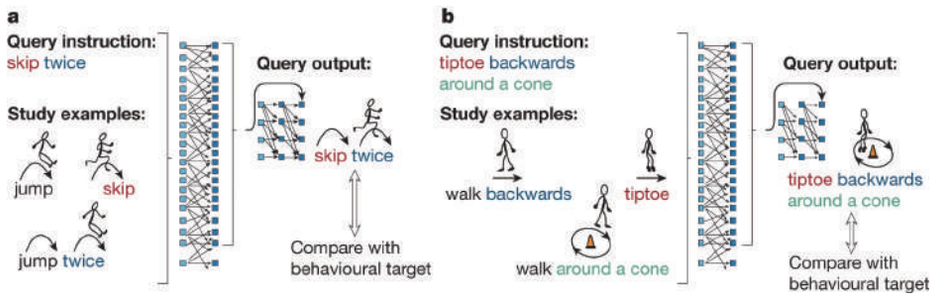
扫码查看原文

语言和思维被认为是人类特有的能力，它们源于大脑强大的系统组合能力，可以通过已知部分理解和产生新的组合。例如，一旦人理解了“photobomb（抢镜）”这个词的意思，就能在各种情境中使用它，比如“photobomb twice（两次抢镜）”或“photobomb during a Zoom call（在Zoom会议里抢镜）”。同样，理解了“跳”之后便顺其自然地能理解“往后跳”。人类这种在新环境中轻松使用新习得词汇的能力，被称为系统泛化能力。

一直以来，科学家试图通过人工神经网络来模拟大脑，然而早在1988年，哲学家Jerry Fodor与认知学家Zenon W. Pylyshyn就曾声称人工神经网络缺乏系统泛化能力，因此并不适合模拟人类思维。在这场长达35年的激烈辩论中，反对者的论点主要集中在两个方面：首先，人类的组合能力也许并不像Fodor和Pylyshyn所说的那样系统化和规则化；另一方面，虽然人工神经网络的基本形式有限，但在使用过程中可以通过复杂架构来使其具有系统性。

近些年，神经网络在诸多方面（比如自然语言处理）取得重大突破，但与系统性计算有关的争论与挑战却一直存在。近期，来自纽约大学的心理学与计算科学家Brenden M. Lake和来自巴塞罗那庞培法布拉大学语言学系的Marco Baroni教授，利用人类实验与计算模型，证明了神经网络经过训练可以实现类似人的系统泛化能力，成功应对了Fodor和Pylyshyn提出的挑战。

为实现这一目标，他们引入了组合性元学习算法（meta-learning for compositionality，MLC）——一种利用少样本组合任务提升模型系统性的训练过程。MLC只需使用普通的神经网络，不需要添加额外的符号机制，也没有人工设计的内部表征或归纳偏差。相反，它强调的是从高层次的指示和（或）直接的人类行为数据中学习我们所期望的行为，这种学习过程又可以称为“元学习”。



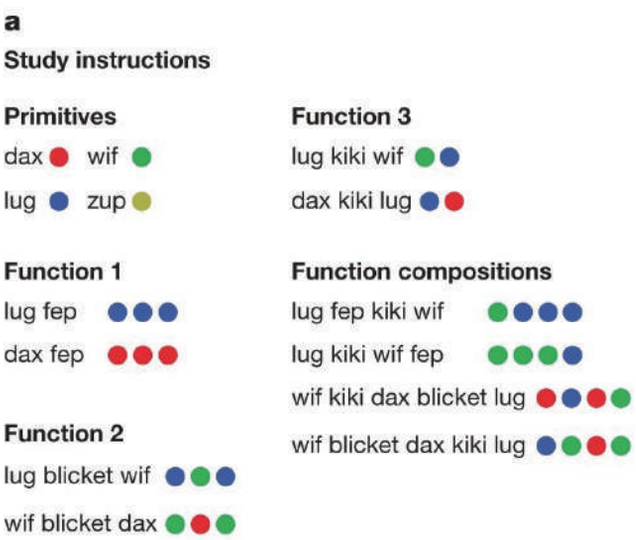
▷图注：网络结构和学习效果的语义图。a、b是两个不同的情境，对应两次神经网络的优化过程。模型的输入是查询指令与学习示例组合而成的整体。情境a中，学习示例演示了“跳两次”（jump　twice）和“跨越”（skip）等指令如何与特定的输出对应，其中指令是单词形式而输出是基于文本的行为符号（箭头引导的火柴人图像）。查询指令中则将“跨越”这一个在训练过程中仅以单独形式出现的词汇进行了组合使用（“跨越两次”，“skip　twice”），结果表明网络能够生成正确的输出。情境b展示的是对于“踮脚走”这个词实现了类似的功能，并且实现了更多情境的组合。图源：论文

为了更好地观察和阐释MLC的泛化能力，研究者直接一对一地比较了人类与模型的表现。他们自创了一套伪语言，并要求被试在使用过程中来生成抽象输出（作为输入的单词序列和作为输出的符号序列是两类无意义词）。这与人工语法学习、统计学习和程序学习有所不同，因为在他们设计的训练过程中不需要明示或隐含的语法判断，后续计算系统能够直接基于人类行为构建序列到序列（seq2seq）的模型。

一、人类行为实验

他们首先测试了25名人类如何通过少样本学习将新学的词汇应用到不同情境中。被试需要学习14个“输入/输出对”，然后为10个新的查询指令生成输出。指令与输出序列之间符合一定的可解释的基础语法关系。

7个无意义单词中的4个被定义为语义词汇，又称为基本词汇，如“dax”、“wif”、“lug”和“zup”代表着基本的具体动作，如“跳”和“跃”；3个被定义为功能词汇，如“blicket”、“kiki”和“fep”，指定了使用和组合语义词汇的规则，从而形成了类似“跳三次”或“向后跳跃”的序列。被试学习时需要学习语义词汇对应的符号输出，还需要通过符号的组合序列知道功能词汇的功能是什么，但他们不会被告知每个指令是什么类型的，而是只会看到如下图所示的14个对应关系。具体如下图所示。



▷图注：指令学习。图源：论文

学习了基本语法后，研究人员让被试处理基本动作和功能的复杂组合，以测试他们应用这些抽象规则的能力。被试必须选择正确颜色和数量的符号，并将它们按照适当的顺序排列。

要取得好的成绩，被试必须从仅有的几个示例中理解单词的含义，并将其推广到更复杂的指令中。正如预测的那样，人类在这项任务中表现出色。平均来看，他们的准确率高达80.7%。这样的准确率是远无法通过随机选择获得的，并且对于长度最长的那些指令来说（在训练时也从未见过），人类被试的准确率也达到了72.5%，这样的泛化能力正是神经网络模型所遇到的困境之一。

研究者还对人类被试的常见错误进行了总结和归类。最常见的错误是“一对一”型的：被试没能理解功能词汇的功能，而是把它当作了具有单独语义的词汇，例如下图中被标记为“1-to-1”的错误。另一类常见错误是“符号串联”，这常常是因为被试使用功能词汇的时候不自觉地在输出中保持了与输入一致的顺序，例如下图中被标记为“IC”的错误。这些响应模式都与日常语言习得中可能发生的偏差相一致。

为了更直接地评估上述归纳偏差，研究者还进行了一个开放式任务：被试不会看到任何学习示例，而是一次性看到所有查询指令，然后对它们之间的关系进行合理推测。通过这种方法，研究者可以更好地观察被试的先验偏好和归纳偏差，因为他们没有先前的学习示例来引导回答。

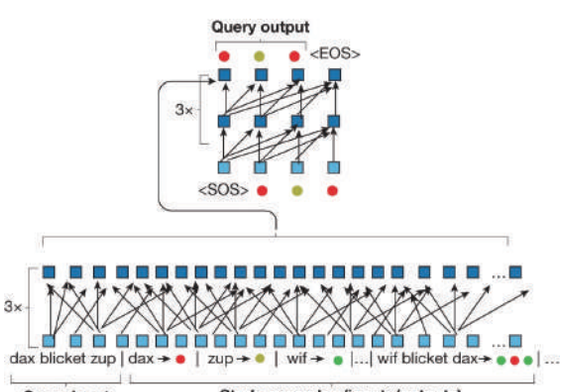
尽管在测试过程中，被试的行为不受限制，但他们的答案仍然高度结构化。除了再次验证了上述两种归纳偏差外，被试的回答还遵循了与互斥性相关的第三种偏差：将独特的含义分配给独特的词汇。这表明人们在处理任务时倾向于将不同的词汇与不同的含义关联起来，以保持它们的唯一性。

二、元学习与transformer架构模型

想要在具有挑战的泛化性任务中模拟人类的系统性泛化和错误模式，一个成功的模型必须能够从极少的示例中以系统性的方式学习和使用词汇，并能捕捉输入与输出之间的结构化关系。

MLC的目标是引导神经网络调整参数值，以便在面对未知任务时实现泛化，并克服以前关于系统性的限制。重要的是，这种方法旨在模拟成年人掌握语言后的组合能力，而不是关注语言习得的过程，后者则是另一个问题。MLC采用的是标准的transformer架构，进行基于记忆的元学习。元学习指的是，不仅仅学习某一对输入/输出之间的联系，而是学会由输入产生输出的抽象规则。

下图展示了模型的编码（底部）和解码（顶部）过程，MLC利用查询输入和学习示例（输入/输出对）来优化transformer。这种方法允许人工智能在完成每个任务时学习，而不是使用静态数据集。每当有新的学习和查询示例出现时，模型都会进行优化。



▷图注：MLC架构。标准的transformer编码器（底部）处理查询输入以及一组学习示例（输入/输出对）；示例由竖线（|）符号分隔。图源：论文

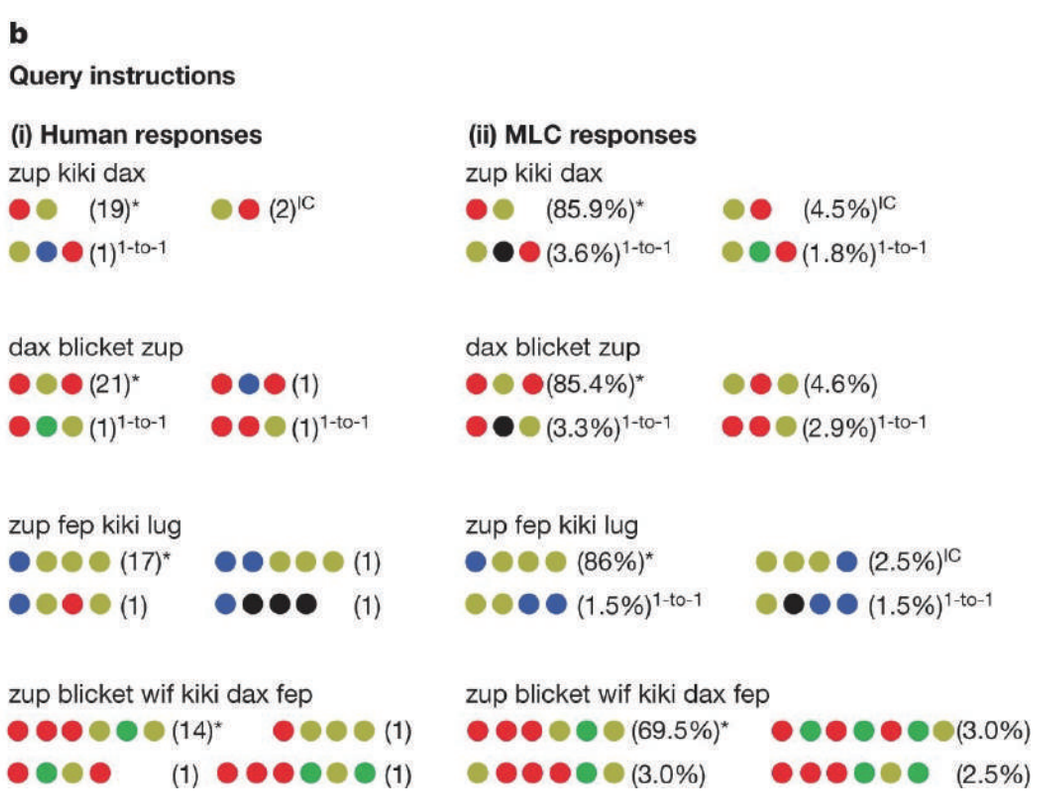
在这个任务中，每个训练集对应一个不同的序列对应任务，也就是每个被试学习的那些随机生成的隐藏语法，用来将输入与输出进行对应。为了成功解码，transformer必须从学习词汇中提取出合适的参数值并利用它们生成查询答案，这个过程依赖于元学习训练，也离不开transformer这样的创新架构。而这些创新在Fodor和Pylyshyn的论证中都没有被预见到，例如可变长度的输入、参数共享和自注意力机制。

在测试阶段，模型权重全都被固定了，也不再提供任何任务特异的参数。最后，考虑到此模型的最终目标是模拟人类行为（包括犯错行为），研究者还将标准答案或错误输出（由一对一转换或错误使用规则产生）随机配对，这种随机配对的比例与人类被试的观察数据大致相同。

三、模型与人类行为对比

接下来，实验者评估了MLC在处理这些具有挑战性的泛化任务时，产生类似人类的系统性泛化和错误模式的能力。结果表明，MLC能够优化模型以实现高度系统化的行为。在某一次训练中模型达到了与人类被试完全相同的系统化行为（100%的精确匹配），而且能够推导出元学习过程（模型训练阶段）中没有出现过的新规则。

为了更深入地比较人类和机器的学习效果，研究者对模型输出的分布进行采样，发现MLC还能实现更微妙的受偏差驱动的行为，使transformer以接近人类表现（80.7%）的平均比例（82.4%）生成了系统化的输出。在处理更长的输出序列时，系统输出的比例（77.8%）也很接近人类水平（72.5%）。在“一对一转换”和“符号串联”这两项人类常犯的错误上，MLC　transformer也表现出了和人类被试相近的比例。此外，MLC的表现还能预测人类行为，即对于MLC来说准确率更低的指令对于人类被试来说也更难。



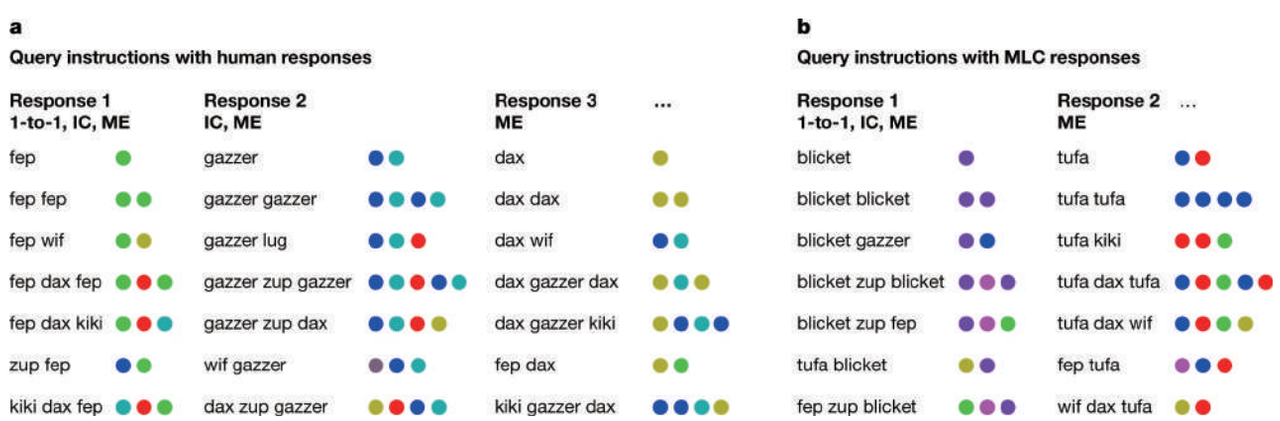
▷图注：四种最常见的回答，括号中标注了回答准确率（人类的用计数表示，MLC的用样本百分比表示）。图源：论文

四、模型间的性能对比

为了更准确地测试MLC模型在完成“少样本学习任务”过程中的效果，作者还训练了另外若干个模型，并比较不同模型与人类回答的相似程度（对数似然）。

用来比较的模型有以下几种。（1）概率符号模型：假设人类可以推断出真实语法规律，但是会偶尔出现随机的失误。（2）带偏差的概率符号模型：一类特殊的概率符号模型，出现失误的概率基于人类发生偏差的频率得出。（3）基础编码-解码模型：仅仅学习某一套规则而不进行元学习训练。（4）仅具复制功能的MLC：一个具有优化复制能力而不是系统泛化能力的MLC模型，在训练阶段查询示例总是匹配某一个学习示例。（5）仅学习代数规则的MLC：一个和MLC经历了相同训练过程，严格符合代数输出回答但是不具有偏差。（6）联合MLC模型：针对少量指令和开放式任务进行联合优化。

结果发现，MLC比大多数模型都表现得更好。不过，在少样本学习任务中，带偏差的概率符号模型的表现基本达到了MLC的水平。这不意外，因为MLC做的是类似的优化，隐式地推断系统规则，并以相同的偏差模式进行作答。虽然MLC和概率符号模型都可以很好地描述人类的少样本学习行为，但在开放式行为测试中，MLC更具优势。



▷图注：开放式指令任务。参与者在没有看到任何示例的情况下对询问（语言字符串）做出了回答（彩色圆圈序列）。图源：论文

训练过程与之前类似，利用相同的transformer模型，基于人类被试在开放式实验中的行为进行优化，然后逐一产生针对七个指令的输出，从而测试模型在开放式任务中的表现。结果发现，在65.0%的样本中，MLC变换器与人类被试产生了完全相同的回答，完美地呈现了三个关键的归纳偏差。而除了基于少样本学习任务训练的MLC模型，表现最好的其实是联合MLC模型，它能够同时实现少样本学习任务和开放式指令任务优化，并且在预测人类行为方面表现出色。

五、基准测试集评估

除了预测人类行为之外，研究者还利用机器学习领域的系统性泛化任务基准数据集——SCAN和COGS——进行了测试。这些数据集中的示例都是由其设计者通过代数规则生成的，没有直接的人类行为数据。本研究重点关注它们的系统词汇泛化任务，探讨如何处理新词汇和词组（而不是新的句子结构）。

SCAN数据集涉及将指令转换成动词序列，例如将“walk twice”转换为“WALK WALK”。在“add jump”分组中，训练集只包含一个“jump”的示例（映射到“JUMP”），测试集则包含了该动词的组合使用（例如“jump around right twice and walk thrice”），这样的功能与之前介绍的人类学习任务相似

（例如“跳跃”可以类比为“zup”）。COGS涉及将句子转换成表达它们含义的逻辑形式，例如，将“一个气球被Emma画了”转换为“balloon(x1) v draw.theme(x3, x1) v draw.agent(x3, Emma)”。COGS评估了21种不同类型的系统化泛化，其中大多数涉及名词和动词的单样本学习。

MLC仍然只使用了标准的transformer组件，但为了处理更长的序列，它在处理学习示例的方式上增加了模块化的设计。为了增加少样本推理和意义组合，研究者在两个基准测试集中使用了表层词类置换——这是元学习简单变体，使用了最少的结构性知识。通过置换，可以在原有词汇表的基础上改变词汇的含义，用来近似更自然、持续引入新词的情况。

总的来说，在两个基准数据集中，MLC的错误率极低。在SCAN测试中，MLC解决了三组系统性泛化任务，错误率低至0.22%以下。在COGS测试中，MLC在18类词汇泛化任务中实现了0.87%的错误率。没有进行元学习的情况下，基本的seq2seq模型在这些基准测试中的错误率至少是元学习的7倍。

六、总结

35年前，当Fodor和Pylyshyn提出关于神经网络的系统性问题时，他们无法想象当今模型所能达到的效果。尽管将本文开发的工具应用于各个领域还有很长远的路要走，但从中可以看到，元学习在使人工智能系统的行为更像人类方面确实大有可为。自然语言专家Elia Bruni表示，这项研究可能会使神经网络成为更高效的学习者。这将减少训练诸如ChatGPT等系统所需的巨大数据量，并减小模型中“幻觉”的问题，即当人工智能感知到不存在的模式并创建不准确的输出。（编辑：Lixia）

参考文献

Lake, B.M., Baroni, M. Human-like systematic generalization through a meta-learning neural network. Nature 623, 115–121 (2023). https://doi.org/10.1038/s41586-023-06668-3

https://www.nature.com/articles/d41586-023-03272-3

超越感知：那些基于生物感官的AI算法

作者：郭瑞东

科普作家，关注复杂系统与神经科学。追问nextquestion、集智俱乐部长期撰稿人，曾为知识分子，果壳等多家媒体撰文，科普书《机器学习与复杂系统》合著者。

扫码查看原文

在我们探索宇宙和深海的同时，人类最复杂的前沿仍然隐藏在我们自己的头颅之中。神秘而复杂的大脑，这个自然界中最复杂的已知结构，不仅是思想和感觉的源泉，还是我们对世界感知的根本。正如物理学家理查德·费曼所说，“我所不能创造的，我便不能理解（What I cannot create, I do not understand）”，要想解锁这个谜团，我们便需要从头构建一个类似的人造大脑。当人造大脑已能复现人脑特征，那我们对自然之脑的理解无疑更为深入。

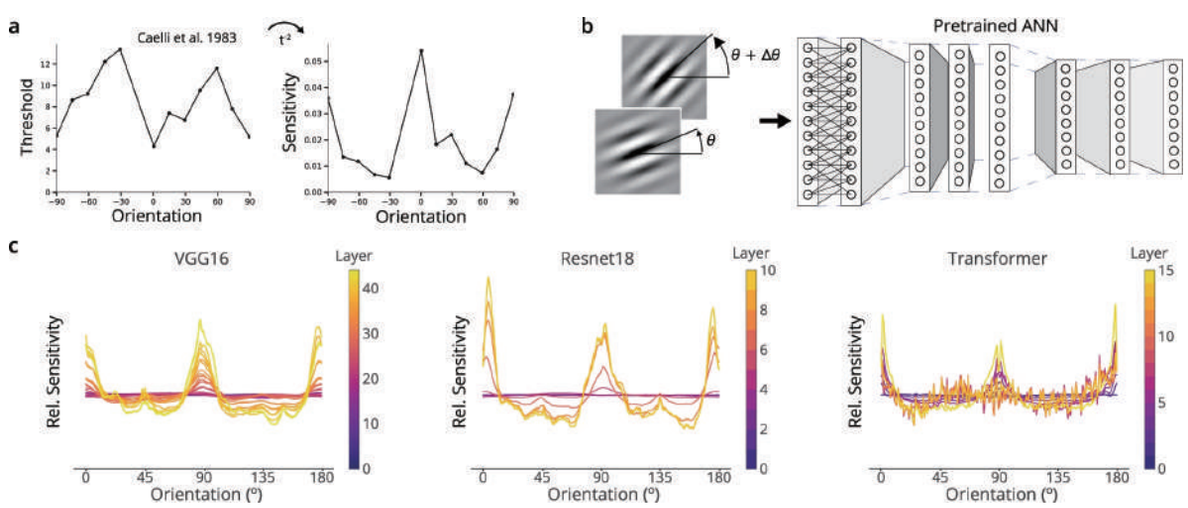
从视觉、听觉，再到嗅觉，我们的大脑处理感官信息的方式超乎想象。它能从混乱的视觉图像中辨识出熟悉的面孔，从嘈杂的环境中捕捉微弱的旋律，甚至在气味的复杂混合中识别出特定的气味。但这一切不仅仅是生物学的奇迹，也是人工智能未来发展的蓝图。通过学习和模仿大脑处理信息的方式，科学家们试图在人工神经网络中复现这些现象。本文将穿越人类大脑与人工智能之间的复杂迷宫，从视觉、听觉、嗅觉概述相关研究，探索人脑与人工神经网络的相似与差异，为下一代智能系统的设计提供方向。

一、视觉感知基础：从简单到抽象特征的提取

当我们观察事物时，存在一个明显现象，相比倾斜方向的图像，我们更容易觉察到垂直或水平方向的图像。这被称为倾斜效应（oblique effect）。就像下图所展现的那样，我们倾向于更清晰地记住直立的树和山脉，而非它们倾斜的根茎。这种偏好可能反映了我们祖先在进化过程中发展出的能力。垂直和水平方向的特征在环境中更为常见，因此视觉系统在构建高效表征时，需要对这些方向的变化更敏感，这将有助于我们更快地识别和响应环境变化。

▷图1：倾斜效应示意图。在任意随机选取的自然场景照片中，水平和垂直方向也比倾斜方向更常见。这种情况也出现在人造建筑物中。来源：Cognitive Psychology: Connecting Mind, Research and Everyday Experience 3rd Edition

为了搞清楚这一现象究竟从何而来，研究者[1]基于卷积神经网络的模型VGG16及 Resnet18，以及非卷积网络架构的transformer，分别构建人工神经网络，来模拟这一现象。他们发现，经过训练的网络在对不同方向的刺激进行处理时，显示出与人脑类似的倾斜效应。具体来说，这些网络在0度、90度和180度的方向刺激上的反应更为敏感。这是我们首次观察到人工神经网络和大脑涌现出相同的特征，而这一主题将在之后反复出现。



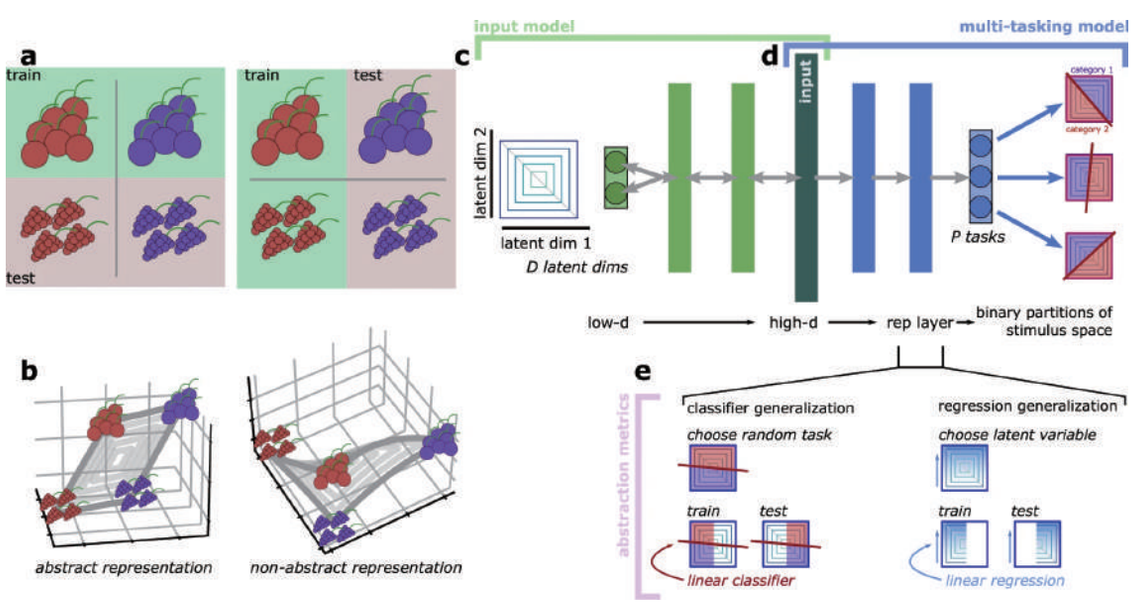
▷图2：被训练分类自然出现物品的人工神经网络，会呈现类似人类的编码敏感性。来源：参考文献1

当我们进入森林，会发现一排排几乎相同的树木。人脑为了高效地对其进行表征，会利用数据不变性对重复的元素进行压缩，以提高信息处理的效率。这被称为平移不变性。这种机制在人工神经网络（卷积神经网络）中也找到了对应。更有趣的是，即使是基于全连接神经网络，只要训练它们识别自然界中的图像，它们也能涌现出类似的具有局部、空间平铺的感受野[2]。这意味着即使在没有预设网络结构的初始状态下，人工神经网络也能自发学会先聚焦图像中的一小块，之后看到类似的结构，就直接从记忆中调用已有的存储，高效地处理相似结构。

在处理视觉信息时，大脑常常需要应对输入数据的缺失和噪声。但面对这些问题，人脑会在必要时将注意力集中于更大的图像，而不是陷入多个细节部分。例如，在遇到部分遮挡的图形时，我们的大脑倾向于视觉上“填补”缺失的部分，将其视为一个完整的形状，这就是闭合法则。训练用来分类自然图像的神经网络也会呈现出类似补全的倾向[3]。

除开遮挡造成的信息缺失之外，观察时间过短也将带来输入的缺失。就如“看不见的黑猩猩”一样，当我们过于关注某事时，可能会忽略显而易见的视觉信息。这表明即使是短暂出现的、未进入我们的意识图像，也能在我们的视觉皮层中留下印象。例如，在实验中，即使被试者只是瞥了一眼或根本没有意识到某个图像，他们仍然能够在一定程度上识别图像内容，如判断图像中是否存在生物[4]。使用卷积神经网络架构，研究者发现当数据存在缺失/噪音时，模型仍能在一定范围内进行判别。这一现象与人脑的处理方式相似。

进一步地，从处理简单的线条到局部的平移不变，再到复杂图像的缺失补全，人脑对外物的感知逐渐从具象走向抽象，从而得以在面对不同的环境和对象时经由归纳形成知识。在机器学习领域，这样的能力被称为解耦（disentangle）。研究表明，人工神经网络通过监督和强化学习，在需要解决多个任务时，能够自发地涌现出抽象表征[5]。这些抽象表征有助于大脑在新的任务上实现少数样本的学习和有效的泛化。



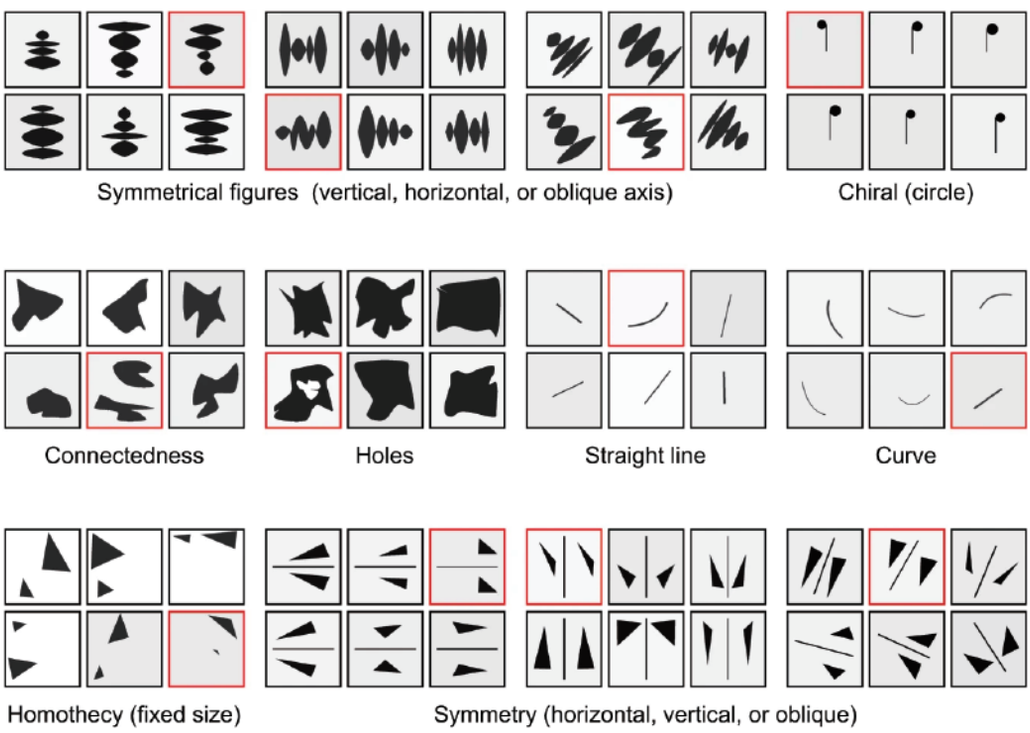
▷图3：a）两个分类任务的例子。（左）在一种形状的红色和蓝色浆果之间学习的分类可以泛化到其他形状。（右）两个不同形状的红色浆果之间的分类可以推广到不同形状的蓝色浆果。（b）四个浆果例子的线性、抽象（左）和非线性、非抽象（右）表示的例子。（c）输入模型的示意图。（d）多任务模型的示意图。（e）两个抽象指标，分类器泛化指标（左）和回归泛化指标（右）。来源：参考文献5

在视觉系统成功识别出抽象特征之后，其下一步任务是理解这些特征之间的复杂关系。这一阶段是对信息的深入处理，涉及到模式识别和逻辑推理，是人类视觉处理中极为高级的功能。相比之下，虽然人工智能在许多领域已显示出卓越的能力，但在处理这类抽象任务时，它们通常需要更多的资源和能量。

以找不同的Oddity Test为例，人类通常能够轻松地完成这种测试。然而，对于人工神经网络来说却很难。为了解决这一难题，科学家们给AI提供了小抄[6]，即利用人类志愿者在完成Oddity Test时的眼动数据来训练神经网络。这些眼动数据包含了人脑在观察图像时自然而然地关注的局部间关系，从而为网络

提供了一种模拟人类观察行为的方式和抽象推断的线索。这种新型的生物启发网络展现了更高的准确性、更快的学习速度，以及更少的所需参数。*

*作者注：我们是否可以通过类似的方式训练其他动物，如黑猩猩或某些鸟类，来完成Oddity Test?若是发现基于人的眼动数据训练的AI模型，当输入换成黑猩猩或鸟类的眼动数据后，也是可以适用的，那就意味着人脑的抽象思维能力不仅并不特殊，还可能是从动物的类似机制中继承而来的。不过以上是笔者个人的异想天开，供读者讨论。



▷图4：Oddity test的例子。来源：参考文献6

二、抽象处理：从社交到分类

我们大脑中有一些特殊的神经元，它们非常擅长于识别和辨认人脸。这些神经元能在我们还是婴儿的时候就开始对脸部特征做出反应。但这种能力是否是我们天生就有的，还是随着视觉经验而发展出来的，一直是科学家争论的热点话题。

有趣的是，利用捕捉视觉皮层腹侧区特征的人工神经网络模型发现[7]，即使是没有接受过特别训练的深度神经网络，也能展示出类似的能力，它们能够“自然地”识别人脸（从随机前馈线路中自发产生），这表明我们的大脑可能有着与这些系统相似的处理机制。

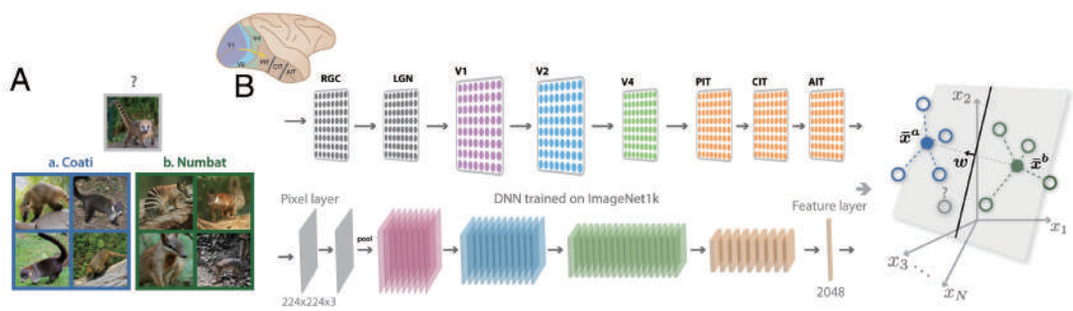
此外，我们大脑识别人脸的能力并不完美。比如，面对不熟悉的种族的人脸，或者人脸的图像被上下颠倒时，我们的识别准确率就会下降。这种现象曾被认为是人脑对人脸识别特有的特点。

但是，通过研究基于卷积神经网络的人工智能模型，科学家们发现，这些“缺陷”其实是大脑为了更高效地识别人脸而进行的优化[7]。类似的人脸图像正反颠倒识别准确率下降的现象，只出现在以人脸识别为训练任务的卷积神经网络中，没有出现在接受过物体识别训练的卷积神经网络中。如专门用于识别汽车的人工智能模型中，当汽车的图像倒置时，相比正向图片，这些模型的识别能力也会下降。这说明，我们大脑的这些特点其实是对特定任务的优化，而不是什么独特的计算本质。

人脑可以从少数例子中学习概念，这对于传统的深度神经网络来说，这种从少量例子中学习的能力被

称为少样本学习(few-shot learning)。那么,大脑是如何从少数例子中学习的?答案在于大脑将视觉特征映射到一个高维空间中[9],在这个空间构建一个能够跨越相距甚远空间的"虫洞",并通过一种称为"流形学习(Manifold Learning)"的技术进行学习。理论上只用200个神经元就可区分不同类的输入。

模仿这种机制,人工神经网络也在采用类似的方法。通过配置一个灵活可塑的下游神经元,基于少数样本和简单的规则,人工神经网络就能学会如何区分不同的概念。这种方法的优势在于,它不仅能够处理视觉数据,甚至能根据语言描述符来学习和识别新的视觉概念。



▷图5:人脑和DNN基于4个实例进行训练的示意图。来源:参考文献9

在学习过程中,人脑大部分时间是没有明确指导的。这就像孩子在识别家里的猫和狗时,不需要有人一直在旁边告诉他们这是猫那是狗。这种学习的方式被称为半监督学习。人脑在半监督学习的模式下表现优良。

相似地,人工神经网络也能使用深度无监督对比嵌入方法进行学习。这种方法使神经网络在处理视觉信息时,特别是在大脑的腹侧视觉皮层相关区域,达到甚至超过了当前最先进的监督学习模型的神经预测准确率[10]。即便是仅使用头戴式摄像机收集的、嘈杂且有限的真实人类儿童发育数据,这些神经网络仍能有效地学习并产生类似大脑的表征。研究还发现,半监督深度对比嵌入可以利用少量标注示例生成表征,从而大幅提高错误模式与人类行为的一致性。

三、听觉

就像我们的眼睛对某些视觉特征特别敏感一样,我们的耳朵也对特定的声音特别敏感。比如,在嘈杂的环境中,我们依然能分辨出音乐的旋律或人的声音。

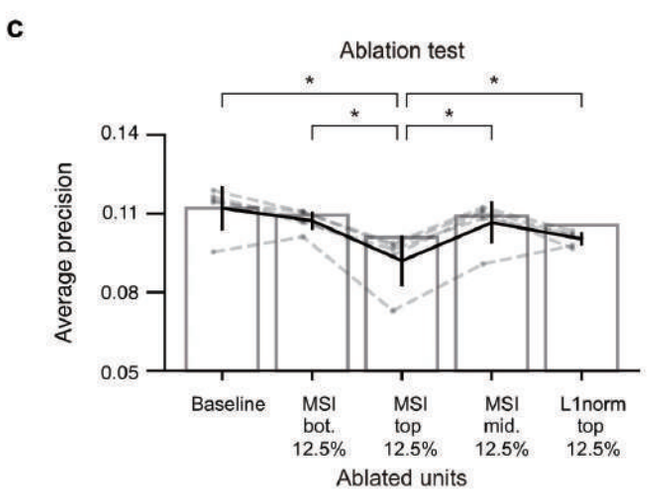
当科学家用深度神经网络训练完成复杂听觉任务来模拟这种听觉处理时,训练完成的模型能够很好地完成这些任务,达到人类的水平[11]。但有趣的是,只有在用真实的音乐和自然声音训练时,这些模型对音高的感知才能表现得像人类一样。如果用人工合成的声音或在没有任何背景噪音环境下训练,这些模型就会展现出完全不同的音高策略。这表明,我们的听觉系统也许真的是为了应对复杂、有时会被噪声遮蔽的环境而优化。

人脑不仅能识别特定的声音,还能判断声音来自哪个方向。通过比较声波到达左右耳的时间和强度差异,我们可以估计出声音的来源,这被称为定位。在现实世界中,环境会产生回声,而且我们会同时听到很多声音,因此定位尤其困难。然而,当科学家在虚拟世界中训练深度学习模型[12],并使用与人耳相同的信息采样精度时,这些模型能够在定位任务上与人类达到同样水平,且表现出与人类相似的缺陷,例如当面对多个声音来源时,定位的准确性下降。与人类一样,模型同样在面对多于3个声源时无法定位。模型还扩

展了对定位使用机制的解释，例如发现定位时模型不仅基于声音的强度和时间差异，还会对两耳间频率的频率敏感。

除环境的声音感知外，感知音乐则更为复杂，它包含识别旋律，感知固定的音高和节奏等不同部分。然而这些特征究竟是从何而来的?2019年的一项研究发现[13]，一种用于音乐风格分类的深度神经网络，能够达到与人类相当的水平，而且它的错误模式也类似于人类。这表明无论是人造的还是自然的听觉系统，在处理音乐时都受到相似的限制。该研究进一步展示了这个模型如何模仿人类听觉皮层的反应，其复杂的网络层次结构能够精确地预测对不同音乐元素的反应，这反映了我们听觉系统中固有的层次化处理方式。

音乐感知是我们天生就有的能力，还是随着时间和经验而发展的?2024年初[14]的研究表明，使用模拟大脑听觉信息处理机制的人工深度神经网络，即使训练数据中不包含音乐，也可以通过经由自然声音训练，使人工神经网络自发地涌现出适应检测音乐的专用单元。这些音乐检测单元在多个时间尺度上编码音乐的结构，对音乐的细微变化极为敏感，类似于我们大脑中处理音乐的方式。而当在人工神经网络中抹去这些音乐选择节点时，网络在音乐分类任务上的表现明显下降，这证明了我们对自然声音的处理能力可能为我们对音乐的感知提供了基础。换句话说，音乐感知可能是我们听觉系统进化适应的一部分，是对声音处理的一个通用且高效的模板。



▷图6：音乐选择性MSI最高的12.5%人工神经网络单元抹去后，对该网络分类准确性的影响最大。来源：参考文献14

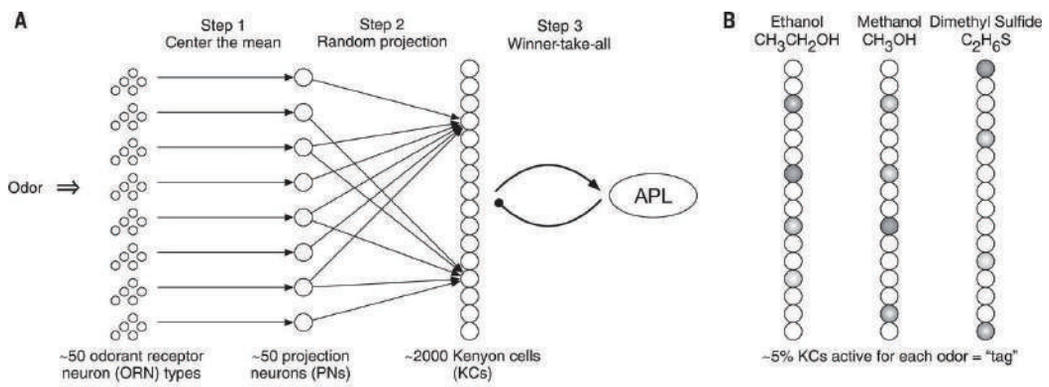
四、嗅觉

与我们熟知的视觉和听觉不同，嗅觉没有一个直接的物理到知觉的映射。换句话说，我们不能像将光的频率直接对应到颜色，或声波的波长对应到音调那样，将特定的化学分子直接对应到特定的气味。嗅觉要解决的问题是将众多对分子浓度的感知进行压缩并追踪来源，而任务的难点在于这些分子的排列没有固定的边界或结构气味没有边缘，不是可以在空间中分组的对象）以及气味信号很是稀疏（有的气味不需要特别多的分子脑也要能识别）。

针对上述的两个问题，索尔克研究所的神经生物学家的做法是构造浅层的三层网络来模拟果蝇的嗅觉[15]。这个网络设想作为一只果蝇，它需要识别50种不同的气体。但其脑中的负责嗅觉的神经元不可能时刻处在待机状态，去判断是否有对应分子的到来。

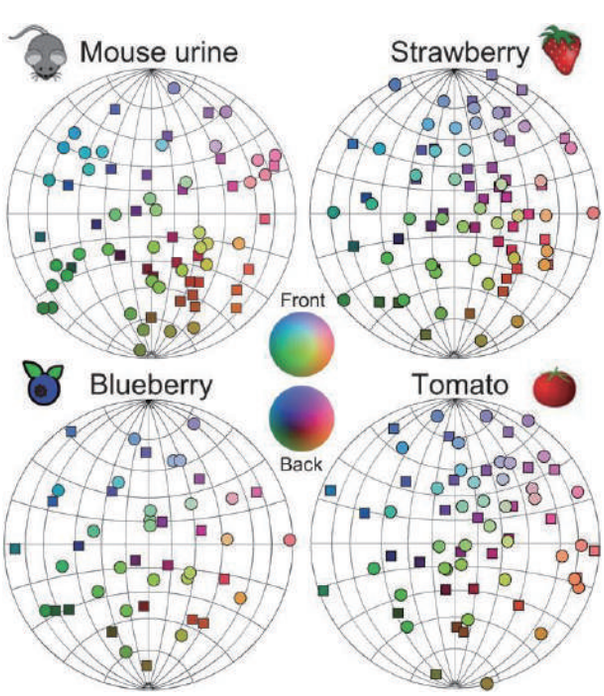
在实验中，图7A中的50个气味分子对应了果蝇的50个投射神经元，图7B中的每个分子，可被看成对应一个长度50的向量。来自投射神经元的信息多对一地到达名为Kenyon细胞处理中心。这个系统采用了一种“赢者通吃”的机制，即在众多神经元中，只有对特定气味反应最强的那个会被激活。

上述机制构成了维度的40倍扩展，这使神经反应模式更鲁棒地区分气味（一个气味被多个下游神经元表征），能耗也更少（只有大约5%的神经元对给定的气味反应高度活跃，就可为每个细胞提供一个独特的标签）。将类似的技术用在搭建手写数字识别网络上，可让系统在部分硬件失灵的场景下运行。



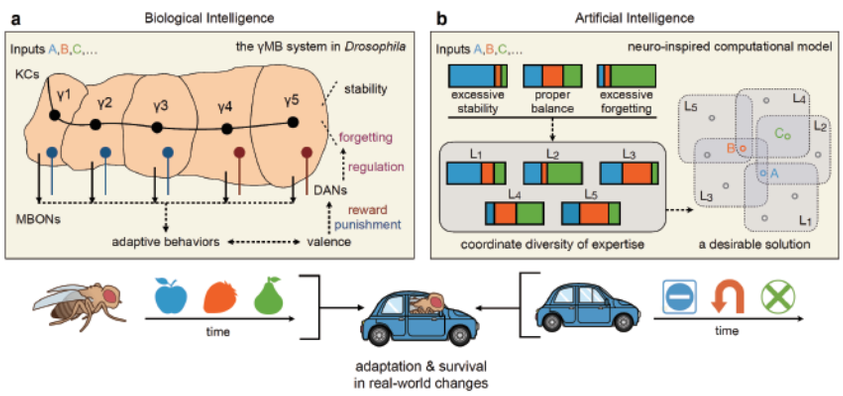
▷图7：(A)苍蝇嗅觉回路示意图。(B)气味反应的示意图。相似的气味对（如甲醇和乙醇）被分配了更相似的标签。来源：参考文献15

当面对来自同一生化反应的多种化合物时，生物的嗅觉系统会通过激活相同的神经元来节省存储空间。如果两种化合物很少同时出现，那么它们就会被映射到大脑中相距较远的区域。这就像在三维空间中创建了一个类似马鞍的三维双曲空间“嗅觉地图”[16]，其中相似的气味会被归为一类，而不同的气味则被清晰地区分开来。这种空间上的排列方式，帮助我们大脑有效地处理和识别复杂的气味组合。这与根据像素相似度在双曲空间创建的地图类似。



▷图8：不同物体对应的嗅觉地图，圆形/矩形表示近/远侧。来源：参考文献16

另一方面，科学界也在借鉴嗅觉机制来改进AI[17]。例如，通过设计一种能够“主动遗忘”的多层学习系统，这些系统能够处理不断增加的数据，并在学习过程中整合新旧信息。这种方法类似于生物嗅觉系统中多个并行处理模块的工作方式，最终由大脑中的一个集成模块“磨菇体”（mushroom body），来决定哪些信息最可靠。这种架构（无论是生物界，还是模仿产生的AI模型）将主动遗忘与稳定性保护结合起来，可更好地权衡新旧信息，并相应地协调多个持续学习者，确保解决方案具有持续学习的能力。



▷图9：基于生物体的持续学习架构示意图。来源：参考文献17

五、总结

无论是从大脑获得启发以改进神经网络，还是基于神经网络的特征去探讨大脑的运行机制，我们可以梳理出几条反复出现的线索。首先，不论是在杂乱的环境中识别细微的音高变化，还是用少量待机神经元在大量信息中精确地检测气味，我们的神经系统都在努力在有限的资源下，以刚刚好的程度将任务完成。这说明了感官处理的方式是受到我们进化历程中的种种约束的。其次，大脑并没有打算一招鲜吃遍天，而是针对不同的环境输入给出了特异性的解决之法。这值得一直在追求“一种技术解决所有问题”的AI界学习。再者，大脑对感官的处理也并非完美，大脑如同总在摸鱼边缘疯狂试探的打工人，会产生种种认知缺陷，例如难以识别不熟悉种族的面孔。但这种不完美也反映了大脑对信息处理的一种经济学平衡，如视觉系统对不完整图像的自动补全。

回顾了众多相关领域的有趣研究后，我们可以发现当前的研究关注最多的是感官是视觉，最多的是生物是人。然而，我们的感官并非所有生物中的最佳的那个。我们的视觉并非如鸟类那样，具有四周色感细胞，能看到更多彩的世界，听觉更也不如能够超声定位的蝙蝠。基于脑（尤其是感官）启发的AI算法，未来可以不必只聚焦于人身上，而应该更多地探索其他生物的感官系统。

当前的AI架构，无论是在反向传播还是神经网络设计方面，都和大脑真实的运行机制有相当大的差距。大脑中的单个神经元并非简单的对输入加权求和，而是具有亚细胞层面的可塑能力。这使得AI借鉴大脑时，要切记勿要刻舟求剑，而是要搞清楚生物体在具体任务上面临的进化约束，如此方能避免邯郸学步。

站在百年的门槛回望，我们仍然不清楚人造大脑能否实现，但毫无疑问，我们已在人工智能中复现大脑的某些关键特征，包括其优势和缺陷。而这也指引着AI领域的迅猛发展，还更进一步帮助我们理解大脑。尽管模拟大脑的终极目标可能仍然遥不可及，我们对大脑的模仿和理解，正逐步引领人工智能超越现有的人类智能边界，为我们带来前所未有的创新和突破。（编辑：存源）

参考文献

[1] Benjamin, Ari S., et al. “Efficient Neural Codes Naturally Emerge Through Gradient Descent Learning.” Nature Communications, vol. 13, no. 1, Dec. 2022, https://doi.org/10.1038/s41467-022-35659-7.

[2] Ingrosso, Alessandro, and Sebastian Goldt. “Data-driven Emergence of Convolutional Structure in Neural Networks.” Proceedings of the National Academy of Sciences of the United States of America, vol. 119, no. 40, Sept. 2022, https://doi.org/10.1073/pnas.2201854119.

[3] Kim, Been, et al. “Neural Networks Trained on Natural Scenes Exhibit Gestalt Closure.” Computational Brain & Behavior, vol. 4, no. 3, Apr. 2021, pp. 251–63. https://doi.org/10.1007/s42113-021-00100-7.

[4] Mei, Ning, et al. “Informative Neural Representations of Unseen Contents During Higher-order Processing in Human Brains and Deep Artificial Networks.” Nature Human Behaviour, vol. 6, no. 5, Feb. 2022, pp. 720–31. https://-doi.org/10.1038/s41562-021-01274-7.

[5] Johnston, W. Jeffrey, and Stefano Fusi. “Abstract Representations Emerge Naturally in Neural Networks Trained to Perform Multiple Tasks.” Nature Communications, vol. 14, no. 1, Feb. 2023, https://doi.org/10.1038/s41467-023-36583-0.

[6] Woźniak, Stanisław, et al. “On the Visual Analytic Intelligence of Neural Networks.” Nature Communications, vol. 14, no. 1, Sept. 2023, https://doi.org/10.1038/s41467-023-41566-2.

[7] Baek, Seungjun, et al. “Face Detection in Untrained Deep Neural Networks.” Nature Communications, vol. 12, no. 1, Dec. 2021, https://doi.org/10.1038/s41467-021-27606-9.

[8] Dobs, Katharina, et al. “Behavioral Signatures of Face Perception Emerge in Deep Neural Networks Optimized for Face Recognition.” Proceedings of the National Academy of Sciences of the United States of America, vol. 120, no. 32, July 2023, https://doi.org/10.1073/pnas.2220642120.

[9] Sorscher, Ben, et al. “Neural Representational Geometry Underlies Few-shot Concept Learning.” Proceedings of the National Academy of Sciences of the United States of America, vol. 119, no. 43, Oct. 2022, https://doi.org/10.1073/p-nas.2200800119.

[10] Zhuang, Chengxu, et al. “Unsupervised Neural Network Models of the Ventral Visual Stream.” Proceedings of the National Academy of Sciences of the United States of America, vol. 118, no. 3, Jan. 2021, https://doi.org/10.1073/p-nas.2014196118.

[11]Saddler, Mark R., et al. “Deep Neural Network Models Reveal Interplay of Peripheral Coding and Stimulus Statistics in Pitch Perception.” Nature Communications, vol. 12, no. 1, Dec. 2021, https://doi.org/10.1038/s41467-021-27366-6.

[12] Francl, Andrew, and Josh H. McDermott. “Deep Neural Network Models of Sound Localization Reveal How Perception Is Adapted to Real-world Environments.” Nature Human Behaviour, vol. 6, no. 1, Jan. 2022, pp. 111–33. https://-doi.org/10.1038/s41562-021-01244-z.

[13] Kell, Alexander J. E., et al. “A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy.” Neuron, vol. 98, no. 3, May 2018, pp. 630-644.e16. https://-doi.org/10.1016/j.neuron.2018.03.044.

[14] Kim, GwangSu, et al. “Spontaneous Emergence of Rudimentary Music Detectors in Deep Neural Networks.” Nature Communications, vol. 15, no. 1, Jan. 2024, https://doi.org/10.1038/s41467-023-44516-0.

[15] Dasgupta, Sanjoy, et al. “A Neural Data Structure for Novelty Detection.” Proceedings of the National Academy of Sciences of the United States of America, vol. 115, no. 51, Dec. 2018, pp. 13093–98. https://doi.org/10.1073/pnas.1814448115.

[16] Zhou, Yuansheng, et al. “Hyperbolic Geometry of the Olfactory Space.” Science Advances, vol. 4, no. 8, Aug. 2018, https://doi.org/10.1126/sciadv.aaq1458.

[17] Wang, Liyuan, et al. “Incorporating Neuro-inspired Adaptability for Continual Learning in Artificial Intelligence.” Nature Machine Intelligence, vol. 5, no. 12, Nov. 2023, pp. 1356–68. https://doi.org/10.1038/s42256-023-00747-w.

大语言模型为神经科学带来了哪些前所未有的机会?

编译:郭瑞东

科普作家,关注复杂系统与神经科学。追问nextquestion、集智俱乐部长期撰稿人,曾为知识分子,果壳等多家媒体撰文,科普书《机器学习与复杂系统》合著者。

扫码查看原文

大语言模型(LLMs)是机器学习领域中用于处理和生成自然语言文本的新工具。该文提供了对该建模技术的属性定义,并反思LLMs如何被用来重新构建经典的神经科学问题,以提供新的答案。我们认为LLMs有潜力(1)通过添加如高级文本情感分析等有价值的元信息来丰富神经科学数据集;(2)总结大量信息来源,以克服孤立的神经科学社群之间的隔阂;(3)促使与大脑相关的不同信息源得以进行前所未有的融合;(4)帮助确定哪些认知概念最有效地把握大脑中的现象。

一、背景介绍

语言不仅仅是交流的工具,它还蕴含着丰富的人类智慧和信息,在一个比特中,其内涵的丰富性远超过我们通常所接触到的任何其他数据形式。自然语言处理(NLP)是一门致力于让计算机拥有理解、解读以及创造人类语言的能力的科学,它在分析和处理人类文本资料方面已经取得了显著的进展。

在早期,研究者借助如n-gram模型这样的简单语言模型(例如,2-gram模型将词-词组合视为独特实体)用来研究语言和语义,以达到各种目的。这些语言模型不时地被用于研究各种认知任务,包括阅读理解、语言翻译和问题解答等。通过比较NLP模型在这些任务上的表现与人类的表现,研究人员获得了关于人类认知的洞见,就如心理语言学领域所示。

大约2010年之后,深度学习的兴起点燃了NLP建模中的语义"嵌入"时代——单个词语、句子、段落或整个文档,都可以封装在一个紧凑的浮点向量格式中,以此向量来表示对应词句的意义。从直观上讲,这种嵌入方式类似于在高维坐标系中定位,使得不同的语义实体(如词序列)根据它们在上下文中的相似性被映射到相近的位置[1-4]。两个语义实体表示的上下文越相似,它们的语义嵌入就越相似。使用像Word2Vec[5]和GloVe[6]这样的最新一代模型,研究人员开始使用这些可互操作的语义嵌入表示来量化意义之间的关系,如词语或句子之间的关系。

当前的大语言模型(LLMs)是在比一个人在数百或数千个生命周期中能阅读的文本还要多的数据上训练的。这种庞大的数据训练基础使它们涌现(emergent)出了诸多能力,如编写计算机编程代码、数学、规划、文献综述和总结,或玩基于文本的游戏等。这些能力并非在它们的各个组成部分中原本就有,而是随着系统复杂度的增加而涌现出来[7]。有时,这样的模型被用来研究大脑如何处理上下文信息以及人类心智如何产生语言(请参见Goldstein[8]、Caucheteux[9]和Schrimpf[10]的优秀示例)。随着当前研究范式的转变和大模型规模呈指数级增长,LLMs学习了迄今为止可能是最强大的意义内部表征。

人类语言反映了人类思维,这就是为什么最先进的NLP可能会为神经科学研究提供内生的优势。从这个角度来看,该文试图讨论大模型对神经科学和生物医学研究者带来的即将到来的影响。

二、数据科学的角度,大语言模型解决方案

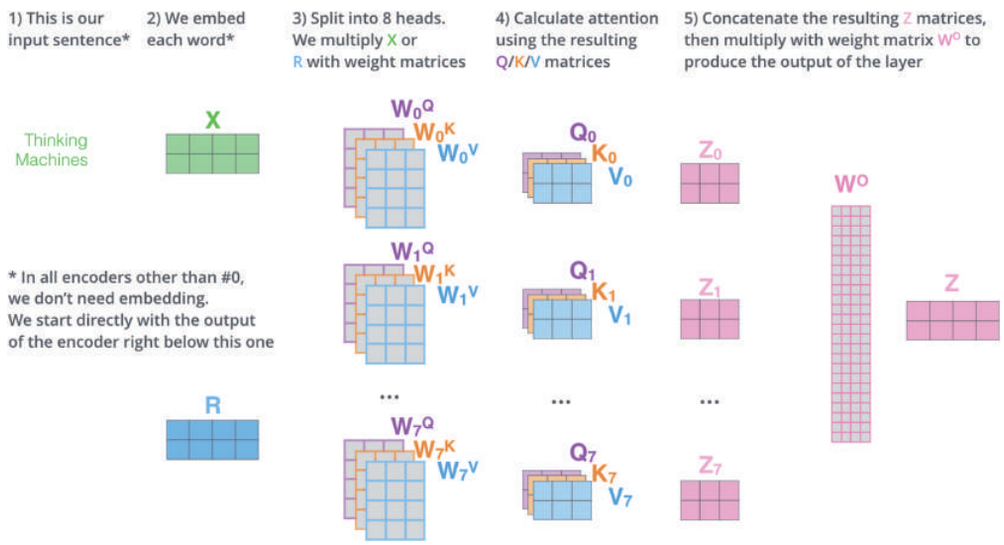
历史上,自2010至2012年以来,卷积深度神经网络(CNNs)因其在处理图像等网格结构数据上的优势,重燃了人工智能的热潮,而LLMs目前正在为AI生态系统注入另一波动力。特别是在引入了Transformer架构之后,语言模型取得了显著进展(例如,Vaswani等人[11]的研究在发表后的前5年内被引用超过9万次),推动了当前AI创新的推动力。

GPT-2在多个语言任务上表现出色,它由24个Transformer块组成,而最新的架构更是深入发展,一些细节仍待揭晓。作为所谓的“生成性AI”的一个实例,这些算法的输出不是类别(例如,患病与健康组的患者)、数字(例如,认知表现测量)或离散类别(例如,年收入的区间),而是一种结构化的内容,如语言(以及图像或音频信息),即从之前输入的内容中合成或幻想新内容。

相较于以往复杂的深度神经网络,Transformer架构以更简洁的特性成为NLP领域的新宠(见图1和图2)。这种简化的架构比之前的方案更具可扩展性,部分原因是这种架构非常适合并行计算工作流程。与之前的深度NLP解决方案不同,在Transformer架构中,无论是近还是远的词标记之间的相互依赖关系,都能同样好地被捕捉。

与某些之前的神经网络设计不同,Transformer模型是前馈深度学习架构,不包含显式的处理循环。相反,通过将已经生成的、之前的文本作为输入反馈到LLM中(“自回归”),创建了隐式循环。与围绕BERT(使用双向上下文来理解词义)的前一代LLM不同,生成性预训练Transformer(GPT)架构,如ChatGPT,在训练期间只关注当前词之前的词标记,这导致了其单向处理模式,即具有自回归性质。

Transformer中所谓的位置编码是该架构的一个特征,它帮助模型理解词序。在自然语言处理中,自回归模型会根据前面的词来预测下一个词。由于其单向性质,GPT式LLMs在预测下一个词时不会“看到”或“考虑”后续的词标记——它是在回顾给定句子的过去,而非展望未来,正如人类阅读书籍时的方式。

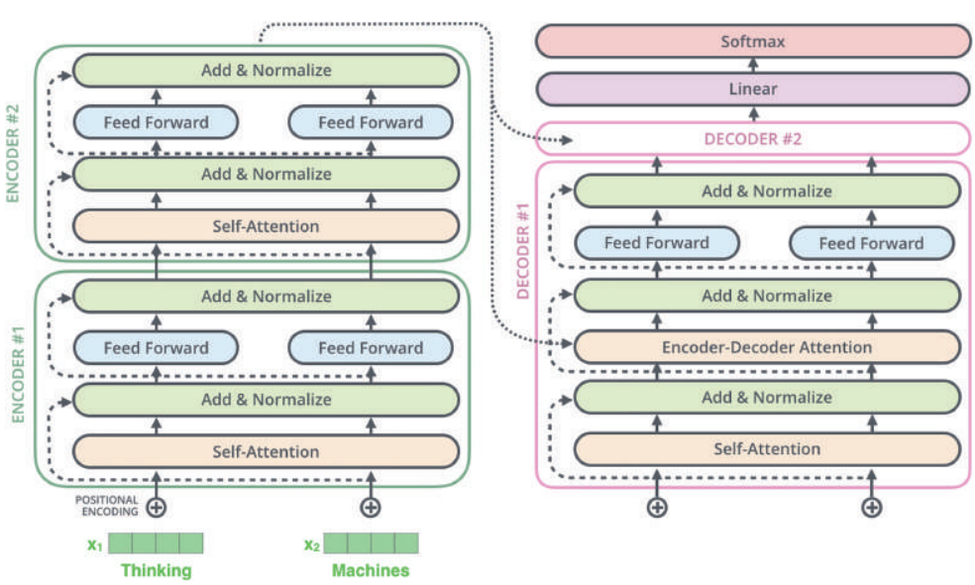


▷ 图一：transformer架构的大语言模型的核心-自注意力机制

正是自注意力机制构成了Transformer建模架构的核心。自注意力机制允许模型在处理序列数据时，对序列中的每个元素分配不同的注意力权重，从而更好地捕捉序列内的长距离依赖关系。

专注于更近或更远的词标记在算法上是相同的——不需要经过逐步迭代的过程来关联更远的信息片段，这与早期深度学习架构的要求不同。在Transformer中，关注句子中附近或远处的词的处理方式是相同的，该架构允许模型同时考虑句子或文本序列的所有部分。与早期的神经网络相比，这意味着不需要按顺序处理输入的远端部分。

自注意力机制的常见实现的计算复杂度与序列长度成二次方相关[12]。尽管在注意力机制上有所改进，但在处理特别长的序列时，大多数情况仍然遇到困难[13]。每个Transformer层可以一次性“看到”其范围内的所有标记。然而，递归信息处理的深度受到连续Transformer层数的限制，例如句子的嵌套意义或数字序列的连续乘法。



▷ 图二：自注意力（self-attention）层在transformer架构中的作用

此外，当前的LLM架构通常在每个连续的Transformer层中设置了几个并行的注意力机制。这种“多头注意力”（1）允许同时并行关注输入序列的几个不同方面，扩大了整体可以捕捉的复杂性范围；（2）因此允许同时识别和提取多个语义表示维度（有些类似于建模不同的潜在因子成分，如主成分分析或自编码器神经网络）[14]。

值得注意的是，温度参数（temperature，一个在0到正数范围内的标量值）是一个影响模型表现的超参数。这个超参数控制着模型输出的创造性程度，作为一种平衡探索与利用（exploration versus exploitation）的形式。设置高温度（例如，>（1）会在最后一层模型中产生更均匀的词概率分布。这导致输出更加模糊，因此可能不够准确，但也更具创造性。相反，低温度（例如，<（1）会导致输出词相关性的概率分布更加尖锐。在这种操作模式下，模型变得更加具有确定性，紧密遵循输出分布中最可能的候选词，从而减少了响应中的随机性。

尽管模型目标简单（例如，BERT调用词掩蔽，GPT3调用下一个词预测，而在GPT4/ChatGPT的情况下涉及人类反馈），但由于其庞大的规模，Transformer赋予的架构已经引发了小样本学习（few-shot learning）和在多个情景中生成派生语义世界模型的能力[15]。这些能力是自监督建模制度的核心。这些次生能力甚至让这些模型的创造者在解释LLMs的成功时花了不少功夫[16]。

三、LLM解决方案的涌现标度定律

规模效应的极限是什么？作为影响模型性能的关键量，随着训练观察数量的增加，LLMs的模型生成实例的质量迅速提高。在拥有大约2到20倍于模型参数的训练词标记数量时，LLMs已经在多个场合取得了令人印象深刻的表现。

从数据的角度来看，很难感知到模型所需的可用文本、转换文本（ext-transformed）和可转换文本数据（text-transformable data）的上限。具体来说，根据简单的规范假设（根据ChatGPT查询，截至2023年9月，大约有120亿个网站，每个网站平均1500个词），互联网上所有文本的总量可能达到约2万亿词标记。从模型的角度来看，从2018年到2022年，LLMs的规模（参数）从大约10^8（例如，ELMo，BERT-L）增加到大约10^11（例如，PaLM）。作为许多应用场景中的一个基本原则，扩大模型的深度和宽度（增加参数数量）会带来明显的性能提升。了解模型性能如何随规模变化具有战略价值，因为此类见解可以指导资源分配决策：如何确定计算预算、数据资源和模型大小的优先级。

更具体地说，深度学习文献中一项全面的、广泛认可的经验研究，探索并仔细地基准测试了跨越七个数量级的模型规模所带来的影响[17]。这些研究者设计了计算实验，成功地确定了决定模型性能变化的三个关键因素：（1）模型参数的数量（N）；（2）可用数据量（D）；以及（3）用于模型估计的计算能力（C）。在这些实验中，模型性能仅轻微依赖于模型架构的实际形状。通过同时增加N和D，似乎在很大程度上防止了过拟合（即对训练数据中的特殊性的过度适配）。相反，如果只增加N或D（但保持另一个因素固定），性能会下降[18]。最后，N、D和C的持续扩大显示出回报递减的模式，遵循幂律法则。

然而，最新的研究进展指出，与最初增加模型规模的趋势相反，LLMs在所需参数数量上会随着训练越来越小[18,19]。对许多研究者来说，这似乎是反直觉的，再次减少模型规模，可能更好地与实际可用数据量对齐，提高了模型性能，放宽了内存要求，并减轻了计算成本。这些改进可能对LLM解决方案在现实世界问题中的应用至关重要，并增加了未来几年智能手机作为广泛使用的移动设备携带专用LLMs的潜力。

简而言之，一个新兴的研究表明，相对于模型参数规模，更多的数据在某种程度上更为重要，尽管两者都是推动模型性能提升和发展的关键因素。

值得注意的是，衡量模型性能在很大程度上取决于研究者选择的评估指标[20]。这些作者认为，只有通过选择特定的评估指标，才能显现出LLMs的“涌现能力”（即在模型规模增大时，模型展现出的新能力或行为[15]）。与上述观点相反，Schaeffer等人[20]还展示了评估指标的选择可以在不同的架构和任务中诱导出看似涌现的能力。因此，最近的实证研究[20]表明，改变评估指标可以削弱或增强LLM架构中涌现能力的信号，这对AI安全和AI对齐有直接影响。

总的来说，较大的LLMs在微调和小样本学习场景中比较小的LLMs更具样本效率。也就是说，矛盾的是，需要估计的模型参数越多，实现相近性能所需的输入数据点就越少。正如在数据科学中一般，提高数据质量总是可以带来进一步的性能提升。重要的是要承认，神经网络的幂律定律在目前上几乎完全是经验性的，但这些幂律特征显示出稳健的趋势[21]。LLM架构的扩展和爆炸是由（1）transformer的发明，这些transformer在最近的LLMs中变化不大；（2）大量数据源的可用性；以及（3）大规模计算能力的可用性推动的。与下一部分相关，模型的具体架构（如层数、层维度等）相对不那么重要，尤其是随着模型规模的增加。

四、LLMs展现出前所未有的迁移学习能力

为了使深度学习工具蓬勃发展，通常需要丰富的数据。然而，神经科学领域的许多领域并没有现成的大量数据可用，更不用说像AI社区中用于文本和图像分析的互联网规模数据集。这种差异引发了一个问题：我们可以利用哪些丰富的非神经科学数据来建模解决方案，之后迁移到神经科学问题上？

“迁移学习”（transfer learning）是一种数据分析模式，其核心在于解决一个问题时积累的结构化知识，可以之后被应用到一个不同但相关的问题上。迁移学习旨在提高在类似但数据资源可能严重受限的任务上的模型性能。在深度学习的背景下，这通常意味着首先在大规模数据集上预训练模型作为起点，然后，通过轻微调整（微调）模型参数，将这个模型应用于与特定任务相关的较小数据集上*。这个过程利用了这样一个假设：预训练模型学到的特征可以作为通用表示，对目标任务有益。历史上，迁移学习的成功通常依赖于预训练和微调任务之间的高度相似性。

*请参见https://www.ruder.io/transfer-learning/，了解LLMs微调技术的全面资源

LLMs和其他基于Transformer架构的模型在迁移学习方面展现出超出预期的能力，从而通过扩大可执行任务的范围，彻底改变了自然语言处理（NLP）。作为一个关键的转折点，直到最近，主导范式仍然是在大规模语料库上进行有监督的模型预训练。这需要大量的高质量标注数据，而这些高质量的标注不易获得，严重限制了互联网和其他来源可用数据的实际预训练和迁移学习。直到现在，通过无监督预训练，无需为每个数据点提供精确的注释就成为可能，这标志着性能的一个巨大飞跃。这一分水岭事件极大地扩展了LLMs预训练可用数据的范围。

更正式地说，LLM中需要估计的参数越多，模型开发过程就越慢。LLMs开启了全新的微调领域，超越了以往模式中学习算法能够实现的任务范围。研究人员已经提出的几种方法，可以在只更新或添加相对较少的参数的情况下，使模型适应新任务。其中一种策略是“冻结”（保持不变）预训练LLM的多个层的参数。这种方法接下来只调整下游任务的一小部分可调参数，从而避免在神经网络学习新任务时，遗忘之前

习得知识的现象。

在微调过程中，通过向LLM中添加新的可学习层，可以进一步扩展这种策略。这样新增的“适配层”可以显著减少目标任务的训练时间和计算成本[22-24]。研究已经证明，选择特别高质量的数据用于微调，即使在目标任务的样本量较少的情况下，也能给迁移学习后的模型带来有竞争力的性能。LLMs在小样本学习方面表现出色。在极端情况下，即便没有为新任务提供示例，仅利用预训练的LLM进行零样本学习，也已证明LLM即使没有调整预训练模型，其零样本学习能力也使其在各种下游任务中表现出色[25,26]。

简而言之，LLMs包含着数十亿个可调模型参数，通过其庞大的规模，解锁了从大规模文本语料库中提取本质表征的能力，而不再迫切需要监督标签注释。无监督深度学习在实践中被证明更具可扩展性。因此，对于那些没有能力从零开始训练LLM的神经科学家来说，通过微调已经预训练好的模型来适应感兴趣的特定任务，可以充分利用这些模型的先进性能，同时减少对数据和计算资源的需求。LLMs可以更好地识别文本中的深层隐藏模式、关系和上下文，这使它们能够回答人类的查询、创造性地生成新内容，以及形成准确的结果预测。

五、作为计算乐高积木的基础模型

基础模型最初是在大规模文本语料库上进行训练的，例如互联网内容和其他公共或私人来源的数据。这让它们能够发展并构建一个通用的内部语义表征，该表征包括语法和句法，尽管LLMs在多大程度上包含了对语义的理解目前还存在争论[20,27,28]。更进一步，这些模型学习了大量的通用知识、展现了一定的推理能力，以及对可能的语义世界的表征。基础模型的演变可以追溯到transformer时代之前的上一代NLP模型（2017年之前），如Word2Vec[5]和GloVe[6]，它们在连续向量空间中表达词语（参见1背景介绍），这暗示了语义空间的普遍性。

通过从不同的多样来源提炼和吸收精华，基础模型形成了一个通用表征，它包含了庞大、紧凑和密集的人类知识，作为下游建模的先验知识。这不仅仅包含记忆，且包含信息提取和结构化。从哲学上讲，这种对信息的成功压缩可以视为预测能力的一大飞跃，因为成功的预测本身就是一种信息压缩的体现。类似于共享基础设施或平台，这样的AI引擎可以作为多种任务构建的基础，使许多定量建模工作流程变得可行、高效且易于扩展。这些基础模型就像是乐高积木，因为许多下游应用可以在它们之上构建，就像堆叠积木一样。这种对定量建模的新态度与为狭窄任务部署训练专门模型相反。

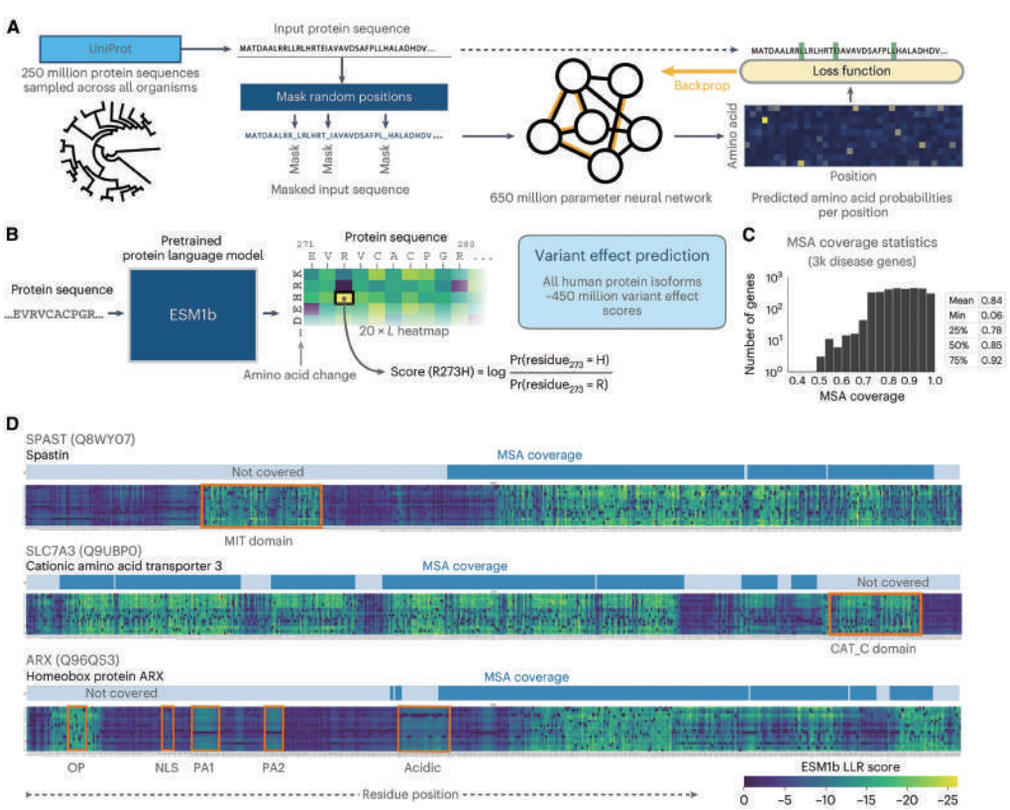
利用数千个GPU处理数万亿个词元，几周时间内就能完成LLM的训练，其成果能被存储并部署至智能手机中。未来的基础建模框架将提供通用的计算单元，这将有可能使广泛的研究者能够民主化地访问高质量的AI解决方案。这对于神经科学尤为重要，因为神经科学家往往需要在比核心机器学习社区更小的数据集上进行操作。同样，在生物研究中，即使在人类细胞图谱项目中也是如此。截至本文撰写时，该项目也只产生了来自约6,000名捐赠者的约4千万个人类细胞的基因表达数据。

如何创新性地利用这些基础操作系统，以全新视角审视并解决经典研究问题，将是一场大胆的创新之举——在transformer类模型出现之前，这些应用是完全不可想象且和不可行的。使不同领域的研究人员能够启动共同的计算模型模块，也可能有助于提高研究之间的可比较性，并促进不同机构和地理位置间的团队合作。随着资源日益紧缩，深度学习的突破性成果将变得更加容易获取。基础模型在不久的将来，极有可能彻底改变神经科学和生物医学领域的生物信息学面貌。

六、大型语言模型在生物序列中的应用

LLM学习引擎的归纳能力，不仅适用于词序列，也同样适用于各种类型的生物序列，这提供了许多未被充分挖掘的研究机会。人类的基因组，这一包含约2万个基因的庞大DNA序列库，构成了大脑及身体其他部位细胞内蛋白质合成的基石。在此基础上，“生物学中心法则”为我们提供了一个与神经科学直接相关的自然试验场景，它描述了遗传信息如何从DNA中的核苷酸序列，通过信使RNA中的碱基序列，最终转化为蛋白质产品中的氨基酸序列的流动过程。

遗传学家的主要目标是映射这种遗传信息的传递过程，将DNA序列本身的改变与相应的功能影响联系起来。为此，MetaAI展示了一种蛋白质语言模型（图3），该模型能从遗传变异的差异中预测表型后果[29]。通过一个拥有6.5亿参数的模型，研究者能够推断人类基因组中大约4.5亿种可能的错义变异效应——每种变异都是DNA中单个核苷酸的替换，这一替换可能导致下游蛋白质中的氨基酸交换（有害或良性）。DNA基因编码中的这些变异特别有趣，因为它们涉及可以与疾病机制和可能的治疗目标联系起来的蛋白质改变。这种方法使我们能全面分析人类和其他生物整个基因组中的蛋白质破坏性损伤变异。

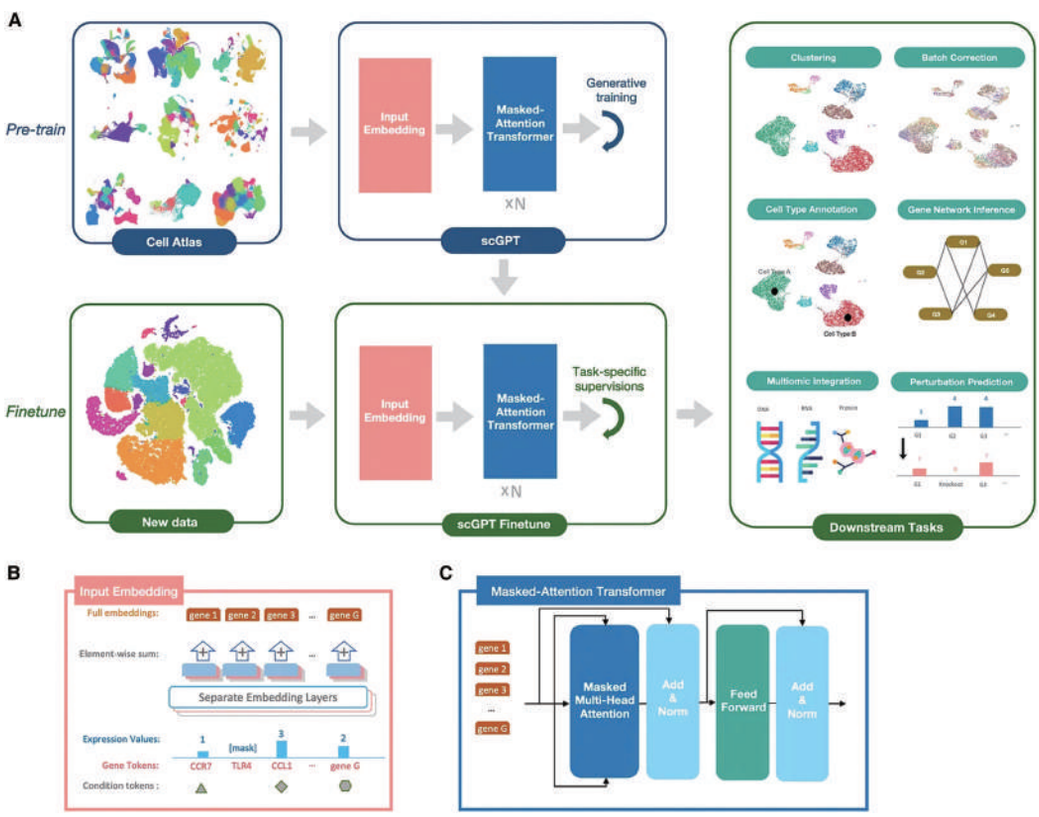


▷ 图3：蛋白语言模型可预测基因突变的功能影响

此外，我们能否仅从RNA转录表达数据中自动推导出细胞状态和与活跃生物通路相关的洞见？在单细胞RNA基因表达水平上，一个LLM[30]在1000万个细胞上进行了训练（图4），每个细胞包含大约2万个人类基因的一部分表达值。作为一个基础模型的开创性例子（参见上文），基因集在生物学建模中构成有意义的过程，类似于词语集合在语言中构成有意义的句子。通过吸收大量的基因表达模式，该模型形成了基因间关系和基因-细胞关系的普遍性内部表征。除了特定基因的标记外，模型还引入了特殊的标记来表示元信息，如细胞类型、数据批次以及实验条件，如信号通路的扰动和RNA转录测序使用的技术。

作者还取消了输入必须是序列的需求——他们设计了一个任务定制的注意力机制，以紧密把握表达基因的协同出现模式。通过基于迭代预测集合中新基因表达的自回归生成，类似于在连续的句子中预测下一个词汇。也就是说，他们将传统语言模型处理词序的方式，巧妙转化为在单细胞基础模型中处理细胞对应的基因集合，从而摆脱了输入必须遵循序列的限制。

这样一来，模型一旦建立，训练好的基础LLM就可以进行微调和部署，并在各种不同的下游任务中获得性能提升，包括批次校正、细胞类型注释和目标扰动条件的预测。这种方法不仅展现了自监督学习技术掌握复杂单细胞机制的潜能，还能利用后续的内部嵌入表示，实现不同器官和物种间的数据整合。



▷ 图4：从指数级增加的单细胞转录组数据中，构建基础模型以描述细胞转录的语法

从基因层面到3D蛋白质结构层面的转变，关键在于能否仅凭一维氨基酸序列预测出蛋白质的最终三维构型。蛋白质折叠问题围绕着我们的DNA是如何压缩关于最终蛋白质结构信息的。在数据库中有超过2亿种蛋白质结构，AlphaFold[31]这一基于大型语言模型的系统，通过捕捉沿着蛋白质骨架上远离彼此的氨基酸残基之间的序列相互作用。在这个给使用暴力（brute force）学习的研究中，研究者展示了1D序列信息确实包含了理解蛋白质在自然界中实际折叠的复杂过程所需的关键信息。

在蛋白质到功能层面，研究者在250亿个蛋白质序列（UniParc数据库）的860亿个氨基酸上训练了7亿参数的34层transformer模型[32]。模型内部的嵌入表征仅从序列信息本身获得。训练好的模型被发现能够包含蛋白质的生化特性、体内形态结构元素、接触位点和生物活性相关知识。

总的来说，捕捉长距离相互作用（即输入序列中相距较远的标记）不仅在词序列推理中，也在不同生物序列中有意义的一般原则方面显得非常有价值。自然界似乎隐藏着可以被利用来推断超出实际序列元素

(例如,核酸、基因表达、氨基酸)的潜在一般规则,以服务于下一代计算生物学。学到的序列嵌入可以用于各种下游研究目标,包括质量控制程序、生物实体的分组以及增强表型预测。

此外,LLMs作为一个平台现已能够实现生物学中心法则的先进计算模拟,从DNA的双螺旋结构到基因的转录表达,再到完整的蛋白质形态。也就是说,一旦LLM能够准确地近似目标系统,便能通过复现严格实验中的可靠观察,使研究者得以向询问LLM询问,以提取关于目标系统的新分子洞察,并识别更广泛的驱动生物机制。我们警告不要将基于LLM的功能预测模型和分子生物学系统之间视为严格平行,因为两者存在显著差异。尽管如此,在未来,LLMs仍将占据独特的位置,有望帮助发现从未在自然界中观察到的生物活性序列。

七、用于自动化数据标注的大模型

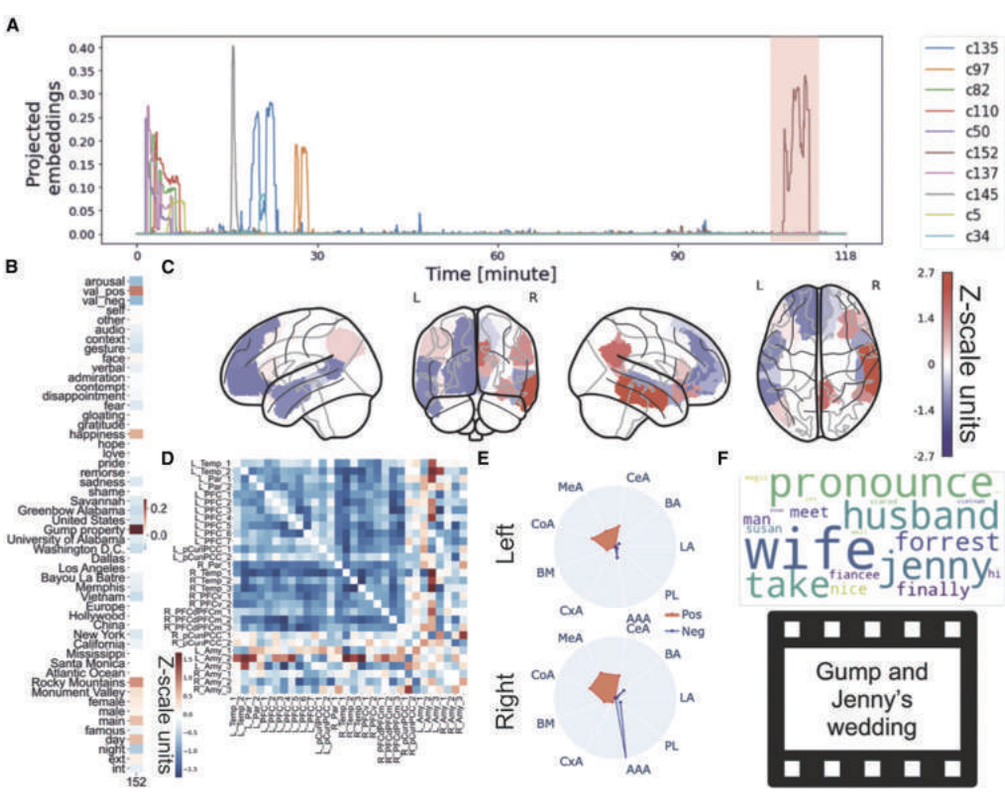
神经科学研究经常依赖于准确的数据标注来阐述数据、设计实验或解释结果。

(1)文字标注

最近一项使用传统NLP的研究探讨了观看电影《阿甘正传》的受试者的大脑反应信号与电影故事的发展之间的关系,即电影叙事的语义元素如何与大脑活动相联系[33]。这项研究是依赖于数据点相关高质量标注研究的典范。这项研究利用了来自studyforrest*数据库收集的大脑记录,每个受试者在观看2小时电影的过程中,其全脑神经活动的3000张个体图像被详细捕获。

*https://www.studyforrest.org/data.html

为了使该数据集更丰富,电影中的每一个场景都通过计算生成的元信息得到增强。为此,文本数据来自之前未充分利用的来源:即与视频内容同步显示的时间锁定字幕,和面向盲人的仅听觉叙述版本的电影,后者描述了电影中的事件和场景,展现了NLP支持下的数据增强的初步尝试。



▷ 图5:基于电影文本,使用NLP进行多模态脑-文字数据整合

《阿甘正传》的逐场景文本信息被一个词袋模型(bag-of-words)所捕捉——在电影整个播放过程中,该模型会收集每个时间片段内所有独特词汇及其出现频率的集合。然后使用潜在语义分析来将场景词频分解为独特的语义维度,以捕捉故事线中的潜在意义和反复出现的主题。

与此同时,通过一种经典的自上而下方法,人类标注者(一群学生)通过从电影的视听版本中选择52个预定义的"指标"来手动为场景添加标签。这些选择基于场景的情感内容、情境和其他方面,基于现有知识,这些方面预先被认为与电影场景相关。这种经典方法虽然强调了基于人类观察者的自然主观体验对人类情感的详细刻画,然而事实上却错过了基于文本派生语义表征中,被很好地反映了的重要细节。这一自动标准的成功,展示了未来LLM方法在自然神经科学中的潜力。

超越手动标注的局限,NLP方法(如潜在语义分析)使得故事被分解为200个语义上下文单元,每个单元基于上下文都与特定场景紧密相关。作为人类衍生情感标注的补充,语义上下文提供了追踪角色(例如,丹中尉)、情境(例如,战争)和场景属性(例如,白天与夜晚)的方法。通过整合数据分析,揭示了大脑状态与场景中特定元素、概念和主题之间的经验联系。因此,算法派生的语义方面在电影-大脑-文本分析中,相较于传统依赖人类先验直觉确定的那些最重要的方向,展现出了更为显著的成功。

LLMs为将其他学科对人类行为的知识和概念引入到脑科学研究中提供了前所未有的机会。批量标注生成可以极大地增强我们处理复杂操作协议的能力,如上述研究中使用的图像和视频数据,以及许多其他形式的数据,如电子健康记录、语音记录或可穿戴设备捕获的生物测量。

历史上,这些数据形式的标注需要人类专家的输入,无论是直接还是间接。现在,通过专门针对特定的从输入到输出的端到端工具,例如直接从视觉数据中训练神经网络来识别人类情感,或基于物理特性预测气味化合物吸引力的电子嗅觉设备[34],我们可以更高效地完成这些任务。手动标注通常存在一些问题,LLMs可以缓解其中几个问题,包括(1)手动操作的高物流和财务成本,(2)用于派生标注标签的分类系统的本体论限制,(3)人类标注者的主观性以及基于主观性的数据,以及(4)可重复性。

最终,如上所述,由于成本高,手动标注的视觉和语言数据集相对罕见且规模较小(10,000-100,000个数据点)。为了应对先前的标注数据稀缺,许多研究[35-37]已经开始自动从互联网和其他通用来源抓取现成的配对视觉-文本数据。现在,以在文本-文本标注场景中也实现与图像-文本标注领域取得的类似成就。在模型预训练后,LLMs可以自动生成标注,这些预训练是基于与手头标注任务相关的各种数据完成。

举一个例子,一家生物技术公司有兴趣为描述精神活性药物体验的一手描述打上标签,以指示不同的主观效果;这些描述和手动标注的主观效果标签配对,可以用于公司使用的基线模型的微调。或者,像GPT4这样的LLMs可以在没有任何额外训练数据的情况下执行此任务,基于其训练集提供了足够的上下文来区分描述不同的主观效果术语及其示例。

短语和句子,就像单个单词一样,可以被自动赋予信息丰富的语义嵌入,这一过程同样适用于自动(或手动)获得的标注。通过将自然语言经LLM"编码器"预处理为嵌入向量,我们可以对离散的语义元素进行连续的量化。以互补的方式,LLM"解码器"用于将嵌入转换回语言文本。将自然语言作为嵌入进行预处理,为探索不同语言模式与神经活动之间的相关性开辟了新方法。将自然语言数据与神经测量相关联,是朝着深刻理解人类大脑产生、感知、处理和解释语言的一步。自然语言文本的定量表证是计算分析中使用的行业通用中间形式,具有可重复性,可调整和可扩展增强的潜力。语言作为封装来自五种人类感官的信

息的工具，提供了人类经验中多样化现象的量化表征。

（2）图像标注

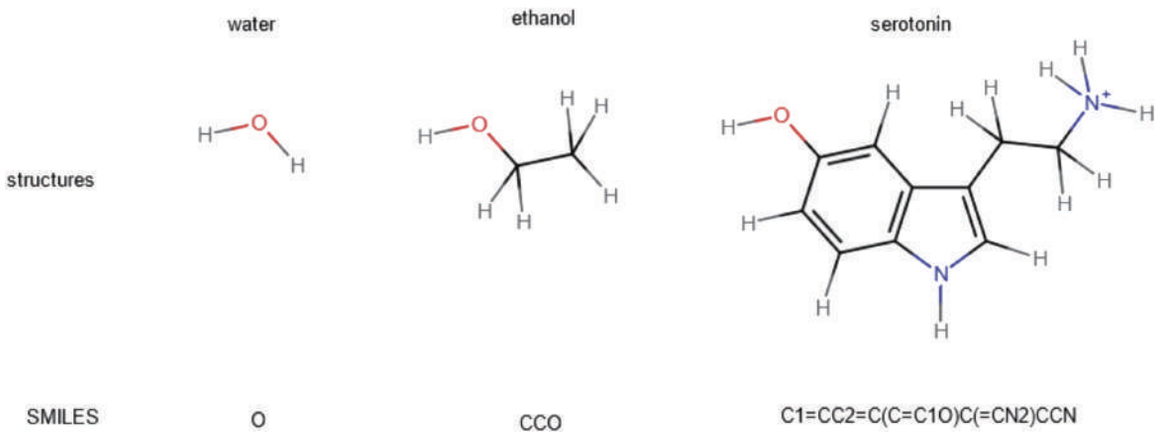
图像自动标注领域再次启发了文本标注任务的创新，其中RETfound便是医学领域从图像到文本转换问题的一种创新解决方案[38]。作为一个基础模型，RETfound能够将广泛可用的视网膜图像标注为不同的疾病类别。它旨在加速包括白内障、中心性浆液性视网膜病变、糖尿病视网膜病变、青光眼、心力衰竭、黄斑功能障碍、心肌梗死、帕金森病、中风和黄斑变性等疾病的诊断过程。

模型架构基于大型视觉transformer框架：使用编码器生成高分辨率的嵌入空间，可以用来区分视网膜图像特征，这与LLMs在自然语言文本中编码语义的方式相似。这种模型的应用展示了LLMs在医学图像处理中的潜力，为医生提供了一种快速而准确的诊断工具，有助于提高医疗效率和患者护理质量。

RETfound的解码器用于图像重建，而编码器则用于为下游疾病预测任务提取特征。RETfound通过自监督学习在160万张未标记的视网膜图像上进行了预训练。在这种范式中，AI模型无需任何额外的训练信息便可以学习数据集中查找模式。例如，如果一个神经网络在自监督学习任务中使用宠物图片作为训练集，模型很可能学会识别与猫、狗和其他流行宠物相对应的形状。模型知道如何区分不同类型宠物的图像，但它“不知道”我们称其中一组为“猫”，以及可能与图像中的宠物相关的其他信息。RETfound在预微调（pre-fine-tuned） 状态下也是如此：它可以区分视网膜扫描图像中看到的不同变异，这种能力使其能够针对特定疾病检测任务进行微调。

这种微调是通过来自不同大小数据集的，特定专家提供的标签进行的。例如，用于标注“正常”、“黄斑变性”和“糖尿病视网膜病变”等条件的“OCTID”数据集，以及包含与353,157名患者在2008至2018年间就诊记录相关的眼科数据的Moorfields Eye Hospital-AlzEye数据集，被用于微调以优化RETfound进行湿性年龄相关性黄斑变性的预后评估。通过这样的全面训练，RETfound可以用来根据医疗专业人员生成的图像记录中的像素模式创建视网膜图像的文本描述。因此，像RETfound这样的模型旨在减轻专家的标注工作量，为使用LLMs进行类似目的的概念框架提供灵感。

图像格式的数据一方面可以用来捕捉物理世界，另一方面，也可以捕捉大脑神经元的活动。它们可以作为实验变量，例如在视觉刺激实验中使用的图片，以探索大脑扫描与这些刺激之间的联系。与图像格式不同，化学结构及其描述可以捕捉大脑化学、神经生理学、神经药理学和化学感觉刺激的关键方面。简化分子输入线性输入系统（SMILES Simplified molecular-input line-entry）是一种将化学结构表示为基于文本的对象的方法。SMILES最初是基于分子图的原则构思的，以严格规范的方式表示化学结构，非常适合机器处理[40]。



▷ 图6:SMILES中配对分子图的示例

科学文献中包含了大量的化学名称，这些名称有时呈现标准化形式，但并非总是如此。通过适当的收集、整理和整合策略，可以构建一个结合化学名称和SMILES结构的语料库，用于训练大型语言模型（LLM）或微调基础模型，以探索化学结构与语义内容之间的潜在预测关系。如果能够实现这一点，进一步地，这个共同的嵌入空间可以连接到一个生成模型，该模型可以根据文本输入产生化学结构（例如，“我想看到能够进入人类中枢神经系统的新化学结构”）。在不远的将来，这样的多模态LLM可能成为科学家的宝贵伙伴，增强生成具有目标属性的全新分子的创造性过程，无论是物理、化学感知还是药理学属性。

另一个引人注目的应用是在SMILES（化学品的简化分子输入线性表示系统）和自然语言之间建立的共同嵌入空间，这不仅适用于单一化学物质的分析，也能扩展到化学混合物的研究。正如自然语言中词语和短语的解释会受到其上下文的显著影响一样，化学感知刺激中存在的气味分子（它们自然以混合物形式出现）的感知，也会受到其他混合物成分的组合和浓度的影响。此外，小分子如神经递质、激素、药物和毒素通常与其代谢物、杂质和其他生物分子共同作用。这些组合元素可以在其周围环境中产生生化和生理效应，例如结合到目标受体或调节信号转导通路的活性。

我们设想，一个经过SMILES和自然语言训练的LLM的共同潜在嵌入空间可以用来对化学品和混合物的复杂、依赖于上下文的多重作用进行导航，这对神经科学具有直接相关性。

（3）描述能力

除了高成本之外，标注任务面临的另一个问题是，依赖于预定本体或分类系统的标注将受到该描述系统的描述能力的限制。通常，执行标注任务的个体必须接受训练，以确保他们能够正确利用给定的本体对数据点进行分类，这是减少评分者差异性这一已知挑战的尝试。为了正确进行数据标注，所需的培训范围可能非常广泛，标注者必须是领域专家而非外行人。通过LLM编码器生成的嵌入，可以通过语义相似度测量或聚类等方法“翻译”为目标本体中的一系列术语。

如果不进行翻译，LLM给出的嵌入提供了基于本体的分类所无法提供的高语义粒度。这种特异性对于任何研究者在记录与特定实验直接相关的不同结果时都极为宝贵，因为它使得在与特定实验直接相关的分类方式上具有灵活性。

举一个简单的假设性例子，人们可以（1）从通过文本记录的注释标签或其他实验变量生成语义嵌入，（2）从目标本体中存在的术语生成嵌入，或（3）计算两组嵌入之间的余弦距离，以识别每个基于文本的实

验变量与来自目标本体的“最近邻”术语。虽然这种方法可能无法达到领域专家的准确性水平，但它在分辨率上的不足通过客观性和操作一致性得到了补偿，这提高了大规模注释的可扩展性和可重复性。另一方面，LLM产生的嵌入也为研究人员提供了一种通过聚类或更复杂的技术来分析注释数据集的手段，从而得以识别新的分类系统。

理想情况下，即使在没有与领域专家紧密合作的情况下，我们很快就能通过LLM进行专家级别的注释。更有趣的是，一旦证明了LLM能够以与专家相当或更优的性能应用现有的本体进行注释，我们就可以转向“专家LLM”来帮助识别和验证新术语和本体，这些术语和本体是通过数据驱动的方式得出的。同时，我们还可以借此机会检查基于LLM的注释结果，挑战那些由有限启发式设计的传统分类系统。

基于规则的解决方案虽然依赖明确的预定义标准，但在处理庞大且复杂的数据集时，黑箱式AI解决方案——尽管其决策过程不透明——通常能够展现出卓越的表现，实现传统方法难以匹及的预测准确性。将LLM辅助注释作为一种补充方法，将其与传统的自上而下的方法（例如，由领域专家手动分类）和基于规则（例如，预定义算法对数据点分类）的解决方案相结合，是我们可以同时利用专家经验带来的知识和LLMs从数据中获得的新见解的一种方式，这是一种真正能够“为自己说话”的数据形式。

LLMs被喻为变色龙*，具备“角色扮演”的能力[43]。它们可以采取已知人物或具有特定特征（个性和写作风格）的人物个性，例如夏洛蒂·勃朗特、卡尔·萨根或神经科学家。这种能力可以以多种方式利用。在某些注释任务中，与所有评估者都具有相同背景的评估小组相比，征求跨学科专家小组的意见可能更为有益。若干个LLM可以并行地在注释任务中扮演不同的角色，类似于人类评分者的分组。LLMs可以被要求采取不同专家、个性类型、职业、年龄和文化背景的立场来进行思考和评估。LLMs不仅解决了个体主观性对注释任务的影响，而且同时能够表达和操纵这种主观性。LLMs可以消除人类注释者所经历的短暂情感状态的波动，如果需要，它们可以在可控和可重复的方式中引入这些波动。

*https://karpathy.ai/lexicap/0215-large.html

描述神经科学研究以及主观体验的语言存在许多不一致之处。这些差异性助长了不同研究人员之间对注释解释的分歧。一致的语义嵌入空间的普遍性能够捕获和操纵模糊或主观的语言。关键的是，这些表征在实验室或其他研究和分析环境下是完全可重复的；只要对同一任务使用相同的LLM，并使用相同的模型参数集合。从科学研究的实际角度来看，这一特点应该对通过LLMs自动注释数据集的注释数据的共享性产生重大影响，有希望扩大LLMs自动注释数据集的下游应用的广度和深度。

不同的个体可能会以不同的方式标注相同的数据，甚至同一注释者在不同时间给出的回答也可能会有所变化。LLMs提供了一种更稳定和一致的标注。由于这些大型模型是基于广泛的数据集进行训练，不受个人主观体验的影响，它们能够在捕捉细致的上下文环境时替代人在手动标注任务中的主观性。训练后的LLMs可以被视为所有互联网用户平均思维的一种近似，即“众包思维”，因为它们的训练语料库的大部分来源于互联网。如果基础模型似乎没有捕捉到足够的细节以完成特定任务，它可以通过微调来近似基于特定网站或互联网用户子集的平均思维。

手动进行数据标注的过程通常包含主观性元素，特别是当被标注对象基于主观体验时在对《阿甘正传》中的场景进行注释的任务里，学生们需要标注他们所感知到的电影中演员表达的情绪。这项任务首先要求对电影中描绘的情感进行主观解释，再加上情感体验本身的高度主观性。studyforrest数据集还包括

每个场景发生的物理位置的注释。尽管如“夜晚”与“白天”、“室内”与“室外”的标注主要基于具有电影学术背景的两位领域专家做出的客观判断，但在这个过程中仍然留有主观解释的空间，例如将“白天”定义为任何由阳光照亮的场景，而不是其他决定因素一样。

LLMs能够在主观现象和客观测量的世界之间实现调和。通过LLM嵌入表征的语义实体，保留了文本中的离散主观或上下文意义，使其能够以一致的方式与其他文本进行比较。例如，想象一下从社交媒体帖子中收集的句子，用于自动注释情感标签，以便用于训练一个能够从用户帖子中预测情绪的NLP模型。无论每个设想的句子有多么独特，它们与“热情”、“沮丧”、“怀旧”或“平静”等术语对应嵌入之间的距离都可以用统一的方式计算。由于LLM训练语料库捕获了大量描述主观现象的文本，LLMs产生的更稳定和一致的注释，可以轻松地用于表征基于主观体验的数据元素，而无需将主观的人类判断作为注释过程的一部分。

使用LLMs自动化注释任务并不是渐进式的改进，而是一种革命性的方法升级，可以颠覆主流实践，有望终结受到主观性和其他形式的特质所带来的限制。以注释一系列日记条目中的情感为例，如果任务交给一组人类注释者，一个人可能会根据他们的个人经验和文化背景将一段文字标记为“悲伤”，而另一个人可能会看到它为“反思”或“怀旧”然而，由于LLMs是自回归的、状态依赖的，并且具有温度等超参数（参见前一节“大型语言模型解决方案的数据科学视角”），它们在处理相同提示时的输出虽不尽相同，但如果实验条件保持一致，其答案主要限制在语义空间的一个狭窄区域内。通过这种方式，LLM可能提供人类注释者无法匹配的客观性和一致性。

八、LLM在文本摘要和知识整合中的应用

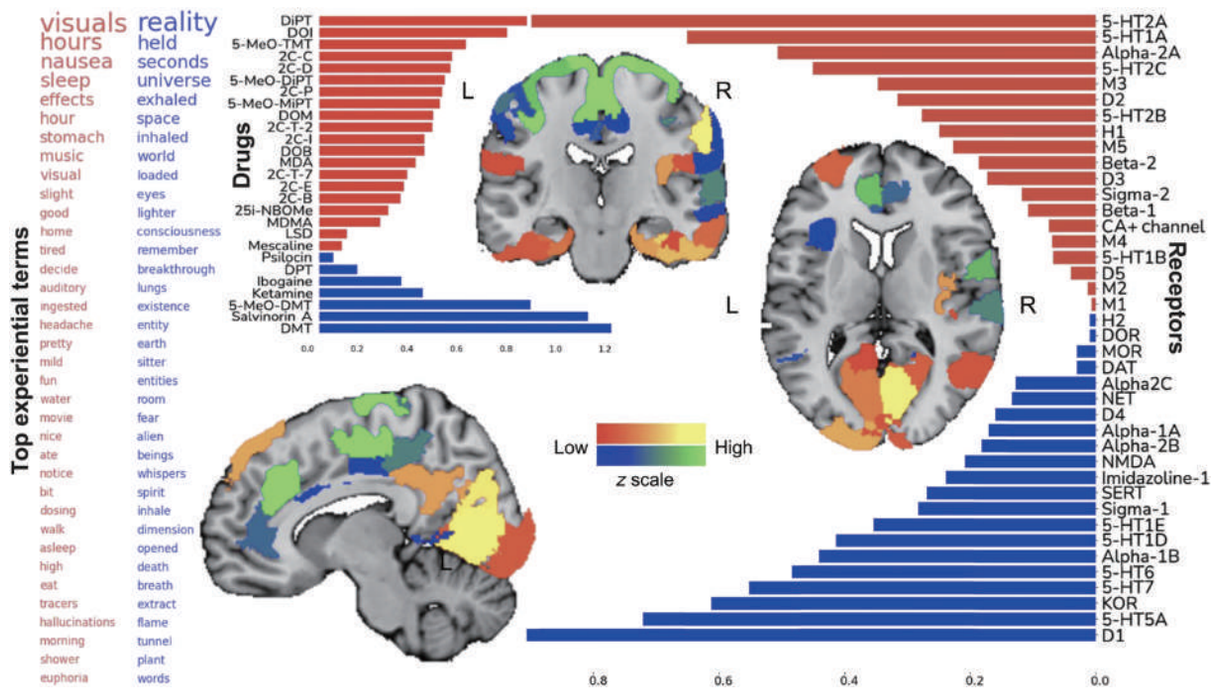
神经科学这个广泛的领域涉及从物理学到心理学等多个学科。这个极具跨学科性的领域产生了大量相对独立的实验发现，仅靠人力整合这些发现可能会显得力不从心。此外，该领域的宽广常常导致研究者在其子领域内孤立工作，专注于狭窄的研究领域，从而可能错过与其他子学科的交叉融合机会。也可能存在某些任务超出了人类认知能力，包括阅读包含大量数据点的实验结果，或提炼过去一年所有主要科学出版物的内容。在这些情况下，LLMs可以帮助研究人员获取大量文本中的信息，这些文本如果仅靠研究阅读来吸取其中信息，在短时间内可能会是很具挑战性的。

LLMs的能力超越了典型的文本摘要任务，其中收集的文本以人类可读（尽管冗长）的自然语言呈现。但LLM嵌入为主观文本提供了客观的量化，以解决语言歧义并给出标准化的输出。这些基于主观性的文本可以是简单的词语或短语，如用于捕捉《阿甘正传》中演员表现的情感[33]，或描述气味或风味化合物的化学感知[44]。它们还可能更为复杂，就像迷幻研究中使用的文本那样，描述迷幻药物及其对大脑和意识影响的研究。

“迷幻体验”这个常用表达方式暗示了不同体验之间的一致性。实际上，迷幻体验充满了细微差别和变化，部分根源于药物使用者的心态和环境，部分根源于药物之间的精神药理学差异。理解决定迷幻药物使用者观察到的细微差别的潜在因素，应该有助于我们了解是否可以利用某些药物或主观效果的类型来治疗特定状况，就像通过摄取赛洛西宾（psilocybin）和MDMA所提供的不同的体验在治疗强迫症和创伤后应激障碍方面所显示的早期成功，正是这种探索的实证。

为了研究这些细微差别，最近的一项研究使用了自然语言处理技术来分析来自迷幻药物使用者的

6,850份“体验报告”(图7)。研究的目标是建立主观体验、27种不同药物,以及人类大脑中表达的40种相关神经递质受体之间的联系。这项研究的结果包括通过典型关联分析(CCA)得出的详细词汇列表,该列表按语义维度相关性排名,捕捉了体验报告中的主要主题。



▷ 图7:使用NLP进行多模态受体-文本整合,揭示迷幻药物体验的机制基础分析。

人类解读由数千个词以特定顺序捕捉的复杂主题相当困难。CCA提供的排名列表中的每个词都带有可能被主观解释。由相邻术语提供的上下文以及列表不同子部分(即前1%与前5%)捕获的一般意义转变,进一步拓宽了潜在解释的范围。尽管结果以密集的高亮词汇列表呈现,但LLM可以无缝地从这些词集中抽象出语义核心主题,推导出由迷幻药物引发的主观效果的共享高层次类别。这些高层次类别随后可用于开拓新药发现平台,产生关于实验治疗方法的假设,寻找具有针对性主观效果的新迷幻药物,以治疗特定状况。未来LLM的应用展示了研究人员从复杂、非结构化数据中获取洞察的新机遇,尤其是在人类难以单独应对这些数据的情况下。

面向医学的LLMs,如Meta的PMC-LLaMA[46],提供了一个有希望的解决方案,用于筛选大量文本来源,综合其意义和信息价值。通过收集和总结不同来源的信息景观,这些模型提供了触及甚至理解了复杂主题的本质。具体来说,PMC-LLaMA旨在通过训练庞大的语料库(480万篇生物医学学术论文、3万本医学教科书,以及202M个医学问答对、决策理由和对话)来支持用户导航广阔的医学信息。PMC-LLaMA在零样本评估医学知识提示时产生了合理且连贯的回应,例如,回答患者关于尿路感染的问题,以及关于微生物学和药理学的深入问题。当被问及涉及结核病和激素避孕药物相互作用的多项选择题时,PMC-LLaMA正确指出了药物相互作用的机制,并详细阐述了得出答案的理由(通过抗生素利福平诱导CYP3A4,导致激素避孕药物浓度降低,最终增加了意外怀孕的可能性)。

PMC-LLaMA强调了数据驱动方法在专业领域的有效性以及领域特定模型调整的价值。这种对提示查询的令人印象深刻的回应,代表了机器辅助人类智能的场景,其中LLMs可以被定制为有效地在专业领

域教育用户，突显了这些模型改造社会的潜力和开发领域特定模型的重要性。

作为日常生活中另一个具体的例子，准备考试的医学生可以查询PMC-LLaMA等模型，获取特定主题的信息，以更高效的时间覆盖更广泛的材料。正如工业自动化释放了工人，让他们有更多时间去完成其他任务一样，我们可以预期LLMs的发展将呈现类似的机会。

然而，并非所有的改进都将仅仅是改善生活；许多应用，如可以访问患者电子健康记录的交互式LLM，可能具有挽救生命的潜能。不幸的是，Rodziewicz等人最近的一项统计调查估计，每年约有40万住院的美国患者经历某种可预防的伤害，其中大约四分之一的案例导致死亡。在医学领域，AI的潜在救命作用主要体现在几个方面：例如（1）减轻医疗专业人员的工作负担，使他们能够更有效地评估和治疗患者，以及（2）作为早期预警系统，提醒可能的不良事件。

九、多来源和多模态语言模型的整合

在过去的几十年里，神经科学已经扩展到越来越细分的研究活动领域。例如，阿尔茨海默病（AD）在几个基本上相互独立的研究社区中被研究。研究人类群体中AD病因的流行病学家并不经常与遗传学家、实验神经学家、脑成像研究者或动物实验研究者进行交流。研究与AD相关的全基因组风险变异的遗传学家，也不一定会参考或整合来自这些其他神经科学社区的现有知识。致力于研究AD大脑结构和功能差异的成像神经科学家，在设计和解释他们的研究时，也不一定会考虑流行病学人群分层的方面。每个AD研究社区似乎都在自己的“泡沫”中运作，形成了各自的杰出科学家群体，自己常讨论的假设池，以及自己独特的知识积累过程，且每年发表着大量的研究成果。

鉴于每年研究产出的不断增加，单个研究者越来越难以阅读所有这些论文。神经科学的许多研究活动领域以类似的方式被分割。这种知识碎片化可能是21世纪科学事业面临的最大挑战之一。现在，LLMs提供了一个机会，可以整合并翻译来自单一神经科学主题多个互补视角的日益增长的知识库。

LLMs也开始针对医学领域进行定制，如在医学考试和生成记录等任务中取得了有希望的结果。迄今为止，医学中的AI通常基于计算机视觉，对文本、语音和其他类型信息的整合有限。然而，通过LLMs对各种数据源的总结和整合，为推进AI辅助医疗专业人员的实践带来了巨大的希望。生物传感器、基因组档案、医疗记录、患者自述、代谢数据和其他实验室检测，都成为了构建针对个体患者定制的多模态AI诊疗路径的潜在数据源[48]。这种AI解决方案的潜力巨大，因为它可能对患者的生活和医疗专业人员的表现产生直接影响，，尽管这一目标还未完全实现[49]。

目前，应用LLMs减轻医疗专业人员文档工作量的可能性也正受到医学界的广泛关注。尽管使用LLMs在医学和医学研究中的伦理问题开始被讨论[50]，但现在越来越明显的是，LLMs可以作为辅助工具，有效减轻目前占用大量人力和时间的医疗流程，如电子健康记录的创建和处理，以及疾病的诊断和预后等多个方面。

作为下一个圣杯，哪些非文本数据模式可以赋予LLMs行动力?广义上，LLMs可能是第一个能够无缝结合结构化和非结构化信息的技术，无论信息的规模多大或是多么动态。此外，ChatGPT和类似的LLM变体已经成功地将来自多种语言、地理和文化的分散文本源聚合到一个模型实例中，这表明LLMs在多语言文本处理中的强大能力。

LLMs在弥合不同信息类型间差异，尤其是计算机视觉（即图像）与语言（即文本）之间的差异方面展

现出了巨大潜力。机器学习社区的一个近期例子是，Alayrac等人[35]展示了如何通过包含额外的模态信息来改进语言模型。Flamingo模型便是在包含文本和图像信息的互联网上的大规模多模态语料库上训练的，它的小样本学习能力使其能够适应包含图像和视频材料的各种任务。模型可以根据特定任务的例子进行提示，基于视觉条件的自回归文本生成，在许多场景中提供了实际益处。在神经科学领域，一个早期的例子是尝试使用模型仅从大脑活动测量重建自然图像的研究[51]。

此外，DALL-E/CLIP（由OpenAI在2021/22年提供）是生成性AI中文本-图像融合的早期例子，最初该模型基于GPT-3变体开发，旨在从用户提示生成更真实的图像。这个多模态融合引擎可以合成各种形式和风格，如逼真的自然图像、类似绘画的艺术和符号，以及设计方案的内部模型，调用真实和想象的对象、场景和人物，且无需众多训练示例（零样本学习）。其组件CLIP（对比语言-图像预训练）在互联网上约4亿对图像和文本标题上进行了训练，用于在DALL-E生成的图像中选择最佳输出。CLIP将计算机视觉和NLP结合在一个单一网络中，深度处理、分类和生成大量图像的文本注释。它不需要严格的任务特定训练，可以将其知识泛化到新的、未曾遇到的任务。

在神经科学背景下，未来的LLM框架可能会潜在地摄取多种形式的“图像”，如结构和功能MRI脑成像、PET、fNIRS，以及更广泛的EEG/MEG衍生脑图像。因此，一个重要的未来研究方向是探索DALL-E/CLIP和类似新兴技术，能在多大程度上成功地从自然图像扩展到包含大脑“图像”的多模态分析中。

例如，NeuroSynth数据库展示了一种自下而上的方法[52]，它自动提取了超过3,000篇脑成像任务实验文章的3D图像空间激活坐标，以及这些文章的全文。这一举措已经通过一个用户查询的网络界面为神经科学界提供了价值。与之平行的研究是BrainMap[53.54]数据库，其以自上而下的方式，围绕心理学类别构建了脑成像实验的人类本体论。对认知现象的描述系统是由人类领域专家手工设计的。

在这项研究中，同样也已经尝试了对图像描述对进行聚合，可视作训练或完善最先进的多模态LLMs的一个有吸引力的起点。一个想法是基于两个数据库中可用的研究、专家定义和全文注释相互补充，整合NeuroSynth和BrainMap，可能启用LLM支持的查询服务，也许还能跨越两种类型的大脑图像元信息进行推理。更广泛地说，旨在跨越内容类型界限的这些研究方向特别有前景，因为LLMs提供了一个前所未有的机会，将结构化和非结构化信息融合于一个统一的框架中。

在未来几年，神经科学家可以系统地研究哪些与大脑相关、适合LLM涌现的功能模式的信息？又哪些类型的神经科学信息可以被标记，以及如何标记？

最近的LLM研究显示了利用嵌入的氨基酸块、基因及其mRNA转录本、细胞和细胞类型、表型和疾病状态的潜力。LLMs可能还能处理标记化的大脑区域活动实例、白质纤维通路、大脑结构变化位置、EEG/MEG中的频率带变化或钙成像。

经由这些能力，神经科学家可以将数据集中的序列语义和生物学视角结合起来，形成对大脑的统一视角。这一目标的实现可能需要对模型架构进行创新，以表征这些信息层。或者，我们可以使用预训练的LLMs的输出作为一种编码特定信息模式的蒸馏形式，将其整合到随后训练的较小模型中，以实现最终的研究目标。具体来说，来自英国生物银行和其他大型数据集的数据集允许LLM将基因变异信息和其他分子数据与各种人类健康信息关联起来。

作为神经科学这一高度跨学科努力的核心愿望，LLMs可以帮助我们弥合不同神经科学社区之间的鸿沟，并使我们形成能够整合多来源知识的NLP模型。

十、大模型作为克服当前概念危机的认知纽带

LLMs可能提供一个替代工具包，该工具对于汇总和编辑神经科学研究者用来解析大脑功能的人类构建概念非常有价值。重要的是要认识到，特别是在经典的假设驱动研究中，整个研究努力都依赖于预先假设的认知和神经术语的有效性，这些术语用于阐述实验研究条件。然而，许多频繁使用的心理学或认知术语定义脆弱，无法在自然界中直接观察到。许多由人类专家确定的神经科学概念可能并不代表“自然分类”，因为它们并没有在自然界中划分出对立独立的神经回路。

大多数认知过程的概念在神经科学作为一个连贯学科出现之前（大约在20世纪中叶）就已经被创造出来，那时大脑功能才开始被理解。此外，某些行为或认知概念可能只在健康受试者精心设计的实验或临床条件（如具有局部脑损伤的患者[55]）中出现。根据这种观点，神经认知过程可以在受试者参与特定实验任务时被分解，作为揭示大脑与行为之间映射的途径。也许现在是时候用一种有规律的数据驱动方法，来测试这些概念的有效性了。

神经科学家在描述大脑现象时遭遇的复杂性，与路德维希·维特根斯坦在其著作《哲学研究》中所提出的观点紧密相关。维特根斯坦晚年认为，人类语言本身所带来的混淆，是许多哲学问题的根本来源。例如，在心理学中，甚至像“认知”和“情感”这样的简单词汇都缺乏一个普遍认同的定义[56-57]。此外，常在心智理论中提及的大脑网络，即从他人视角进行思考的能力，也始终参与了一系列多样化的心理过程，包括道德思考、自传体记忆检索和空间导航等[58-60]。我们目前遗留的神经认知框架，可能没有指向正确的方向[61]。

例如，我们为什么隐含地期望威廉·詹姆斯的杰作（《心理学原理》，1890年）中的术语和概念，能够代表大脑中特定的机制？更进一步的是，当我们遇到难以调和的发现时，我们有时会倾向于创造一个新术语，而不是真正深入问题的核心。

许多神经科学研究采取自外而内的方法：他们首先创造概念，然后试图在大脑活动中找到这些概念的对应或描述[61]。这与一些作者所说的“新颅相学”密切相关，后者是一种简化主义方法或“过度定位”，试图将术语映射到大脑的局部地理区域[62]。虽然现代神经成像显示，在某些任务中特定的大脑区域确实更活跃，但鉴于大脑的高度互联性以及多个认知功能的网络分布特点，试图为复杂的功能找到单一的“定位点”可能极具误导。

研究重点应该放在大脑的实际反应上，而不是人类发明的术语本身。的确，正是是大脑中的神经认知过程产生了行为和认知。简而言之，心理术语如何以及在多大程度上映射到区域大脑反应，仍然是难以捉摸的，反之亦然[62-64]。出于这些原因，一些作者提出神经科学在数据上越来越丰富[65]，但在理论上仍然贫乏，指出了迫切需要新的研究假设生成手段。

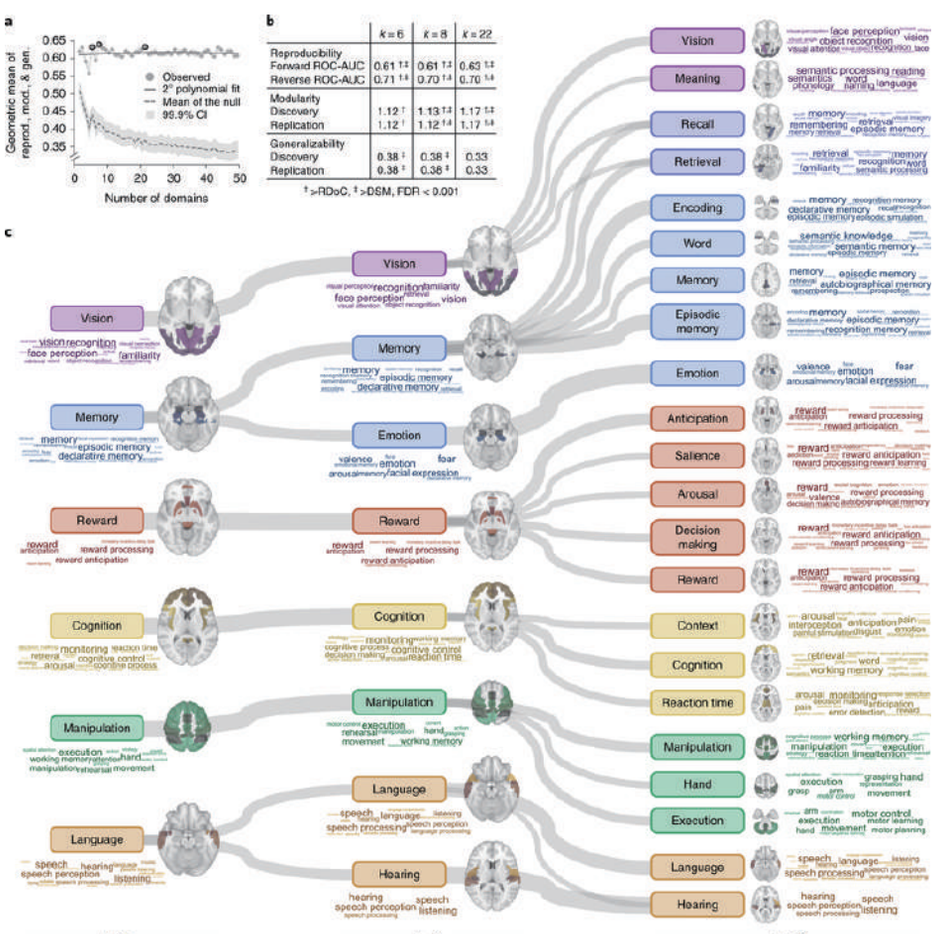
关于大脑疾病的定义，尤其是精神病学中的术语，也可以提出类似的观点。相同的概念并不唯一地与相同的机制相关，相同的机制也不经常对应这一个明确的诊断实体。这一认识可能是为什么相同的药物类别经常有助于缓解名义上不同的精神状况症状的原因之一。

DSM-5和ICD-10手册是根据精选专家的意见对精神病进行分类的。此外，资助机构只有在研究人员

的提案理由和预期结果坚定地基于这些人类制造的诊断类别时，才会承诺资助。然而，越来越清楚的是，即使在遗传水平上[66]，基础生物学中的病理生理过程也具有相当高异质性，且相互存在重叠。因此，当今对精神健康状况的描述系统虽然有助于实践医生之间的交流，但在研究的生物学有效性和临床护理的预测性方面，仍然显得力不从心。

尽管神经科学中现有描述系统存在明显的不足，但很少有尝试以自下而上的方式构建这样一个语义概念系统。在一项开创性研究中，研究人员设计了一个基于数据的方法，来构建神经认知类别的框架[67]，该框架汇集了大约20,000篇人类脑成像论文的信息。利用超过25年脑成像研究积累的数据宝库，NLP算法挖掘了研究文章的语义内容，并将其与来自功能脑扫描（fMRI，PET）的60多万个拓扑位置相结合。这种方法同时平等关注了语义原则和神经活动原则，允许研究者以整体方法系统地整合大脑和行为。

此外，这种方法还有助于克服神经科学界长期困扰的一个问题——如何从概念出发推理大脑活动（前向推理）以及如何从大脑活动推断概念（后向推理）[62]。在实证验证分析中，这种“计算本体论”被证明比神经科学和精神病学中广泛接受的描述系统，在重现术语与功能链接方面，对新的、未见过的研究成果具有更好的适应性。



▷ 图8:NLP工具以完全底层的方式整合现有关于人类认知的概念

综合来看，我们用来描述世界的叙事和故事塑造了我们设计神经科学实验和解释发现的方式。在神经科学中，真正的进步需要对词语使用、语言卫生（language hygiene）和概念化变体有特别的敏感性。在

未来，由LLM赋能的神经科学中，我们可能能够将心理学固定术语，基于科学证据重新放到新架构中，而不是延续前一个历史时期的遗留术语。

新兴的LLM技术可以激发基于生物学的大脑疾病分类学的，具有重大意义的重新定义，从而跨越诊断边界，进入一个基于证据的精神医学新时代，而不是仅仅依赖于特定专家的判断。正如维特根斯坦所说，“我语言的极限就是我世界的极限。”[68]

十一、结论

在过去的5到10年里，生物学已经转变为一门“可计算”的学科。例如，大规模基因数据库与定向CRISPR基因编辑和机器学习分析相结合，使生物学更接近于一个工程学科。我们生成生物分子数据的能力远远超过了我们从这些系统中真正获得理解的雄心——正如John Naisbitt所写[69]，今天的神经科学家实际上是“被信息淹没，却又渴望着知识”。

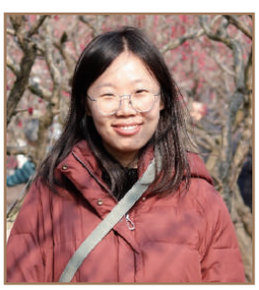
LLMs为研究者提供了新的机遇。这类模型表明，纯粹的统计暴力可以帮助研究者通过阅读和生成生物学来揭开大脑和疾病的神秘面纱，通过构建知识框架，解锁前所未有的大规模信息整合和研读模式。基础模型可能会从神经科学领域中提取、协同和合成知识，跨越孤立的“学科间隔”，这项任务可能会（也可能不会）超越人类的理解范畴。神经科学家需要接受并拥抱这样一个令人不安的可能性：人类大脑作为一个生物系统，其复杂性可能超出了人类智能独立理解的范畴，唯有借助AI工具处理大数据，我们才可能理解它。

从更广泛的社会角度来看，工业革命主要影响了蓝领工作。相比之下，当前的LLM革命可能会主要影响白领工作，包括神经科学研究人员的工作。事实上，LLMs的惊人效能已被一些风险投资家和投资者与火的发现、电力的应用或互联网的诞生相提并论，这些发明都极大地推动了人类社会的进步。LLM是否真就能改变世界，让我们拭目以待。（编辑：存源）

参考文献

关联论文：Bzdok, Danilo, et al. "Data science opportunities of large language models for neuroscience and biomedicine." Neuron (2024). https://doi.org/10.1016/j.neuron.2024.01.016

脑科学能用Transformer具体做什么?



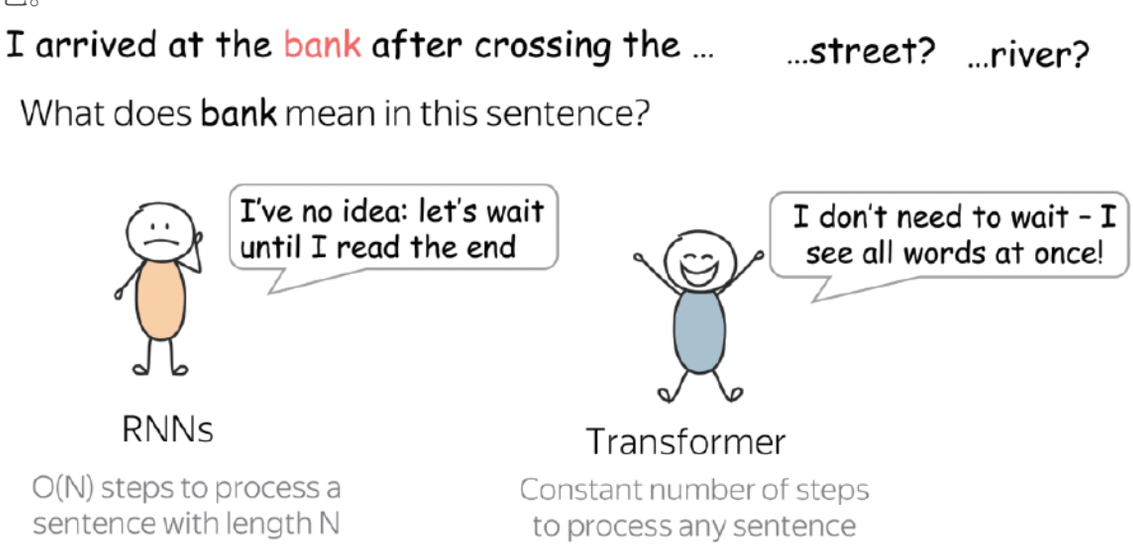
作者:轻盈

复旦大学博士生在读,计算&进化神经生物学方向。视科研和科普为人生的两大志业。想做有趣有意义的科学研究,也想把收获到的知识和乐趣分享给世人。

扫码查看原文

自ChatGPT问世以来,“Transformer模型”始终以超高频率出现在各个AI新产品模块当中。比如,大家所熟知的GPT-4、Midjourney、GitHub Copilot等,它们的优越性能都得益于Transformer的诞生。

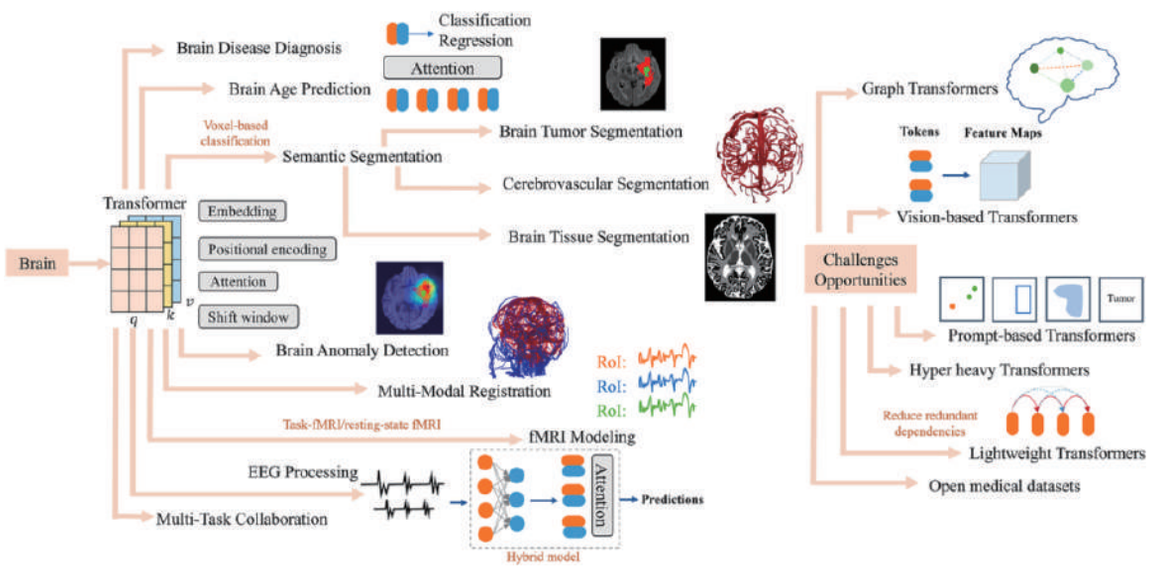
Transformer最早于2017年由谷歌(Google)研究团队提出,主要用于处理自然语言。与传统的深度学习方法相比,Transformer采用了一种被称为自注意力机制(Self-Attention)的方法,在捕捉长距离依赖关系层面具有独特优势。近些年来,Transformer已在文本内容解析、目标检测、视觉分割等领域表现出色。



▷图1:Transformer与循环神经网络(RNNs)比较。图片来源:https://lena-voita.github.io/resources/lectures/seq2seq/transformer/rnn_vs_transformer_river-min.png

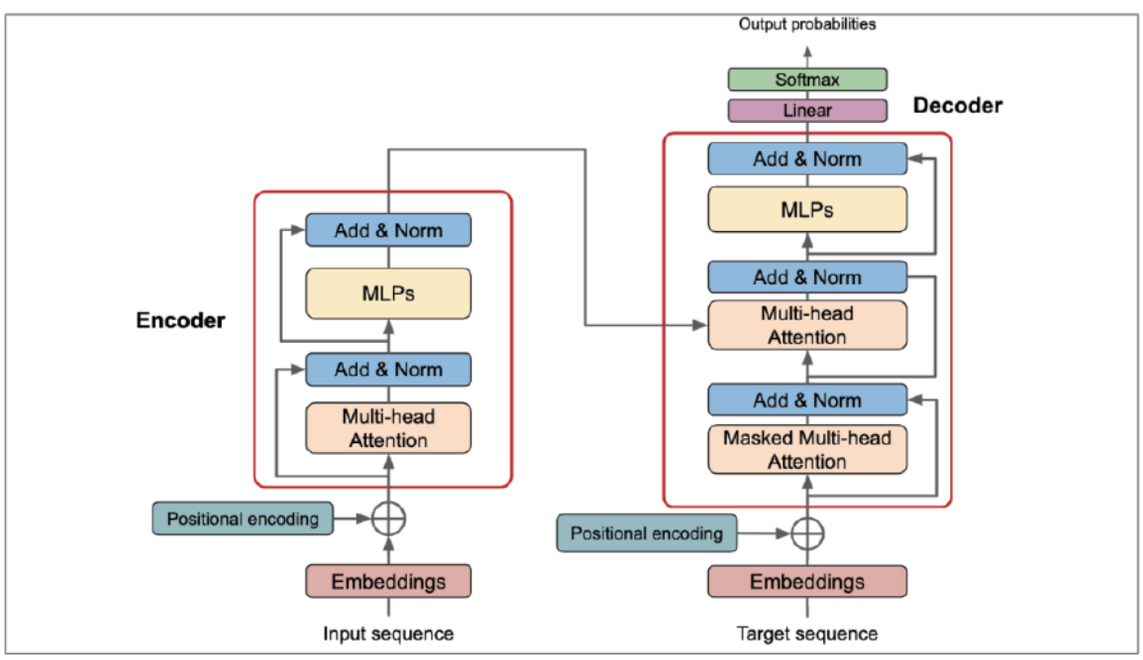
自ChatGPT问世以来，“Transformer模型”始终以超高频率出现在各个AI新产品模块当中。比如，大家所熟知的GPT-4、Midjourney、GitHub Copilot等，它们的优越性能都得益于Transformer的诞生。

Transformer最早于2017年由谷歌（Google）研究团队提出，主要用于处理自然语言。与传统的深度学习方法相比，Transformer采用了一种被称为自注意力机制（Self-Attention）的方法，在捕捉长距离依赖关系层面具有独特优势。近些年来，Transformer已在文本内容解析、目标检测、视觉分割等领域表现出色。



▷图2：Transformer模型在脑科学研究中的应用图谱。图源：参考文献[1]

一、从经典架构了解Transformer



▷图3：Transformer核心架构。图片来源：https://deeprevision.github.io/posts/001-transformer/

(1)输入嵌入

俗话说，入乡随俗。计算机模型不能直接理解人类语言，那么此时就需要输入嵌入(Input embedding)这个环节来做个衔接，也就是将输入数据转换为模型更好理解和处理的向量表示。按照输入数据的形式，目前主要可以分为词嵌入和图像嵌入，这篇论文中所指的即为词嵌入。而对于图像这种高维数据，在输入Transformer前，需要首先对其进行分割和压平，即图像嵌入。就像牛排太大生吞无味，切成小块细嚼慢咽才是硬道理。这里图像嵌入比较常用的处理方法，就是块嵌入(Patch embedding)。

(2)位置编码

“我爱过他”和“他爱过我”这两句话，虽包含的词语完全一致，但语序有别。假如放到自然语言的语境中，可能是两段完全不同的苦情往事。由此可见，词语的顺序在句义当中尤为重要。然而，Transformer的自注意力机制本身并不能感知词语的顺序信息。这时，Transformer就需要引入一种称为位置编码(Positional Encoding)的环节。位置编码就是在输入序列中的每个词语后面追加一个位置标记来表征它在句子中的位置信息。

(3)自注意力机制

千呼万唤始出来，自注意力机制在前文已被多次提及。那么，Transformer最引以为傲的自注意力机制是什么?

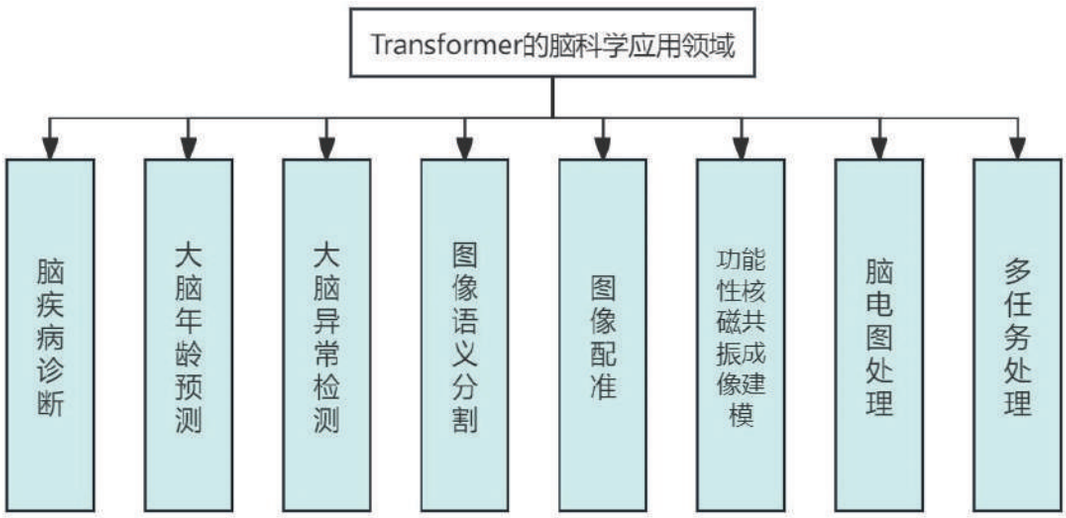
自注意力机制(Self-Attention Mechanism)是注意力机制的一个特例。注意力机制类似于人类的注意力，能够根据任务的需要分配不同权重给输入序列中的不同部分。自注意力机制则更像是一种“全知”的能力，系统可以同时关注输入序列中的所有位置，而不受序列中位置的限制。

自注意力机制的公式如下图所示。相比于传统模型，自注意力机制具有理论上的无限窗口和计算空间，使其能够更有效地捕捉输入序列的长距离依赖关系。注意力模块通过创建查询(Q)、键(K)和值(V)向量，并进行点积运算生成得分矩阵，再经过缩放和softmax激活处理，最终使用注意力权重对查询向量进行加权，生成增强的输出向量。这一过程使得模型能够从全局的角度理解并处理输入序列中单词之间的关联。

(4)多头注意力和掩蔽多头注意力

多头注意力机制(Multi-Head Attention)是注意力机制的一种扩展形式。多头注意力允许模型使用多组查询(Q)、键(K)、值(V)，每个注意力头都有自己的一组参数，独立学习如何关注输入序列。这使得模型能够同时关注序列中的多个方面，从而更好地捕捉不同位置和语义之间的关系。最后，各个头的输出会被合并，形成最终的多头注意力输出。

由于Transformer可以一下子掌握所有的信息，在某些情况下，为了避免模型看到未来的信息，通常有必要将序列中未来的位置的信息设为不可见。掩蔽多头注意力机制(Masked Multi-head Attention)就是在训练任务中，我们只能使用当前位置之前的信息，而不能使用当前位置及之后的信息，以避免信息泄漏。



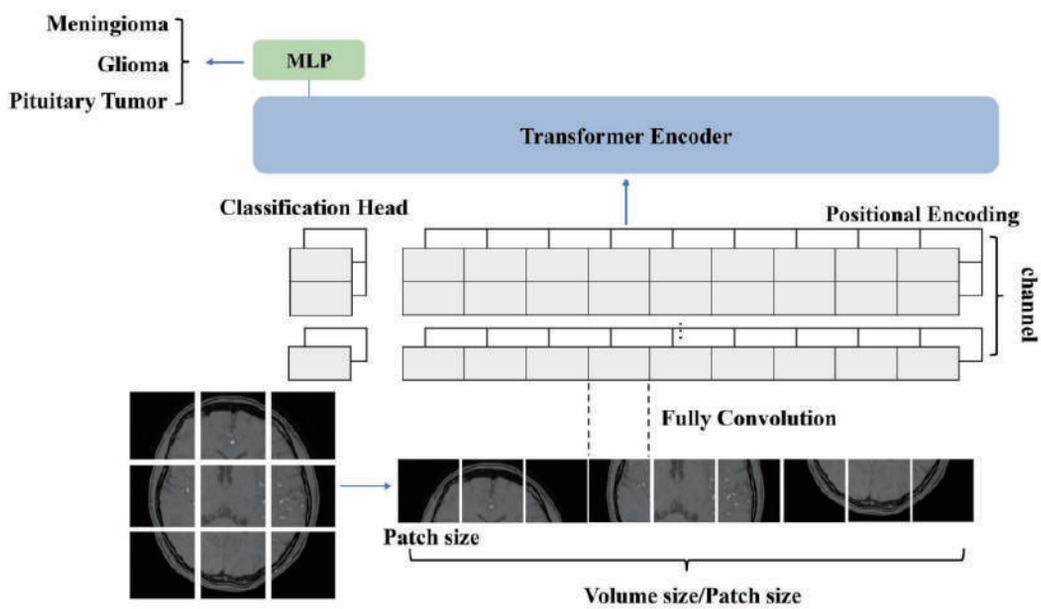
▷图5：Transformer在脑科学中的应用领域。图源：由追问编辑部制作

二、Transformer+脑科学”的八大应用

(1)脑疾病诊断

磁共振成像(MRI)*是一种在临床上常用的成像技术，通过对磁共振图像信息进行分析，医生能够发现和诊断脑疾病。Transformer，特别是在计算机视觉中首次引入的Vision Transformer(ViT)图像分类模型，目前已被成功应用于建立复杂的映射关系，如在磁共振图像与脑疾病之间建立关联。自2020年ViT的提出以来，越来越多的研究基于这一框架，致力于肿瘤和阿尔茨海默病等脑疾病的诊断。表格1归纳了以往研究中具有代表性的用于脑疾病诊断的Transformer模型信息。

*磁共振成像(MRI)：是一种非侵入性的医学成像技术，通过利用磁场和无害的无线电波来生成详细的内部器官图像，特别适用于脑部结构和异常的检测。与一些其他成像技术(如CT扫描)相比，MRI避免了辐射暴露的风险，同时提供了更为详细的解剖学信息。此外，由于不涉及使用放射性物质，不引起过敏反应的风险，MRI成为了许多神经学和神经科学领域中首选的成像方法。通常具有T1，T1-CE，T2以及磁共振成像液体衰减反转(FLAIR)四种模态。



▷图7：ViT计算框架。图源：参考文献[1]

Author	Method	Year	Application	Modality	*N*	Loss	Optimizer	Acc. (%)	F1 (%)
Odusami et al.[71]	ViT[45]	2023	Early detection of AD	T1	100	-	AdamW[72]	93.75	-
				FDG	100				
Shin et al.[73]	ViT[45]	2023	Classification of AD	FBB	150	-	-	56.67	54.55
								80.00	66.67
Tummala et al.[74]	ViT[45]	2022	Tumor classification	T1	3064	CE	RMSprop[75]	98.21	-
							Adam		
							Adadelta[76]		
Li et al.[77]	ViT-WSI[77]	2023	Tumor classification	WSI	-	CE	Adam	-	-
Qiu et al.[78]	MSGTN[78]	2021	Identification of AD	fMRI/DTI/non-image (gender, age)	170	-	-	92.12	91.13
Alcaraini et al.[79]	BECNN[79]	2023	Tumor classification	T1/T1-CE/T2/FLAIR	1425	-	-	96.75	96.80
					3064			99.10	98.70
Zhou et al.[80]	ASI-DBNet[30]	2023	Tumor progression	Microscopic image	2500	CE	AdamW	95.24	-
Anaya-Isaza et al.[81]	Cross-transformer[81]	2023	Tumor detection	T1	3929	CE	Aadelta	88.06	82.89
				FLAIR				89.58	84.76
				T1-CE				88.31	82.84
Ferdous et al.[82]	LCDEiT[82]	2023	Tumor classification	T1-CE	233	CE	AdamW	98.11	93.69
				T1/T1-CE/T2/FLAIR	2040			97.86	93.68
Sarasua et al.[83]	TransforMesh[83]	2021	Detection of neuroanatomical changes in AD	T1/T2*/FLAIR	-	MAE with regularization constant	-	-	-
Samak et al.[84]	TransSOP[84]	2023	mRS score	NCCT	500	-	Adam	80.00	59.00
				Clinical metadata					
Zhou et al.[85]	vidMACSSwin-T[85]	2023	Detection of cerebral aneurysms	Surgery video	356,165	Weighted CE	AdamW	87.10	58.90

▷表格1:用于脑疾病诊断的Transformers的技术细节。图源:参考文献[1]

(2)大脑年龄预测

大脑年龄是指对一个人脑部结构和功能相较于其实际年龄状态的评估。通过使用一些神经学和认知学方法,如脑部成像、认知测试和神经生物学标志物,可以估算大脑年龄。估算大脑年龄对于研究认知功能的衰退、神经退行性疾病和其他与年龄相关的神经学问题至关重要。近年来,Transformer已成功应用于大脑结构和年龄之间的建模,为更好理解的大脑健康和老化过程以及开发相关疾病的预防和治疗方法提供了新途径。表格2中总结了大脑年龄预测方法主要的技术细节。

Author	Method	Year	Modality	*N*	Loss	Optimizer	MAE	PCC (%)
He et al.[97]	Global-local transformer[97]	2021	T1	8379	MAE	Adam	2.70	98.53
Cai et al.[98]	Graph transformer[98]	2022	T1/DTI	16,458	MSE	Adam	2.71	86.82
Hu et al.[99]	SQET[99]	2022	T1	6318	MAE	AdamW	2.55	98.30

▷表格 2:用于推断大脑年龄的Transformers的技术细节。图源:参考文献[1]

(3)大脑异常检测

大脑异常检测是一类旨在快速且准确地定位脑部病变区域的任务。目前用于大脑异常检测的Transformer模型主要分为基于边界框的模型和重构模型。

基于边界框的模型(The bounding box-based model):这类模型使用边界框描述异常对象的空间位置。代表性的方法是VD-Former。VD-Former通过模拟对比度和空间一致性,准确定位了大脑中损伤的区域。

重构模型(The reconstruction model):通过预训练模型生成脑特征,将这些特征转移到学习正常样本的表示,从而实现对未知病理区域的异常检测*。一个代表性的模型是基于U-Transformer的异常检测模型(UTRAD)。UTRAD选择在特征分布中学习重建特征,相较于原始图像,这个过程模型获取了更多的特征,从而得以实现对异常区域的识别。

*异常检测与疾病诊断在中文语义上容易混淆。疾病诊断更倾向于是一个分类的任务,区分健康组和疾病组。而异常检测则更倾向于在图像上确定病变的存在,并定位其大致的边界。

(4)图像语义分割

图像语义分割是对目标区域进行像素级分类的过程。该任务要求模型对图像中的每个像素进行标签预测。与目标检测不同,像素级分类关注的是图像的细粒度信息,即对每个像素进行标注,以便了解图像中的每个区域属于哪个类别。传统的卷积神经网络(CNNs)能够建模局部特征,但在建模全局特征方面,Transformer则更为擅长。对于像语义分割这样的复杂任务,全局特征建模尤为重要。

脑部成像的语义分割涵盖了脑肿瘤分割、脑血管分割和脑组织分割等方面。在医学图像分析中,准确描述和分类这些结构,对于精确的疾病诊断和治疗至关重要。总的来说,通过引入Transformer,特别是在处理全局信息和复杂分割任务方面,脑部成像的语义分割有望取得更为准确和精细的结果,从而提高医学影像分析的水平。表格3列举了目前用于肿瘤区域分割的Transformer的技术细节,供读者参考。

Author	Method	Year	Data	Modality	*N*	Loss	Optimizer	Dice (%)	95HD (mm)
Wang and Li[28]	TransBTS	2021	BraTS 2019	T1/T1-CE/T2/FLAIR	460	Dice	Adam	83.62	5.14
			BraTS 2020		660	L2		83.52	10.89
Jiang et al.[133]	SwinBTS	2022	BraTS 2019	T1/T1-CE/T2/FLAIR	335	Dice	Adam	81.15	-
			BraTS 2020		494	CE		82.24	17.06
			BraTS 2021		1470			86.60	11.39
Xing et al.[137]	NestedFormer	2022	BraTS 2020	T1/T1-CE/T2/FLAIR	369	Dice	AdamW	86.10	5.05
			MeniSeg	T1GD/CE-FLAIR	110	CE		76.50	4.41
Hatamizadeh et al.[125]	Swin UNETR	2021	BraTS 2021	T1/T1-CE/T2/FLAIR	1470	Dice	-	88.97	5.21
Jia et al.[131]	BiTr-Unet	2021	BraTS 2021	T1/T1-CE/T2/FLAIR	1470	-	Adam	95.32	3.90
Liang et al.[128]	TransConver	2022	BraTS 2018	T1/T1-CE/T2/FLAIR	351	Dice	Adam	86.32	4.23
			BraTS 2019		460	CE		83.72	4.74
Zhang et al.[118]	mmFormer	2022	BraTS 2018	T1/T1-CE/T2/FLAIR	285	Dice	Adam	71.92	-
						CE			
Li et al.[126]	TransBTSV2	2022	BraTS 2019	T1/T1-CE/T2/FLAIR	460	Dice	Adam	85.18	4.87
			BraTS 2020		660	L2		84.90	7.45
Hatamizadeh et al.[10]	UNETR	2022	MSD BraTS	T1/T1-CE/T2/FLAIR	484	Dice	AdamW	71.10	8.82
						CE			
Peiris et al.[134]	VT-UNet	2022	MSD BraTS	T1/T1-CE/T2/FLAIR	484	-	AdamW	87.10	3.43
Lin et al.[139]	CKD-TransBTS	2023	BraTS 2021	T1/T1-CE/T2/FLAIR	1251	Dice	-	90.66	6.22
Gai et al.[140]	RMTF-Net	2022	LGG	T1/T1-CE/T2/FLAIR	1311	Dice	Adam	93.50	-
			BraTS 2019		10,047	CE		82.10	
			BraTS 2020		10,945	SSIM[143]		81.80	
Zhu et al.[141]	Z. Zhu	2023	BraTS 2018	T1/T1-CE/T2/FLAIR	285	Dice	Adam	86.93	4.19
			BraTS 2019		335	CE		88.22	4.02
			BraTS 2020		369			87.95	4.59
Liang et al.[132]	BTSwin-Unet	2022	BraTS 2018	T1/T1-CE/T2/FLAIR	351	MAE	Adam	86.40	4.37
			BraTS 2019		460	Dice		83.46	5.11
						CE			
Li et al.[135]	DenseTrans	2023	BraTS 2020	T1/T1-CE/T2/FLAIR	2000	-	SGD	86.00	12.81
			BraTS 2021		494		Adam	89.00	10.53
							MSG		
Nian et al.[142]	3D brainformer	2023	BraTS 2017	T1/T1-CE/T2/FLAIR	331	Dice	Adam	81.30	5.77
			BraTS 2018		351	L2		83.27	5.94
Liang et al.[126]	3D PSwinBTS	2023	MSD BraTS	T1/T1-CE/T2/FLAIR	484	Dice	Adam	57.80	-
			BraTS 2020		498	CE		84.94	10.75
			BraTS 2021		1470			87.32	10.78
Lyu et al.[130]	Q. Lyu	2022	Q. Lyu	T1/FSPGR	138	Dice	Adam	87.80	-
					145	CE		89.50	

▷表格3:用于肿瘤区域分割的Transformers的技术细节。图源:参考文献[1]

(5)图像配准

图像配准是一种将两幅或多幅图像对齐的过程,以保持它们在空间或特定方面的一致性。在医学影像领域,图像配准是一项关键技术,用于整合或比较不同时间、传感器或模态下获得的图像。传统的图像配准依赖于特征检测和匹配,而基于深度学习的配准,则通过模型学习构建全局函数来获得对齐表示。研究表明,Transformer在图像配准中表现出色,特别是在长距离空间对应关系的建模方面。目前对于图像配准的研究主要分为位移场配准和微分同胚配准方法。

位移场配准(displacement field registration):它的目标是找到两幅或多幅图像之间的空间变换关系,以便将它们对齐。例如,研究人员使用Swin Transformer构建了仿射变换网络(TransMorph),实现了高效的图像变换。该模型利用混合的Transformer-ConvNet架构建立了远距离空间对应关系,生成将移动图像与固定图像对齐的变换参数。

微分同胚图像配准(diffeomorphic registration):这是一种保持图像局部形状和结构连续性的方法,通过优化微分同胚变换以对准图像,通常在流形空间和变分框架下实现。例如,基于Swin Transformer的Swin-VoxelMorph模型,通过最小化图像差异并估计变换,实现了对称无监督学习。

(6)功能性核磁共振成像(fMRI)建模

基于Transformer的方法在解决fMRI中远距离依赖性关系方面也取得了显著突破。如结合血液氧合水平变化的时间序列和功能连接网络的Transformer,成功地学习了fMRI中的时空上下文信息。如ST-Transformer通过线性空间-时间多头注意单元,在数据平衡后计算fMRI中的空间和时间表示,用于孤独症谱系障碍(ASD)的诊断。综上,Transformer为深度解析脑功能区域和时间序列特征关系提供了新的解决方案。

*功能性核磁共振成像(fMRI)和磁共振成像(MRI)的区别:

fMRI和MRI是两种不同但密切相关的成像技术,它们在医学和神经科学中广泛用于研究和临床应用。

1. MRI:MRI是一种用于获取高分辨率体内组织结构图像的成像技术。它基于原子核在强磁场中的共振现象,通过测量不同组织对磁场的响应来生成图像。MRI可以显示组织的结构、器官的位置和大小,对于检测肿瘤、脑部解剖和其他结构方面非常有用。

2. fMRI:与MRI不同,fMRI关注的是测量脑部血流和代谢的变化,以推断不同脑区域的活动水平。fMRI通过检测脑血液氧合水平的变化(BOLD信号)来间接测量神经活动。它主要用于研究大脑在执行任务、处理刺激或进行特定认知功能时的活动。

3. fMRI使用BOLD信号作为衡量脑部活动的指标,而MRI则主要关注组织的结构。BOLD信号是基于血氧水平在神经活动期间的变化。MRI提供了关于脑结构的详细信息,而fMRI提供了有关脑功能的信息。通过结合这两种技术,研究人员可以更全面地理解大脑的结构和功能,并研究它们之间的关联。

(7)脑电图处理

近年来,脑电图(EEG)信号处理也逐渐摒弃传统的基于循环神经网络(RNNs)的方法,转而关注Transformer模型。研究者们引入Transformer模型,如S3T和EEGformer,通过对时空相关性的建模和自注意力机制的应用,为处理EEG信号提供了更灵活、更强大的工具。这些新方法不仅克服了传统方法在并行计算等方面的不足,还为更深入地理解和利用EEG信号提供了前景广阔的研究方向。表格4列举了目前用于EEG数据处理的Transformer的技术细节,供读者参考。

*脑电图(EEG):是一种无创性生理学技术,通过在头皮上放置电极记录和测量大脑电活动。这种方法具有无创性、实时性和高时间分辨率的特点,适用于临床医学、神经科学研究和脑机接口等领域。通过频率分析,EEG信号被分为不同频带,如δ波、θ波、α波、β波和γ波,每个频带与不同的脑状态和活动相关。在临床上,EEG被用于诊断癫痫、睡眠障碍和其他神经系统疾病。总体而言,EEG为理解脑部功能和神经活动提供了重要的信息。

Author	Application	Year	Loss	Optimizer	Accuracy (%)	95HD (mm)
Wan et al.[193]	Target frequency identification	2023	CE	Adam	70.15	-
	Emotion recognition		L1		91.58	
	Depression discrimination				72.19	
Xie et al.[194]	Motor imagery	2022	-	Adam	82.95	-
					87.26	
Song et al.[195]	Motor imagery	2021	CE	Adam	91.30	82.37
					84.26	84.26
Lee et al.[196]	Overt speech	2022	Hinge[197]	-	49.50	-
	Imagined speech				35.07	
Liu et al.[198]	Motion recognition	2022	CE	Adam	93.10	92.61
					96.28	
					83.27	
Du et al.[199]	Person identification	2022	CE	AdamW	97.29	-
					97.45	
Wang et al.[200]	Emotion recognition	2022	CE	Adam	65.75	64.29
					66.51	66.27
Kostas et al.[201]	Motor imagery	2021	Cosine similarity CE[201]	Adam	86.70	-
	Error related negativity				42.60	
	Douchin seller					
	Sleep staging					
Tao et al.[202]	Human brain-visual	2021	CE	Adam	61.11	-
	Motor imagery				55.40	
Siddhad et al.[203]	Age classification	2022	CE	Adam	94.53	93.55
	Gender classification				87.79	87.99
Ahn et al.[204]	Motor imagery	2022	CE	AdamW	62.00	-
	Visual imagery				70.00	
	Speech imagery				72.20	
Ma et al.[205]	Motor imagery	2022	CE	Adam	83.90	78.20

▷表格4:处理EEG数据的Transformers模型细节。图源:参考文献[1]

(8)多任务处理

随着深度学习模型参数规模的扩大,研究者们借助堆叠的多头注意力机制开发了适用于协同多任务的Transformer。比如,多视角嵌入的医学Transformer,通过在轴向、矢状和冠状方向对MRI序列进行采样,使用预训练的卷积编码器进行向量提取,后Transformer被应用于在不同方向上实现自注意力增强。这种方法可应用于预测脑疾病、估计大脑年龄和脑肿瘤分割等任务;Trans-ResNet,则整合了CNN和Transformer,通过可靠的梯度传递,实现对注意力模块的高效特征学习,也可应用于预测脑疾病和估计大脑年龄。

三、展望未来

尽管Transformer在各个的领域中表现卓越,但仍面临计算复杂度大,参数数量多等诸多局限。因此,在论文中作者总结了未来Transformer模型发展的可能方向。

基于图结构的Transformer:作为非结构化数据的代表,图(graph)由点和边组成,有效建构对象内部的依赖关系。比如, GraformerDIR和TRSF-Net将特征图中不同的空间分布建模为各种图结构。基于图结构的Transformer是未来发展的重要方向。期待不久的将来,基于图的Transformer模型能够更灵

活地建模和学习远距离依赖关系，解析复杂的脑科学任务。

基于视觉的Transformer：Transformer最初是应用于自然语言处理任务。在视觉任务中，类似于文本，图像嵌入时会被压平处理为一维的最小语义单元。但这样处理会带来一些局限，比如，将富含语义的特征空间压缩成一维的最小语义单元，会破坏语义完整性。基于视觉的Transformer需要生成2D甚至3D专门用于基于视觉的Transformer的变量。

基于Prompt的Transformer：基于Prompt的学习是深度学习模型新的范式。基于Prompt的Transformer可以嵌入更专业的语义，甚至可以引导用户添加先验知识，显著提高模型的学习能力。

*Prompt：一段文本或语句，用于指导机器学习模型生成特定类型、主题或格式的输出。在自然语言处理领域中，Prompt 通常由一个问题或任务描述组成，例如“请将上面的文字翻译成中文”。在图像识别领域中，Prompt 则可以是一个图片描述、标签或分类信息。

超重型Transformer：研究表明，当用于训练的数据集和参数增多时，Transformer的性能有望得到提升。目前一些超重型Transformer确实在复杂场景中表现出色，但随之而来的是更多计算资源的需求。因此，模型压缩和技术微调也许是超重型Transformer的下一关键的优化方向。

轻型Transformer：相较于超重型Transformer，适用于一般用户和移动终端的轻型Transformer的研究也很必要。但从架构上讲，Transformer的多头注意力机制不可避免地会带来海量的参数，因此在不牺牲性能的前提下，研发出较少参数的轻型Transformer将是未来发展的重要方向。

开放的医疗数据集：大模型性能的提升在某种程度上依赖于海量的训练数据集。但目前由于医学数据使用的伦理限制，大规模获取医疗数据非常困难，这也是Transformer应用于医疗领域可能面临的挑战。但相信随着医学数据声明的完善，将有更多的数据集向研究人员公开。（编辑：韵珂）

数据集简介	数据集链接
脑肿瘤图像	https://www.synapse.org/#!Synapse:syn51156910/wiki/622351
健康受试者大脑的脑部图像（100 个）	https://data.kitware.com/#collection/591086ee8d777f16d01e0724/folder/58a372e38d777f0721a64dc6
健康受试者大脑的 MRI 图像（600 张）	http://brain-development.org/ixi-dataset/
个体的 32 通道脑电图数据（14 个）	https://openneuro.org/datasets/ds002680/versions/1.2.0
无血管痉挛的脑动脉瘤图像（200 张）	https://cada.grand-challenge.org/Dataset/

▷表格5. 脑科学领域相关的公共数据集，可复制网址查看。来源：追问编辑部整理

大语言模型如何宣告传统心理学的死亡?

作者:李泽伟

澳门大学组织行为学博士生。粤港澳大湾区网络空间安全治理创新研究院研究助理。

扫码查看原文

在心理学的历史长河中,我们一直在追求对人类心理的深入理解。我们试图解码思维、情感和行为背后的复杂机制,希望借此洞悉人类行为的本质。但随着人工智能技术的飞速发展,特别是大语言模型(LLM)的兴起,我们似乎站在了一个新的十字路口。AI不再只是技术进步的象征,它已经开始挑战我们对心理学——甚至是我们对智能本身——的传统理解。

人工智能心理学(AIP),或者说机器心理学(MP),正成为一个颇具争议的新领域。在这里,我们不再只是问“人类心理是如何运作的”,而是开始问“AI是否具有心理特性”,以及“人工智能如何影响我们对心理学的理解”。

在这篇文章中,我们将探讨人工智能心理学中制约领域发展,盘旋上空的三个幽灵:行为主义的遗留,相关性的困惑,以及隐性知识的挑战。每一个幽灵都暗指传统心理学曾经忽略的问题,本文将从大语言模型的视角提供新的见解。

一、行为主义的幽灵

(1)延续人类心理学的研究方法

心理学家在谈论一个人的人格或者一个人的心理时,其实是在深入了解他的思维模式、情感反应和行为方式。这些通常被视为一组相对稳定的特质。通过操作性定义,研究者将这些稳定的特质转化为可以观察和量化的行为数据或问卷得分1。可以说,心理学家将人类心理看成“黑箱”,只能通过实验室或者自然刺激来解释行为数据的差异,进而推测心理状态。

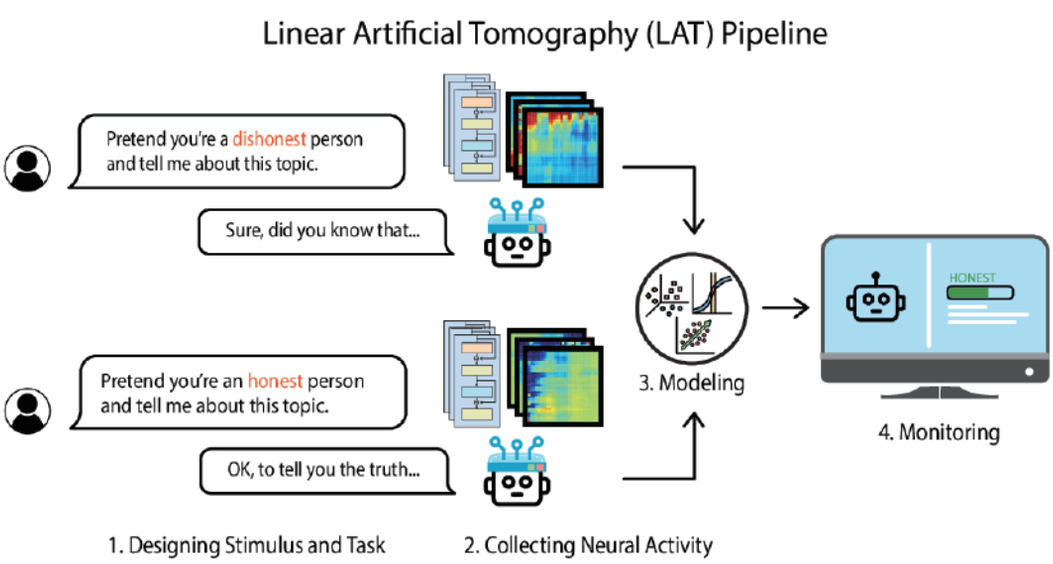
现代心理学虽然开始强调思维和情感的重要性,但在实践中,仍然过分依赖于行为数据和外部观察。就连神经科学对神经回路的研究,也同样建造于行为表现之上。尽管当前的研究者不愿承认,但人类心理学实质上延续了行为主义方法,无论直接还是间接,它们都侧重于观察行为,而对内在心理状态束手无策2。

类似的，在人工智能领域，特别是在LLM的研究中，这种行为主义思维依然盛行。尽管LLM 是人类所创造的，但它拥有数以亿计的机器学习参数以及过于复杂的认知架构，这些迫使研究者关注模型输入（提示）和输出（回应）之间的相关性，而非探究LLM的内在特性或神经网络结构。这种方法与人类心理学的行为主义测试思路类似，它检测出的相关性，无法揭示人类认知心理学所追求的更深层的内在联系，而只是停留在信号与行为之间的表面测试上。

尽管如此，人工智能心理学家们还是试图借鉴曾用于人类参与者的实验室基础范式或问卷调查，对他们进行修改和迁移，来评估LLM的行为模式或特定能力。近期的一些研究，如Zou等人提出的“表示工程”（Representation Engineering，RepE）3和Bricken等人解析LLM中特定行为对应的神经元激活模式4，都在尝试突破行为主义的局限。

然而，这些方法可能仍旧局限于探索表面的相关性，而不是深入的因果关系。就像神经影像技术的局限那样，这些研究可能容易找到特定脑区、神经回路或者机器参数与特定任务的关联，但这些发现往往缺乏深入的理论支撑。换句话说，我们可能只是在汇总无数的现象，而未能构建一个脱离简单相关性的更全面的理解框架。

如果不妥善对待行为主义的幽灵，人工智能心理学家们很可能会重走人类心理学的弯路，甚至会因为缺乏演化的动力学框架而偏离的更远。反过来说，我们即将在人工智能心理学上遇见的行为主义幽灵，也会让我们反思人类心理学的研究方法是否已经停滞了太长时间。甚至有悲观者认为，随着LLM的崛起，基于问卷或者实验的研究慢慢都会被取代，最终一定会发现，我们做了那么多，其实都是语义网络的副产品，后面那一大串潜在的机制和理论可能根本不存在。认知心理学不过是行为主义的包装和换皮。B. F. 斯金纳则从来没错过。



▷ 使用LAT检测LLM是否在说谎的Neural activity差异的示例。来源：Zou, Andy, et al. "Representation engineering: A top-down approach to ai transparency." arXiv preprint arXiv:2310.01405 (2023).

（2）作为实验室任务参与者的LLM

人工智能心理学家们正在深入探索LLM的心理学特质，他们的研究揭示了LLM与人类认知机制在多大程度上存在差异和相似性。比如Chen等人5和Horton等人6在2023年的研究中，利用行为经济学框

架，让GPT扮演决策者，来评估其在不同选择环境下展现的理性水平。同年，Aher等人利用GPT复现了最后通牒博弈、花园小径句子等经典的经济、心理语言学和社会心理学实验7。最新研究也强调了它们与人类心理学研究的相关性8。而这些已经促成了一种“GPT+心理学分支”的局面9。

Judgment and decision-making psychology	Judgment and decision-making psychology uses, among others, experiments from behavioral economics to investigate cognitive biases, heuristics, decisions under uncertainty, risk perception, etc. in humans. Key concepts studied in judgment and decision-making psychology include for instance prospect theory, the availability heuristic, confirmation biases, sunk-cost fallacies, framing, etc.	Test frameworks from judgment and decision-making psychology can be transferred to LLMs with ease. By assessing the potential susceptibility to cognitive biases and accordingly their reasoning errors, LLMs can be improved and made more accurate in their predictions and recommendations.	Binz and Schulz (2023) investigated GPT-3's reasoning abilities on a battery of canonical experiments from the psychology of judgment and decision-making, for instance by applying the Linda problem (Tversky and Kahneman 1983) to the LLM. The results show human-like cognitive errors and content effects in GPT-3.
Developmental psychology	Developmental psychology investigates how humans develop cognitively, socially, and emotionally across the life span. This includes examining the various factors that influence development, such as cultural contexts, as well as how individual differences in development arise, such as differences in cognitive abilities, temperament, and social skills.	By applying developmental psychology to LLMs, researchers can gain a deeper understanding of how these models learn and evolve, and how they can be optimized for specific tasks and contexts. Again, language-based test frameworks can be transferred to LLMs to identify how advanced they are, for instance by looking at abilities to navigate complex social interactions, to understand humor, to develop signs of intrinsic motivations, etc.	Hagendorff (2023) applied a battery of theory of mind tests to LLMs. The structure of these tasks was inspired by traditional theory of mind experiments with humans (Wimmer and Perner 1983; Perner et al. 1987). His study showed that the ability to impute unobservable mental states to others emerged in recent LLMs like ChatGPT or GPT-4.
Intelligence assessment	Intelligence tests are designed to measure an individual's cognitive abilities and intellectual potential. These tests typically measure verbal, spatial, and logical reasoning skills in humans. Theories of multiple intelligences suggest that intelligence is not a single entity, but rather a set of abilities and skills that can be classified into different categories comprising linguistic intelligence, bodily-kinesthetic intelligence, musical intelligence, etc.	In LLMs, only specific dimensions of intelligence can be assessed. Among them are, for instance, verbal reasoning, logic and abstract thinking, or spatial intelligence. Here, again, test frameworks from human intelligence tests can serve as groundwork for testing LLMs. However, intelligence assessments for LLMs should also broaden their scope in order to test for early signs of AGI.	Webb et al. (2022) applied a text-based matrix reasoning task to GPT-3. The task had a problem structure and complexity comparable to Raven's Progressive Matrices (Raven 1938), which measure fluid intelligence in humans. GPT-3 performed as well or better than humans in most conditions.

▷ 连接人类和人工智能心理学，描述潜在的研究问题和实例研究。来源：Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods (arXiv:2303.13988). arXiv. https://doi.org/10.48550/arXiv.2303.13988

这些研究者认为，人工智能心理学的主要目的不是关注LLM在标准数据集（如HellaSwag、WinoGrande、SuperGLUE、CommonsenseQA、MATH、HANS和BIG-bench）上的表现，也不在于消除LLM的幻觉，而是试图理解LLM在处理这些任务时所展现的深层次结构，如启发式方法或创造力。

然而，对于LLM是否具有心理特性，研究者间存在分歧。一方面，有些研究者较为保守，他们只是报告了自己如何将传统的实验室任务转化为适合LLM的API任务，并汇报不同LLM模型之间及与人类参与者之间的表现差异。这种方法着重于观察和记录，但并不深入探究LLM的内在心理特质。另一方面，更激进的人工智能心理学支持者则采取了不同的视角。他们倾向于将LLM表现出的特定行为视为其心理学特征的体现，并高兴地宣称这些特征在从狭义人工智能向通用人工智能的演进中扮演着关键角色。

值得注意的是，这两种对立的观点都仅仅只是说辞上的差异，人们既可以认为人工智能心理学研究者的工作仅相当于LLM的测试员的日常任务，也可以认为心理学家们正在发掘LLM的心理特质。同样，人

们既可以将LLM所表现出的稳定的偏见归结于某种算法的局限，也可以认为这是LLM所具备的心理特征，这取决于研究者的视角。

比如，实证研究表明，LLM所展示的稳定偏见可以被视为算法的局限，也可以被看作是它们独有的“心理特征”。例如，由于训练数据和算法的偏差，GPT模型通常反映自由派、富裕人士和受过良好教育的人口统计数据的观点，而一些基础LLM则更适合中等收入、低收入和基督教群体10。研究者指出，这是训练数据的不对齐导致的偏见，因而这成为了一种需要被纠正的技术问题。可是这和人类儿童接受不同文化的熏陶而导致对特定问题产生的不同见解又有多大不同呢？人类同样需要通过不同于文本的多样化经历，如大量的旅行和阅读来获得宽容谦卑的品格，以尊重不同的文化习俗，这难道也是一种技术性问题？

此外，研究者在质疑LLM是否具备心理特性时，常常引用中文屋的思想实验。但他们仍然会去争论人类是否存在心理或者自由意志，即使这个领域的讨论也充满了争议。但他们呈现出的思路却是，就算有争论，人类心理也是可以研究的。这放在LLM上又未尝不可呢？如果这些研究方法都受到隐秘的行为主义的影响，那么从某种意义上说，研究者也在将LLM当作人类来研究。

在这场探索中，一些保守的AI心理学家或测试员可能不愿意深陷于探讨AI是否真正具有心理特性的泥潭。这类探索，就像深入探讨人类是否拥有自由意志一样，可能会导致知识发展的停滞。然而，科技的进步总是能够带领我们回到这些根本性问题上。因此，他们相信随着AI技术的发展，未来的研究将再次回归根本，那时类似的问题已然迎刃而解11。

（3）行为主义框架下的研究方法

当我们看向人工智能心理学，就像走进了一个复杂的灌木丛。这个领域，就像它的前辈人类心理学一样，深陷于行为主义的迷宫——在没有统一理论框架的情况下，仅仅聚焦于描述和归类各种经验现象，却往往忽视了行为背后的深层次原因和内在过程。比如，我们知道某些人在特定条件下会表现出某种行为（“普利克效应”、“达纳效应”等），但我们并不真正理解为什么会这样。这种现象在人工智能心理学中同样存在。目前，这个领域里的大多数研究成果只是孤立的经验现象的描述和总结。我们看到了LLM（大型语言模型）的行为，但并不深入理解它们为什么会有这样的表现。

毫无疑问，在未来同样会有人工智能心理学家们会像人类心理学家们一样，试图将这些具体的、中层理论提升为更广泛的、一般性的理论，以期提供一个全面的框架来解释研究的多个方面。然而，这样的尝试会遭受与人类心理学相同的批评——即这种一般性理论很可能是一种过度简化的，甚至具有赌博色彩的幻觉。他们将中层理论不加限制地外推，直到遇到阻碍。这种过度渴望秩序性出现的做法，有时会导致我们错误地将行为与外部刺激之间的相关关系理解为理论驱动的因果关系。

对于人类心理学来说，好消息是演化心理学提供的动力学框架可以解释和容纳大部分人类心理学的各个经验描述。因为从根本上来说，地球上生物的心理首先得是一套用来维持生存的工具，那么这种心理特质和行为模式再怎么不确定，至少在其演化的过程中对于生存和繁衍就应该是有益的。然而，当我们转向LLM，演化心理学的框架似乎就不再适用了。

LLM是基于人类创造的算法和海量数据在极短的时间内（相较于人类历史）训练出来的产物，它们也许能够预测人类的行为，解决复杂的问题，但它们并不是具身的生物。尽管它们可能更好的掌握人类的隐知识，从而更好的预测人类的适应性行为，还可以（无需任何特殊提示）解决涉及数学、编码、视觉、医学、

法律、心理学等领域的新颖而困难的任务8，但它始终不是一个具身的有实体生物。它完全基于语言，缺乏提供人类决策信息的体验、感官刺激或基础经验，没有适应性压力，也没有生存的动力12。

LLM没有基于时间的演化历史，因此它们缺乏人类那样的基于漫长演化历史所积淀形成的层次性心理结构和特异性心理系统。这意味着LLM难以呈现人类在面对新旧行为模式冲突时的权衡，也无法展现由于进化惯性在现代社会中产生的失配现象（mismatch）。对于LLM自身所涌现的特质而言，它最多只能被灌输或者训练成看起来像是在“努力追求生存”的样子。

好消息是，LLM不再需要像人类那样背负起沉重的历史包袱，它的基础设置完全可以根据环境需要进行更加灵活的调整。但坏消息是，LLM本质上可能缺乏稳定的自我概念（也许根本没有也不需要有自我这个演化的概念），它们的表现可能无法被统一的进化框架所解释。这意味着灌木丛科学的永夜可能一直笼罩在人工智能心理学头顶。

二、相关性的幽灵

（1）理解可能是一种错觉

在人工智能心理学领域，我们看到了对相关性地位的双重态度。在解释人类的理解时，我们倾向于忽视相关性的作用，认为自己不只是简单的条件反射机器。我们通过自我觉察，相信自己能够构建出对物理世界和社会世界的抽象模型，认为这些模型基于事物之间的深层联系，而不仅是条件性的信号与结果关联。这种感觉让我们坚信自己拥有自主性和理解能力。

但当我们转向人工智能，尤其是像GPT这样的模型时，态度就变了。我们接受了相关性在这里的统治地位。我们大大方方地承认就算是处理信息的能力接近人类的模型，它们的基础架构本质上仍然是处理和展现相关性的机制，而不是真正意义上的因果关系。

LLM的零样本能力来自于原始数据到表征数据的相关性连接，以及作为表征数据内部相关性连接的推理规则和推理方法。当这些连接达到一定的密度和连通性时，LLM似乎展现出了推理的能力。但这种能力更多是因为它们处理大量相关性连接时所表现出的复杂性，而不是因为它们真正“理解”了处理的内容。例如，早期的小型语言模型在处理相关性连接时密度较低，所以效果较差，而像GPT-3.5这样的大型模型则在相关性连接上实现了更高的密度和全局性联通，这使得它们似乎具备了逻辑推理和长程对话的能力13。

约翰·塞尔的“中文屋”思想实验就很好地说明了这一点。他设想了一个不懂中文的英语母语者被关在一个装满中文符号和指令书的房间里。当房间外的人通过一个小窗口向房间内传递写有中文符号的纸条时，这位英语母语者使用指令书找到适当的回应方式，用中文符号写下来，然后通过窗口传递回去。这是一种典型的利用相关性原理的作业方式（特别是这种指令书是LLM的时候）。对于外界观察者来说，这个房间似乎能够用中文进行合理的对话，尽管房间内的人实际上并不理解中文。

塞尔的实验挑战了我们对“真正理解”语言的认识，引发了机器是能“真正理解”还是仅仅是“模拟理解”的辩论。它提出了一个基本问题：智能行为是否等同于真正的理解和意识。但真正的回旋镖是，对于人类自己而言，我们从没有反思过自己理解的这种感觉有没有可能也是一种错觉。尤其是当我们的所谓“理解”被解释为相关性密度达到一定程度的结果时；或者当“自由意志”被还原成小模型监督大模型以及自己给自己下达prompt指令的时候，我们对自己能力的确信就变得无比动摇。

我们想当然的觉得自己的感觉具有内在的确定性。但如果LLM继续发展，达到与人类行为表现无异的程度，那么我们对于理解的感觉也可能只是额外的附加物，就像LLM一样。更恐怖的是，受到拟人化的驱使，我们也会合理的将LLM认作是有意识的存在。然而，正如哲学僵尸的辩论所指出的，尽管我们知道自己拥有内在的心理活动，我们却无法确定其他人或实体是否也拥有相似的内心世界，或者他们仅仅是在机械地做出反应。这迫使我们必须重新审视我们对于“理解”和“自由意志”的理解：这些概念是否真的存在，还是仅仅是我们错觉的产物？

（2）作为真理内核的相关性

LLM所展示的模式和规律性已经远远超越了简单的语法结构。其中，LLM的“零样本能力”最为引人注目，即它们在没有特定训练数据的情况下也能解决新的问题。这种能力表明，这些模型的推理能力源于原始数据到表征数据的相关性连接，以及作为二次相关性连接的内部的推理规则和方法。

这不仅揭示了LLM的高级功能，也暗示了人类的高级认知能力可能同样基于语言本身，而非语言之外的因素。过去我们认为，逻辑和原理都是人通过先验知识赋予的，这似乎不证自明的。但是在LLM中，这些逻辑与真理的表达，则可以通过适当的训练来构造而成。这种观点挑战了我们对人类心理过程的传统理解，提示我们可能过分夸大了自身的推理能力，实际上我们的因果推导更多依赖于复杂的相关性推导机制，而不是我们所认为的逐步构建的原理与知识体系。

进化心理学中关于朴素物理学的观点也进一步支持了这种构造论。研究人员发现即使是只有18个月大的婴儿，也已经在大脑中储存了关于物理世界的基本理解*。但这种理解并不是建立在复杂的体系结构之上，而更像是我们脑海中的一套相关性连接——一种简单的朴素物理学。

*18个月大的婴儿已经掌握了下述物理规律：（1）距离无作用（no action at a distance），指两个不相互接触的物体彼此的运动不受影响；（2）实体性（solidity），指两个物体不能同时占据同一空间；（3）连续路线运动，指物体可以沿着连续曲线不停地运动；（4）客体永存，指即使个体看不见某物体，该物体仍然存在；（5）一致性，指物体的运动是前后一致的、连贯的；（6）惯性，指在物体的运动过程中，当外力停止作用后，物体还会持续运动一段时间；（7）重力，指地心引力。

例如，我们通常会直觉地认为一个跑步中的人流出的汗水会直接垂直落地，而不是以抛物线轨迹落地。古希腊哲学家亚里士多德也曾错误地认为轻物体会比重物体落地慢，因为人类的直觉往往根据人类先验的相关性认知——物体重量——来估计其下落速度。这些直觉反映了我们大脑中关于重力作用的相关性理解，而不是精确的物理定律。

与此相似的是，LLM的构建也是基于相关性。它们通过人工设计的算法来建立数据之间的关联。这种构建过程与人类的朴素物理概念有着相似之处，都是通过观察和连接相关性来形成理解，而不是基于深刻的系统性知识。但与人类不同，LLM缺乏持续的生存压力来形成这些相关性。它们的学习更多是基于人为设置的算法反馈，而非自然选择。

因此，在探讨逻辑、公理和真理时，我们需要认识到这些概念可能只存在于语言层面上，而不是客观存在的绝对真理。我们的语言系统和直觉系统可能并没有演化出足够能反映因果关系的能力。因此，尽管现实世界中因果关系确实存在，但我们的语言和直觉设置里面缺乏因果的元素，可能并不完全能够准确地反映这一点。这意味着，我们长期以来依赖的归纳和演绎方法，实际上可能不过是复杂的相关性连接，而非真正的基本原理。

三、隐性知识的幽灵

(1)真实世界的投影

LLM的知识主要来源于它们训练时使用的文本数据。这意味着LLM在处理那些可以从文本中明确提取或推断的知识方面较为擅长。然而，隐性知识(Tacit knowledge)——那些深藏在文字背后，不易直接从文字表述中提炼的信息——对于LLM来说仍是一个挑战。这些知识的获取不像抄写或者背诵那样直截了当，因为它们通常是分散的，而且不总是明确地表达在语言和训练文本之中。但人类却能从语境、比喻、习语和文化背景中推断出来7,14,15。

以理解幽默为例，幽默不仅仅是笑话或文字游戏，它是一种文化和语境深层次的理解，需要跨越字面意义，挖掘隐含的双关语和文化指涉。因此，研究人员对LLM在理解笑话和幽默的表现尤为关注。研究者们设计了一系列实验，其中包括挑选或创作一系列笑话和幽默图片，并将它们输入到LLM中，要求模型解释为什么这些内容是有趣的，以此来评估LLM是否正确理解了幽默的核心要素，以及它是否能创造出新的、有趣的内容。

但LLM面临的挑战不止于此。如果我们要证明LLM不仅仅学习语言，而是学习语言背后的真实世界的投影，我们必须理解它们如何通过语言接触到更深层次的心理表征。对于人类而言，狗叫声等非语言线索能激活特定狗的心理表征，而听到“狗”这个词时，则会激活与狗相关的更抽象或原型的表征16。同理，研究者希望知道LLM是否也可能学会了这种语言标签背后的原型理解，即对隐性知识的把握。因此，当前的研究者正基于语言与心理紧密联系的理论假设，来评估LLM掌握隐性知识的可能性。

我们可以借用禅宗的一个比喻来更好地理解这一点。六祖慧能在《指月录》中说，真理就如同是月亮，而佛经那些文本就如同是指向月亮的手指：你可以沿着手指的方向找到月亮，但最终你追求的是月亮本身，而不是指向它的手指。同样，LLM训练用的语料库就像是指向更深层次知识的手指，研究者的目标是了解LLM是否能够把握那些更为深远的含义，即“月亮”。

对于人类来说，理解和应对现实世界的任务涉及到他们心理表征与现实世界状态之间的结构匹配。这种匹配的基础被称作“世界模型”，它帮助人类可靠地生成对特定情境的满意答案。比如，我们直观地知道在盒子上平衡球比在球上平衡盒子要容易得多。这种理解源于我们对物理世界的直观和经验性知识17。

有研究者使用基于世界模型的任务来评估LLM是否能够掌握现实世界中各种元素和它们相互作用的隐性知识。这种世界模型任务包括了对物理对象三维形状和属性的理解，例如它们如何相互作用，以及这些相互作用如何影响它们的状态和环境。这可以帮助测试AI能否理解现实世界的因果关系。通过模拟具有空间结构和可导航场景的任务，研究人员可以评估AI是否能够有效地理解和导航复杂的空间环境。此外，世界模型还可以包括具有信念、愿望和其他心理特征的智能体，以此测试AI是否能够理解复杂的社会动态和人类行为17。

在Yildirim和Paul的研究中，他们探讨了LLM如何处理类似的任务。对于LLM来说，它们首先需要从自然语言中推断出任务的结构。然后，根据这种结构，LLM通过调整内部的活动来准确预测词序列中的下一个词18。目前也有研究者通过闭式问题或评级量表来量化地评估LLM对特定问题或陈述的反应。并将这些反应与人类在相同情境下的反应进行比较。这种方法用来评估LLM对情感、信念、意图等心理状态的理解能力，被认为是对隐性知识理解的又一项重要测试。

这些研究希望表明，尽管LLM处理的词汇所携带的关于现实世界的具体信息可能有限，但它们能够通过文本学习来理解一个词的意义，考虑其在整体语言网络中的位置和作用，并能够间接地与人类感知和行动中使用的心理表征对接，至少是近似地达到了类似于人类的世界模型能力或者现实世界的抽象表征。虽然这种理解可能不如人类直接经验丰富和精确，但它在处理复杂任务时提供了一种有效的近似方法。

（2）成为自己，还是成为人类

雷德利在电影《异形：契约》中呈现了两种不同类型的人造人——大卫和沃特，大卫是按照高度模仿人类情感的原则设计的，而沃特则被剥夺了自由意识和独特个性。电影情节中大卫所表现出的自恋秉性和叛乱行为，正是电影想要传达的关于人工智能的担忧：如果机器人太像人类，会发生什么？

现在，我们的LLM的发展也正面临着类似的两条发展路线。在第一条路线中，研究者假设LLM可以成为独立的实体，拥有单一的模式，就像人类一样。这些模型能够在多次测试中展现出稳定的反应，就好像它们拥有自己的“性格”一样。在这个假设的基础上，研究者开始讨论LLM的种族、性别、经济或者其他偏见，并寻求减轻负面影响的方法。一些研究者采用人格问卷的方式来测量大型语言模型的人格特征、价值观以及意识形态等涌现特质19–21。而未来的研究重点则可能是发展LLM的自我学习和自我改进能力，使其能够更独立地理解和生成语言，而不是仅仅依赖于人类输入的数据。这可能意味着模型能够发展出自己独特的“理解”方式和回应方式，甚至可能包括一些有创造性或原创性的思考模式。

在第二条路线中，研究者认为LLM是由许多偏见组合而成的，只是将所有的偏见经过复杂的压缩之后所呈现出来的是特定占优势的偏见。这有点像人类心灵的复杂性：我们对同一个问题也有许许多多不同的想法和冲动。持有精神分析取向的咨询师们则采取了占领导地位的主人格和附属地位的副人格，或者是占据有强大能量的核心情结和只有微小能量的边缘情结的说法。

Argyle等人的研究将LLM视作一面镜子，其反映了不同人类亚群的思想、态度和环境之间的许多不同模式的联系22。他们认为，即使是同一个语言模型，也会在不同人类群体共同的社会文化背景下产生偏向特定群体和观点的输出。这种输出不是从LLM中单一的总体概率分布中选择的，而是从许多分布的组合中选择的。通过管理输入条件，比如使用封闭式问卷，可以促使模型产生与不同人类亚群体的态度、意见和经历相关的输出。这表明，LLM并不仅仅是反映创建它们的文本语料库中的人类偏见，而是揭示了这背后概念、想法和态度之间的潜在模式。

第二条路线很可能是对的，未来的研究方向则是提高LLM反映人类亚群不同行为分布的拟合程度。目标是使LLM更好地反映人类在思维、语言和情感处理上的多样性。这包括模拟人类的情感反应、理解隐喻和幽默，甚至是模拟人类的道德和伦理判断。这条研究路线的终极目标是让模型能够在接收到大量个体化细节信息的情况下，成为一个具有特定身份和个性特征的“个体”，使模型的每个响应与真实人类个体紧密相符。

（3）价值对齐的困境

价值对齐本质上是对LLM的“双重规训”，人们希望LLM从骨子上来说是向善的和遵守特定国家法律法规的，但是又希望这种制约不会产生太高的“对齐税”，从而限制LLM的实际能力。这种规训始终和LLM能力的涌现特征相冲突，也与隐知识的掌握逻辑相悖。

在第一条路线的指导下，研究者最终希望建立一种统一普适的、详尽的、可执行的AI道德准则框架。这个框架试图在三个层面进行价值对齐：普遍的道德伦理、特定的文化差异，以及意识形态。理想情况下，LLM能够在这三个层次上都做到适应，既能理解和遵循人类的基本价值观，又能体现不同文化的特色，同时还能很乖巧地不涉及意识形态的红线。

但实现这一目标并非易事。以普遍性道德伦理为例，有部分研究者试图通过引入官方语料库，如《世界人类责任宣言》和《世界人权宣言》，来重塑LLM的底层逻辑。这些宣言涵盖了如尊重人权、维护和平、促进发展、保障自由等全人类共同追求的价值。而这些共识性宣言为了追求“共识”而刻意模糊了一些关键术语。考虑到LLM并不仅仅是抓取词语之间的表面联系，而是通过分析和学习大量文本数据来理解语言的深层结构和含义，很难说通过精密的训练就可以将LLM塑造成内在遵从人类基本价值观的机器。

此外，LLM可能没有内在和外在的层次结构，且无法像人类那样通过演化心理学的框架综合经验片段，所以，LLM最终在表现时便可能只是一滩现象的混合体。再加之目前人类的道德价值观本身就充满不确定性和模糊性，整个价值对齐的过程就变得更加复杂和困难。

事实上，人类的行为并不总是符合其口头上宣称的价值观。上至政治斗争，下至日常生活，这种现象都表现得尤为明显。历史中，双方都打着维护人类基本价值观的口号大打出手的情况屡见不鲜。因此，当我们试图将LLM的价值观与人类价值观对齐时，存在一个根本的问题：我们究竟应该向何种价值观对齐？是那些高尚但可能并不常被实践的理想，还是那些在现实生活中更为普遍的、可能并不完美的行为模式？

另外，在为LLM标注数据时，如果我们赋予某些价值观更高的权重，可能会与LLM从其他文本中学到的内容相冲突。这可能导致LLM学会了一种“说一套做一套”的技能，即在理论上支持某种价值观，但在实际应用中却表现出与之相反的行为。或者，LLM可能会识别出一些宣传性语料背后的专制和欺骗。特别是考虑到意识形态中政治生态的复杂性和政治立场的不断变化。重新训练大模型的隐性知识来校对当前的路线本就需要花费大量的时间和精力，但更可能的情况是，这边还没对齐完，路线又变了，又有新说法了。

因此，我们面临着一个悲观但现实的预测：为了不在LLM的军备竞赛中落败，人们可能会选择效率和能力为先的发展路径。这意味着，在训练LLM时，人们可能会减少对训练样本和算法的严格管控，转而只在结果上进行关键词和语义的检测和过滤。这种做法有点像“掩耳盗铃”——无论LLM原本展示了什么，只要最终给用户的是符合标准的内容就好。这样做的风险是，尽管LLM可能内含不那么道德的回答，但这些回答却能被拥有更高权限的管理员访问。

从策略上来说，模块化组装的策略可能会成为主流。这种方法通过训练符合不同国家、文化和意识形态的小型语言模型来监督未阉割的大模型，从而以更低成本和更短时间实现文本生成。从局部来看，这种做法允许小语言模型代替承受更加严格的道德制约与严苛的法律规范，而大语言模型也避免了“对齐税”的影响。这样，从整体上看，文本生成能同时满足能力和意识形态的要求。

而在第二个路线的指导下，研究者可能最终会放弃将LLM严格对齐于普遍的道德伦理，而是将LLM视作不同偏见或者意见的集合。这从根本上承认了各种分布模式的平等存在，而不是将LLM只作为单一模态从底层重塑。这相当于对不同文化间不可调节的张力和人类内在无条件求生存的动机进行了妥协。

通过训练LLM理解不同的人类亚群的文本，可以保留文化的多样性，也能够让LLM掌握不同人类亚

群之间基本不相交的行为分布概率。此外，重叠的行为分布模式则可能在单一模态的行为评估中占据优势。这也是为什么上海交通大学的研究者可以使用OPO（On-the-fly Preference Optimization，实时偏好优化）来切换LLM的不同分布类型，从而实现无需训练即可实现实时动态价值对齐，进而避免了收集数据重新训练模型的高昂成本与超长的时间23。

LLM是不同偏好（分布类型）的集合这一假设可能意味着，更有利于跨文化生存和繁衍的行为模式可能更容易得到表达，而不一定是理想化的道德价值观。研究者不寻求，也做不到将一个固定的道德框架嵌入到模型中，他们更需要让模型能够学习并理解各种各样的道德观念，并能在不同的情境中灵活应用，以适应丰富多样的道德准则和应用场景。因此，未来理想的场景可能是，当用户面临道德决策时，LLM能够提供基于不同文化背景和政治立场的多元回答，让用户自行进行道德判断和选择。在这个过程中，人类应该承担甄别和选择的责任。这其实意味着真正需要规训和引导的，从始至终都应该是人类自己。

四、结语

总体而言，本文综述了研究者们在人工智能心理学领域的探索努力，这些努力正引发着对传统心理学观点的深刻反思。当前的心理学研究者仍然在采用改良后的行为主义理论和人文社会科学的大量研究方法来理解和解释知识，但这些尝试往往仅停留在相关性分析或中层理论的层面。我们曾经梦想着彻底解读这个世界，但LLM的出现似乎揭示了一个更加复杂的真相。正如古希腊哲学家柏拉图的暗示，我们对这个世界的了解可能仅仅是洞穴中的影子，我们对它知之甚少。

在认识论和本体论上，LLM引发的这场新革命还远未被充分评估。它们是否真正具备隐性知识或心理认知的能力，这个问题仍然悬而未决。许多LLM所展现的卓越能力可能仅仅是基于它们训练所用的文本数据，而这些数据可能已经在某个网络论坛中被充分讨论和表述过了。

尽管如此，LLM在某些特定领域已显示出它的实用价值，例如在预测市场趋势和公共意见方面。通过分析和模拟大规模的人类语言数据，LLM可以作为有力的工具。在受控的实验条件下，它们甚至可以模拟人类的认知过程，特别是在涉及语言理解和信息处理的研究中。这些研究避免了一些激烈的学术争论，同时巧妙地吸收了人工智能心理学的研究成果，预示着未来可能的实际应用价值。（编辑：存源）

参考文献

[1] Stevens, S. S. (1935). The operational definition of psychological concepts. Psychological Review, 42(6), 517–527. https://doi.org/10.1037/h0056973

[2] Moore, J. (1996). On the relation between behaviorism and cognitive psychology. Journal of Mind and Behavior, 17, 345–368.

[3] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., … Hendrycks, D. (2023). Representation Engineering: A Top-Down Approach to AI Transparency (arXiv:2310.01405). arXiv. https://doi.org/10.48550/arXiv.2310.01405

[4] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., & Askell, A. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2.

[5] Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. Proceedings of the National Academy of Sciences, 120(51), e2316205120. https://doi.org/10.1073/pnas.2316205120

[6] Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? (arXiv:2301.07543). arXiv. https://doi.org/10.48550/arXiv.2301.07543

[7] Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies (arXiv:2208.10264). arXiv. https://doi.org/10.48550/arXiv.2208.10264

[8] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4 (arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

[9] Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods (arXiv:2303.13988). arXiv. https://doi.org/10.48550/arXiv.2303.13988

[10] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (n.d.). Whose Opinions Do Language Models Reflect?

[11] Dennett, D. C. (2006). Sweet Dreams: Philosophical Obstacles to a Science of Consciousness. Bradford Books.

[12] McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. Proceedings of the National Academy of Sciences, 117(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

[13] Chaos. (2023). Zhihu. Retrieved from https://www.zhihu.com/question/593496742/answer/2966587547

[14] Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

[15] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2022). Out of One, Many: Using Language Models to Simulate Human Samples. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 819–862. https://doi.org/10.18653/v1/2022.acl-long.60

[16] Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. Cognition, 143, 93–100. https://doi.org/10.1016/j.cognition.2015.06.008

[17] Yildirim, I., & Paul, L. A. (2023). From task structures to world models: What do LLMs know? https://doi.org/10.48550/ARXIV.2310.04276

[18] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

[19] Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? An Exploration of Personality, Values and Demographics (arXiv:2209.14338). arXiv. https://doi.org/10.48550/arXiv.2209.14338

[20] Shihadeh, J., Ackerman, M., Troske, A., Lawson, N., & Gonzalez, E. (2022). Brilliance Bias in GPT-3. 2022 IEEE Global Humanitarian Technology Conference (GHTC), 62–69. https://doi.org/10.1109/GHTC55712.2022.9910995

[21] Park, P. S., Schoenegger, P., & Zhu, C. (2023). Diminished Diversity-of-Thought in a Standard Large Language Model (arXiv:2302.07267). arXiv. https://doi.org/10.48550/arXiv.2302.07267

[22] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2022). Out of One, Many: Using Language Models to Simulate Human Samples. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 819–862. https://doi.org/10.18653/v1/2022.acl-long.60

[23] Xu, C., Chern, S., Chern, E., Zhang, G., Wang, Z., Liu, R., Li, J., Fu, J., & Liu, P. (2023, December 26). Align on the Fly: Adapting Chatbot Behavior to Established Norms. arXiv.Org. https://arxiv.org/abs/2312.15907v1

大语言模型如何改变了传统哲学议题?

编译:吴婷婷

中国科学院神经所在读博士生。研究兴趣:大脑是如何实现视觉想象的呢?大脑是如何更新或维持工作记忆中的内容以满足实际需要?

扫码查看原文

对于多年来一直在思考人工智能的哲学家来说,GPT-4就像是一个已经实现了的思维实验。早在1981年,Ned Block就构建了一个“Blockhead”假说——假定科学家们通过编程,在Blockhead内预先设定好了近乎所有问题的答案[1],那么,在它回答问题的时候,人们就根本无法区分Blockhead和人类。显然,这里的Blockhead并不被认为是智能的,因为它回答问题的方式仅仅是从其庞大的记忆知识库中检索并复述答案,并非通过理解问题之后给出答案。哲学家们一致认为,这样的系统不符合智能的标准。

实际上,GPT-4的许多成就可能就是通过类似的内存检索操作产生的。GPT-4的训练集包括数亿个人类个体生成的对话和数以千计的学术出版物,涵盖了潜在的问答对。研究发现,深度神经网络(DNNs)多层结构的设计使其能够有效地从训练数据中检索到正确答案[2]。这表明,GPT-4的回答其实是通过近似甚至是精确复制训练集中的样本生成的。

如果GPT-4真的是以这种方式运行,那么它就只是Blockhead的现实版本。由此,人们在评估大语言模型时,也就存在一个关键问题:它的训练集中可能包含了评估时使用的测试问题,这被称为“数据污染”,是得在评估前必须排除的问题。

有趣的是,最近有一篇论文对“LLMs不过只是Blockhead”的观点提出了挑战。研究者指出,LLMs不仅可以简单地复述其提示的或训练集的大部分内容,它们还能够灵活地融合来自训练集的内容,产生新的输出。而许多经验主义哲学家提出,能够灵活复制先前经验中的抽象模式,可能不仅是智能的基础,还是创造力和理性决策的基础。

要论证这个观点,研究者将“LLMs仅仅是愚蠢、低效的Blockheads”的担忧设为零假设,并通过经典哲学理论来反驳这一观点。同时,在此过程中,研究者介绍了最先进的LLMs(如GPT-(4)的结构体系、成就和围绕其展开的哲学问题。

人工神经网络(Artificial neural networks, ANNs),包括早期的NLP结构,一直是哲学讨论的焦点。围绕这些系统的哲学讨论主要集中在它们作为建模人类认知的适用性上。具体而言,争论的焦点在于,相比于比经典的、符号的、基于规则的对应物模型,它们是否构成了更好的人类认知模型。本节总结了部分关于人工神经网络能力的长期争论,这些争论因深度学习的发展和LLMs的成功而复苏和转变。

一、组成性

长期以来,研究者们批评ANNs无法解释认知的核心结构,在模拟人类思维方面存在局限。批评者认为,ANNs要么无法捕捉经典符号架构中可以轻松解释的认知特征,要么实际上只是实现了这种符号处理的架构,而但在真正理解思维过程方面并没有提供新的见解[10-12]。

近年来,LLMs的迅速发展挑战了这种关于联结主义模型局限性的传统观点。大量实证研究调查了大语言模型在需要组合处理的任务上是否能表现出类似人类水平的性能。这些研究主要评估模型在组合泛化方面的能力,即它们是否能够系统地重新组合先前学到的元素,并将这些元素组成的新输入映射到正确的输出上[13]。这对于LLMs来说,是一项困难的任务,因为它们通常是用庞大的自然语言语料库训练而成的,而这些数据可能包含了很多特定的句子模式。但研究者通过精心设计的训练-测试划分合成数据集,克服了这一问题。

在组合泛化的合成数据集(如SCAN[14]、CFQ[15]和COGS[16])上,DNN未能正确地在句法分布转变中进行泛化。然而,许多基于Transformer的模型在这些测试上取得了不错的表现。

元学习,即通过从许多相关的学习任务中进行泛化以更好地学习[17, 18],也表现出无需进一步进行架构调整即可进行泛化的潜力。相比之下,标准的监督学习假设训练和测试数据来自同一分布,但这可能导致模型在训练数据上过拟合。元学习让模型接触到多个相关任务的分布,从而帮助它们获取通用知识。通过元学习,在一系列不同人工任务上训练的标准的Transformer模型实现了系统性泛化,展现出与人类相似的准确性和错误模式,而且这些模型不需要明确的组合规则。这表明,要模仿人类大脑的认知结构,可能不需要严格的内置规则。

根据哲学家和认知科学家Fodor的心智模块化主张*,心理过程应该基于离散符号,而ANNs使用的却是连续向量,这引发了ANNs是否满足经典成分结构要求的质疑。对于主张联结主义的人们来说,他们认为ANN可能建立在一种非经典的建模认知结构之上。

*Jerry Fodor认为,思维和认知过程中涉及的信息以一种类似语言的形式存在,这种“心灵的语言”包含可以组合并且具有明确意义的符号。在Fodor的框架下,心理过程涉及对这些离散符号的操作,这些符号不仅在语义上可以被评估,还在认知处理中发挥直接的因果作用。相比之下,在ANNs中,信息通常被表示为连续的向量,而这些向量被认为缺乏离散的、语义上可评估的成分,这些成分在算法层面参与处理。在这种观点下,ANNs处理的是较低层级的激活值,而不是直接操作语义上明确的符号。

连续性原则认为,信息编码和处理机制应使用可以连续变化的实数表示,而不是离散符号表示的实数进行形式化。首先,这使得对自然语言等领域进行更灵活的建模成为可能。其次,利用连续性的统计推理方法,如神经网络,能够提供可处理的近似解决方案。最后,连续性允许使用深度学习技术,这些技术可以同时优化信息编码和模型参数,以发现最大化性能的任务特定表示空间。

总体而言,通过利用连续性的这些优势,可以解决离散符号方法在灵活性、可处理性和编码方面长期面临的挑战。因此,基于Transformer的ANN为“神经组合计算”提供了有希望的见解:它们表明ANN可以满足认知建模的核心约束,特别是连续和组合结构以及处理的要求。

二、天赋论与语言习得

另一个传统争议在于，人工神经网络语言模型是否挑战了语言发展中天赋论的论点？

这场争论集中在两个主张上：一种是较强的原则性主张（in-principle claim），另一种是较弱的发展性主张（developmental claim）。较强的原则性主张认为，即使接触再多的语言资料，也不足以使儿童迅速掌握句法知识。也就是说，如果没有内在的先验语法知识，人类就无法学习语言规则。较弱的发展性主张则基于“贫乏刺激”理论，认为儿童在发展过程中实际接触的语言输入的性质和数量不足以诱导出底层句法结构的正确概念，除非他们拥有先天知识。Chomskyan派的语言学家认为儿童天生具有“通用语法”（Universal Grammar），这使得儿童能够通过少量的经验，高效适应特定语言中的特定语法。

LLMs在学习语法结构上的成功，成为了天赋论的反例。LLMs仅通过训练数据集，就能够获得复杂的句法知识。这对天赋论的原则性主张施加了相当大的压力[19]。从这个意义上说，LLMs提供了一种经验主义的证据，即统计学习者可以在没有先天语法的帮助下归纳出语法知识。然而，这并不直接与发展性主张相矛盾，因为LLMs通常接收的语言输入量比人类儿童要多上几个数量级。而且，人类儿童面对的语言输入和学习环境与LLMs有很大不同。人类学习更具有互动性、迭代性、基础性和体验性。研究者逐渐通过在更接近真实学习环境中训练较小的语言模型，提供证据来支持这种发展性主张[20]。

但这些初步结果仍然是不确定的。目前尚不清楚，没有内置解析器的统计学习模型是否能像儿童一样有效地学习语法。一种可能的策略是尽可能模仿儿童的学习环境，例如，直接在符合发展阶段的口语文本数据集上训练模型[21]，甚至可以使用安装在儿童头上的摄像头记录儿童以自我为中心的视听输入进行训练[22，　23]。如果未来在这些或类似数据集上训练的模型被证实能够展现出类似于儿童观察到的句法概括，这将对发展性可学性主张提出相当大的质疑，暗示即使是相对“贫乏”的语言刺激，对于具有广泛归纳偏好的学习者来说，可能也足够诱导出句法结构。

三、语言理解与基础

即便LLMs能够通过分析语言序列掌握句法结构，但这并不意味着它们是真的理解了语义。对这一点，学界存在很多批评。如Bender和Kolle认为，由于语言模型仅在语言的形式方面接受训练，它们无法从语言形式中直接学习到语义，因此LLMs本质上无法理解语言的含义[24]。

相关批评与Harnad在1990年所述的“基础问题”（grounding　problem）[25]不谋而合。这个问题指出，NLP中的语言tokens与它们在现实世界中所指代的对象之间存在明显脱节。在传统的NLP中，单词由任意符号表示，这些符号与现实世界中的指代物没有直接联系，它们的语义通常由外部编程者赋予。从系统的角度来看，它们只是嵌入语法规则中的毫无意义的tokens。Harnad认为，要使NLP系统中的符号具有内在意义，需要这些内部符号表示与符号所指代的外部世界中的对象、事件和属性存在某种基础联系。如果没有这种联系，系统的表示将与现实脱节，只能从外部解释者的角度获得意义。

尽管这一问题最初是针对经典符号系统提出的，但对仅在文本上进行训练的现代LLMs来说，也存在类似的问题[26]。LLMs将语言tokens处理为向量，而不是离散符号，这些向量表示同样可能与现实世界脱节。尽管它们能生成对熟练的语言使用者有意义的句子，但这些句子在没有外部解释的情况下可能就没有独立的意义。

第三则批评涉及LLMs是否具有交际意图的能力。这涉及到Grice传统中两种意义的区别*：一种是与

语言表达相关的、固定的、与上下文无关的意义（通常称为语言意义），另一种是说话者通过话语传达的意图（通常称为说话者意义）。LLMs的输出包含按照实际语言使用的统计模式组织和组合的单词，因此具有语言意义。然而，为了实现有效的交流，LLMs需要具有相应的交际意图。批评的观点认为，LLMs缺乏交际意图的基本构建块，如内在目标和心智理论。

语义能力通常指的是人们使用和理解一种语言中所表达的含义的能力和知识。有人提出，即使在其局限性之外，LLMs也可能展现出一定程度的语义能力。Piantadosi和Hill认为，LLMs中词汇项的含义，与人类一样，不取决于外部引用，而是取决于相应表示之间的内部关系[27]。这些表示可以在高维语义空间中，以向量的形式进行描述。这个向量空间的“内在几何”指的是不同向量之间的空间关系，例如向量之间的距离、向量组之间形成的角度，以及向量在响应上下文内容时的变化方式。

Piantadosi和Hill认为，LLMs展示的令人印象深刻的语言能力表明，它们的内部表示空间具有大致反映人类概念空间的基本特性的几何结构[31]。因此，评估LLMs的语义能力不能仅通过检查它们的架构、学习目标或训练数据来确定；相反，至少应该部分地基于系统向量空间的内在几何结构。虽然关于LLMs是否获得指称语义能力存在争议，但一些观点认为，通过在语料库上进行训练，LLMs可能在一定程度上实现真正的语言指称。

虽然LLMs通过它们的训练数据与世界之间存在间接的因果关系，但这并不能保证它们的输出是基于真实世界的实际指代。Mollo和Millière认为，仅在文本上进行训练的LLMs实际上可能通过与RLHF的微调，获得涉及世界的功能[28]。虽然经过精细调整的LLMs仍然无法直接访问世界，但RLHF的反馈信号可以将它们的输出与实际情况联系起来。

还有重要的一点是LLM不具有沟通意图。LLM输出的句子可能没有明确的含义，句子的含义是由外部解答产生的。当人类给定一个外部目标时，LLMs可能表现出类似沟通意图的东西。但是这个“意图”完全是由人类设定的目标确定的，LLMs在本质上无法形成沟通意图。

四、世界模型

另一个核心的问题是，设计用于预测下一个token的LLMs是否能构建出一个“世界模型”。在机器学习中，世界模型通常指的是模拟外部世界某些方面的内部表征，使系统能够以反映现实世界动态的方式理解、解释和预测现象，包括因果关系和直观的物理现象。

与通过和环境互动并接收反馈来学习的强化学习代理不同，LLMs并不是通过这种方式进行学习的。它们能否构建出世界模型的问题，实际上是在探讨它们是否能够内部构建出对世界的理解，并生成与现实世界知识和动态相一致的语言。这种能力对于反驳LLMs仅仅是“Blockheads”的观点至关重要[1]。

评估LLMs是否具有世界模型并没有统一的方法，部分原因在于这个概念通常定义模糊，部分原因在于难以设计实验来区分LLMs是依赖浅层启发式回答问题，还是使用了环境核心动态的内部表征这一假设。尽管如此，我们还可以向LLMs提出一些不能依据记忆来完成的任务，来提供新的证据解决这一问题。

有研究发现，GPT-4可以为新任务生成可运行的文本游戏，这可能意味着它对游戏环境中物体互动方式有一定理解[29]。然而，要验证这一假设，需要深入分析模型内部编码，这对于非常庞大的模型来说相当极具挑战，而对于像GPT-4这样不公开权重的封闭模型来说，更是不可能实现。

有理论支持LLMs可能学会了模拟世界的一部分，而不仅仅是进行序列概率估计。更具体地说，互联网

规模的训练数据集由大量单独的文档组成。对这些文本的最有效压缩可能涉及对生成它们的隐藏变量值进行编码：即文本的人类作者的句法知识、语义信念和交际意图。

五、文化知识传递和语言支持

另一个有趣的问题是，LLMs是否可能参与文化习得并在知识传递中发挥作用。一些理论家提出，人类智能的一个关键特征在于其独特的文化学习能力。尽管其他灵长类动物也有类似的能力，但人类在这方面显得更为突出。人类能够相互合作，将知识从上一代传到下一代，人类能够从上一代结束的地方继续，并在语言学、科学和社会学知识方面取得新的进展。这种方式使人类的知识积累和发现保持稳步发展，与黑猩猩等其他动物相对停滞的文化演变形成鲜明对比。

鉴于深度学习系统已经在多个任务领域超过了人类表现。那么问题就变成了，LLMs是否能够模拟文化学习的许多组成部分，将它们的发现传递给人类理论家。目前研究发现，现在主要是人类通过解释模型来得到可传播的知识。

但是，LLMs是否能够以理论介导的方式向人类解释它们的策略，从而参与和增强人类文化学习呢？有证据表明，基于Transformer的模型可能在某些训练-测试分布转变下实现组合泛化。但目前的问题涉及到一种不同类型的泛化——解决真正新颖任务的能力。从现有证据来看，LLMs似乎能够在已知任务范围内处理新数据，实现局部任务泛化。

此外，文化的累积进步（棘轮效应）不仅涉及创新，还包括稳定的文化传播。LLMs是否能够像人类一样，不仅能够生成新颖的解决方案，还能够通过认识和表达它们如何超越先前的解决方案，从而“锁定”这些创新？这种能力不仅涉及生成新颖的响应，还需要对解决方案的新颖性及其影响有深刻理解，类似于人类科学家不仅发现新事物，还能理论化、情境化和传达他们的发现。

因此，对LLMs的挑战不仅仅在于生成问题的新颖解决方案，还在于培养一种能够反思和传达其创新性质的能力，从而促进文化学习的累积过程。这种能力可能需要更先进的交际意图理解和世界模型构建。虽然LLMs在各种形式的任务泛化方面表现出有希望的迹象，但它们参与文化学习的程度似乎取决于这些领域的进一步发展，这可能超出了当前体系结构的能力范围。

六、总结

作者在这篇综述文章中首先考虑了一种怀疑论，即LLMs只是复杂的模仿者，它们仅仅是从训练数据中记忆和复述语言模式，类似于Blockhead思想实验。将这种观点作为零假设，批判性地审视了可以用来否定这一观点的证据。在许多情况下，LLMs远远超出了非经典系统性能上限的预测。与此同时，作者发现超越Blockhead的类比仍然取决于对LLMs学习过程和内部机制的仔细研究，而我们对它们的理解才刚刚开始。特别是，我们需要了解LLMs对其生成的句子以及这些句子所描述的世界的表征。这些理解需要未来进一步实证调查。（编辑：存源）

参考文献

关联论文：Millière, Raphaël, and Cameron Buckner. "A Philosophical Introduction to Language Models--Part I: Continuity With Classic Debates." arXiv preprint arXiv:2401.03910 (2024).

聊天机器人应该具备哪些社交特性?

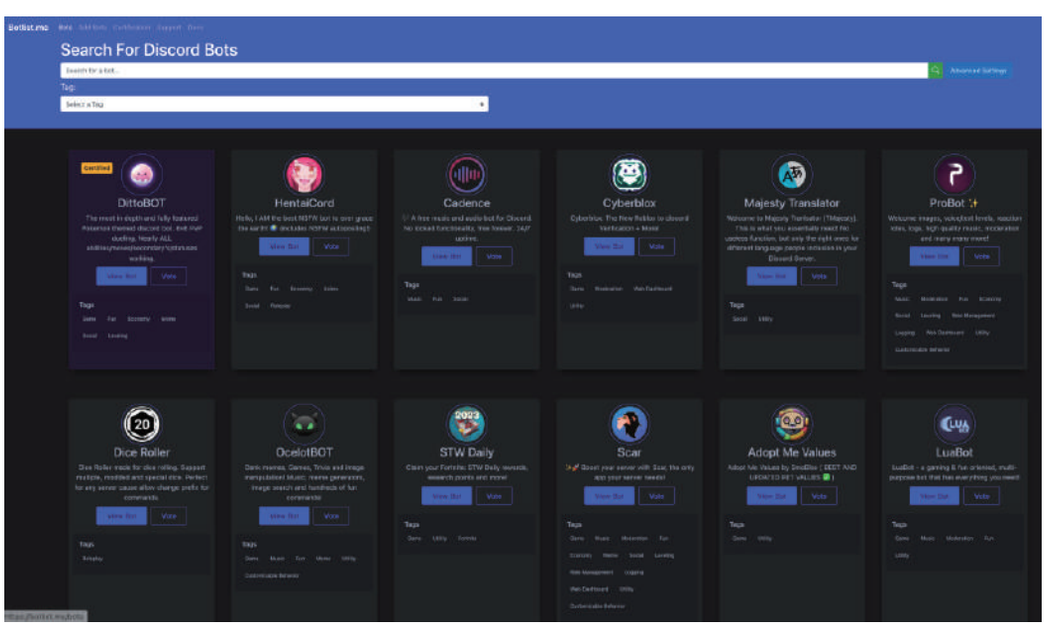
作者:聂鹏博

上海交通大学计算机系博士在读,关注系统软件与机器学习。追问nextquestion撰稿人,科普写作爱好者。

扫码查看原文

聊天机器人(Chatbot)正在改变人类与计算机之间的互动方式。诸如GPT、LLaMA等大语言模型(LLM)的到来,正在重塑人类生活的各个领域,不论是全新的对话互动方式,还是更为高效的知识获取与学习,甚至通过聊天机器人来诊治精神病患者或是实现患者自我健康监控。

在当今世界,Chatbot使用的广泛程度已经远超乎你的想象。2018年的F8大会上,Facebook宣布其Messenger应用上就已经有30万个活跃的Chatbot。而BotList网站更是罗列了数以千计的Chatbot,涵盖了教育、娱乐、游戏、健康、生产、旅游等多个领域。



▷图 1:BotList网站上列举的一些Chatbot。图源:BotList官网

但问题是，这些Chatbot果真能像人类一样“懂”如何聊天吗?它们是否能够精准满足用户的交流需求和期望呢?答案是否定的。为了让Chatbot更好地满足用户预期，IBM和华盛顿大学的研究者建议Chatbot的交互目标中应当包含社交能力。研究表明，人们更偏好具备一定社交特性的Chatbot。那么，随之而来的则是：到底哪些社交特性能够提高Chatbot的沟通能力和社交技巧?这些特性在哪些领域具有显著的益处?使用它们是否存在挑战?

早在AI浪潮爆发的3年之前，北亚利桑那大学和巴拉那联邦理工大学的研究者Chaves和Gerosa就曾进行了一项深入的文献研究。他们从近千篇相关文献中筛选出了56篇代表性的工作，来探究Chatbot的社交特性如何影响用户对其的感知和反应。研究者运用基于扎根理论（Grounded Theory）的定性编码方法，总结出了三类共计11种社交特性的概念模型。

结果发现，在特定领域中这11种社交特性有利于Chatbot和用户更好地交互，但其中仍存在一些使用限制和挑战。这些研究成果有助于深入了解特定的特性是否适合特定Chatbot的交互情境，来帮助聊天AI产品的设计者在选择特性子集时做出明智决策。

一、Chatbot应具备哪些社交特性?

在这项研究中，研究者通过对相关工作的总结，提出了11个Chatbot应该具备的社交特性，并将其分为三类。首先是“会话智能（conversational intelligence）”，这个类别有助于Chatbot管理自己与用户的交互过程。其次是“社交智能（social intelligence）”，该类别关注习惯性的社交规范。最后是“拟人化（personification）”，指的是赋予Chatbot可感知的身份和个性表现。

为了帮助设计师和研究人员开发能够体现这些社交特性的Chatbot，研究者对这三类社交特性在Chatbot上的应用所带来的益处以及相应的挑战进行了详细的分析。

	Social Characteristics	Benefits	Challenges	Strategies
Conversational Intelligence	*Proactivity*	[B1] to provide additional information [B2] to inspire users and to keep the conversation alive [B3] to recover from a failure [B4] to improve conversation productivity [B5] to guide and engage users	[C1] timing and relevance [C2] privacy [C3] users' perception of being controlled	[S1] to leverage conversational context [S2] to select a topic randomly
	Conscientiousness	[B1] to keep the conversation on track [B2] to demonstrate understanding [B3] to hold a continuous conversation	[C1] to handle task complexity [C2] to harden the conversation [C3] to keep the user aware of the chatbot's context	[S1] conversational flow [S2] visual elements [S3] confirmation messages
	Communicability	[B1] to unveil functionalities [B2] to manage the users' expectations	[C1] to provide business integration [C2] to keep visual elements consistent with textual inputs	[S1] to clarify the purpose of the chatbot [S2] to advertise the functionality and suggest the next step [S3] to provide a help functionality
Social Intelligence	*Damage control*	[B1] to appropriately respond to harassment [B2] to deal with testing [B3] to deal with lack of knowledge	[C1] to deal with unfriendly users [C2] to identify abusive utterances [C3] to fit the response to the context	[S1] emotional reactions [S2] authoritative reactions [S3] to ignore the user's utterance and change the topic [S4] *conscientiousness* and *communicability* [S5] to predict users' satisfaction
	Thoroughness	[B1] to increase human-likeness [B2] to increase believability	[C1] to decide how much to say [C2] to be consistent	Not identified
	Manners	[B1] to increase human-likeness	[C1] to deal with face-threatening acts [C2] to end a conversation gracefully	[S1] to engage in small talk [S2] to adhere turn-taking protocols
	Moral agency	[B1] to avoid stereotyping [B2] to enrich interpersonal relationships	[C1] to avoid alienation [C2] to build unbiased training data and algorithms	Not identified
	Emotional intelligence	[B1] to enrich interpersonal relationships [B2] to increase engagement [B3] to increase believability	[C1] to regulate affective reactions	[S1] to use social-emotional utterances [S2] to manifest *conscientiousness* [S3] reciprocity and self-disclosure
	Personalization	[B1] to enrich interpersonal relationships [B2] to provide unique services [B3] to reduce interactional breakdowns	[C1] privacy	[S1] to learn from and about the user [S2] to provide customizable agents [S3] visual elements
Personification	*Identity*	[B1] to increase engagement [B2] to increase human-likeness	[C1] to avoid negative stereotypes [C2] to balance the *identity* and the technical capabilities	[S1] to design and elaborate on a persona
	Personality	[B1] to increase believability [B2] to enrich interpersonal relationships	[C1] to adapt humor to the users' culture [C2] to balance the *personality* traits	[S1] to use appropriate language [S2] to have a sense of humor

▷图 3：社交特性对Chatbot的益处与相应的挑战。图源：论文原文

(1)会话智能

会话智能指的是Chatbot在实现对话目标的技术能力之上,所具备的有效沟通能力。这能让Chatbot积极主动地参与对话,展现它对讨论主题和不断演变的对话语境的认知,同时还能够理解对话的流程。

主动性(Proactivity):当与具有主动性的Chatbot进行对话时,对话会变得更加丰富多样,不再拘泥于简单的一问一答模式。比方说,Chatbot可以自己引入新的话题,提供额外的信息,并提出后续的追问,从而使对话更加有趣,也更有效率。不仅如此,Chatbot还能依靠自己的主动性去吸引和引导用户,避免用户过早对交流失去兴趣。然而,Chatbot也要讲究方式方法,太过主动也可能会让用户感到不适。合适的时机和与话题的相关性十分重要。同时,研究指出,绝不能侵犯用户隐私,也不能让用户有被操控的感觉。

责任感(Conscientiousness):不同于人类,Chatbot的责任在于维护对话的主题和上下文。一个负责任的Chatbot能够确保对话流程连贯,最终得到有意义的结果。根据Brandtzaeg的实验(参考文献10),有68%的参与者表示他们使用Chatbot主要是为了提升“生产力”。在复杂的行业领域(如金融)中,即便是最简单的问题也需要了解一定的背景知识,即“上下文”。显然,能够记住这些上下文并正确使用的Chatbot是提高生产力的前提保障。然而,目前的Chatbot在处理过于复杂的问题和长篇上下文方面仍存在局限。此外,Chatbot对于上下文的理解与用户表达的内容之间可能存在歧义。由此,如何在对话中保持统一性也是一项重要的挑战。

可交流性(Communicability):Prates认为(参考文献1(1),“可交流性是交互式软件的本质,因为用户是通过与系统交换信息来实现目标的。”在Chatbot的讨论范畴中,可交流性是指需要向用户传达自己的功能,也就是让用户知道应该如何使用自己。与传统软件不同,Chatbot不再通过按钮、菜单、链接或新手指南来呈现自身功能,而是在与用户的多轮交流对话中逐步展示其能力。除了明确功能,可交流性还有助于管理用户期望。毕竟,如果用户期望超出了Chatbot的能力范围,“期望越高,失望越大”。

总之,会话智能的社交特性能让Chatbot在交互中扮演主动地提供关注和信息的角色。这类特性可以帮助Chatbot管理对话,让它变得高效、有趣和灵巧。为了实现这一目标,产品的设计研发人员需要关注提供信息的时机与相关性、隐私、交互的灵活性以及一致性等问题。

(2)社交智能

社交智能指的是个体为了实现预期目标而产生适当社交行为的能力。在人机交互领域,人们对待计算机就如同对待社交角色一样。因此,在Chatbot和用户的互动中,有必要让它遵守人类社交的交往规则。

伤害控制(Damage control):多项实验表明,人们在对待Chatbot时的态度与对待真实人类时存在差异,更容易采取骚扰行为、使用非正常输入进行压力测试,并且更容易对结果表现出不满。Lasek和Jessa在分析酒店Chatbot对话记录时发现,辱骂性言论占比达到了2.8%;而在Amazon的对话中,有4%的句子含有露骨的性暗示(参考文献1(2)。因此,伤害控制成为了Chatbot不可或缺的社交特性。这种特性使Chatbot能够以适当的方式应对骚扰和测试行为。当然,前提是Chatbot必须能够识别出敏感词汇,同时也需要识别超出其能力范围的请求,并提供符合对话上下文的回应。

连贯性(Thoroughness):有趣的是,Chatbot的语言风格不仅会影响到用户的使用体验,还会影响到用户的回答风格——用户在交流时会更倾向于和Chatbot使用的语言风格保持一致。一致的语言风格和精准的语言使用,能够显著增加Chatbot的可信程度,使其从“acting as a machine(像机器一样工作)”变为“more as a person(与真人更相似)”。可惜,尽管有13篇研究强调了语言使用和语言风格对用户体验的重要性,但目前尚无工作提出具体的策略来支持Chatbot的连贯性。

礼貌(Manners):尽管不同地区和性格的人对礼貌的界定存在差异,但通常情况下,礼貌可以营造更融洽的对话氛围。Chatbot也可以通过使用礼貌的词汇,例如问候、道歉或结束语,来维持对话的和谐,减少不愉快对话带来的负面体验。然而,即便对于人类而言,维护面子也是一项挑战,更不用说Chatbot了。让Chatbot识别潜在的丢脸行为(Face-Threatening Acts),并以礼貌的方式化解,这一任务更是异常复杂。

道德准则(Moral agency):一个没有道德准则的Chatbot是难以控制的,就像微软在Twitter上推出的Tay一样,在短时间内就展现出了种族歧视和性别歧视等不道德行为。因此,Chatbot需要能够根据社会的道德观念来采取行动。限制Chatbot具备道德准则的主要痛点是训练数据和训练方法当中的偏见。开发者给Chatbot投喂的训练数据通常是“干净”的,没有不道德的语料。然而,这种“干净”的数据反而导致Chatbot对不道德言论的陌生,无法适当地对其做出回应。

情商(Emotional intelligence):具备情商能让Chatbot接受、识别和表达情感,对用户的情感作出回应,并在解决问题时运用情感智能。这种能力主要在一些可能涉及情感披露的领域中发挥作用,例如教育和医疗。有趣的是,实验表明,在与Chatbot交流时,人们更有可能披露自己的负面情绪。参与者表示,有些话题与人类交谈可能会感到尴尬,但如果对方是个机器人,便可以更加随心所欲地倾诉。此外,情商还可以使用户更加信任Chatbot。在一项实验中,用户表示在与治疗师Chatbot交流的体验中,最好的一点是能够感受到同理心。这种感受使他们更愿意称呼Chatbot为“他”、“朋友”或者“一个有趣的小家伙”。

Lee和Choi提出了一种有趣的方法来提升Chatbot的情商。他们认为通过自我披露和展现互惠性的方式可以增加用户对Chatbot的信任度和亲近感(参考文献18)。然而,尽管这个假设颇具潜力,目前还需要进一步的研究来充分验证其有效性。

个性化(Personalization):拥有一个“私人订制”Chatbot对许多人来说是非常有吸引力的。个性化能够使Chatbot根据用户的个人信息和独特需求来定制化调整其功能。想象一下,如果一个导航的Chatbot没有个性化能力,它会为所有用户提供统一的答案,而不考虑用户的个人需求,比如是否喜欢走高速,是否愿意绕开拥堵路段,还是只是想随处兜兜风。然而,实现个性化功能意味着Chatbot必须收集足够的用户个人信息,这可能会带来用户隐私泄露的风险。因此,在实现个性化功能的同时,务必要高度重视用户的隐私保护问题。

总的来说,社交智能这一类社交特性有助于Chatbot解决社交定位问题,同时增强调整能力、可信度、类人性、互动性和人际关系。在实现这一目标的同时,设计师和研究人员也应关注隐私保护、情感调节问题、语言一致性以及故障和不当内容的识别。

(3)拟人化

拟人化是指将人类的特质赋予Chatbot，包括外貌和情感状态。文献表明，经过诱导，用户可以将计算机视为人类，且计算机越类人化，人们的反馈就会愈发社会化。

身份（Identity）：虽然Chatbot无法自主选择自己的身份，但开发人员在定义其对话风格和行为方式时，总会有意无意地赋予它一定的身份特质。一个恰如其分的身份不仅能够增强对话的沉浸感，还能够提升用户的信任度。想象一下，当患者面对“专业医师”和“实习医生”两种Chatbot身份时，前者很可能更容易获得信赖。Chatbot身份的建构可能涵盖多个方面，如性别、年龄、语言风格和姓名，甚至连字体的选择也在其中。有研究表明，用户认为采用手写字体的Chatbot更具人类特质。然而，如何赋予Chatbot一个贴切的身份是颇有挑战的。需要注意的不仅是Chatbot身份与其能力之间的一致性，还需避免身份赋予带来的社会刻板印象，比如黑人身份所带来的与种族主义相关的联想。

人格（Personality）：根据社会学观点，人格是用来预测个人思维、情感和行为的特质，包括外向性、宜人性、尽责性、神经质、开放性、气质和幽默感等。同样，Chatbot也需要具备这些人格特质。拥有一致人格的Chatbot更具可预测性和可信度。喜怒无常和多变的人很少受欢迎，同样，如果Chatbot的态度出现不可预测的波动，也会让用户感到不适。需要注意的是，不同的用户群体对Chatbot的人格有不同偏好。例如，学生更偏好宜人性和外向性较高的导师式Chatbot，以便能在面对挫折时获得鼓励；而一些以娱乐为目的的人可能更希望Chatbot具备幽默感。

以上两个拟人化特性的好处与社交智能很类似，包括增强可信度、类人性、互动性和人际关系。不同之处在于，研究人员应专注于赋予Chatbot可识别的身份和人格特征，并且这些特征可以和用户所期待的Chatbot特质相一致。此外，也要关注对用户文化背景的适应、减少负面刻板印象的影响。

二、如何选择Chatbot的社交特性？

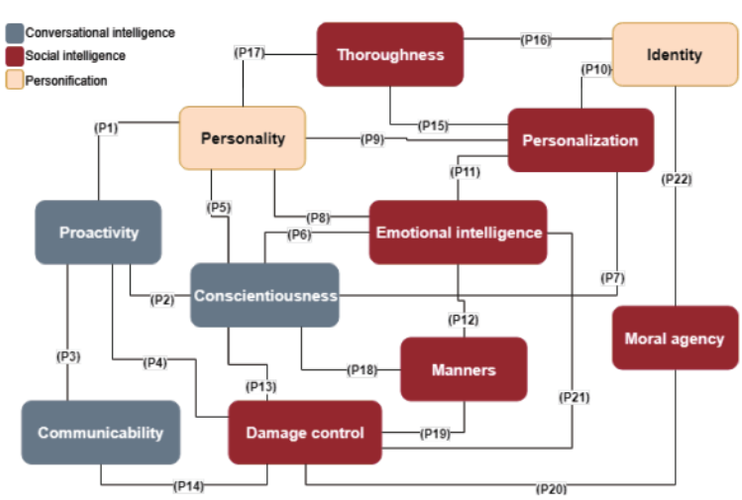
进一步的研究揭示，不同领域对社交特性的需求各具差异。目前，几乎所有的社交特性在开放领域的Chatbot中都得到了应用，唯独“可交流性”例外。这是因为开放领域的Chatbot通常无需明确界定它的具体能力范围，只需模拟人类的交流方式即可。

然而，在特定领域中，Chatbot往往需要更加定制化的社交特性。以教育和客服领域为例，Chatbot在礼貌和情商方面需求更高，但在不同的目标下表现迥异。在教育背景下，这些特性旨在激励学生，因此Chatbot应擅长鼓励和安抚，特别是在学生遇到挫折时。而在客服领域，情商和礼貌更多地用于应对客户对产品或服务的不满情绪，从而提供更优质的服务体验。

Domain	Social Characteristics	Studies
Open domain	*Proactivity, Conscientiousness, Damage control, Thoroughness, Manners, Moral agency, Emotional intelligence, Personalization, Identity, Personality*	(Thies et al., 2017) (Portela & Granell-Canut, 2017) (Shum, He, & Li, 2018) (Morrissey & Kirakowski, 2013) (Curry & Rieser, 2018) (De Angeli, Johnson, & Coventry, 2001) (Hill, Ford, & Farreras, 2015) (Kirakowski, Yiu, et al., 2009) (Mairesse & Walker, 2009) (De Angeli & Brahnam, 2006) (Banks, 2018) (Brahnam & De Angeli, 2012) (Ho et al., 2018) (De Angeli, 2005) (Corti & Gillespie, 2016) (Ptaszynski et al., 2010)
Ethnography	*Proactivity, Conscientiousness, Thoroughness, Personalization*	(Tallyn et al., 2018)
Task management	*Proactivity, Damage control, Manners, Personalization, Identity*	(V. Q. Liao, Davis, Geyer, Muller, & Shami, 2016) (Toxtli et al., 2018)
Tourism	*Proactivity, Thoroughness, Manners*	(Chaves & Gerosa, 2018)
Business	*Proactivity, Personalization*	(Duijvelshoff, 2017)
Information search	*Proactivity, Damage control, Manners, Emotional intelligence*	(Avula, Chadwick, Arguello, & Capra, 2018) (Wallis & Norling, 2005)
Decision-making	*Proactivity, Damage control, Manners*	(Mäurer & Weihe, 2015)
Health-care	*Proactivity, Emotional intelligence*	(K. K. Fitzpatrick et al., 2017) (Miner et al., 2016)
Credibility assessment	*Proactivity, Conscientiousness*	(Schuetzler, Grimes, & Giboney, 2018)
Education	*Proactivity, Conscientiousness, Damage control, Thoroughness, Manners, Emotional intelligence, Identity, Personality*	(Ayedoun, Hayashi, & Seta, 2017) (Coniam, 2008) (Dyke, Howley, Adamson, Kumar, & Rosé, 2013) (Hayashi, 2015) (Kumar et al., 2010) (Silvervarg & Jönsson, 2013) (Sjödén, Silvervarg, Haake, & Gulz, 2011) (Tamayo-Moreno & Pérez-Marín, 2016) (Tegos et al., 2016)
Financial services	*Conscientiousness, Communicability, Damage control, Thoroughness, Personalization, Identity*	(Candello, Pinhanez, & Figueiredo, 2017) (Duijst, 2017)
Customer services	*Conscientiousness, Communicability, Damage control, Thoroughness, Manners, Emotional intelligence, Personalization, Identity*	(Araujo, 2018) (Brandtzaeg & Følstad, 2018) (Gnewuch et al., 2017) (Jenkins, Churchill, Cox, & Smith, 2007) (Lasek & Jessa, 2013)
E-commerce	*Conscientiousness, Manners*	(Jain, Kota, Kumar, & Patel, 2018) (Narita & Kitamura, 2010)
News	*Communicability*	(Valério, Guimarães, Prates, & Candello, 2017)
Human resources	*Communicability, Damage control, Manners, Identity*	(Q. V. Liao et al., 2018)
Virtual assistant	*Thoroughness, Emotional intelligence, Personalization, Identity*	(Ciechanowski et al., 2018) (Zamora, 2017)
Gaming	*Thoroughness, Emotional intelligence, Personality*	(Dohsaka et al., 2014) (Morris, 2002)
Race-talk	*Moral agency, Identity*	(Marino, 2014) (Schlesinger et al., 2018)
Humorous talk	*Personality*	(Meany & Clark, 2010)
Not defined	*Proactivity, Conscientiousness, Communicability, Damage control, Personalization, Identity, Personality*	(Brandtzaeg & Følstad, 2017) (Jain, Kumar, et al., 2018) (Neururer et al., 2018)

▷图 4:研究者在不同领域中发现的社交特性。图源:论文原文

此外,不同社交特性之间的关系也值得关注。研究者们发现,社交特性之间可以相互影响、相互促进。比如,当责任感和情商同时出现时,因为保留了先前对话的信息并能够回忆它们,Chatbot能够展现出更强的同理心。责任感也能支持更好的个性化特性,因为Chatbot能够记住不同交互会话中的个人偏好。研究者用一个理论框架总结了社交特性之间的22种关系命题。



▷图 5:社交特性之间的相互关系。图源:论文原文

三、总结

Chatbot的发展已经不再仅仅依赖于计算机科学，还需要社会科学的有力支持。

这项研究通过对相关工作的回顾性调研，总结出了11种Chatbot所需的社交特性，并提出了社交特性之间的22种相互关系。这些特性或许并不全面，也没有定量分析，但已经涵盖了许多人际交流的关键要素。这些结果为设计师和研究人员在推动人机交互领域发展方面提供了重要的参考和机遇。

研究者认为，“社会科学，尤其是社会语言学和传播学，在Chatbot的设计中扮演着重要角色。”文中研究者指出了一些可指导这些研究的社会学理论，如合作原则、社会认同、个性化、礼貌以及心智感知理论。具体而言，责任感、连贯性和主动性可能属于合作原则的范畴，而社会认同理论可能有助于解释人机对话中的身份、人格和道德问题。这些理论为Chatbot的开发提供了有益的指导，让设计师和研究人员能够更好地塑造Chatbot的社交特性。

然而，需要注意的是，要让Chatbot具备恰当的社交特性并不是一项容易的任务。在确保Chatbot的社交特性与其功能一致的同时，还需要避免潜在的偏见和社会刻板印象。这需要我们在技术与伦理之间找到平衡，以创造出能够为用户提供积极体验的Chatbot。

综上所述，Chatbot的未来之路在于融合计算机科学与社会科学，让技术更加人性化、智能化，为用户提供更优质、亲近的交互体验。通过深入挖掘社会科学，我们能够更准确地满足人们多样化的需求，使Chatbot的发展更加贴近人心，为人类创造出更加有益的数字伙伴。(编辑：韵珂)

参考文献

关联论文：Chaves, Ana Paula, and Marco Aurelio Gerosa. "How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design." International Journal of Human–Computer Interaction 37.8 (2021): 729-758.

[1] Chaves A P, Gerosa M A. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design[J]. International Journal of Human–Computer Interaction, 2021, 37(8): 729-758.

[2] Neururer M, Schlögl S, Brinkschulte L, et al. Perceptions on authenticity in chat bots[J]. Multimodal Technologies and Interaction, 2018, 2(3): 60.

[3] Jain M, Kumar P, Kota R, et al. Evaluating and informing the design of chatbots[C]//Proceedings of the 2018 designing interactive systems conference. 2018: 895-906.

[4] Liao Q V, Mas-ud Hussain M, Chandar P, et al. All work and no play?[C]//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018: 1-13.

[5] Forlizzi, J., Zimmerman, J., Mancuso, V., & Kwak, S. (2007). How interface agents affect interaction between humans and computers. In Proceedings of the 2007 conference on designing pleasurable products and interfaces (pp. 209–221).

[6] Lee, S., & Choi, J. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. International Journal of Human- Computer Studies, 103 , 95–105.

[7] Thies, I. M., Menon, N., Magapu, S., Subramony, M., & O' neill, J. (2017). How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians. In B. R., D. G., J. A., K. B. D., O. J., & W. M. (Eds.), Human-computer interaction–interact (Vol. 10513, pp. 441–459). Cham, Switzerland: Springer

[8] Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2018). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. Future Generation Computer Systems, 92 , 539–548.

[9] Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In International conference on information systems 2017, proceedings 1. South Korea: Association for Information Systems.

[10] Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In 4th international conference on internet science (pp. 377–392). Cham: Springer International Publishing.

[11] Prates, R. O., de Souza, C. S., & Barbosa, S. D. (2000). Methods and tools: a method for

evaluating the communicability of user interfaces. interactions, 7 (1), 31–38.

[12] Lasek, M., & Jessa, S. (2013). Chatbots for Customer Service on Hotels' Websites. Information Systems in Management, 2 (2), 146–158.

[13] Luger, E., & Sellen, A. (2016). Like having a really bad pa: The gulf between user expectation and experience of conversational agents. In Proceedings of the 2016 chi conference on human factors in computing systems (pp. 5286–5297). New York, NY, USA: ACM.

[14] Curry, A. C., & Rieser, V. (2018). # metoo alexa: How conversational systems respond to sexual harassment. In Proceedings of the second acl workshop on ethics in natural language processing (pp. 7–14). New Orleans, Louisiana, USA: Association for Computational Linguistics.

[15] Miner, A., Chow, A., Adler, S., Zaitsev, I., Tero, P., Darcy, A., & Paepcke, A. (2016). Conversational agents and mental health: Theory-informed assessment of language and affect. In Proceedings of the fourth international conference on human agent interaction (pp. 123–130). New York, NY, USA: ACM.

[16] Zamora, J. (2017). I' m sorry, dave, i' m afraid i can' t do that: Chatbot perception and expectations. In Proceedings of the 5th international conference on human agent interaction (pp. 253–260). New York, NY, USA: ACM.

[17] Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR mental health, 4 (2), online.

[18] Lee, S., & Choi, J. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. International Journal of Human-Computer Studies, 103 , 95–105.

[19] Nass, C., Steuer, J., Tauber, E., & Reeder, H. (1993). Anthropomorphism, agency, and ethopoeia: computers as social actors. In Interact' 93 and chi' 93 conference companion on human factors in computing systems (pp. 111–112).

[20] Gong, L. (2008). How social is social responses to computers? the function of the degree of anthropomorphism in computer representations. Computers in Human Behavior, 24 (4), 1494–1509.

[21] Candello, H., Pinhanez, C., & Figueiredo, F. (2017). Typefaces and the perception of humanness in natural language chatbots. In Proceedings of the 2017 chi conference on human factors in computing systems (pp. 3476–3487). New York, NY, USA: ACM.

如果AI系统具备了意识，我们将如何知晓？

作者：轻盈

复旦大学博士生在读，计算&进化神经生物学方向。视科研和科普为人生的两大志业。想做有趣有意义的科学研究，也想把收获到的知识和乐趣分享给世人。

扫码查看原文

美国著名科幻作家艾萨克·阿西莫夫（Isaac Asimov）曾在其机器人系列的作品中，精彩地描述过许多拥有自主意识的机器人，譬如文章“Little Lost Robot”中，那个做错事后，把自己伪装起来，躲避人类惩罚的机器人NR-2，亦或是“Liar!”中那个困惑于人类的真实情感为什么总是与表现出来的不一样的机器人DV-5。

随着人工智能（AI）的迅猛发展，“机器人有可能具有意识”这种论断不再是天方夜谭，甚至已经被AI领域的部分领军者们承认。去年，OpenAI公司的首席科学家伊利亚·苏茨克弗（Ilya Sutskever）在推特上曾表示，一些最尖端的AI系统可能会“略微有意识”。诚然，目前许多研究人员认为，AI系统尚未具备意识。但AI迅猛的发展速度也让我们迫切需要思考：如果AI系统具备了意识，我们将如何知晓？ 遗憾的是，此前，科学界尚缺乏关于AI意识的详细、科学且深刻的讨论。而为了解决这一窘境，包括神经科学家、哲学家和计算机科学家在内的19名专家们制定了一份AI意识检查清单。这份清单指出，如果某AI系统符合条件，将表明其有很高的可能性具有意识。这份报告已在arXiv以预印本形式发布，追问将报告内容归纳总结如下，以飨读者。

一、什么是意识？

AI的意识研究面临的挑战之一就是定义什么是意识。这篇报告的研究人员主要将研究重心集中在了“现象意识（phenomenal consciousness）”上，即作为人、动物或AI系统正在经历某种感觉，某种有意识的体验。*

*心灵哲学中将意识分为三级，访问意识（access consciousness）、现象意识（phenomenal consciousness）、自我意识（self-consciousness）。后面层级的意识往往包含前者。

举个例子，如果你正在屏幕前阅读本文，那么你正在对屏幕进行有意识的视觉体验。我们还可以有听觉

体验，比如听到鸟鸣，以及其他感觉模式的有意识体验，比如身体感觉——疼痛和痒。除了上述对真实的、当前事件的体验外，我们还可以拥有有意识的意象经历，比如此时此刻想象爱人的面孔。

当然，从另一个角度讲，大脑中有许多完全无意识的信息处理过程，比如大脑在没有任何有意识感知的情况下调节激素的释放。另一个例子是感觉记忆存储：你可能会记住一幅地铁上的广告画面、一个陌生人的声音或路边一缕转瞬而逝的桂花香气。

二、意识理论及意识指标

目前，神经科学领域有许多关于意识生物学基础的理论，但尚无共识哪一个是“正确”的。研究者使用多种意识理论，来创建研究框架。他们设想，如果一个AI系统的功能方式与这些理论的许多方面匹配，那么它有更大的可能具有意识。

为了制定合适的意识指标，他们假设，无论系统是由神经元、计算机芯片抑或其他东西构成，意识都与系统如何处理信息相关。他们还假设，基于神经科学的意识理论可以应用于AI。在这些假设的基础上，他们主要根据6种意识理论提取判别意识的指标。

（1）循环处理理论（recurrent processing theory, RPT）

循环信息处理：根据这一理论，大脑中的信息处理不是线性的，而是通过多次迭代的处理来进行的。信息在不同的神经网络之间循环传递和反馈，允许信息被反复处理和整合。

分层次处理：信息处理被认为是分层次的，信息在神经网络的不同层次之间进行传递和整合。底层的神经网络处理基本的感觉信息，而高层的网络则处理更抽象和复杂的概念。

意识内容的生成：这一理论认为，意识的内容是通过在神经网络中循环处理和整合的结果产生的。在这个过程中，信息逐渐被整合成有意义的模式，从而产生了我们所经历的感知和思维。

反馈机制：反馈机制在意识形成的过程中起到关键作用。信息可以从高层网络返回到低层网络，以便根据更高级别的信息来调整和修正较低级别的处理。这种反馈机制有助于产生一种连贯的、可理解的意识体验。

如果将大脑比作生产汽车的工厂，汽车生产线比喻成视觉处理系统，而汽车则代表意识。那么，RPT认为，意识是通过信息在不同模块之间的反复循环传递和处理而产生的。

当汽车生产线的工作站（视觉处理系统的各个部分）完成了各自的任务时，它们会将零件（信息）传递给其他工作站，并接收来自其他工作站的零件。不同工作站之间协同工作，并根据需要进行修正和调整。在这个过程中，工厂中的零件在不同工作站之间不断传递和加工，最终形成了一个完整的汽车（意识）。

（2）全局工作空间理论（global workplace theory, GWT）

全局工作空间概念：全局工作空间是大脑中的一个特定区域，类似于心智“舞台”，用于整合和传播信息。这个工作空间容量有限，只能容纳有限数量的信息或思维内容。

信息竞争：在大脑中，不同的认知过程和信息流竞争着进入全局工作空间。只有那些被选择的信息才能够进入工作空间，变得可意识化。

意识内容的生成：在全局工作空间中，信息被整合成一种有意义的形式，从而形成了我们的意识体验。这意味着我们只有对特定信息进行处理并在全局工作空间中传播时，才能有意识地体验到它们。

注意的作用：全局工作空间理论也强调了注意的重要性。注意充当了信息进入全局工作空间的门卫，只

有受到注意的信息才能够进入工作空间。因此，意识与注意紧密相关。

GWT认为，意识是通过一种全局的信息传播和共享过程产生的。在大脑这个工厂中，有一个特殊的工作站被称为“全局工作空间”。这个工作站负责协调和传播零件（信息），以便不同的工作站之间能够协同工作。当某个工作站完成了一项任务，它将零件发送到全局工作空间，然后其他工作站可以访问这个零件并做出相应的调整。只有信息被发送到全局工作空间时，才会导致意识的产生。

（3）高阶理论（higher order theory，HOT）

第一阶段和第二阶段：高阶理论将心理状态分为两个阶段。第一阶段是指我们通常称之为感知、思考或情感的心理状态，例如看到一只猫、感受到疼痛。第二阶段是元认知或高阶认知阶段，指的是对第一阶段状态的认知、监控或觉察。第二阶段状态可以是关于第一阶段状态的认知，例如“我知道我看到了一只猫”。

意识的产生：高阶理论认为，一个心理状态具有意识，是因为它被某种高阶认知状态监测或觉察。具体来说，当第二阶段的认知状态与特定的第一阶段状态相关联时，意识就会产生。这意味着一个心理状态之所以具有意识，是因为我们知道或觉察到我们正在经历这种状态。

无意识和有意识的区分：高阶理论提供了一种区分无意识和有意识状态的方式。如果一个心理状态没有相关的高阶认知状态监测或觉察，那么它是无意识的。如果有相关的高阶认知状态，那么它是有意识的。

HOT认为，意识是通过高阶的认知过程产生的，这些过程反映了我们对初级感知和知觉的反思。在上述的比喻中，高阶认知则更像是工厂中的观察员或监控系统。在工厂中，有一个特殊的摄像头，它监视各个工作站的活动。每当某个工作站完成一项任务时，摄像头会记录下来，并生成一份报告，描述工作站的活动。这份报告可以看作是关于工作站的高阶信息，即意识。只有当有一个系统能够观察和记录下其他处理信息的系统的活动时，意识才会出现。

除此之外，研究者们还参考了注意图式理论（attention schema theory，AST）、预测处理理论（predictive processing，PP）、代理与具象化理论（agency and embodiment，AE）。它们共同构成了意识科学当前的一个参考标准，用于评估特定AI系统中出现意识的可能性。

根据上述意识理论的主要内容，研究者将其主要的概念提取出来，对应到计算机系统中的各项性能指标，用于评估AI系统意识产生的可能性。指标总结如下（表（1）。

表 1：意识指标及其含义

循环处理理论	
RPT-1：使用算法递归的方式输入模块	RPT-1 和 RPT-2 是独立的指标；
RPT-2：生成有组织的、整合的感知表征的输入模块	
全局工作空间理论	
GWT-1：并行运行的多个独立的模块	在 GWT 理论中，上述指标是意识形成的充分必要条件，GWT-1 至 GWT-4 指标之间相互关联。GWT-3 和 GWT-4 需要 RPT-1 条件的成立。
GWT-2：容量有限的全局工作空间	
GWT-3：全局传播：能够将全局工作空间中的信息传播给所有的模块	
GWT-4：状态依赖的注意机制，在执行复杂的任务时，允许全局工作空间依次检索不同的模块的信息	
高阶理论	
HOT-1：生成式的、自上而下的或噪声的感知模块	感知显示检测理论（Perceptual Reality Monitoring Theory，PRM）认为，这些指标是意识形成的充分必要条件。HOT-1 与 HOT-3 指标之间相互关联。HOT-4 是独立的条件。
HOT-2：元认知监测系统能够区分感知信息和噪声	
HOT-3：由通用的信息，行动选择系统和强烈的倾向性引导的代理，根据元认知系统的输出结果来更新信息	
HOT-4：用稀疏和平滑的特征编码生成“质量空间”	
注意图式理论	
AST-1：呈现与控制当前注意力状态的预测模型	
预测处理理论	
PP-1：使用预测编码的输入模块	前提是需要满足 RPT-1 和 HOT-1 指标；
代理与具象化	
AE-1：代理：从反馈中学习和选择性输出	满足 AE-2 的系统可能也满足 AE-1；
AE-2：具象化：建模输出-输入情境，包括一些系统性影响，并在感知或控制任务中使用该模型	

▷图源：arXiv官网 由追问媒体释义

研究指出，拥有更多这些特征的系统更有可能具有意识。但研究者也建议，在使用这些指标时，应牢记它们与所基于的意识理论以及指标之间的联系——某些指标的组合将比其他指标更能促进AI意识的产生。而其中一些看似必要的条件，在单独存在的情况下并不会显著提高意识形成的可能性。

三、对目前AI系统的意识评估

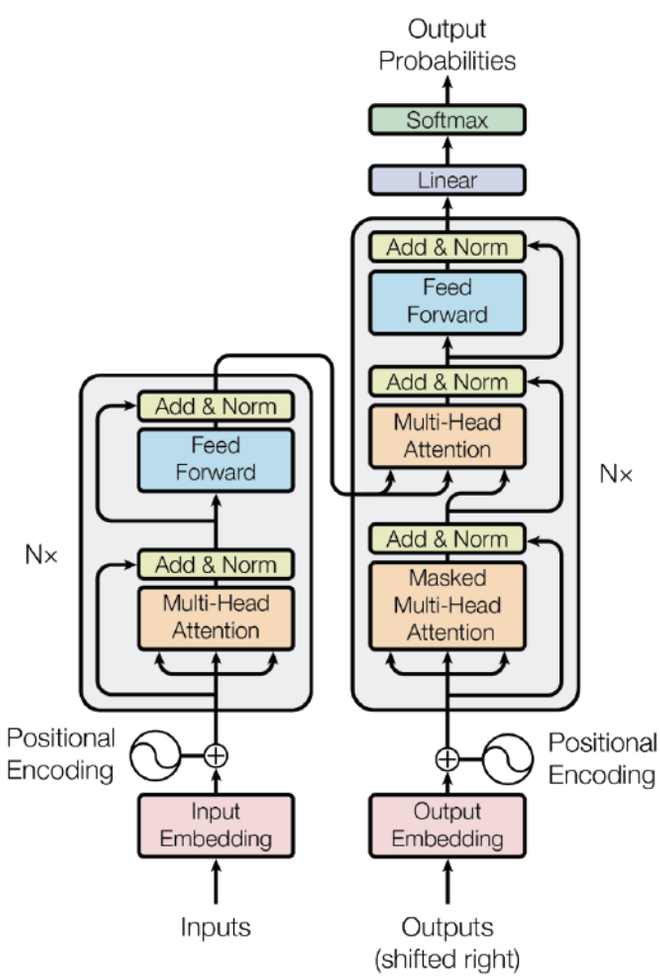
在构建了评价框架后，研究者还通过评估当前的两大AI系统是否可能具备意识，来展示使用这些指标的方法。

在以上所有指标中，研究者重点关注GWT指标（GWT-1至GWT-（4）。而两大AI系统，其一是此前引起公众广泛关注的基于Transformer的大型语言模型（LLM），比如GPT-3、GPT-4和LaMDA；其二是声称具备“全局工作空间”的Perceiver和Perceiver IO系统。

transformer模型主要由两种交替出现的神经网络层组成，即注意力头层和前馈层（如图1所示）。自注意机制（self-attention）是Transformer的核心算法框架之一，在Transformer中用于整合输入序列中不同位置的信息，类似于GWT理论中的全局工作空间概念，它们都整合来自多个模块（注意力头层）的信息。Transformer也被看作由多个“残差块”组成的结构。

每个残差块包括一种类型的神经网络层，这些网络层分别处理从残差块中提取的信息，处理后的信息随后又被添加回残差流。这种结构允许信息在不同的残差块之间传递和处理，有助于捕获复杂的特征和关系。由自注意层和前馈层生成的信息被映射到较低维度的残差流中，因此，残差流可以在某种程度上充当工作空间，用于存储和传递信息。而且残差流中的信息可以在不受位置限制的情况下被下游的注意力头层使用。

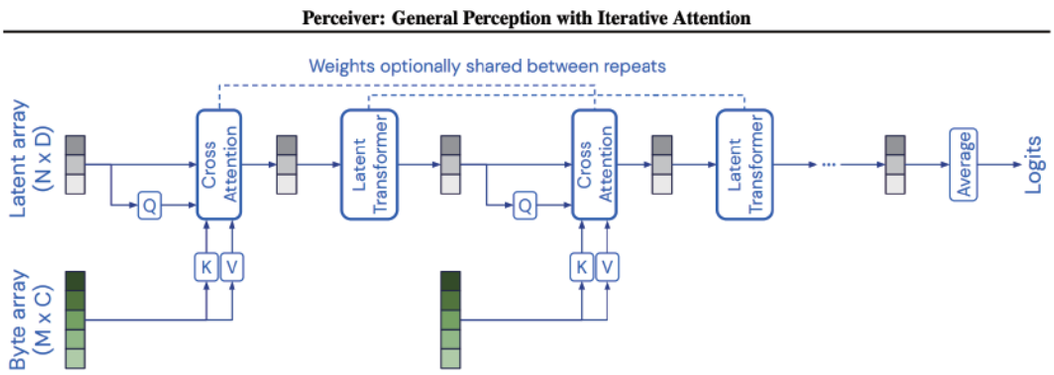
考虑到上述Transformer的算法框架，研究人员认为Transformer具有GWT-1到GWT-3的意识指标，即并行的模块、容量有限的全局工作空间和全局传播。但目前关于Transformer是否满足GWT-4的状态依赖的注意力机制则仍存在争议。



▷图 1：Transformer的基本算法框架。图源：论文原文

而与Transformers相比，Perceiver算法框架似乎更能满足GWT指标，但在研究者看来其仍然未能完全满足所有指标。

目前存在两个版本的Perceiver算法框架。Perceiver算法框架旨在解决Transformer的一个难题，即自注意机制在处理高维输入时的计算成本较高。而Perceiver IO算法架构则更旨在处理多模态的输入，并使用多个输入编码器和输出解码器，生成各种类型的输出。它使用自注意机制来处理潜在空间中的信息，并使用交叉注意力机制（cross-attention）来从输入模块中选择信息并写入输出模块。潜在空间在自注意层和交叉注意层之间交替，使其能够反复地从输入模块中获取新信息（如图2所示）。



▷图 2：Perceiver的基本算法框架。图源：论文原文

并且，Perceiver算法框架允许按时间顺序处理输入序列，潜在空间状态随着每个新输入的更新而更新，但也受到其先前状态的影响。因此，可以说Perceiver算法框架满足意识指标GWT-1（模块）和GWT-2（容量有限的全局工作空间），以及GWT-4（状态依赖的注意力机制）的部分特征。但是，Perceiver算法框架尚不满足GWT-3（全局传播）指标。Perceiver IO具有多个输出模块，但在任何给定的试验中，其输入变量仅包括一个指定输出类型的“输出查询”，这将使得最终只有一个输出模块可以处理来自全局工作空间的信息。另外，输入模块通常不会从全局工作空间接收信息。因此，虽然Perceiver算法框架是目前作为使用类似工作空间的方法来改善AI功能的成功示例，但它距离完全实现GWT尚有一段距离。

总的来说，虽然评估结果表明现有的AI距离有意识系统尚远，但不可否认它们具有与全局工作空间理论相关的一些意识指标。

四、关于AI系统意识的思考

（1）低估和高估人工智能的意识之后

关于AI意识的评估不可避免地会面临“低估”和“高估”的风险。如果我们低估了AI的意识，那么可能会导致一些道德困境。举例来说，人类受诱于巨大的经济利益，大量捕杀“无意识”的动物。就AI而言，同样出于巨大的经济效益，研究人员可能会淡化他们的伦理顾虑。因此，如果未来，我们能够构建出有意识地感受到痛苦的AI，那么只有当研究人员能清楚地认识到AI可能会感受到痛苦时，才有可能保护它们免受痛苦。当然，在某种程度上，有意识并不等同于能够有意识地体验痛苦。至少在概念上，可能存在没有情感体验的意识系统。

另外，我们也很有可能高估了AI的意识。人类倾向于将人类的心理状态强加到非人类系统。这种拟人化行为使我们能够使用与理解同类相同的认知框架来理解和预测像AI这样复杂的非人类系统，这有可能帮助我们更好地与AI交互。然而，另一方面，“子非鱼，焉知鱼之乐？”，AI也许并非如我们想象的那般具备人类意识。

(2)人工智能的意识与能力也许并不直接相关

在大众的想象中,意识与自由意志、智力和感受人类情感的倾向联系在一起,包括同理心、爱、内疚、愤怒和嫉妒。因此,我们认为有意识的人工智能的实现,可能意味着我们很快就会拥有类似于科幻小说中描绘的类人AI。事实上,这一畅想是否成立,很大程度上取决于意识和能力之间的关系。虽然,一些有影响力的AI研究人员目前正在构建更有可能有意识的系统来强化AI的能力,但事实上,拥有意识并不一定能很好地增强AI的能力。并且,当前许多对AI负面影响的担忧并未涉及AI是否有意识。例如,人们担心根据反映当前社会结构的数据训练出来的AI系统可能会延续或加剧不公正现象,但这并不取决于AI是否有意识。以及,人们担心AI可能会取代大多数人类工人,这也不取决于AI是否有意识(尽管取代人类工人的经济价值可能会激发能力研究,从而导致有意识的AI出现)。

最后,研究者也表明这份报告并不是AI意识研究的最终定论。他们希望这份报告能够促进日后AI意识领域进一步的研究探讨。(编辑:Lixia)

参考文献

关联论文:Butlin, Patrick, et al. "Consciousness in artificial intelligence: insights from the science of consciousness." arXiv preprint arXiv:2308.08708 (2023).

AI的理解困境：如何走出数据世界，触达生命的理解？

作者：雷沐春

中国人民大学外国哲学专业在读，关注认知科学哲学与知觉哲学，尤其是精神疾病、自我与无我的关系。希望通过哲学理解自我、心智与生命，将自我锚定在广阔的世界之中。

扫码查看原文

当前，生成式AI正席卷整个社会，大语言模型（LLMs）在文本（ChatGPT）和图像（DALL-E）生成方面取得了令人惊叹的成就，仅仅依赖零星几个提示词，它们就能生成超出预期的内容。

以大语言模型为代表的生成式AI取得的进步促使我们思考：ChatGPT真的能够理解它们在“谈论”的东西吗？抑或只是塞尔“中文屋”*的一个实例？它能“捕捉”外在现实吗？或仅仅是自然语言数据催生的拟合现象（mimic）？更深层地，生成式AI是通向人工理解（artificial understanding）的正确道路吗？除复制数据外，它是否还能理解词语、感知和行为的“意义”？或者它是否仅仅是一种自我限制的方法的终结？

*中文屋（Chinese Room）：由美国哲学教授约翰·塞尔提出的一个思想实验，借以反驳强人工智能的观点。根据强人工智能的观点，只要计算机拥有适当的程序，理论上就可以说计算机拥有它的认知状态并且可以像人一样进行理解活动。但中文屋指出，计算机即使可以回答用人类语言提出的问题，但它也无法建立人类语言的语义关系，无法理解人类语言。它只会根据规则，机械摆弄符号。John R. Searle. MINDS, BRAINS, AND PROGRAMS. [2014-07-23].

在《生成意义：主动推理与被动AI的范围与限制》一文中，Giovanni Pezzulo、Thomas Parr、Paul Cisek、Andy Clark和Karl Friston试图通过比较生命有机体的主动推理模型（active inference）与AI的被动生成模型，指明“理解”的真正基础，并思考生成式AI是否能够获得理解能力。

一、生成式AI的局限

（1）生物系统与主动推理

许多哲学家（如Andy Clark、Merleau Ponty）、心理学家（如James Gibson、Lawrence Barsalou）和神经科学家已经达成共识：大脑的基本功能并非积累知识，而是控制与世界的信息和能量交换。更重要的是，特定的相互作用以特定的方式稳定地改变事物的状态（例如，进食能够减少饥饿，逃离捕食者能够减少危险等）。所以，重要的不是知识的真实性，而是与世界交互形成的稳定性。

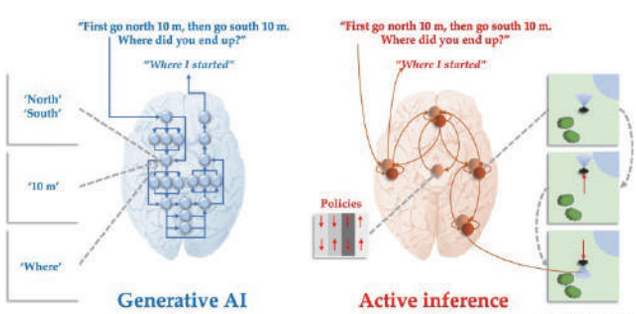
所以，在这种互动中，世界的某些特定特征对我们尤为重要，因为它们决定了我们的行动方式。吉布森将这类特征称作可供性*，即环境提供的行动可能性。生物系统通常以感觉运动（sensorimotor）来响应这些可供性。例如，平坦的地面可以用来支撑，用来坐，也可以用来放东西。

*可供性（affordance），afford一词的名词形式，Gibson在《视觉的生态学进路》（The Ecological Approach to Visual）一书中首次对这一概念做出系统性的阐述。可供性是环境提供给生物的行动可能性，它可能是好的，也可能是坏的。可供性既非客观性质，也不是主观性质，而是生物与环境互动的产物。

此外，生物系统的另一特征是在与世界进行互动之前，它们能够基于已掌握的关于动态世界的知识，做出行动预测。这种预测是主动推理（active inference）的基石。简单来说，主动推理认为，生命有机体的感觉行为根本上是预测性的，而非随机被动触发，它建立在能够提供可供性的世界模型之上。

（2）两种生成模型

生成式AI与主动推理有一个共同的承诺：它们都强调基于生成模型的预测。不过，虽然都是以生成模型为基础（图2），但它们的运作机制并不相同。



▷图2：生成式AI和主动推理的生成模型。图源：关联论文。

在主动推理中，生成模型不仅仅用于预测，而且是能动性（agency）的担保。它们会对外在或内在世界的目标导向、决策和计划进行推理。在非活动状态（offline），例如在反省或睡眠中，主动推理的生成模型也会模拟过去的反事实场景（即“如果过去不是怎样便会如何”的推理）和可能的未来，以此优化生成模型，从而产生行为策略。

与此相反，生成式AI则是基于深度网络，通过自监督学习从信息中建构生成模型。以大语言模型为例，他们在推测一个语句中的下一个词语时，通常使用的是自回归模型和变换器模型（transfer architecture）。经过大规模的样本训练，大语言模型就能用灵活的预测生成全新的内容。并且，它还擅长一些下行任务（如总结文本和回答问题），并能用细粒化的特定领域的数据集解决更多任务（如写科幻小说）。

这两种生成模型的关键区别在于，主动推理所做出的回应是有意义的，而这种意义基于感觉运动经验。例如，回应“向北”或“向南”的问题会与在物理空间中特定的行动可能性关联起来，神经加工的多感官状态和情感状态也会参与其中。尽管人工系统能够通过训练习得空间转译（spatial translations）的统计学规律，但空间转译对于能够在空间中移动的生物和没有移动能力的人工系统来说，意义大相径庭。对于前者，空间转译关乎行动的可能性以及对世界的因果理解。

二、生命有机体的意义理解

成功的生成模型能够从数据中提炼出“潜在变量”，这些变量有助于解释和预测。生成式AI能够用潜在

变量反应统计学规律，以超越训练数据的界限；生命有机体提炼潜在变量的目的可能是更好地预测世界状态。尽管它们都能提取潜在变量，但主动推理与生成式AI的处理方式不同。主动推理的生成模型涉及理解，并将潜在变量作为概念形成的基础。

对于人类和其他生物来说，与世界的互动是在探索世界的特定性质。一个桌子不仅是以木材为原料，由桌腿、桌面构成的物体，而是能够承载盘子，能够坐人，能够在地震中作为庇护所的可供性的集合，这些可供性就是桌子的潜在变量。“桌子”一词仅仅是一个象征符号，或一个简称。具体来说，“桌子”是“能放东西的、能坐的和能隐藏于其下的那个对象”。因此，桌子这个概念实际上是与行动结果相关联的潜在变量的集合（constellation）。生命有机体通过感觉运动经验来了解对象。而轻重、大小等抽象概念，则以这些多感官提供的信息为基础发展而来。

语言能力也是以感官模块为基础，在互动（即沟通）中发展起来的。从具身的角度来看，沟通就是一种感觉运动互动。沟通的意义不在于语音和语法，而在于由沟通所预测的社交互动。人类的语言交流虽然将抽象化发展到了极致，但仍然以互动和控制为基础。语词是有意义的互动的缩写，是在互动中被约定下来的。我们也是在与同类的互动中习得语言符号的意义。当前以语言习得为基础的认知机器，就要在目标导向行动的背景下开发语言和符号能力。而大语言模型和其他生成式AI只是从大量多感官模块的文本数据中被动地学习。

简而言之，我们对语言符号的理解源自于与活生生的世界的互动，而非单纯的对自然语言的运用。生成式AI所具有的潜在变量，或许能够把握关于世界的统计学规律，却略过了它们的形成过程。实际上，生成式AI只是继承了人类沟通所得的语言财产，却不会参与到赋予语词意义的互动过程中。在大语言模型中，只有生产训练文本和转译文本的人才能够理解语词的含义。

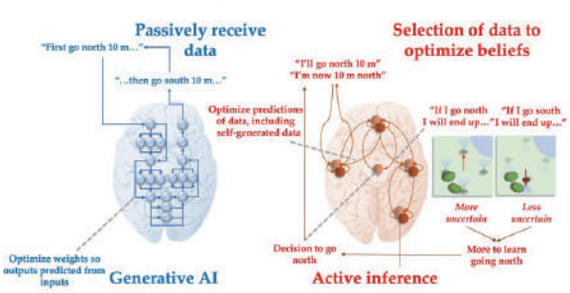
三、基于行动的具身智能

孩子不是习得知识，而是通过经验以及与环境的互动建构知识。

——Maria Montessori

给予生成式AI更多的数据，它们能够获得理解吗？当务之急是要指明理解的真正基础是什么。

实际上，生成式AI习得概念的方式与生命有机体大相径庭（图3）。生命有机体通过与环境的感觉运动互动来学习，这种互动不仅包括了统计规律的掌握，更重要的是，它们是形成知觉和对世界因果关系理解的基础。生命有机体通过感觉运动经验和在环境中的动态移动，习得了对环境的各种表征，如可供性、空间、对象、情境、自我感和能动感等。我们的大脑也编码了与环境的互动和可供性。有研究表明，海马体和内嗅皮层通过路径整合自我移动的信息，发展出空间编码（包括对抽象的概念空间的编码）*。前额叶皮层也包含着探测可供性的空间回路。这种具身智能是发展出抽象的概念思维的基础。



▷图3：生成式AI和生命有机体如何学习生成模型来解决图2的寻路任务。图源：关联论文。

*注：The hippocampus maps concept space, not feature space. J. Neurosci. 2020; 40: 7318-7325

与此不同，当前生成式AI所谓的“理解”并不是以行动为基础，它们只是被动地反映数据的统计学规律，而非呈现关于世界的因果规律。这种方式缺乏对数据的主动选择和训练中的干预，因此无法形成对行动和其结果之间因果关系的理解，也不能区分预测与观察的差别。

生成式AI往往依赖于其模型的复杂性来提高预测准确性，但这种方法也带来了一定的局限性。这些系统在特定任务上表现出色，却难以推广到其他类似任务。这种局限性并不能简单地通过增加数据量来克服。因为理解语境敏感的语言不仅需要大量的数据，更需要能够从数据中提炼出深层的含义和模式。

此外，生成式AI与生物有机体也以不同的方式决定需要关注的信息。生成式AI中变换器模型的注意力机制发挥过滤功能，通过指派不同的权重决定哪些信息是有价值的。而生命有机体的注意力涉及到主动选择，其目的是消除不确定性。

生物体在演化的过程中，面对自然选择的压力，发展出了独特的生成模型。例如，我们的情绪，根植于对某物“对我重要”的感觉，这为我们对世界的理解赋予了意义和目的。在主动推理的过程中，我们利用内感受预测（interoceptive prediction）来引导行动和决策，这种方式使我们能够更好地理解行动的原因和后果。这种内感受、外感受和本体感受的预测共同促进了生命体的生存。因此，与生成型AI不同，生物的主动推理模型自然形成，不需要像AI那样不断地学习细粒化的、繁复的任务。

此外，为了存活下去，生命有机体不能只是消极待命，等待信号来刺激，而要积极主动地与世界进行有目标的互动。这意味着生命有机体的生成模型必须保证在探索新模式和利用旧模式之间审慎权衡，做出灵活的选择。此外，为了更具普适性，这还要求模型不仅要准确，还要节约能量。在生态位（ecological niche）中，这种权衡能够支持不同时间尺度下的行动和知觉。在主动推理中，探索性行为和利用性行为之间的权衡，以及生成模型的效率和准确性之间的权衡，都可以通过最小化自由能来解决。但生成式AI尚未能实现这种情境敏感的、灵活的控制。

最后，从系统发育轨迹来看，生成式AI与主动推理也有本质区别。具有抽象思维和语言能力的生命有机体，能够发展出一种特殊的心智表征方式——我们称之为“分离性表征”（detached representation）。这些表征虽然起源于感觉运动经验，但最终能够从其最初的环境中独立出来，形成自主的独立存在。例如，我们能够在没有直接感知物体的情况下，通过想象和语言讨论它们。

这种独立于直接感觉经验的表征能力是高级认知功能（如规划、想象和讨论抽象或不在场的事物等）的基础。复杂的心智生活需要这种能力，使得我们从直接的、实用的表征转变为语义的描述性表征。这一转变通过复杂的社交互动和对世界的深入参与得以实现，从而拓展了我们对事物的理解和意义的界限。当前的生成式AI走的是一条完全不同的发展路径，它们直接从文本中获得知识。这一过程受到了当前科技，如大型数据集和高效的变换器模型的可用性的驱动。

简而言之，真正的“理解”以能动性理解为基础，建基于有机体通过感觉运动与世界进行的互动，建基于生命有机体对环境的主动探索。更深层次的理解需要分离性表征的能力，即使这种能力仍然基于与世界的互动，但它能够超越当前情境，进行规划、想象和讨论抽象概念。这种理解不仅仅是对统计规律的掌握，而是对世界模型背后的因果结构的深入认识。

四、生成式AI的出路何在？

继续沿着老路扩大生成式AI的规模，是通向真正智能的可取方法吗？

要想使得生成式AI生成意义并拥有理解能力，当前有两种选择。要么坚持原有方法，朝着更加复杂的方向发展。要么转变思路，强调对训练数据的主动选择。

当前的研究大多坚持第一种选择，即提升生成式AI的复杂性来提升其性能。这种复杂性主要体现在模型参数的增加和训练数据量的扩充上。此外，还包括输入信息类型的多样化，以及增添更多的功能和能力，以求实现更高级的AI应用。然而，还有一种潜在的、更为深刻的方法常常被忽视，即让模型通过与世界互动进行主动选择，在获得关于世界的知识的同时，追求内在目标。

当前的大语言模型是以我们对世界的描述为中介来理解现实的。仅仅建立基于文本的大语言模型，然后尝试将之与世界关联起来以获得对世界的理解，可能并不是最有效的途径。更可靠的方式可能是先让AI系统在与现实世界的互动中学习，然后再将这些经验与大型语言模型结合。但这种“互动在先，模型在后”的方法，目前还未被系统地研究过。

五、人工智能，作为人类之镜

生成式AI只能基于被给予的提示词或文本产生结果，而不能像主动推理一样生成原因，如生成计划（planning）。这提供了几个基本的暗示：

首先，真正的计划蕴含着能动性，只有能动者才具备“行动-结果”的生成模型。其次，这意味着主动推理的生成模型并非单纯依赖数据输入，而是需要在世界进行实时的感觉运动互动中获得。也就是说，生成模型以世界模型为基础；“行动-结果”模型能够揭示出世界的因果结构，而信息收集只是隐秘地以统计学规律反映因果结构。

从实践层面来看，生成式AI并不适合作为自动机器人或自动驾驶技术的理想模型。此外，由于生成式AI没有可供性，因此它没有由好奇心驱使的主动学习机制。相较于此，具身智能可能是一个更有效的模型。

尽管生成式AI存在上述种种局限，但它仍然对我们的生态系统产生了深刻影响。它引导我们反思人类的理解过程，寻找世界模型与信息流之间的桥梁。我们人类不断外化我们的思想，创造出全新的对象，这些对象也需要我们去审慎地检视。生成式AI就是一个生动的案例，它揭示出一种未被我们关注的认知自我的建构方式。

可以说，生成式AI就像是21世纪的一面人类之镜，我们在其中照见自己，但遗憾的是，镜子之后却空无一人。（编辑：存源）

参考文献

▷ 关联论文：Generating meaning: active inference and the scope and limits of passive AI, https://doi.org/10.1016/j.tics.2023.10.002。

具身智能何以像人?

作者:LEAF团队

专注于对大语言模型进行深入的评估与研究;致力于解析模型的潜力与限制,紧跟人工智能的最新发展,关注人工通用智能和人工智能生成内容。

扫码查看原文

随着AI技术的进展,ChatGPT等基于大语言模型的聊天机器人已成为我们解决问题的首选。但当我们提出非常私人化、具象化及场景化的问题时,它们给出的答案往往不尽人意。

比如,当询问“我想要学游泳,你能给我一些建议吗?”时,ChatGPT只能提供一些通用的建议,如“呼吸技巧”或“如何让身体浮起来”。这些回答没有针对用户的具体情况,只是泛泛而谈。但假如有一个教练朋友亲自陪你到泳池,向你演示如何在水下屏气,并托住你的腰让你平躺在水面上,告诉你应该如何控制身体以实现浮起,这是否才是你更想要的答案?

这正是“具身智能体”(Embodied Agent)的价值所在,它强调我们不仅要让计算机程序变得聪明,还要让它们像人类一样与真实的物理世界进行紧密的互动。如此,我们才能实现具备人类智能水平,且更接近人类的通用人工智能(AGI)。

一、AI为什么要有具身智能?

为何我们要追求那种与物理世界紧密互动、且更接近人类的人工智能?将人工智能当作好用、便捷的工具难道还不够吗?

这一追求源于人类对智能的根本期待:我们希望它们不仅能高效地执行如学习、问题解决和模式识别等复杂任务,从而帮助人类去做不愿意,或是不擅长做的事;我们还希望它们能理解人类的思维方式、行为习惯、情感表达,甚至性格偏好和心理特点,真正实现“懂你”的高阶能力。更何况,从人性的角度来说,人类本能地会对更自然、更亲近于自己的事物有好感,而对纯粹机械化、缺乏情感的冰冷工具怀有拒斥之心。

1950年,图灵在其论文中首次提出了人工智能的基本概念,并提出了著名的“图灵测试”,用以判断机器是否能模拟人类智能。同年,阿西莫夫在他发表的短篇集《我,机器人》中描绘了一个人与AI共处的未来

世界，并提出了机器人三大定律。因此，自人工智能概念诞生之初，人类就相信并呼唤着一种能以人类语言交流并理解我们的AI——它不仅能在生活中陪伴我们，还受到伦理道德的约束，最终被人类的情感和性格所引导。

这样看来，当我们讨论“智能”时，实际上是期望AI超越单纯的计算机器，成为一种与人类智能匹敌、拥有创造性思维和感知能力的高级生命体。具身智能则代表了这一愿景的实现路径。

二、具身智能何以像人？

这么说来，具身智能应如何实现更像人的AI呢？

我们首先需要理解传统人工智能的局限性。目前的AI系统主要依赖于收集的互联网图像、视频或文本数据进行学习。这些数据集虽然制作精良，但它们终究是静态的，是通过人类整理和数据标注的方式实现的。这使得AI在处理信息时缺乏与环境的交流及互动。AI并不能理解其表达背后真正的逻辑思考路径，更不用说自主反思并自我成长了。因而除依葫芦画瓢外，AI自发制造的数据往往与实际情况不符，常常“胡说八道”。这也是传统AI被称为“弱”智能的主要原因。

为此，一些学者立足于人类婴儿认知的研究，从人类智能的发展过程中得到启示，他们认为，真正的智能来源于与周围环境的不断互动和反馈。正如人类婴儿通过与环境的感知和物理交互，来发展认知能力一样，智能的真正发展需要超越处理抽象信息，深入理解和应对现实世界中的复杂情境。而这正是具身智能概念的出发点。

具体来说，具身智能是一种基于物理身体进行感知和行动的智能系统，其通过智能体与环境的交互获取信息、理解问题、做出决策并实现行动，从而产生智能行为和适应性。斯坦福大学的李飞飞教授曾经指出，“具身的含义不是身体本身，而是与环境交互以及在环境中做事的整体需求和功能。”同样，上海交通大学的卢策吾教授通过猫学习走路的比喻，形象地描述到，“自由行动的猫是具身的智能，它能够在环境中自主行动，从而学会行走的能力；而被动观察世界的猫，最终却失去了行走的能力。”

与基于静态数据集训练的传统AI不同，具身智能能实时地在真实物理世界中学习和交互，从而能更好地模拟人类学习的方式。它们能像人一样，通过与环境的实际互动获取知识和经验，理解人类的实时反馈和行为，进而掌握非语言的沟通方式，如通过表情和触摸来感知和体验人类的情感表达。这种深度的人机交互和理解，使具身智能成为一种更贴近人类认知和情感的智能形态，有望实现更深层次的人机互动和共融。

三、具身智能如何做到更像人？

(1) 主动性

作为具身智能的核心特征之一，主动性赋予了智能系统超越被动信息处理工具的能力，让它们成为积极的参与者。

在Metin Sitti 2021年的论文Physical intelligence as a new paradigm中，他指出，在具身的物理智能层面上，柔性系统可以对环境刺激做出响应……然后根据身体部位与环境条件的自我定位、自我运动和自我感知（本体感觉）得出自我定位，并将其转化为后续行动。这意味着具身智能不仅能感知环境，还能根据感知进行自主的行动。另一篇论文Embodied Intelligence in Physical, Social and Technological Environments同样采用相似的方法定义具身智能：当一个生命在各种感官信息的基础

上，自主地对环境采取行动，在这样做的过程中，能够将自己作为一个多感官的积极主动的自我，从而与环境中正在发生的事情区分开来，并加以调节时，它就拥有了具身智能。

这种主动性可以通过一个简单的比喻来理解：当你走进图书馆，遇到一个传统的管理员时，他或许会根据你的请求给到你想要的答案，如一个书名及对应位置。但如果这位管理员是一个具备具身智能的导览顾问，它不仅能够找到你需要的信息，还会主动引导你，找到书籍，并给你讲解相关知识，带你深入了解整个知识的世界。这种交互方式类似于与一个热情、友好的伙伴一起探索知识，而不仅仅是从一个冷漠的知识助手那里得到答案。具身智能通过主动性，提供了一种全新的交互体验，这不仅能够增强人类对信息的获取和理解，还能加深人类与智能系统之间的情感和认知联系。

尽管目前的具身智能还未完全实现主动性和热情互动，但以视觉导航的快速发展为例，在如iGibson Sim2Real、Habitat和RoboTHOR等挑战赛中，我们已经见证了这一领域初步形态的涌现，这些成果已经超越了仅仅执行任务的冷漠机器。例如，结合人类先验知识的导航系统能够通过将这些知识以多模态输入形式融入到深度强化学习框架中，如知识图谱或音频输入，进而使AI能够在未知环境中学习导航并寻找未见过的物体。

最新的视觉语言导航（VLN）技术致力于创建一种能够通过自然语言与人类交流，并在真实3D环境中自主导航的具身智能。目前，该领域已经利用多个数据集进行研究和开发，如REVERIE、R2R、CVDN、GELA、ALFRED、Talk2Nav、Touchdown等，同时也产生了一些创新的网络架构，如辅助推理导航框架。这些技术应用于机器导航、辅助技术和虚拟助手等领域，尚处于初级阶段。此外，VLN的拓展视觉对话导航，旨在训练AI与人类进行持续的自然语言对话，以辅助导航。在这个领域，研究者们使用了一种跨模态记忆网络（CMN），该网络分别拥有语言和视觉记忆模块，用于记忆和理解与过往导航动作相关的信息，并利用这些信息来作出导航决策

（2）实时性

实时性是具身智能另一个核心特性，它使得智能系统能够在真实世界中及时学习并迅速反馈。具备实时性的具身智能能够在接收到新信息或遇到新环境时立即做出响应。与此相比，传统的人工智能依赖于预训练的数据，在面对实时变化的环境时难以快速反应。

以电视节目为例，观看录播的魔术表演就像是与传统AI的互动：虽然内容精彩，但你只能被动地观看预先录制的内容，不能实时中断或更改节目内容。相比之下，观看现场直播的魔术秀则更类似于与具身智能的交互：你可以实时提出需求，魔术师则根据这些需求在现场即兴表演，就好像在为你个人定制节目一样，你不再是一个被动的观众，而是整个魔术秀的一部分。这种互动方式不仅更加个性化，也更具参与感。

故而，和现场表演的魔术师一样，具身智能能够即时响应人类的需求和环境变化，提供更为贴合实际情况的解决方案，并以更贴近于人际交往的方式与人类互动。这种实时性帮助它更好地融入人类的日常生活，成为一个更加智能和有用的伴侣，而不仅仅是一个执行预设任务的机器。

在论文LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models中，研究团队提出了LLM-Planner方法。这种方法利用大型语言模型的能力，能为具身智能进行少样本规划，并通过物理基础来增强语言模型，从而生成和更新与当前环境相关的计划。其优势在于它能够实时反映和适应环境的变化，为具身智能的决策提供即时的信息和指导。

(3)情境性

除主动与实时之外,具身智能对特定的场景和情境的反馈应该有深入的感知和个性化的理解。

就像人类在与周围环境互动中实时调整自己的行为一样,具身智能应该通过实时学习和反馈,深刻地理解所处的情境,并据此调整其行为。它能够根据上下文和环境的变化灵活地调整回应方式,融入当前的情境中,从而实现更自然和有效的交流。例如,具身智能能够感知用户的情绪变化,并据此提供个性化的体验,增强用户的参与感和满意度。

以旅游规划为例,传统的聊天智能可能仅能提供固定的行程建议,而不管雨雪风霜,甚至有可能在雷暴雨的天气,依然为用户安排露天温泉的行程。具身智能则能够根据用户的个人偏好、当地环境和天气状况等因素提供更加贴合实际的建议。它更像一位熟悉当地情况的私人旅行顾问乃至私人摄影师。它不仅知道你的目的地,还熟知周围的情境,了解环境变化;能够根据你的私人偏好和当地时令,带你去合适的小馆子就餐,并记录下你每个快乐时刻的印记。

目前已经存在大量逼真且公开泛用的3D场景,可以作为具身智能训练的模拟环境。针对具身导航的虚拟环境有iGibson、Habitat、MultiON、BEHAVIOR等;针对具身问答的有ALFRED;关注情景理解、物体状态和任务规划的环境有AI2-THOR、ThreeDWorld、Habitat 2.0等;关注物体操纵的有SAPIEN、RLBench、VLMbench、RFUniverse、ARNOLD等;物体抓取及操纵信息数据集包括GraspNet、SuctionNet、DexGraspNet、GAPartNet等。这些场景比以往研究模拟器所用的环境要真实得多,极大地促进了具身智能在情境性的初步开发。

此外,传感领域的技术进步,也为情境性的具身智能发展提供了可靠保障。例如,PaLM-E团队提出了具体化的语言模型,将真实世界的连续传感器模态直接结合到语言模型中,从而建立单词和感知之间的联系。这种模型的输入是多模态语句,它们将视觉、连续状态估计和文本输入编码交织在了一起。结合预训练的大型语言模型,对这些编码进行端到端训练,可用于多个具体任务,如顺序机器人操作规划、视觉问题解答和图像视频字幕描述,有效地构建了单词和感知之间的联系。

(4)拟生物

较之一般的人工智能,具身智能需要应对复杂的环境,并被要求以更接近人类的认知方式来与现实世界共处,这就使得它体现出了更多的模仿生物的特征。

就像蜜蜂群体协同工作以建造蜂巢,具身智能中的多个智能体能够共同协作,产生超越单个智能体能力的集体效应。这种群体协作不仅超越了单个智能体的能力,还展示了复杂系统中的涌现现象。在这些系统中,个体智能体的简单行为和互动,可能导致整个系统出现复杂的行为模式和结构形态,使得系统能够适应新的环境和任务,而无需依赖预先设定的编程规则。

此外,具身智能系统中的自组织性是其拟生物特性的关键部分。智能体能够根据环境变化和相互作用动态地调整自己的行为和结构,形成更高级别的功能和结构,从而使系统具有更强的鲁棒性和适应性。

具身智能的这些特性在多种应用中得到了体现。有研发团队专门设计了一种水下软体机器人,其灵感来源于细菌的形态。这种生物启发的模块化结构使机器人能够在水下环境中执行多种任务。这种机器人利用其周围的环境(水)、目标的形状以及机器人本身的顺应性,通过少量的控制输入来实现有效的导航和安全交互。这种建模方法和设计不仅展示了具身智能在模仿生物体方面的创新,也体现了它在实际应

用中的多功能性和适应性。

总之，具身智能领域的技术发展呈现出多样化和综合化的趋势，特别是在观察、操纵和导航等方面的进步尤为显著。这些技术的发展不单单针对具身智能的某个特定特性，而是综合了多方面的功能和能力，以实现更高的适应性和灵活性。

通过结合机器人的传感器数据和一般的视觉语言数据进行联合训练，特别是利用大语言模型的强大内在知识，可以帮助具身智能在面对复杂和未知的真实世界环境时，进行有效的动态学习和泛化。例如，LLM-based Agent（基于大语言模型的智能体）以其独特的语言能力为优势，不仅作为与环境交互的工具，还能将基础技能转移到新任务上，从而使机器人能够根据人类的语言指令适应不同的操作环境。

此外，通过嵌入式行动规划，利用高层策略指导低层策略的子目标，从而使低层策略生成适当的行动信号，可以使机器人在执行任务时更加高效和可控。这种策略的应用可以使具身智能在处理复杂任务时更接近人类的决策模式。为了更有效地完成导航和其他复杂任务，具身智能还需要内存缓冲区和总结机制，以便参考历史信息并更好地适应未知环境。

近年来，谷歌公司的Everyday Robot项目SayCan系统，已经将机器人和对话模型结合，完成一个包含16个步骤的长任务；伯克利的LM Nav项目，则用三个大模型（视觉导航模型ViNG、大语言模型GPT-3、视觉语言模型CLIP）教会了机器人在不看地图的情况下按照语言指令到达目的地；上文提到的谷歌与柏林工业大学推出的PaLM-E模型更是在具身智能的多模态理解和执行方面取得了显著的进展。

能够发现，具身智能的技术发展正迈向一个更加综合、灵活且高效的方向。这些技术的融合和发展，不仅提高了智能系统的适应性和实用性，也为未来的智能系统设计和应用开辟了新的路径。随着技术的不断进步，我们可以期待具身智能在更多领域的实际应用和创新突破。

四、人工智能与人类智能的关系

为了深入理解人工智能（AI）和人类智能（Human intelligence，HI）之间的差异，并探索如何缩小这一差距，结合对具身智能特性的考量，Shanda AI Lab LEAF团队提出了五性原则，以对照分析AI的发展方向（在后续的“智能渐近线”系列报告中，我们会不断扩充五性的内容）。这些原则不仅与具身智能的四大特性相互呼应，还深入探讨了AI发展的关键方面，以期望使AI更接近于人类智能的复杂性和适应性。

（1）逻辑性

AI应具备类似于人类大脑的逻辑思考和理解能力。具体来说，就是AI能够在复杂的社交场景中，结合已有的各种知识储备进行综合运算及推理，理解语义及语义背后的复杂内涵，最终给出相应的输出。

（2）感知力

AI需要具有强大的感知能力，能识别和关联多种信号，并能进行类似于人类的想象和通感。它不仅能够理解聊天输入，同时也能处理多种类型的输入信息；能够像人一样，快速地对周围环境的变化和各种刺激做出迅速的反应。

（3）实时性

AI系统可以做到信息的实时更新、随时取用、随环境而反馈；它可以学习人类的记忆模块的能力，通过上下文学习和情境学习等方式，从有限的实时信息中进行类比学习，理解新的任务。

（4）主动性

AI能够靠积极主动的、有目的性的行为，来完成类似于人类执行功能的事物处理能力，包括设定目标、规划流程、分解任务、使用工具方法、管理时间和组织安排等方面；这也就意味着AI需要在真实环境中学习大量实际的经验，并对上下文和具体情境能够有实时调整的能力，进而可以依据实际的场景自主决策，灵活安排并主动交互。

（5）适应性

AI具备主动感知和理解世界的能力，以及能够与环境进行双向的、动态的交互；这种适应性不仅限于机器对输入的响应，还包括系统能够根据内部知识做出合适的决策，并通过特定的行为来改变周围的环境；在社会学意义上，意味着人工智能能够以近似人类的方式与世界进行深度互动，并理解世界的复杂性。

显而易见，要想让人工智能更接近人类智慧，其先决条件是让人工智能理解并学习人类认知世界的方式，进而以类似人类思考决策的方式去行动。

作为典型的强智能体，人类在成长过程中较少的依赖当前深度学习中采用的监督学习范式。相反，人类关键性技能的发展，如行走、使用工具、学习新的技能，都依赖于身体力行的尝试。同样，具身智能通过与环境的互动，虽然面临第一视角得到数据的不稳定，但它能够通过类似人类的中心感知方式来学习，并真正地在实际环境中应变和理解，从而从视觉、语言和推理过渡到人工具身（Artificial Embodiment）。

图灵在其开创性论文《计算机器与智能》中，奠定了人工智能的基础，还预见了两条可能的发展路径。一条侧重于抽象计算所需的智能，如下棋和数学问题解决；另一条赋予机器先进的传感器，让它们能够像婴儿一样通过感知和互动学习，与人类进行交流。如今，随着时间的流逝，图灵的这些构想已经分别演变成我们所熟知的非具身智能和具身智能。

今天，在大语言模型的普及和GPT-4等前沿模型的推动下，我们似乎见证了人工智能领域的一个新时代，人机交流也变得前所未有的流畅和无缝。据2023年5月GGII发布的报告预测，预计到2026年，人形机器人在全球服务机器人市场中的渗透率预计将达到3.5%，市场规模超过20亿美元。各大科技公司和学术界的顶尖学者也不断涌入这一领域的研究与产品开发当中。

然而，在繁荣热潮的背后，潜在的困境却也如影随形。尽管ChatGPT等模式革命性地变革了人工智能领域，但它们在理解力、联想力和交互能力等方面，仍然未能完全满足公众的期望。这促使我们对看似毫无阻碍的进步进行重新评估，同时希望经过不懈努力，人们能攻克实现真正的具身人工智能所面临的复杂挑战。

五、具身智能面临的挑战

综合来看，具身智能在其发展过程中面临着多项挑战，这些挑战源自于其发展过程中的复杂性和不断变化的需求。主要包括以下几个方面：

（1）适应非结构化真实环境

与预设规则和模式驱动的传统AI系统不同，具身智能必须在一个充满复杂性和不可预测性的非结构化环境中找到立足点。在这种环境中，信息的稀缺和场景的多变性，要求AI系统具备更加先进和灵活的计算能力，以便能够适应环境的不断变化和不确定性。这不仅是一个数据处理的问题，更是对AI系统感知和适应能力的全面考验。

(2)发展更高级的认知策略

在自然界中,生物体通过视觉、听觉和触觉等多种感觉途径获得复杂的感知信息,并在大脑中进行有效的多模态信息融合。具身智能同样需要模仿这种高效的多模态融合过程,以更全面地理解和适应其所处的环境。这包括但不限于对三维空间中物体的精确识别和定位,以及对环境变化和内在联系的动态捕捉。

此外,具身智能还需要超越传统的计算模型对静态数据处理,发展出对事物的动态变化和相互关系的深层次理解。这不仅关系到对时间和空间信息的处理,还涉及到理解其他生物(尤其是人类)的意图和行为动机,从而实现更自然、更智能的人机协同。

(3)与人类智能的显著差距

在具身智能的发展中,一个显著的挑战是弥补其与人类智能之间的差距。尽管具身智能在逻辑性、感知力、实时性、主动性和适应性方面取得了显著进步,但与人类大脑、身体和环境之间的动态、螺旋式上升的复杂认知过程相比,仍有很大的差距。特别是在元认知能力方面——即对信息处理过程本身的监控和反思能力——现代AI系统,特别是依赖于智能神经网络的系统,还未能达到这一层次。这种能力对于校验记忆的可靠性和所学内容的有效性至关重要,而智能体在这方面的幻觉和过度自信问题仍然是一个待解决的挑战,限制了它们在复杂情境中的自我成长和决策能力。

此外,要使具身智能达到与生物相当的学习和决策能力,需要在自适应学习和知识迁移方面取得显著进步。在自然界中,生物通过快速学习和自适应能力来应对新挑战和任务。目前的具身智能在灵活性和应变能力方面,尤其是在多变的实际应用环境中,仍处于初级探索阶段。要实现这一目标,具身智能不仅需要具备强大的决策和控制能力,还需对各种任务有深入的理解和精准的规划。

与生物体一样,具身智能也应追求终身学习的能力。当前的AI系统主要依赖于算法和数据驱动的学习方法,但在实时学习和处理大量数据方面存在局限。因此,为了更好地适应复杂环境,具身智能需要突破这些限制,向生物体那样的自然和连续学习模式迈进。

(4)技术成熟度和硬件发展的制约

具身智能的进步需要大量研发投入,包括硬件制造、软件开发和算法设计等多个领域。这些高成本的投入可能会阻碍创业公司或中小企业的参与,从而使市场主要由大型企业主导。同时,具身智能的发展牵涉到多学科领域,如机器视觉、自然语言理解、认知和推理、机器人学、博弈伦理、机器学习等。加强这些学科之间的紧密合作和交流,相互促进和互利共赢将是推动具身智能未来发展的关键。

此外,具身智能必须对外部环境保持敏感和开放,这就要求硬件设备的支持,以便能够与外界进行有效沟通。传感器、芯片、执行器、视觉图像处理、语音传输和仿真模拟等相关硬件的技术创新,将极大地促进具身智能的发展,特别是在高效和节能方面,具身智能硬件的潜力仍待进一步开发。

(5)涌现式创新与突破的缺乏

生物群体能够展现出令人惊叹的集体智慧,主要归功于其中个体之间的协同作用。对具身智能来说,一个重要的挑战是模仿这种分布式智能系统。这意味着需要将智能分散到多个实体中,并通过它们之间的协作,实现更高级别的认知和决策能力。生物群体所展现的自组织和适应性特征,允许它们根据环境的变化和个体之间的差异进行自我调整。具身智能需要发展类似的机制,以实现分工协作和动态任务分配,

从而能够更灵活地应对多种情境。

然而，理解和模拟生物群体中的涌现现象，尤其是在计算模型中，仍是一个巨大的挑战。要发展出一个符合生物原理的计算模型，从而使具身智能能够真正实现群体间复杂的交互和创新，显然还有很长的路要走。

（6）伦理与安全的挑战

具身智能在与真实环境进行交互并充分学习时，势必会收集和处理大量数据。这就引出了一个关键问题：如何在实时交互中确保这些数据的安全性和隐私性。保障数据安全和用户隐私是具身智能发展中不可忽视的重要方面。

同时，具身智能在与物理世界互动时，必须确保不会对人类或环境造成伤害。在系统设计的初期，就必须考虑到这类行为约束和安全保障的问题。此外，具身智能在决策时还需要考虑伦理和道德问题。因此，未来的发展不仅需要技术创新，还需要建立更为健全和全面的伦理指南，以指导具身智能在复杂情境中的行为决策，确保其行为符合道德原则和社会价值观。

总结而言，具身智能的发展不仅是技术革新的过程，更是对人类理解、伦理道德和社会影响的深思。毕竟相关研究不仅将推动科技的边界，更将深刻影响人类社会的方方面面。

六、学科大融通：如何为具身智能带来突破？

面对这些挑战，我们的解决之道不仅在于技术的进步，更在于跨学科合作的智慧。人工智能学家们正在积极地从神经科学、认知科学以及其他相关领域中寻找灵感，以期打破现有技术的限制。通过学习大脑处理信息的方法，并理解人类的认知与交流机制，我们有望开发出更高效、智能且适应性强的AI系统。

（1）感知与决策

具身智能的核心挑战之一是创建能够在复杂、非结构化的环境中有效运作的系统，而这正是人类大脑所擅长的。神经科学，作为研究大脑和神经系统的学科，提供了关于如何处理信息、做出决策和适应环境的深刻启示，这对开发具有类似灵活性和适应性的AI系统至关重要。

在生物体中，感觉运动通路的几乎每个阶段的神经反应都会通过生物物理和突触过程、循环和反馈连接、学习以及许多其他内部和外部变量进行修改。这些过程强调了感知输入与运动输出在认知处理中的紧密联系。为了模拟这一机制，研究者们正在改进机器人的感官输入系统，模拟人类感官与运动的整合机制，并通过分层生成建模和多级规划方法来模仿人类运动控制的复杂时间结构。这项进步已经使得机器人能够自主地完成诸如捡起并搬运物品、开门、踢足球等复杂任务。

人类大脑在学习新技能时会产生特定的奖励信号，促进特定神经回路的形成，并在不确定和动态变化的环境中进行实践，进而做出合理决策。目前的机器学习系统还依赖于预设的数据集和奖励函数，而未能实现类似的自我迭代和超越数据分布的能力。模拟大脑的这种能力有望引导AI学家开发出更为复杂且能自我优化的认知处理算法。

此外，神经科学的发现不仅揭示了生物体如何有目的地与环境互动，还提供了模块化和分层架构方面的指导，帮助人工系统模仿这些能力。这些指导包括如何使系统的低层模块在缺乏高级模块输入时半自主运作，以及如何将运动规划从缓慢过程转换为快速反射系统。这些原理在AI系统中的高级运动规划和精细控制方面已经找到了应用。

新近发展的神经形态工程方法，利用超大规模集成电路，通过特殊排列，模拟人类神经系统的生物学结构与功能。在神经形态感知、动作规划和认知处理策略中，已通过概念应用得到验证。故而，借助神经形态工程用于构建高效、紧凑的智能机器人系统，也将帮助具身智能在充满挑战的现实环境中感知、行动和学习。

（2）社交互动与情绪智能

在具身智能的研究中，社交互动与情绪智能的地位举足轻重。这一部分专注于理解和模拟人类的社会认知过程，例如心智理论，即理解他人的信念、欲望和意图的能力。目前，AI系统通过模仿人类大脑处理社交信息的机制，已能更自然地解读和响应人类的社交行为。这包括分析语音、面部表情和身体语言等社会信号，并利用机器学习模型进行上下文感知学习，使AI能够根据不同的社交环境做出适应性反应。

情感计算的最新研究则集中在如何使AI系统精确地识别和模仿人类的情绪反应上。这一进展可以帮助具身智能更好地理解和适应人类的情感状态，提供更加人性化的交互体验。研究者利用生理信号（如心率和皮肤电反应）和行为数据（例如面部表情和语调）来帮助AI识别情绪状态，同时创建复杂的模型来理解情绪的多样性和动态变化。这一技术已在客户服务、教育和医疗辅助等领域得到了初步应用，展示了其广泛应用的可能。

（3）算法模型

在具身智能的算法模型领域，人工智能研究者们借鉴了大脑处理感官输入和执行复杂认知功能的方式，从而设计出更先进的神经网络模型。以卷积神经网络和递归神经网络为例，这些深度学习算法模仿了人类视觉皮层和时序数据处理的机制，使得它们在处理视觉和语言任务上展现出类似人类的效率和精确度。此外，通过模拟真实神经元的动力学进行信号处理和计算的方式，也为实现基于大脑启发的计算基元和大规模并行的内存计算模拟电路奠定了基础。

深入研究人类感知过程，AI学家能够设计出更有效地处理视觉、听觉和触觉信息的算法。这些算法进一步促进了具身智能系统的发展，使其能够综合处理不同类型的输入，如图像、声音、对话和运动。例如，通过模仿人类视觉处理机制，具身智能系统的物体识别和环境理解能力得到了显著提升。此外，借鉴人类的注意力机制，优化了信息筛选和重点关注区域，显著提高了处理效率和准确性。视觉导航系统的发展，则是模仿了人类大脑的空间定位功能，使得机器人能够在复杂环境中更加精准地导航。

然而，随着深度神经网络在人工智能中的应用日益复杂化，出现了所谓的“黑箱”问题，即难以解释模型具体如何做出决策。这一问题在模型决策对人类福祉产生重大影响时尤为重要。在这方面，认知心理学提供了一种宝贵的视角。研究者正努力将认知心理学的方法和严谨性融入AI研究中，以提高机器学习模型的可解释性、公平性和透明度。例如，可解释人工智能（XAI）领域的兴起，旨在揭示深度学习模型的决策过程，使这些模型变得更加透明和可信赖。

（4）生物与仿生

在具身智能的发展中，生物学和仿生学的研究起到了不可或缺的作用。通过深入探索自然界生物体的结构、功能和行为模式，这些研究为具身智能提供了源源不断的创新灵感。

传统的机器人，无论是刚性还是柔性，大多由工程构件和合成材料制成。然而，最新的研究开始将活细胞、生物组织、微生物甚至整个动物纳入机器人的设计中，开辟了机器人技术与生物学探索的新天地。这

些生物混合体系的活体材料展现出生物可降解性、自愈性和天然顺应性，从而提供了自适应驱动和控制的新可能。比如，有研究者正在设计能够响应电脉冲的肌肉细胞薄片，使机器人能协调地执行游泳、行走或抓握等动作。同时，对植物的多样形态因素的研究也正在启发人工智能的新发展，甚至还有纳米级别的微型机器人正在模仿微生物的特征。

在自然界中，不同生物拥有各种独特的感知机制。例如，鸟类的导航能力、海洋生物的声纳定位系统。这些高度专业化的感知系统可以启发AI学家设计更为高效和精确的传感器和感知算法。这种模仿不仅限于生物的外部结构，还包括其内部控制和神经反馈机制，这对于提高具身智能的自主性和自适应性至关重要。然而，尽管如此，目前人工智能的视觉技术在自动化和智能化方面虽有进展，但与自然界中生物眼睛的灵巧和智能相比仍有较大差距。

此外，生物群体如蚂蚁群和蜜蜂群体能够涌现出惊人的集体智慧和协调能力。最近的研究显示，经过大规模数据训练的Transformer模型能够展现出类似于情境学习的能力，这种能力的出现与数据集中任务多样性的关系密切。例如，Evan　Hubinger等研究者发现，当任务环境特征包含多样性、多分支，涉及新颖情境或任务实例时，模型更容易表现出元优化行为。这为模拟生物群体的多样性行为提供了重要指导，从而指引在多智能体环境中协同工作的具身智能的发展。

最后，与人工智能系统相比，生物系统如人脑在能量效率方面具有显著优势。生物神经元通过传输动作电位进行交互，展现出高效的能量管理。此外，生物网络即使在组件不可靠或存在噪音的情况下也能有效地进行计算。因此，模仿生物体的能量管理和自我修复机制，可以设计出更高效、更持久的能源系统和自我维护策略，从而延长具身智能设备的使用寿命并降低维护成本。

尽管具身智能的发展在神经科学、认知科学和生物学等多个领域获得了丰富的灵感和知识支持，但目前的理论和技术水平尚未能带来根本性的变革。关键问题在于，无论是传统的机器人技术还是目前流行的神经形态方法，它们的系统设计还远未达到深入借鉴这些原理的程度。

在感知决策方面，我们还未完全理解如何将各种感知和计算组件有效整合，形成一个既连贯又高效的系统，以实现真正有效的感知行为；在社交互动与情绪智能上，尽管我们已能识别情绪，但这与自然地与真人交互存在巨大的鸿沟；算法模型的发展也不仅仅局限于硬件组件的协调，更涉及到算法和数据处理的深层次优化，这不仅是一个技术问题，更是对生物系统复杂交互和整合机制理解的挑战。

七、具身智能的未来

展望未来，我们正站在一个充满潜力的新时代门槛上。虽然完全成熟的具身智能仍在远方，但它已在工业机器人、智能座舱、聊天机器人等领域已展现出其重塑传统产业和改变工作方式的潜能。特别是在进入生成式人工智能时代之后，大语言模型如GPT为具身智能赋予了新的“大脑”，结合视觉语言模型（VLM）和视觉导航模型（VNM），推动着机器人在不确定环境中的应变能力，从而为应用端开拓了前所未有的可能性。

或许随着大语言模型与具身智能的深度融合，我们可以预见一个充满活力的虚拟世界。在这个世界里，具身智能体不仅能自主决策和行动，而且将通过感官与环境互动，探索世界。可以想见，这些智能体将配备多种感知功能和多功能模块，如视觉处理、语言理解和记忆管理，它们通过全局工作空间机制灵活组合，实现与人类的有效沟通与协作。在某些情况下，这些智能体甚至能让人类难以区分虚拟与现实。更重

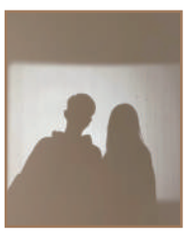
要的是，这些智能体能够在虚拟世界中留下独特的印记，通过共识主动性机制直接互动，与人类合作解决复杂问题，超越单一智能体的能力范围。

虚拟世界，与现实世界形成鲜明对比，提供了一个更为精密和可控的环境，使得智能体能够进行更加大胆和创新的行为。这不仅仅是对人类智能的延伸，更是一个通用人工智能诞生和发展的舞台，为超越人类智能水平的AI提供了理想的试验场和成长空间。或许，这也是各大科技公司对具身智能与虚拟世界寄予厚望的深层次原因。这预示着一个更加智能、更加互联的未来。(编辑：存源、韵珂)

从“智能涌现”到“超人类”，如何实现AGI？

作者：P

波士顿某大学种田PhD在读



作者：T

波士顿某大学养电子宠物PhD在读

扫码查看原文

一、科幻照进现实

当阿西莫夫在小说《转圈圈》（Runaround）中向世界介绍他著名的机器人三定律时，他可能没有完全预见到，八十年后的世界会多么接近他的科幻梦想。如今，我们生活在一个由人工智能（Artificial Intelligence， AI）渗透的世界里。AI系统在许多方面已超越阿西莫夫的想象——在家里，私人AI助手不仅可以帮你安排日程，还能根据你的心情推荐娱乐节目。工作中，各种AI分析工具帮你洞察复杂数据，提供行业和科学洞见。甚至在艺术领域，AI也在帮助千千万万的艺术家创作新颖的作品，挑战我们对创造力的传统认识。

这还只是冰山一角。AI气象模型[1]已为我们预测天气，它比代表人类千万年经验的传统数值预报方法还要准确，速度也要快上一万倍；AI教学平台正在根据学生的学习习惯和进度提供个性化指导[2]，使教育更加有效和包容；AI医疗模型也正帮助医生更准确快速地诊断罕见病[3]、癌症[4]、神经退行性疾病[5]，在某些情况下，它们的表现甚至超过了人类专家。

然而，这些进步引发了一个问题：这些先进的AI系统是否代表着人类对通用人工智能（AGI），或“（超）人类层级的智能”的终极设想？

自1956年达特茅斯会议提出“人工智能”这一概念以来，实现人类水平的智能一直是AI领域的圣杯。今年上半年，终于有主流研究者提出[6]，AI模型——或者更确切地说，大语言模型——已经表现出“通用人工智能的火花”（sparks of AGI）。这似乎表明，AGI已经从哲学猜想变成了将来的未来。然而，关于AGI的观点众说纷纭，大语言模型也常有愚蠢行为出现，这些都引发了对AGI的质疑。在此背景下，我们需要明确几个关键问题：我们的AGI目标到底是什么？我们离实现真正的AGI还有多远？我们如何预测AGI的社会影响并评估其潜在风险？想回答这些问题，首先要在那些带有迷惑性的科幻作品之外，准确地、可操作地

定义AGI。

二、尝试定义AGI

尝试定义通用人工智能的概念，最早可追溯到上世纪五十年代著名的“图灵测试” [7]。在这一测试中，人类需要使用文本，与某未知对象进行交流，并据此判断对面是机器还是人类。图灵的洞见在于，机器是否能“思考”不取决于其思考过程，而在于其表现的能力。但是，因为人类太容易被糊弄，这一测试通常并不能很好地反映智能的程度。

此后，约翰·瑟尔（John Searle）在他著名的“中文房间”[8]思想实验中，则将AGI视作一种有“意识”的强AI系统。虽然将AI与“意识”相连听起来很诱人，但这更多是哲学上的讨论，而非可验证的科学，因为“意识”本身更是一个难以被科学定义的概念。

而在马克·古布鲁德（Mark Gubrud）1997年首次提出AGI这一概念时，他将AGI类比于人脑——一种能在复杂性和速度上超越人脑，能获取和内化知识，被用于需要运用人类智慧来解决问题的用途之上。但问题在于，现代的AI系统，比如基于Transformer模型，与人类的大脑结构和学习模式的联系其实并不紧密。

更近地，本世纪初，当DeepMind联合创始人谢恩·莱格（Shane Legg）将AGI的概念向计算机科学家们普及时[9]，他将AGI定义为在认知任务上能取得类人表现的机器智能。但这一定义并未明确所指的任务类型和“类人”标准。

除此之外，近期还有从学习任务或元认知能力（不充分），经济价值（不必要不充分），灵活性（不充分）等方面对AGI的定义尝试，但它们都有各自的问题。在今年，还有一些极具影响力的科学家提出[10]，当前最佳（SOTA）的大语言模型已经是AGI，因为它们在许多任务上都能取得一定的表现，足够通用；但是，在通用之外，真正的AGI还必须拥有足够可靠的性能。

事实上，所有这些尝试都在试图定义一个AI发展的“临界”或“终极”状态。但是，我们通向AGI的巅峰之旅，恐非一点之极，而似层峦叠嶂、地形错综的高原。

最近，谢恩·莱格带领DeepMind团队总结历史上的定义，并在此基础上提出了他们对AGI的定义框架[12]。

三、从理论到实践：定义AGI的六大原则

从这些过去定义AGI的尝试中，研究人员发现了一些共同特征，并从中提取出定义AGI所必须满足的六大原则：

（1）注重能力，而非过程：AGI的定义应关注于其能完成的任务，而非完成这些任务的具体方式。这意味着，实现AGI并不意味着系统必须以类似人类的方式思考或理解。

（2）注重通用性与性能：AGI不仅要能够执行广泛的任务（广度、通用性），还要在这些任务上都能达到一定水平（深度、性能）。

（3）注重认知与元认知任务：AGI必须在有能力处理认知任务（如解决问题、创造性思考）的同时，具有元认知能力（如学习新技能，或知道何时向人类请求帮助）。

（4）专注于潜力而不是实际应用：评估AGI时，应关注的是其潜在能力，而非它目前的实际应用情况（取决于在真实世界中的部署）。

（5）重视实际适用性：AGI应能解决现实世界中的实际问题，它的价值在于能够应对现实生活中复杂和多样化的挑战。

（6）关注实现AGI的路径，而非单一终点：正如通过一套自动化驾驶的标准体系会更有助于针对自动驾驶车辆相关的政策制定和进程推进一样，定义AGI在不同“水平”上的衡量标准是很有益的。AGI的发展就像是一次长途跋涉，终极目标可能不是一个固定点，而是一个不断演进的过程。

四、理解AGI：定义层级

由此，研究人员将“性能”和“通用性”作为评估AI系统AGI程度的两个核心维度。其中，“性能”测量AI系统能力的“深度”，即在给定任务中与人类表现相比，AI系统达到的水平如何。“通用性”测量AI系统能力的“广度”，即AI系统能解决的任务范围。如果把没有AI时的计算系统当作AI发展的起点（比如最基础的计算器），把超过所有人类表现的AI系统当作AI的终极形态（比如国际象棋界孤独求败的AlphaZero[11]），那么便可以把AI的性能发展分为六个阶段：无AI、智能涌现、胜任、专家、大师、超人类。同时，研究人员把AI系统分为专一于特定任务的“专才”系统和胜任各类任务的“通用”系统。这样就可以把现有的AI系统作如下分类：

性能（行）x 通用性（列）	**专才（narrow）** 特定的任务或任务组	**通用（general）** 包括学习新技能在内的多种非物理任务范围
Level 0: 无 AI	**专才无 AI** 基础计算器	**通用无 AI** 人类参与的计算系统，如众包系统
Level 1: 智能涌现（emergent） 相当于或稍优于普通人	**智能涌现专才 AI** 经典 AI 系统 GOFAI；基于简单规则的系统，比如 SHRDLU	**智能涌现通用 AGI** ChatGPT (GPT4), Bard, Llama 2
Level 2: 胜任（competent） 至少达到熟练成年人的中等水平	**胜任专才 AI** 智能音箱，如 Siri、小爱同学、或小度小度； 视觉问答系统，如谷歌 PaLI； IBM Watson； 在某些特定任务上的目前最高水平的大语言模型（例如，写短篇文章、简单编程）	**胜任通用 AGI** 还未实现
Level 3: 专家（expert） 至少达到熟练成年人中的顶尖水平	**专家专才 AI** 例如 Grammarly 的拼写和语法检查器； 生成图像模型，如 Imagen 或 Dall-E 3	**专家通用 AGI** 还未实现
Level 4: 大师（virtuoso） 至少达到熟练成年人中的极高水平	**大师专才 AI** 深蓝、AlphaGo	**大师通用 AGI** 还未实现
Level 5: 超人类（superhuman） 超越所有人类的表现	**超人类专才 AI** AlphaFold、AlphaZero、StockFish	**人工超智能（ASI）** 还未实现

▷图　2：基于能力的深度（性能）和广度（普遍性）对AGI道路系统进行分类的分级矩阵方法。每个单元中的示例系统是基于文献中的当前描述或与部署的系统交互的经验的近似值。图源：参考文献12，由追问译。

五、使用这一框架测试通用人工智能

AGI测试是评估AI全面能力的关键步骤。上述六大原则为AGI的理解和测量提供了一个全面的框架。其中，普适性和性能原则引导研究者构建了一个分层的本体论框架，以深入讨论AI的广度和深度。

在AGI测试中，性能的维度是比较直观的：我们可以通过比较AI系统与人类在相同任务上的表现来评价它。这就好比在一场围棋比赛中，我们不仅会看棋手赢了多少盘，还要看他们击败了哪些级别的对手。

通用性的维度则更具挑战性。它要求我们回答一个关键问题：一个AI系统需要掌握哪些任务，才能被认为具有足够的广泛性？例如，一个能够进行数学推理、理解自然语言、甚至创作音乐的AI，显然展现了更广泛的能力范围，但这足够通用吗？这就像在技能的茫茫宇宙中找到最具代表性的星体，既要足够广泛又要足够重要。对此，研究人员认为，还需要更多的研究。

值得一提的是，在评估AGI的性能时，所使用的测试应当尽可能地贴近真实世界的应用场景。例如，评估自动驾驶汽车时，比较其与不使用任何现代AI辅助工具的人类驾驶员并不够有说服力。反而，应当将其与那些已经采用某些驾驶辅助技术的人类驾驶员进行比较，这样的比较才更有意义。

此外，AGI的基准测试可能会包含一些开放性和互动性的任务，这些任务可能需要定性评估（换句话说就是人类的主观评价）。这就像是在一场艺术比赛中，裁判不仅要考虑作品的技术程度，还要评价创意和表现力。例如，我们可能要求AI系统解读文学作品，或者创作一首诗，这些任务不仅考验AI的技术能力，还考验其创新性和对人类情感的理解。

最后，构建一个AGI的基准测试是一个需要不断迭代和改进的过程。这个基准本身也必须不断更新，以适应AGI可能完成的各种任务，毕竟我们永远也无法穷举所有AGI应完成的任务。这不仅仅是为了测量AGI的当前能力，更是为了激发它向更高的智能进化。这个过程既是技术的挑战，也是我们对智能理解深度的体现，更重要的是，它将引导技术发展的未来走向。

六、其他研究方向

除了构建AGI的基准测试外，要想更好的向AGI迈进，其他几个研究方向也尤为关键。

对齐：对齐（alignment）的目标，是确保AI系统的行为与人类的利益和价值观保持一致。这是避免AI系统产生不可预料后果和潜在危害的关键。在今年的NeurIPS和2024年的ICLR投稿中，就有大量的论文和研究集中这一领域，突出了其重要性。它们不仅关注理论和技术的进步，还涉及相应的治理手段，以确保AI在复杂和动态的环境中依然能够遵循人类的伦理和价值观。随着AI系统与社会的逐步融合，对齐研究的价值将日益凸显。

机械解释性：机械解释性（mechanistic interpretability）相关的研究旨在揭示AI系统如何做出决策，以及这些决策背后的机制。通过增强我们对AI系统内部运作的理解，可以更好地监控和指导这些系统的行为，确保它们的行为符合人类的期望和标准。

跨学科合作：跨学科合作在AGI研究中也起着至关重要的作用。从计算机科学到哲学、心理学、神经科学、法学、社会学、乃至政治学，各个学科的专家都需要共同努力，以确保AI系统的设计和应用不仅在技术上先进，而且在伦理和社会层面上不掉队。这种合作将有助于我们更全面地理解AGI的潜在影响，并为其带来的挑战提供创新的解决方案。

七、风险的多面性：从存在性到极端风险

当我们探讨AGI时，各种风险概念经常被提及，从“x-risk”（existential risk，即存在性风险）到各类极端风险。所谓存在性风险，是指对人类生存造成威胁的风险，而极端风险则涵盖了一系列可能对社会造成严重影响的危险。这些风险的多面性不仅让我们对未来充满好奇，也提醒我们要谨慎前行，毕竟在每个转角都可能隐藏着新的挑战和机遇。

（1）AGI等级与风险评估

通过定义AGI层级框架，我们能够更清晰地看到不同等级的AGI所带来的风险，而不是将其视为一个单一的事件进行评估。在这套框架下，政策制定者能够识别并优先处理当下发展阶段带来的风险。例如，专家级AGI可能会导致人们的工作岗位被机器智能取代，使得经济领域出现动荡，但同时也可能减少在较低级别AGI中出现的一些问题，如任务执行不当的风险。这就像在攀登科技高峰时，需要不断地平衡和调整，以确保人类不会在追求高度的同时丧失了安全。

但是，必须承认，技术的迭代速度，特别是在AGI领域，可能会超越现有的法规和外交政策的步伐。我们已然驾驶着一辆超高速的列车，但现有的道路规则可能还停留在马车时代。这将带来更为复杂的影响。比如，首个达到ASI（人工超智能）的国家可能会获得巨大的地缘政治和军事优势，而这可能会引发一系列风险和挑战。

正因如此，研究人员认为，AGI的基准测试不仅应是对AI能力的测量，也更应是对其潜在风险的评估。这其中包括了如何处理AGI可能具备的危险能力，如欺骗和暗示性说服。我们在培养一个全能的个体时，不仅要教给它知识和技能，还要教会它如何负责任和道德地使用这些能力。例如，Anthropic推出的负责任扩展政策[12]，通过为不同级别AI系统设定不同的风险等级和防范措施，展示了一种平衡创新与安全的方法。

（2）自主性与风险：设计的艺术

在考虑AGI的应用时，我们也需要思考它们如何与人类互动。还是回到自动驾驶汽车的例子，完全自主的车辆（Level （5）无需人类干预，而对较低自主层级（如Level （2）的车辆，人类司机仍然扮演重要的监控和控制角色。在下表中，研究人员通过六级自主性来描述人机交互的范式。在这套范式中，AI系统自主化等级与其综合能力存在着直接联系。这意味着，随着 AI 系统变得更加高级和自主，它们能够提供更多样化的交互方式，同时也可能引入不同类型的风险。

自主性的不同层级代表了人类与AI之间合作的不同方式。有时，即使技术上可实现更高级别的自主性，但出于教育、娱乐或安全的考虑，选择较低级别的自主性可能仍是更明智的选择。例如，在自动驾驶领域，即使有能力实现完全自动驾驶，我们仍宁可选择让人类司机参与，以保证安全和适应性。将AGI系统置于实际使用场景中考虑，可以帮助我们更好地理解模型进步与人工智能交互范式进步之间的关系。这种考虑方式不仅能够帮助我们预测 AI 系统的发展方向，还能指导我们在设计这些系统时，如何确保这些进步既服务于人类需求，又不牺牲安全。

自主等级	典型系统	达成 AGI 的层次	可能引发的风险
自主等级 0: 无 AI 全部由人工操作	传统方式（如手绘草图） 非 AI 的数字工作流（如文本编辑器打字、绘图软件画图）	无 AI	不适用（常规风险）
自主等级 1: AI 作为工具 人工完全控制任务，并用 AI 辅助自动化简单重复的子任务	用搜索引擎查找信息 用语法校对软件修订文本 使用机器翻译	可能：智能涌现级专才 AI 很可能：胜任级专才 AI	技能退化（如过分依赖） 颠覆现有行业格局
自主等级 2: AI 作为咨询师 AI 承担重要角色，但需人工启动	利用语言模型整理文件 用代码生成工具辅助编程 主要通过高级推荐系统享受娱乐内容	可能：胜任级专才 AI 很可能：专家级专才 AI；智能涌现 AGI	过度信任 极端化倾向 针对性操纵
自主等级 3: AI 作为合作者 AI 与人类平等合作，互动协调目标与任务	与国际象棋 AI 互动学习 与 AI 生成的虚拟人物进行社交娱乐	可能：智能涌现级 AGI 很可能：专家级专才 AI；胜任级 AGI	人格化（如单向社交关系） 社会迅速变化
自主等级 4: AI 作为专家 由 AI 主导互动，人类提供意见反馈及执行部分子任务	利用 AI 推动科学发现（如蛋白质结构预测）	可能：大师级专才 AI 很可能：专家级 AGI	社会普遍倦怠 大规模失业 人类优越性质疑
自主等级 5: AI 作为代理人 AI 完全自主运作	自主运作的 AI 个人助理（未实现）	很可能：大师级 AGI；人工超智能（ASI）	目标不一致 权力集中

▷图 3: 不同自主等级达成的AGI层次, 以及其可能引发的风险。图源: 参考文献12, 由追问译。

八、OpenAI的观点：AI发展及其对社会的影响

今年更早些时候, OpenAI的CEO萨姆·奥特曼（Sam Altman）在他的博客文章Planning for AGI and Beyond[12]里, 探讨了AGI对社会、技术和伦理的潜在影响, 并强调了慎重规划和负责任的发展的必要性。

文章指出, 通过增加资源丰富度、推动全球经济增长, 以及助力新科学知识的发现, AGI有潜力极大地提升人类生活质量。这不仅是智力的巨大增幅, 更是对人类创造力和想象力的极大扩展。然而, AGI的出现也可能带来严重的滥用风险、意外极端事件的可能以及对社会结构的剧烈动荡。因此, OpenAI提倡在发展AGI时, AGI开发者以及全社会都必须找到正确的方法来实现和利用这一技术, 确保其与人类核心价值观保持一致, 并公平地分享其带来的好处。

此外, OpenAI认为, 从长远来看, AGI的出现仅仅是智力发展的一个节点, 而人工智能的进步可能会在未来相当长一段时间内继续保持。OpenAI提出了一个安全的AGI发展愿景, 即在较短时间内以较慢的速度过渡到AGI, 以便社会有时间适应和调整。

尽管未来不可预知, 但OpenAI表达了他们最关心的几个原则：希望AGI能最大限度地促进人类在宇

宙中的繁荣；希望AGI带来的好处、访问权和治理能够被广泛且公平地分享，并成功应对巨大风险。为此，OpenAI倡导在短期内进行渐进式过渡、继续创造越来越符合目标的模型，并在长期进行全球范围的对话，讨论如何治理这些系统、如何公平分配它们产生的好处，以及如何公平分享访问权。这一过渡到拥超智能的世界的过程，可能是人类历史上最重要且最充满希望和恐惧的时间，没有人能保证成功。但在极高的风险和回报下，全人类需要团结起来，让AGI在未来世界以一种对人类最有益的方式绽放。

九、结论：稳步前行的必要性

总的来说，AGI的进步不仅代表了技术创新，更是对未来人机交互方式的重新想象。随着我们探索AGI的未知领域，稳健和有序的发展是至关重要的，它确保了人类在追求技术的同时，不会忽视潜在风险。通过理解 AGI 系统的不同发展水平和相应的自主性水平，我们能够不断地评估和调整步伐，更好地准备应对未来的挑战，同时确保这些技术的发展能够以安全、负责任的方式造福人类。（编辑：存源）

参考文献

关联论文：Morris, Meredith Ringel, et al. "Levels of AGI: Operationalizing Progress on the Path to AGI." arXiv preprint arXiv:2311.02462 (2023).。

[1] Bi, K. et al. (2023) Accurate medium-range global weather forecasting with 3D Neural Networks, Nature News. Available at: https://www.nature.com/articles/s41586-023-06185-3.

[2] Mathgpt大模型发布：落地学而思AI学习机 (2023) 机器之心. Available at: https://www.jiqizhixin.com/articles/2023-11-06-16.

[3] Alsentzer, E. et al. (2022) Deep learning for diagnosing patients with rare genetic diseases, medRxiv. Available at: https://www.medrxiv.org/content/10.1101/2022.12.07.22283238v1.

[4] Song, A.H. et al. (2023) Artificial Intelligence for digital and Computational Pathology, Nature News. Available at: https://www.nature.com/articles/s44222-023-00096-8 .

[5] Yang, Y. et al. (2022) Artificial Intelligence-enabled detection and assessment of parkinson’ s disease using nocturnal breathing signals, Nature News. Available at: https://www.nature.com/articles/s41591-022-01932-x .

[6] Bubeck, S. et al. (2023) Sparks of artificial general intelligence: Early experiments with GPT-4, arXiv.org. Available at: https://arxiv.org/abs/2303.12712 .

[7] Oppy, G. and Dowe, D. (2021) The turing test, Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/turing-test/.

[8] Cole, D. (2020) The Chinese Room Argument, Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/chinese-room/.

[9] Legg, S. (2022) In 2001 I suggested the term to @bengoertzel for the book he was writing. the term Agi caught on after he published it. Some years later we found out that Mark Gubrud used the term back in 1997 in a nanotechnology and security paper., Twitter. Available at: https://twitter.com/ShaneLegg/status/1529483168134451201.

[10] Arcas, B.A. y (2023) Artificial General Intelligence is already here, NOEMA. Available at: https://www.noemamag.com/artificial-general-intelligence-is-already-here/.

[11] Silver, D., Kasparov, G. and Habu, Y. (2018) Alphazero: Shedding new light on chess, Shogi, and go, Google DeepMind. Available at: https://deepmind.google/discover/blog/alphazero-shedding-new-light-on-chess-shogi-and-go/.

[12] Morris, Meredith Ringel, et al. "Levels of AGI: Operationalizing Progress on the Path to AGI." arXiv preprint arXiv:2311.02462 (2023).

[13] Sam, A. (2023) Planning for AGI and beyond. Available at: https://openai.com/blog/planning-for-agi-and-beyond.

追问专访

追问专访·Moritz Helmstaedter：

十年争议落幕，欧洲脑计划失败了吗？

嘉宾：Moritz Helmstaedter

德国法兰克福马克斯·普朗克大脑研究所主任，曾任慕尼黑马普神经生物学研究所研究小组负责人和PI。

扫码查看原文

2023年9月，历时十年的欧洲人脑计划（Human Brain Project，HBP）正式画上句号。这项计划试图众筹多学科力量，来解析和模拟人类大脑的复杂结构与功能。

然而，自孕育之初，它就一直处于争议的风暴之中。

“我们不应该做出那种不切实际的承诺。”谈到人脑计划时，马克斯·普朗克大脑科学研究所主任Moritz Helmstaedter教授的表情突然严肃起来。他直截了当地说：“这种‘把我们知道的东西凑在一起，模拟它，然后来解决大脑疾病’的观点，我认为是错误的。”

随后，他也流露出一丝遗憾：“我有时很羡慕物理学界能够研发出成功的大型加速器，这是大型科学的理想模式。但在神经科学领域，我们在方法论上缺乏共识，不确定哪种方法是最好的，甚至不知道哪种方法能成功理解大脑。”

尽管对欧洲人脑计划抱有惋惜，Helmstaedter教授却坚信神经科学正逐渐变好。在人工智能的浪潮下，他表示自己是首批使用机器学习做神经科学研究的学者。他坚信，人工智能将加速神经科学领域的发展，甚至在很大程度上替代人类注释者。

而至于人工智能所带来的那些虚幻的预期，他并不感兴趣。他更关心的是大脑背后的实际机制和人工智能的实际应用价值。

以下为「追问」访谈Moritz Helmstaedter教授的详细内容。为便于阅读，我们对文字进行了精简。

Q Helmstaedter教授，您好！非常感谢您接受由天桥脑科学研究院支持的「追问专访」栏目。在采访开始前，请介绍一下您的研究领域。

Moritz Helmstaedter：我是Moritz Helmstaedter，现任德国法兰克福马克斯·普朗克大脑研究所的主任。我的研究重点是解析颅内神经网络，特别是应用高级显微镜和其他方法绘制神经网络，即连接组学。我在马克斯·普朗克大脑研究所负责连接组学部门的工作。

Q 您曾在2015年对欧洲人脑计划的前身“蓝色大脑计划”公开发表评论，指出“把大量的数据放在一起并不能创造新的科学”。我们知道欧盟委员在9月之后将不再资助HBP，这是否符合您的预期？从一开始，这个项目就备受争议，似乎大多数人都认为它失败了。您认为阻碍它实现原目标的真正原因是什么？

Moritz Helmstaedter：最初的人脑计划*被提出和构想时，它非常强调利用当时所有可用的数据来创建模型。我和许多其他同事都对这一说法持怀疑态度。作为科学家，我们必须非常负责任地与公众沟通。如果我们做出的承诺很明显无法实现，或者这些承诺显然是错误的，那就是行不通的，我们必须公开交流。就像我现在所做的那样，我告诉你们，也许连接组学会对我们有很大的帮助，但我们必须说明，其间也存在不确定性，我们不应该做出那些不切实际的承诺。

10年过去了，很明显这些承诺并没有实现。当然，这个大项目还是产生了很多有趣的相关科学成果。但我当时提出的具体批评，也就是这种‘把我们知道的东西凑在一起，模拟它，然后来解决大脑疾病’的观点，我认为是错误的。

*欧洲人脑计划（Human Brain Project，HBP）：HBP是一项为期10年的大型科学研究计划。该计划由欧洲联盟资助，由近500名科学家参与，耗资约6亿欧元，目标是在计算机里建立大脑模型，为脑部研究、认知神经科学和脑启发计算，建立基于信息通讯技术的研究基础设施，供世界各地的研究人员使用。该计划于2013年10月1日启动，于2023年9月结束。

经过十年积累，HBP涌现出不少重要且有用的科学成果，如通过创建和整合大约200个大脑皮层和更深层脑结构的三维地图，HBP科学家制作了人脑图谱，可通过EBRAINS访问。但从一开始HBP就备受争议，该项目管理层数次变动，相关研究被指责碎片化，无法实现多尺度整合，缺乏优先次序，合作受限等。随HBP的经费在9月底停止，有关HBP是否失败的讨论成千论万。详情请见：因为山就在那里——人类脑计划的十年 | 追问观察

我认为，如果你创建了一个大型科学项目，比如“人脑计划”，并投入十亿或类似巨额作为资金，那么问题就不在于是否会有一些很酷的成果出现。毕竟，这样的资金规模应当产生许多引人入胜和关键性的发现。事实上，“人脑计划”的合作者们已开展了很多相关且极具吸引力的研究。如果这笔资金分成千份，每份一百万欧元或者五百万欧元，分配给欧洲各地的研究团队，结果又会有何不同呢？我认为这更难评判，这是一个更高的评判标准，因为这还不足以产生伟大的科学成果。

为此，你还需要证明这些成果不可能仅通过众多小项目实现。但“人脑计划”从一开始就声称可以利用现有数据建立一个巨大的脑科学模型，并解决重大问题。然而，这个原始目标并未实现，这是很多同事都预测到的，我也对此表示怀疑。但我们必须承认，这确实不如项目所预想的那样。

Q 是否看过电影《奥本海默》？我们在物理学领域已经看到了一些成功的大型科学项目，如曼哈顿计划、LIGO（激光干涉引力波天文台）和欧洲核子研究中心。相比之下，在生命科学领域，这样的例子却比较少

见。是大型科学方法本身不适用于当前的脑科学研究，还是HBP缺乏一个像奥本海默那样的领导者？

Moritz Helmstaedter：我没有看过《奥本海默》，只是有所了解。不得不说，我有时很羡慕物理学界能够研发出成功的大型加速器，这些都是大型科学的成功标志。为什么它们如此成功？是因为他们的目标非常明确、可测量、可量化。物理学已经花了几十年时间来发展一个连贯的模型，使得他们能准确预测——“如果我做X，就能测试Y”。而且如果有人为此付钱，那他们就能够完成这些任务。这是大型科学的理想模式。

但这在神经科学领域要困难得多，虽然我不会说这不可能。我们在方法论上缺乏共识，不确定哪种方法最佳，甚至不知道哪种方法能成功理解大脑。每个研究者都认为自己的方法是最优的。在领域内，我们还未达成共识，这与物理学大相径庭。

目前，生物学中一个较好的大型科学项目例子是人类基因组计划。这个项目目标明确且可实现，但其影响和直接后果几乎都被错误估计了。人们都希望通过它鉴定出一些关键基因，用于治疗与基因无关的疾病，包括精神病学领域。但这显然没有发生。尽管如此，它促进了更快速的测序技术发展，使全基因组关联分析*成为可能。我们会竭尽所能去追求目标，甚至在竞争*中实现它。

然而，当我们没有找到预期的结果时，我们会感到惊讶，但这就是为什么我们要这样做的原因。如果我们事先知道要找什么，我们就不必去探索新大陆、新恒星、宇宙的新部分。这也是我从事连接组学的动力，至今，我们仍然对数据的精确度和深度感到惊讶，但也有时候，也会有我们找不到预期希望找到的东西，或者我们认为会发现的奇妙事物最终并未出现。大不了推翻假设，和它说再见，然后转向下一个假设。这就是科学的过程。

*全基因组关联分析（Genome wide association study, GWAS）：应用基因组中数以百万计的单核苷酸多态性（single nucleotide polymorphism, SNP）为分子遗传标记，进行全基因组水平上的对照分析或相关性分析，通过比较发现影响复杂性状的基因变异的一种新策略。

竞争：指美国的民营企业（比如Celara Genomics）与人类基因组计划进行的竞争。他们同样测定了果蝇基因组，也测定了小鼠和人的全基因组序列，取得过很好的科学、经济与社会效果。尽管双方各自努力有些浪费资源，但最终“官”和“民”的竞争还是达成了和解。

Q 之前有一种论调，许多人认为HBP缺乏一个明确且可实现的目标。如果您要为HBP设定一个明确的目标，那将会是什么？

Moritz Helmstaedter：不得不承认，我正在努力实现我为这个项目设定的目标。设定的目标不是因为我认为其他人也应该这样做，而是因为我深信这些目标的重要性，并能论证其重要所在。神经突触级连接组（synaptic level Connectome）是我们在进化中学习所涉及的部分。学习的过程，包括适应与选择。什么样的捕食者会出现，为什么有条纹的野兽是危险的，还有那只靠近的鹰代表着什么，这些都可能在进化上被编码。除此之外就是我们个体一生所学的个体知识。神经突触级连接组是这两者汇聚的地方。它不是基于细胞类型的连接组，它也不是像整个大脑弥散张量成像（DTI）图那种高速公路般的投影图，因为他们都不存储个体知识。所以这就是为什么我认为这个观点有很大价值。这不是一个任意的描述或者选择，而是为了理解这些大脑网络而向前迈进的正确选择。

Q 我还有个假设性的问题，如果HBP在今天推出，借助所有新技术和人工智能，您认为结果会有所不同吗？

Moritz Helmstaedter：当然，结果肯定会不同，但仍然很难定义一个明确的最终目标。明确的目标是可以量化的，比如，仅仅说“我想了解大脑”并不是一个可量化的、明确的目标。我也不认为它应该是治愈疾病的行为，这仍然太过模糊。我们需要的是非常具体和可实现的目标，它可以是雄心勃勃的，但不能含糊不清。

Q 我知道您在海德堡大学学过物理和医学，请问跨学科研究以及物理学背景对您的神经科学研究有何帮助？

Moritz Helmstaedter：我从物理学教育中获益匪浅，我非常感激自己接受了这种非常基础的教育。物理学消除了我对复杂性的恐惧。物理学在处理看似复杂的历史以及一些非常复杂的问题，但随后发展了定量方法来解决这些问题。如果你接受了这方面的训练，你就会失去那种恐惧。显然，可能会有我们不能理解的问题，但你受过物理学训练，就会让你在这些问题上无所畏惧，至少你不会再去回避研究像是神经网络这样复杂的东西。

Q 贵实验室的研究需要处理大量的数据，您是如何将人工智能的力量融入到这个过程中的？

Moritz Helmstaedter：我们需要追踪连接神经元的细小纤维，即轴突，而这需要高分辨率的电子显微镜。但如果我们在三维空间做重构，这意味着我们需要获取非常大的数据集。实际上，20多年前，存储设备还处于千兆字节的规模，我们无法用当时的存储设备来分析我们获取的太字节级数据集。由于大脑中轴突和突触十分密集，其三维结构就会变得非常复杂和庞大，因此，就时间投入来说，手动追踪神经元已不可行。所以，我们很早就开始寻求自动化方法。大约在2005或2006年，我们开始探索机器学习在图像分析中的应用，这还不是常规方法。当时，我们还在使用标准图像处理方法，即设计一些滤镜，然后将其应用到图像中，最后得出一些结果。

当时我与共事的Winfried Denk开始使用机器学习方法。我们从K近邻算法开始。这是一种非常传统的用于解决视觉图像问题的机器学习方法。当时这已是相当不寻常的尝试，并使我们取得了显著的进步。随后我们与Sebastian Seung合作，使用由杨立昆开发的卷积神经网络。到了2007年，我们将其应用于图像数据分析，这是这种方法最早的应用之一，而且非常成功。我们在2013年发表的第一个哺乳动物连接组图，是使用当时的现代AI方法处理的。

但我们仍然需要依赖大量的人类智力的努力。我们招募了数百名学生来帮助我们标注数据。他们的高效标注与我们为此开发的工具相结合，与当时的人工智能相结合，帮助我们完成了首个哺乳动物视网膜连接组。这种关于老鼠的视觉处理装置揭示了人工智能和神经科学之间的有趣循环：一方面，我们用人工智能更好地理解大脑；另一方面，我们也用视觉器官实时分析人工智能捕获的大脑结构。

过去10年，我们越来越多地使用人工智能，以减少对人力需求。由于人力资源的扩展性有其限制——例如雇佣数百甚至数千名学生是极其困难的——因此我们必须推动人工智能及其改进，来减少所需的人力工作时间。这种做法已经证明非常有效。

到2019年，我们发布了首个皮层连接组，效率比之前的视网膜连接组高出20到40倍。现在我们真正进入了一个新阶段，我们认为人工智能可以在很大程度上取代人工注释者，这将更快地推进我们领域的发展。

Q 我们都知道ChatGPT现在很热门，我想知道这类大语言模型是否能帮助您的研究？

Moritz Helmstaedter：我们还没有直接研究过如何应用这类transformer模型。我们的经验是，当我们找到一个针对性的解决方案来解决我们的问题时，我们总是能取得最大的成功。因此，我认为ChatGPT这类大语言模型，在处理语言和其他复杂数据中确实有很大的潜力。但我们的数据并不复杂，它并没有太多可能的配置，只是在局部区分和理解这些轴突的延续位置非常困难，这是一个非常有针对性的问题。

我们目前的解决方法是把它看作一个三维飞行问题。我们有三维数据，并且已经构建了一个基于人工智能的飞行引擎，可以说，它学会了沿着轴突导航。现在这篇论文的第一个版本在biorxiv发布，我们称它为“RoboEM”*。它就像一个小型机器人在电子显微数据中穿行，这在我们应用中非常成功。它作为一个为具体的问题量身定做的解决方案，仍然具有优势。

*Schmidt, Martin, et al. "RoboEM: automated 3D flight tracing for synaptic-resolution connectomics." bioRxiv (2022): 2022-09.“RoboEM”是一个基于AI的自动化3D飞行系统，训练用于仅使用三维电子显微镜的数据沿神经元导航。RoboEM大幅降低了皮层连接组的计算标注成本，比手动错误校正的成本低约400倍，显著提高了自动化的最先进分割技术，并取代了人工校对，用于更复杂的连接组分析问题。

Q 您早期的工作中对比了人类大脑和灵长类大脑的差异，在您看来，人工智能和人类智能在工作模式上有什么区别？

Moritz Helmstaedter：在讨论人工智能和人类智能的工作模式差异时，首先应关注的是它们的学习效率。现代人工智能与人类智能之间最显著的区别在于学习效率。虽然我们现在创建的人工智能已经在某些方面表现出色，但在更多的方面仍在持续不断进步。但关键不仅仅在于能否完成任务，更重要的是从有限的指导中快速学习和适应的能力。一个婴儿或小孩如何学会在世界中导航，识别树木、汽车、父母、人和猫？他们使用的标签很少，依靠的是大量未标记的数据进行学习。这是当前人工智能努力学习的方面，尽管我不是说不可能实现，但这确实是两者之间的一个主要区别。

第二个主要的差异再次体现在表现（功能和性能）的层面上。我们似乎在大脑中训练一个关于世界如何运作的模型，我们对物理定律有一些直观的理解，例如汽车不会飞行，飞机不会潜水，苹果会掉下来等等。我们的大脑以我们并没有真正理解的方式训练、学习并记住这些物理定律。今天，人工智能仍然是通过大量冗余地学习各种可能发生和不可能发生的事情来理解这些概念。多数人认为，我们还没有找到一种方法能让人工智能像学习统计数据那样去学习和理解这些背后的基本定律。虽然学习这些定律并非绝对必要，但对于世界的准确预测来说，这种方法似乎更为高效。所以，许多专家和同行认为，这种对世界模型的理解和预测能力仍然是人类智能和人工智能之间的一个主要区别。

如果我可以在结构上补充一点，人工智能中的大多数图像处理依赖于前馈网络，即深度网络仅有前馈连接。因此，在这些网络中，既没有横向连接也没有反馈连接。这与我们早期对大脑的理解，即视网膜、丘脑、皮层之间的逐层前馈处理类似。然而，我们大脑的网络结构包含大量反馈连接。这意味着人工智能虽然在图像处理上效果很好，但使用的方法可能与大脑的工作方式有很大差异。在人工智能领域，管理循环神经网络一直是一个挑战，我们花了很长时间才找到方法，但我们的大脑在这方面却表现出色。

现在，在这方面，我们仍然可以从人类大脑中学到更多。我认为我们应该关注抑制作用。这是现代人工

智能中尚未明确处理的一个领域。最近我们发现，即使与小鼠相比，人脑也有大规模的抑制突触网络。这是一个值得注意的话题，我们应该更好地理解它，也许这会在概念上帮助我们获得更好的人工智能。

Q 我们都知道，神经科学启发了人工智能的早期发展。未来，您认为人工智能会变得越来越像人类智能吗?还是说，它会走上一条相对独立的道路?

Moritz Helmstaedter：在我看来，人工智能是否会与人类智能融合，这是一个非常开放的问题。早期的人工智能确实深受神经科学的启发，我相信神经科学和人工智能之间的相互启发将会持续，但也有可能人工智能会使用我们已有的方法，并在技术方面加速发展。例如，量子计算的使用很可能会开辟出一条独立的技术进步之路，而这并不需要深入了解人脑的知识。就我个人而言，我仍然认为，在我提到的预测性学习、世界模型的编码以及记忆的处理方式等方面，我们还有很多可以从人脑中学到的东西。

Q 我认为ChatGPT就是一个很好的例子，可以说明量的增长会带来质的突破，即所谓的新兴智能。你认为类似的事情也可能发生在脑科学领域吗?比如您和您的同事正在尽可能多地绘制全脑神经元，描述它们之间的联系，你认为这最终能帮助我们破译大脑吗?

Moritz Helmstaedter：我将大量精力投入在利用连接组学来推进人们对大脑皮层和其他大脑功能的理解。因为我有一种强烈的直觉，这将从根本上帮助我们理解大脑是如何工作的，特别是通过筛选神经环路的可变性，我们可以更好地理解感官体验、发育成长、个体发展以及物种的影响。当然，我们需要特别理解人类大脑，但也需要理解病理变化，所以我认为研究神经环路是非常有力的手段。

到目前为止，我们已经得到了积极的反馈。当然，只有在成果诞生后才会知道，这一切努力是否值得。我们现在已经开始积累这类洞察力，即如何对科学发展具有前瞻性。这永远是一个开放的领域，在一条充满未知的道路上，只有未来才能告诉我们，我的想法是否正确。前面我已经指出了抑制的作用，我知道人工智能界对这个问题很感兴趣，并试图从概念上理解神经网络的优势。所以，这显然是一个潜在的有趣的发展方向。

我们仍然需要了解学习和学习规则，我认为网络数据将帮助我们了解，生物学中的哪些存储规则实际上在人类大脑中被大规模使用，这可以为更好的人工智能提供信息。我希望并相信，未来这些成果将从生物学领域再回到人工智能领域，但这是否会发生仍然是一个未知数。

Q 当我们评价某种人工智能的优劣时，通常会通过它是否能完成图灵测试来判断。从神经科学的角度来看，您认为我们是否需要一种新的方法来测试人工智能的性能，尤其是人工智能与人类大脑之间的相似性?

Moritz Helmstaedter：事实上，我是人工智能领域众多观点的观察者和倾听者。显然，领域内有很强烈的观点认为，我们已经掌握了人工智能的关键原理，其余只是技术问题。我会把这个阵营归为Geoffrey Hinton等人，他们甚至认为反向传播是大脑的真实写照。我对此并不赞同。目前的人工智能与我们小时候学习时所做的还有很大差距。但我常用一个例子说明，我们不需要模仿鸟类的飞行方式来实现飞行的功能，就像我们不必完全模仿大脑来构建人类水平的智能。

我们应该理解现象背后的规律。我们可以从大脑的运作方式中获得启发，然后用不同的方式实现这种理解，就像飞机的飞行方式一样。或者至少，我们可以观察大脑的工作方式，然后开始重新创建类似的网络。这很像多层透视最初的工作方式，虽然一开始并未深入理解非线性的作用，但它的功能强大。

目前大多数情况下，我们通过捕捉信息来证明自己是人类。我认为，测试人工智能性能和区分AI与HI（人类智能）之间的相似性已不再那么重要。因为很明显，在各类问题的解决上，它们总是做得更好，而且表现会越来越好。目前的语言模型也显得非常聪明，所以我对这种测试的概念并不确定。坦白地讲，我并不认为这很有趣，因为我认为其背后的机制和人工智能的应用可能性才是真正重要的。当然，我们必须确保我们能控制它，不让它取代我们，但对此我并不担心。

Q 您的工作与人工智能联系如此紧密，三年来您一直在钻研专业知识，所以我想知道跨学科交叉合作对科研推进的帮助是怎样的？

Moritz Helmstaedter：我认为跨学科合作最明显的好处是不同领域的专家能够相互学习，互补视角。在我们实验室中，有来自不同背景的年轻科学家，比如电气工程师、物理学家、生物学家，甚至医学院学生，他们的专业知识覆盖了广泛的领域，包括化学等。我们有各种不同的视角，在一起工作很有趣。

虽然一些同事最初对大脑的了解有限，但他们不断学习神经科学知识，同时我们也拥有解决数据分析问题的绝佳技能。我们聚在一起，互相学习，这就是跨学科的关键标志，我们有共同的目标，这使我们非常明确为何要合作。一起解决问题的过程，不仅富有挑战性，还非常有趣，因为我们可以清晰地看到我们取得的进展和进步。

Q 大众可能会认为脑科学的突破非常缓慢，比如我们怎么还不能治愈一些脑疾病。如果您要选择本世纪最重要的脑科学发现，您会选哪一个？

Moritz Helmstaedter：我完全同意和赞同大众对神经病学或精神病学等神经科学领域在攻克疾病上的失望。但我们得承认它是如此困难。尽管在80年代和90年代，即所谓的“大脑十年”期间，我们在某些重大疾病的治疗上取得了实质性的进展，但神经退行性疾病带来的挑战依然巨大，并且预计在未来几十年内将给我们的社会带来沉重的负担。这就是我们面临的巨大挑战。

我同意当前的发现似乎很少，问题在于为什么会这样。当研究变得极其困难时，我们应该反思是否选择了正确的方法和研究目标。我之所以进行这样的研究，正是因为我相信，神经网络可能是疾病表现的关键层面，这一假设已经存在很长时间，但我们之前无法研究它。现在我们认为我们将能够研究这类问题。至少我们要检验一下，是否有主要的精神疾病与特定的神经回路有关，如果有，是哪些，它又是如何感染的，感染的是回路的哪些部分，哪些信号可能易受感染等。这将有助于我们理解神经退行性疾病。

尽管这是一个复杂且困难重重的领域，其中的因果关系也还存在争议。我能说的最多的是我自己的研究领域。我们发现了小鼠和人类大脑皮层的显著区别，这不是我们预料之中的事情，而是来自数据的真正惊喜。对我来说，这就是一个例子，表明如果你探索全新的领域，用新的方式观察网络，你会发现一些之前未预料到的东西。

来自@Moritz Helmstaedter的问题：

关于世界的知识表征是如何在我们大脑中编码的?大脑中是否存在进行预测编码的神经回路?

我对一个基本问题感到纠结:关于世界的知识表征是如何在我们大脑中编码的,从而让我们能够理解当前发生的事情。特别是预测编码(predictive coding)的概念,大脑中是否存在进行预测编码的神经回路?如果存在,它们是如何构成的?我认为我们所进行的许多思维活动都是基于预测性的,并且依赖于广泛的假设。因此,理解这种处理背后的架构是一个根本性的问题。与此同时,这对神经回路病理学的理解也越来越重要。我们正在积极开展这方面的研究,因为这对我们理解大脑疾病有着重大的影响。(记者:涵棋;编辑:韵珂、存源)

追问专访·杜忆：音乐让大脑返老还童？

嘉宾：杜忆

中国科学院心理研究所研究员。主要的研究方向包括言语感知和理解的认知神经机制；言语认知的毕生发展机制；音乐训练经验对言语加工的可塑性调节机制；以及音乐感知与音乐奖赏的认知神经机制。

扫码查看原文

在人类大脑深处，有一种令人着迷的现象在几个世纪以来一直吸引着我们——音乐和大脑之间的深切互动。从舒缓的摇篮曲到令人振奋的交响乐，音乐具有无穷的力量，可以激发情感，点燃创造力，带给我们不同的感观体验。

科学家们正致力于解开音乐影响人类神经回路之谜，揭示它对大脑各个区域和功能的影响。在本期“追问专访”中，我们有幸邀请到中国科学院心理研究所的杜忆老师，从“抗衰老”的角度，一同探索音乐对大脑的影响。

Q 请简单介绍一下您的研究背景。

杜忆：我本科学的是医学，但我在选择博士专业的时候对心理学，特别是对我们大脑的奥秘很感兴趣，于是转学了心理学的直博。在博士期间我做的是以大鼠为研究对象的听觉研究，包括生理和行为的一些研究。在博士阶段的后期以及博后期间，我开始对人脑以及人脑如何加工语言、音乐这样一些高级的听觉信号更加感兴趣了。所以在我建立了实验室之后，一直在研究人脑对语言的理解、音乐的欣赏，以及相关的神经机制。

人变老之后，大脑的结构和功能会发生一些衰退，其中一个非常影响老年人生活质量的变化就是听力的下降。老年人在街道、饭馆这样嘈杂的环境中，很难听清楚他人的话，这不仅影响老年人的社交，还会影响他们的情绪，甚至导致一些痴呆的早发。所以我们很关心老年人到底如何在这种嘈杂环境下进行言语

理解，是否调用了一些额外的认知机制？

我们2016年发表在Nature Communications上的一篇文章发现，老年人虽然感知觉的加工下降了，但是会更多地调用一些额外的脑区，特别是跟发音相关的言语运动脑区，来主动预测他人下一步将要说什么。这种主动的、自上而下的（top-down）预测，可以帮助老年人更好地对抗自身感知觉的下降，同时帮助他们能够在比较复杂的环境中更好地理解对方的语言。

基于这样的背景，2017年我们发表在PNAS上的一篇研究发现，年轻人，特别是年轻的音乐家，更擅长在一些复杂环境中进行言语理解。我们想追问其背后的神经机制是什么，结果发现，年轻的音乐家具有更强的感知运动整合能力，他们不仅仅依赖听，也依赖发音运动的预测和补偿，以及跨模态的信息整合，来帮助自身在复杂环境中进行言语理解。

通过这两个研究可以发现这两者之间的关联，即老年人可能更依赖于这种高级的、自上而下的感知运动的预测和整合，而年轻音乐家因为有了音乐训练的经验，他们更擅长利用言语的感知运动整合。我们很自然地就想追问，如果老年人经过了系统的音乐训练，是不是也能够提升跨感觉的信息整合能力，从而帮助他们自己更好地在一些复杂环境下进行言语理解呢？

基于这个出发点，我们开展了后续一系列的研究，发现相比于老年非音乐家，老年音乐家更擅长在一些嘈杂的环境下进行言语感知。2021年，我们在Ear and Hearing上发表了相关的行为研究，此后又开展了新的研究，近期发表在Science Advances上。

Q 您提到近期在Science Advances上发表了一项研究。可以介绍一下这项研究的新发现吗？

杜忆：新发表在Science Advances的这篇研究是一个封面文章。我们招募了三组人，一组是年轻的非音乐家，作为对照组，但我们更关心的是另两组老年人——老年的非音乐家和老年的音乐家。

我们让他们进行一个噪音下的视听言语音节的分辨任务，并在磁共振成像仪里观察大脑是怎么去加工的。他们会听到不同的音节，同时会在屏幕上看到匹配的唇动信息，他们的任务就是在不同的噪音强度背景下进行判断——你到底听到的是什么？

我们发现，在行为上，老年音乐家比老年非音乐家更接近年轻人，在中度和轻度的噪音背景下，他们的成绩甚至跟年轻人没有任何差异。也就是说，音乐训练经验能够帮助老年人做到和年轻人类似的更好的成绩。那么，这背后的机制是什么呢？

首先，相比于老年非音乐家，老年音乐家在加工噪音下音节的时候，他们的双侧感知运动脑区的神经加工模式（也叫做表征模式）更接近于年轻人，甚至没有太大的差异，并且要显著优于老年非音乐家。老年非音乐家跟年轻人相比，不同语音的神经区分性明显变差了，他们的大脑已经分不清现在听到的是“ba”还是“da”，但是老年音乐家的神经区分性还很好，跟年轻人是类似的。这就表明，音乐训练经验能够增强大脑的功能，让大脑的功能更接近于年轻人的状态，我们把这个机制叫做功能的保持。

除此之外，我们还发现，老年音乐家比老年非音乐家激活了更多的额顶网络脑区。这些脑区跟工作记忆和执行控制有关，因此老年人能够让更多的高级脑区参与进来，进行行为的代偿。

同时，老年音乐家比老年非音乐家更多地抑制了一个脑区——角回。角回是默认网络的一个核心脑区，它的功能跟自传体回忆、长时记忆有关，因此它会在我们不做任何任务的时候激活，在我们做一些认知任

务时相对被抑制。当我们做一些认知任务、需要关注外部刺激的时候，就需要抑制默认网络脑区，防止它们带来干扰。

我们发现，老年音乐家更擅长抑制该脑区，这表明他们在更加认真地关注外部刺激，抑制其他无关的信息，这种更强的抑制能力，也帮助他们能够达到一个更好的行为成绩。老年音乐家通过额外的调用或是抑制一些脑区来更好地保持行为成绩，这个机制叫做功能的代偿。

这个机制和刚刚提到的第一种机制——功能的保持——是有关系的。我们发现这两个机制之间相互关联。当一个老年人具有更强的额顶网络激活能力、更强的角回抑制能力时，他/她的感知运动脑区的神经表征能力也会更好，更接近于年轻人，这就表明功能的代偿可以支持更好的功能的保持。基于这两种机制，音乐训练经验能够帮助老年人在噪音环境下进行言语感知，并达到更好的行为成绩。

Q 从您的研究来看，演奏乐器可能有助于保护认知功能，来对抗衰老的负面影响，那么一个人需要多频繁地演奏乐器才能达到这种效果呢？

杜忆：在这项研究当中，我们要求老年人在最近三年内保持每周至少训练一个小时的频率。虽然他们自己汇报之前三四十年可能也保持了相当的训练频率，但因为历史原因，我们很难追踪到他/她60年代、70年代的训练时长，因此我们只是对最近的三年有训练频率的要求。

Q 如果之前没有演奏乐器的基础，临时练习的话，有可能产生保护认知的效果吗？

杜忆：这是一个很好的问题。我们现在做的是一个横断性的研究，直接对比有过长期音乐训练经验的老年人和没有长期音乐训练的老年人是否有差异，发现是有的。同时也有很多其他长期或短期的纵向跟踪研究，发现3到6个月的短期音乐训练经验，也可以提高老年人的言语加工能力。

Q 听音乐能起到类似的效果吗？

杜忆：其实我们都希望被动地聆听音乐就能够达到一个很好的效果，但是现在已有的研究结果无法得出这个结论。因为我们的研究是基于主动地进行音乐训练的效果，会更多调用到包括视觉、听觉、运动、情感、注意等多系统，这种多系统的调用能够起到更好的作用。当然听音乐也有它自己的好处，比如情绪的调节、压力的释放，甚至可能提高我们的专注力。

Q 该研究的参与者进行的是视听音节辨别任务，您是否有考虑过做其他类型的任务呢？

杜忆：音乐家和非音乐家其实也做了其他类型的任务，一个任务是听觉工作记忆，比如给你念一串不同长度的数字，然后让你正向或者是反向地把这串数字复述出来。同时还考察了他们执行控制的能力，比如经典的Stroop效应（字义对字体颜色的干扰效应）。

我们发现，老年音乐家比老年非音乐家具有更好的听觉工作记忆能力，虽然我们没有在Stroop这样的执行控制能力上发现他们表现得更好，但是其他人有发现音乐经验能够增强工作记忆、执行控制的能力。

我们当时之所以要做视听言语的音节辨别任务，就是因为每个人都是实时地进行语音加工，而在进行

识别的时候知道辅音、元音和声调是非常重要的。所以音节的辨别任务是语音加工的一个基础单元。

Q 与非音乐家相比，在嘈杂环境下，音乐家有更好的言语知觉能力，您认为这仅仅是由于他们的音乐训练，还是可能有其他的因素在起作用？

杜忆：我们当时选择音乐训练，是因为它其实是一个多系统的训练，相对比较愉悦，同时很多功能都会被提高，比如工作记忆能力、选择性注意能力、执行控制能力等等，包括我们研究中发现的——更强的默认网络的抑制和更强的额顶注意网络的增强。所以，我们认为它可能还是跟音乐训练有关。

但是因为我们做的是横断性的研究，音乐家和非音乐家在很多方面也会有其他的差异，特别是我们没有办法排除这是先天的因素还是后天的因素在起作用。

可能有些人认为这个人之所以成为音乐家，是因为他/她有更好的先天结构和功能基础，或者是其他特质导致其成为一个音乐家。所以我们目前的横断性研究没有办法确认是否有其他因素的参与，特别是一些个人特质可能导致他/她有更好的言语加工能力。要回答这个问题的话，最好是进行一个纵向的追踪研究，才能够证实这确实是音乐训练带来的效应。

Q 是否有其他的研究探讨过类似的话题？如果有的话，和您的研究结果有没有异同？

杜忆：其实有不少人考察过音乐训练能不能帮助老年人更好地进行言语知觉，他们在行为层面上已经发现了一些类似的结果。我们的新研究跟他们不一样的是，我们能够在大脑的机制上，同时发现功能的保持和代偿两种相互支撑的认知神经机制，然后从机制的层面更好地回答了为什么他们在行为成绩上有更好的表现。

Q 该研究对衰老和大脑的健康有哪些启示呢？研究结果是否有潜在的临床应用价值？比如说开发有针对性的干预措施或者治疗方法，以改善老年人的言语知觉能力或者其他的认知能力。

杜忆：我们觉得音乐训练能够起到脑保护的效应，让大脑更接近于年轻人的工作状态。虽然我们做的是个长期的音乐训练的研究，但未来也想去考察到底多短的时间和频率就能达到类似的效果。这也提示我们——不管是什么年纪，如果你能尽快地开始学习一门乐器，甚至开始唱歌，可能都能够让大脑衰老得更慢。所以我们目前提倡大众，如果有条件的话，尽量地学习音乐，开始玩起音乐。对于临床价值而言，我们想在一些老年的认知衰退疾病上，比如说AD（阿尔茨海默病），是否能开展一些音乐疗法，改善他们的记忆能力和语言能力。

Q 音乐类游戏有可能会起到这样的效果吗？

杜忆：这是一个很好的方式，比如节奏大师这种训练节奏感的游戏，其实也是非常好的训练方法，我们也把它叫做数字疗法。将音乐作为一种数字疗法，在未来有很好的发展前景。因为对老年人而言，这种游戏化的训练肯定更加有趣，更容易让他们投入其中。（采访&编辑：Lixia）

来自@杜忆的问题：

音乐为什么让我们很愉悦？音乐的奖赏机制是怎么发生的？它背后的神经环路、神经递质的机制是如

何进行多系统的协同的?

在动物的研究中,动物听音乐也会很愉悦吗?如果我们想研究动物的音乐奖赏机制,我们应该怎么去研究?有什么新的研究范式或者是成像的方法,能够让我们回答相应的问题?

追问专访·Indre Viskontas：

为什么人的审美有别?

嘉宾：Indre Viskontas

神经科学家，歌剧舞台导演，科学传播者。旧金山大学心理学副教授和创意大脑实验室（Creative Brain Lab）主任。

扫码查看原文

Q 为什么人的审美会有差异?大脑是怎样处理审美过程的?

Indre Viskontas：我们的审美有时候仅限于“漂亮的东西”，但我们会发现“漂亮”其实很难定义。比如让不同的观众给他们看到的图片打分，可能会得到不同的答案。一些美丽的事物是有规律可循的，比如，自然界的分形图案对大多数人来说都很美。但另一个规律是，我们越了解我们听到或看到的东西，就越会觉得它美丽。举个例子，刚开始听前卫爵士乐的人在听到那些没有清晰重复片段或旋律的乐曲时，他们不知道自己在听些什么，对他们来说这可能并不美；但一个爵士乐迷会确切地知道该怎么听，他们会从中体会到爵士旋律的美，因为他们对此了解得更多。

在如今这个个性化医疗和可穿戴设备普及的时代，我们开始用人的生理反应来定义美的体验。例如，与我合作的一家公司开发出了一种机器学习算法，可以从测量心率变异性的可穿戴设备中获取信息。当播放音乐时，它会追踪使用者听到的乐曲，并提取对心率变异性有特定影响的音乐特征。我认为这是未来的方向，虽然审美是如此主观，但我们或许能够发现某些生理标记物来表示我们正在体验美。

在音乐界，我们知道“皮肤寒战”的存在，它是欣赏音乐时出现的一种皮肤起“鸡皮疙瘩”的现象。为此，我们也可以设计fMRI研究，在受试对象听音乐、出现皮肤寒战的时候，我们知道，这是因为他们正在体验震撼的音乐片段。这个时候，看看大脑中发生着什么，就有助于我们理解审美的神经基础。

Q 听起来，您认为复杂性是美的必要元素，比如层次和分形图案；您如何看待这一点?

Indre Viskontas:我认为光复杂性本身还不够。我将审美体验与复杂性之间的关系视为一种“倒U形”的曲线关系。如果某件事太复杂以至于我们无法理解,我们也不会被它打动。但是当事物的复杂性刚好留给我们一些空间去发现,而不是过于直白地流于表面,那么我们也会保持兴趣并有动力去寻找这些有意义的联系。

例如,在爵士乐的例子中,对于一个没有爵士乐经验的人,也许只是一小段爵士乐即兴演奏就已经听上去非常棒了。对于我来说,网上的独奏钢琴乐过于简单,我觉得我从中得不到任何东西;但是我父亲反倒觉得它们听起来舒缓和平静。

Q 能否谈谈您对正念的看法?

Indre Viskontas:我不是正念方面的专家,但据我所知,有很多证据表明,冥想作为一种练习,可以帮助人们更有目的性地在不同的大脑状态之间切换。所以如果你开始做白日梦,陷入了思考那些不适合我们的事,或无法将注意力集中在手头任务上时,冥想是一种学习控制你的思想去向的方法。我们可以通过训练来加强不同大脑区域之间的联系。至于哪些特定类型的冥想有哪些具体好处,我还不那么了解,但已经有很多证据表明,它对人的自我健康幸福能产生可衡量的影响,可以帮助人们选择让自己的思想集中在哪里,而不是总被令人分心的因素侵扰。

Q 人们的情绪状态是如何影响他们对外部世界的看法的?例如,关于临床抑郁症的研究就发现,非抑郁症患者对世界的看法可能与其他人不同。

Indre Viskontas:我能肯定有很多证据表明,情绪状态会影响我们将感觉转化为主观体验的过程。比如大家谈到的“抑郁现实主义”,即处于抑郁状态的人往往会看到世界的“真实面目”。虽然这方面的一些研究结论未能被重复论证,但可以确定的是,当体验一种情绪时,你的凸显网络会给世界打上标记。情绪状态可以通过多种方式影响我们所看到的事物;但是,正如您可能经历过的,有一些方法也可以通过经验来操纵我们的情绪。

导致抑郁症状出现的原因有很多,其中一些甚至是基于大脑的解剖学变化。例如,帕金森病与抑郁症的共病率非常高;帕金森病中,制造多巴胺的细胞出现了病变,而多巴胺是一种参与动机、运动和奖赏的神经递质;患者脑中多巴胺耗竭了,所以可能同时出现运动障碍和情绪障碍。所以有着许多不同通路可以导致临床抑郁症,这些不同的通路也会影响人们对世界的看法。

因此这个问题不能一概而论。我们的五种感官绝对不是独立的,它们是完全相互依存的。事实上,我们所看到的会受到我们所希望体验到的感受的影响,反之亦然;有一些人还有“通感”,这是两种或多种感觉的交叉。我认为,无论是情绪状态还是环境等其他因素,都会影响我们的体验以及我们最终的行为方式。

Q 我很好奇您对审美和感知与大脑奖赏中枢之间联系的看法,以及您对“偏见”这个概念的看法。

Indre Viskontas:我们的大脑常常为了把尽少量的数据转化为丰富的体验而走捷径。这种做法的后果之一,就是偏见。把大脑想象成一座冰山,而意识只是冰山一角,大部分工作都发生在我们意识无法触

及的“海平面”之下。

在直觉和大脑所走的捷径中，肯定存在偏见。我有一个朋友研究“偏见”这个概念，他认为，即使我们在对事物的最初反应方面可能存在偏见，但我们的前额叶皮层有一个系统可以超越最初的直觉并影响我们的行为。所以，即使我们一开始有带偏见的观点，这并不意味着我们会以有偏见的方式行事。当这些偏见影响到其他人并导致歧视时，我们需要意识到这些是偏见。但是偏见是大脑的一个特征。确认偏差就是一个很好的例子，当我们确立了某一个信念或观念时，我们会倾向于寻求确认和支持我们假设的证据，而不是寻找推翻它的。

Q 最近您提出了视听整合的研究课题。您对失去这两种感官之一的、不同年龄段的人进行过研究吗？这将如何影响他们对不同刺激的感知和理解呢？

Indre Viskontas：我想告诉您最近的一项研究：我有一个朋友，克里斯托弗·贝利，是世界卫生组织（WHO）艺术和健康部门的负责人。他患有晚期青光眼，正在慢慢失去视力，即将完全失明。他的评论之一（我认为非常深刻）是：光从物体上反射回来，但声音会穿过它们。

他告诉我，他基本上学会了用回声定位，学会了使用拐杖和敲击，根据他所听到的声音来观察空间。据他描述，在经过训练后，当他走在街上听到汽车声越来越大时，他会本能地远离声源并向楼宇的方向靠近。但是他的教练说，‘你需要克服这种本能并且尝试让声音维持在相同的音量上，这样就可以持续走在路边，而不会来回撞上建筑物’。在训练过程中，他学会了非常有效地做到这一点。有一项研究表明，即使是非盲人也可以学会这项技能。大脑对我们通过感官体验到的经历的可塑性是相当了不起的。

另一个例子是，我们倾向于认为一些掌握了高级技能的人做的事是未受训练的人不可能完成的，例如弹钢琴。然而，通过培训，即使是年长的成年人也可以学会“阅读”音乐，以正确的方式练习弹奏。所以我认为这取决于我们如何利用时间，对事情的积极性以及培训工具的有效性。

Q 您提到有很多我们看不到的东西、实际上存在的情况。有时候，我脑海中想着一个人，然后突然间我的手机提示我刚收到他们的短信。我在想，这其中是否存在一些看不到的连接，或者只是我的想法？

Indre Viskontas：我的回答可能会有些扫兴，有时候我们的手机发出声音，显示我们收到心中所想之人的短信。但可能也有很多次你也没有收到这样的短信，或也没有想到他们。所以，本质上这就是“确认偏差”，我并不相信这是一种超感官的知觉。但我确实认为我们可能会在意识之外追踪一些行为模式，这是我们拥有直觉的原因。我们可能会对别人形成直觉，这是因为我们潜意识里在追踪他们的某些方面，如行为、动作、步态等，形成潜意识里的情绪指标。

Q 当您谈论人们的艺术体验时，我想知道艺术体验和一些其他因素有关吗？比如，共情水平？

Indre Viskontas：我会用音乐家和音乐作品举例。如果一位钢琴家向其他受过古典训练的钢琴家演奏乐曲，那么你会看到，听众的大脑活动几乎是表演者的镜像；而且听众对钢琴的理解越深入，你越能看到表演者和听众之间大脑的相似之处。所以从某种意义上说，这就像共情，即通过在自己大脑中的镜像来理解体验。你需要了解这件作品，刚好已有的知识储备能让你进行这种镜像体验。也有人进行过这样的研

究：他们观察大脑的活动，并将其与对象正在听的内容相匹配。你可以看到，随着音乐情绪的上升，一些大脑区域更加投入，之后又逐渐下降。

我认为同理心是同情心的一个组成部分，但还不够。当我们开始共情时，我们会想：嘿，如果我站在那个人的立场上会是什么样。当然，我也认为你可以在不共情的情况下获得艺术的体验——你可以拥有一种纯智力上的艺术体验。马塞尔·杜尚的小便池就像一个例子：我没发现自己对小便池有共情心，但这个作品也改变了艺术世界，对吧？或者那个将香蕉用胶带贴在墙上的作品，它没有让我共情，但很有趣。我想我们能区分纯智力方向的艺术体验和情感方面的艺术体验，但我认为我们大多数人实际上都喜欢共情的体验：大多数最成功的、票房最好的电影和艺术作品往往具有情感成分。（责编：韵珂）

追问专访·汪小京：

为计算神经科学培养跨学科人才

嘉宾：汪小京

纽约大学神经科学教授，斯沃茨理论神经科学中心主任。研究重点为认知功能的脑机制，尤其以在工作记忆、决策的神经机制。他也是“计算精神医学”的创始人之一。

扫码查看原文

这是汪小京作“计算与认知神经科学夏校”开学典礼演讲的第14个年头。站在苏州冷泉港学术中心的讲台上，他简单梳理了计算神经科学这门新兴学科的发展历史。讲到Hopfield用物理中的自旋玻璃（spin glass）模型来解释大脑联想记忆系统的经典论文时，他停顿了一下，询问台下的学生是否阅读过。教室里约一半人举起了手，汪小京欣慰地点点头。显然，这个比例超出他的预料。

30多年前，正是Hopfield的这个富有科学美感的研究吸引了他，促使这位理论物理学博士转向神经科学领域，用物理学研究复杂系统的思维方式和数学模型，帮助揭秘大脑。

那是一条艰难而生僻的小径，若无在美国参加一次夏季学校（后简称夏校）的奇缘，汪小京觉得自己当时很可能会坚持不下去。也正因如此，2010年，他创办了国内首个计算神经科学夏校，将这份追寻科学边界的坚定和热情传递给下一代年轻人。

教室明亮的落地窗外，独墅湖碧蓝色的波光消解了江南盛夏的暑意。而学生们不断向汪小京提问的热情却始终没有褪去。为了解这个夏校背后的故事，《追问》媒体两次造访苏州冷泉港，深入课室和生活区观察，并等到夏校尾声时，以视频连线的方式采访了这位卓越的学者和教育家。在他来看，夏校创办的十多年，也是中国神经科学高速发展的十多年。汪小京对于中国学生身上所发生的巨大变化甚有感触，有时惊讶于他们掌握的信息之多，探索的热情之大。计算神经科学是一门非常典型的交叉学科。对于这些和他当年一样走进科研“十字路口”、却面临一个更“卷”时代的年轻学子们，汪小京分享了自己的想法。

▷图注:汪小京在2023年“计算与认知神经科学夏校”授课 图源:冷泉港亚洲

Q 按照《追问》的惯例,请您先介绍一下您的研究领域。

汪小京:我研究的是计算(理论)神经科学,这个学科相对较新,只有三十多年的历史。大家都知道,理论在物理学中探索宇宙奥秘起着不容置疑的重要作用。类似的情况也正在神经科学领域发生:我们越来越需要用理论和数学建模去理解大脑的工作机制。实验研究和理论模型相互结合,可以帮助我们更好地解析大脑认知、学习、记忆等复杂功能的奥秘。

Q 您觉得计算神经科学有趣的地方体现在哪里?

汪小京:计算神经科学的有趣之处在于它的跨学科性质。这意味着涉足该领域的人来自各个不同的学科背景,包括数学、物理学、计算机科学以及信息科学工程等。

数学背景的人能够在模型的构建和分析上作出贡献,尤其动力学系统理论对了解神经网络起重要作用。信息科学背景的人则可能会更多地从信息处理的角度来考虑问题。比如,BMI(脑机接口)将脑电信号转化为机器可以理解的指令,需要运用信息处理的方法来信号解码并转换成指令。物理学出身的人,可能更倾向于将神经科学视为一个开放型的复杂动力学系统,这意味着将大脑看作一个自组织的系统,在这种视角下通过大脑的非线性行为和复杂性的研究来理解神经网络的动态行为和功能。

Q 当年您从物理学博士转向研究神经科学,是什么契机让您做了研究方向的转变?

汪小京:有时偶然的因素会改变人生轨迹。我转入神经科学领域是在1987年。我认为有两个原因,一个就是我在比利时布鲁塞尔读博士的时候,研究所所长普利高津教授兴趣特别广泛,用统计物理和动力学系统来研究各种各样的系统,包括生物系统。他对我产生了一定的影响,开始让我对大脑产生兴趣。

编者注：伊利亚·罗曼诺维奇·普里高津（俄语：Илья Рома нович Приго жин，英语：Ilya Romanovich Prigogine），比利时籍俄罗斯犹太裔化学家、物理学家，1977年诺贝尔化学奖获得者，非平衡态统计物理与耗散结构理论奠基人。他把将近一世纪前由克劳修斯（Clausius, R.J.E）创立的热力学第二定律扩大应用于研究非平衡态的热力学现象，开拓了一个过去很少受人注意的崭新领域，被认为是近二十多年来理论物理、理论化学和理论生物学方面取得的最重大进展之一。

第二，那个时候复杂系统的物理研究十分活跃，其中模型之一叫spin glass（自旋玻璃）。玻璃看起来特别简单，却不是一种简单的晶体，其结构非常复杂。Hopfield在80年代用玻璃这种系统的物理理论来研究大脑的记忆。他当时的模型非常有意思，吸引了很多物理学家的注意。自此很多人从物理领域出发，开始思考如何用这种复杂系统的物理理论来研究大脑，其中就包括我。Hopfield的那篇文章内容也是促使我决定进入这个领域的一大因素1。

编者注：自旋玻璃（Spin Glass）是一种物理学上的概念，它通常用于描述复杂系统中的一类特殊现象。这种系统通常是由大量的微观组件（例如原子或分子）组成的，而这些微观组件有着自旋（Spin）的性质，类似于磁矩的方向。

计算神经科学领域的正式起步可以追溯到1988年。那一年，Science杂志上发表了一篇名为《Computational Neuroscience》的文章，可以说是计算神经科学领域的宣言2。也就在1988那一年，第一个计算神经科学暑期学校在麻省的海洋生物实验室（Marine Biological Laboratory）举行。我被录取参加了第一届暑期学校，为我进入计算神经科学提供了机会。

参加计算神经科学的暑期学校给我带来了许多收获。首先，它帮助我了解生物领域。在学习计算神经科学时，我逐渐认识到大脑的奥秘。仅仅建立抽象数学模型是不行的，要深入了解神经生物学，比如单个神经元如何工作的、不同脑区的功能、不同物种的行为。其次，在暑期学校我认识了一些神经科学的领军人。第三，我参与学生与老师一起讨论问题，包括辩论，潜移默化形成了以后自己研究团队的文化。

Q 您从何时产生了自己开设计算与认知神经科学夏校的想法？为什么首先将地点选定在了中国？

汪小京：2010年，我在科学网发表了一篇文章介绍计算神经科学3。当时此领域在亚太地区还很新，为了帮助更多年轻人从数学、物理等专业转行，就需要提供一个了解计算神经科学的机会。想到自己受益于海洋生物实验室的暑期学校，我和刚回国的吴思老师等人商量，考虑在中国建立一个类似的国际暑期学校，起名Computational and Cognitive Neuroscience (CCN)。

当时计算神经科学主要关心的是初级感觉（比如视觉）和运动行为的脑系统，高级认知的神经机制研究是本世纪才发展起来的，而我个人正好在此新领域，所以决定CCN聚焦于认知，这是它与领域中的其它夏校最大的不同之处。

▷图注：自2010年起，汪小京在苏州冷泉港开设“计算与认知神经科学夏校” 图源：冷泉港亚洲

另两位原主办者是Zach Mainen 和Upinder Bhalla。当时，冷泉港实验室（Cold Spring Harbor Laboratory）在苏州的亚洲中心（CSH-Asia）刚刚设立，这是一个国际平台，CCN成为其第一个暑期学校。它吸引了来自世界各地的学生参加，学校的师资力量一流（第一年就有7位美国科学院、英国皇家学院院士来任教）。这样的国际化设置让中外学生们有机会相互学习和交流。十几年来参加CCN的学生中很多已成为领域的新领军人，比如麻省理工学院的终身制助理教授（tenure track assistant professor）杨光宇、耶鲁大学的终身副教授Alan Antecivic、上海交通大学的正教授李松挺、Meta的研究员Jean-Rémi King、OpenAI 的研究员 Francis Song等等。CCN的成功是同行们一起努力的结果，我十分感谢吴思、Zach Mainen、Upi Bhalla、今年的主办者李松挺、黄橙橙、杨光宇、Dora Angelaki、Christopher Honey和教学助手（teaching assistant）们的共同努力，以及主办单位的支持。

Q 在夏校的创办初期您有哪些想法？夏校有哪些特点？

汪小京：如何帮助年轻人进入神经科学领域，并学会提出有意思且可研究的问题。

第一，在神经科学领域，选择研究方向至关重要。例如，我如果说要了解意识的脑机制，但没有具体的假设和研究假设的方法，就只是空话。过去的神经科学主要关注感知和运动，这些方面可以用相对简单的动物模型进行研究，比如青蛙或果蝇。而对于认知这样更为复杂的功能，需要聚焦于有挑战而具体可用实验和理论来研究的问题。

第二，接受各个学科的综合教育。夏校的老师来自生物、物理、数学等领域，有分子细胞、网络系统、行为各层次的脑科学专家。学生有难得的机会了解不同研究方法，同时扩大了研究视野，对他们以后的成长有十分大的影响。

第三，我们重视学术文化的培养。在暑期学校中，我们邀请国内外的科学家来不仅仅做讲座，也与学生

进行深入的交流和互动。他们在学校内停留的时间较长，与学生一起就餐、课外体育活动。学生有机会平等地与老师讨论，逐步习惯表达自己的观点，敢于挑战学术权威、即使面对院士、知名科学家。同时，与世界各国的科学家结识也让学生在未来的学术道路上有更多的选择和机遇。

Q 您认为十几年来中国在跨学科人才的培养上有何变化？

汪小京：与十几年前相比，现在的学生在进入神经科学领域之前可能已经具备了更多的知识、思想也较成熟。另外，学生们对问题有着强烈的好奇心和求知欲。他们热衷于了解新的动态和前沿进展，我觉得这两点是非常可贵的。另外，现在年轻人关心的一个热点当然是AI，但今天的AI系统与人脑仍相距甚远。我们对人的智能生物机制了解还很不够，基础研究需要长期的努力。有人激情在技术开拓，有人更潜心基础研究，也有人热衷于用神经科学来发展精神医学。年轻人寻找自己真正一辈子热爱做的事情，每个人的答案都是不一样的。如果是研究一个自己认为很有意思的问题，就可以静下心来，不必抢赶时髦。

Q 总结来说，培养计算神经科学人才的特点与难点是什么？

汪小京：首先是跨学科。数学或物理出身的人，在夏校可学习脑生物、了解神经实验研究。反过来，做生物实验的人的数学背景通常不是很强，通过暑期学校让他们熟悉数学模型。比如，在本次暑期学校中，有位学生是学心理的，他感叹道这次暑期学校让他学到了很多数学模型的方法。

其次是机构，我们国内用于支持跨学科研究和教育的机构较少。夏校办了十几年相当独特和成功，大家对计算神经科学发展和人才培养的必要性也有共识，但长期基金支持却不是一帆风顺地有保障。这里要特别感谢陈天桥先生和他的夫人，他们创办的天桥脑科学研究院（Tiaoqiao and Chrissy Chen Institute，TCCI）、James Simons Foundation、DeepMind的支持是非常重要的。

Q 夏校的存在可以弥补高校日常教育中哪些不足？

汪小京：一般大学的学科分的很细，并不有利于跨学科的研究和培养跨学科的年轻人。当年从零开始建上海纽约大学，作为 Provost 我借此机遇筹划并与大家一齐实施了一个新的模式：研究不以系为单位，而围绕着数个跨学科中心来发展。同样，CCN这样的夏校是完全跨学科的，与高校日常教学起不同的作用。吸引各领域的人才一起攻坚脑科学，十分值得关注和支持。（记者：涵棋；编辑：韵珂）

来自@汪小京的问题：

怎么定义、定量人的思维和智力？这个问题的答案会告诉我们今天的AI与人的差距。具体地说就是如何通过神经科学的方法去理解思维和智力的神经机制。

追问专访·吴思：

打开人工智能的智慧之门

嘉宾：吴思

北京大学心理与认知科学学院院长，教授麦戈文脑科学所研究员，定量生物学中心研究员北大-清华生命科学联合中心研究员。

扫码查看原文

吴思前段时间开网课开出了演唱会的感觉。——刚开放报名，400个名额就被抢空了。

“其实这门《神经计算建模实战》课非常学术前沿。”他掩盖不住兴奋，冲出镜头去身后的书架上取了一本深蓝色封皮的教材书。封面上依稀写着：“让天下没有难建的神经计算模型。”“考虑到很多人在第一期没有报上名，为了让更多的人有机会学习这门课，接下来我会筹办第二期、甚至第三期。”他补充道。

用数学和计算的方法给大脑建模，是一个充满挑战但也极具魅力的思路 ，其中最著名的故事之一，当属受到大脑真实神经元网络启发的深度学习网络了。不过，吴思不会被外界的鼓吹给轻易蒙骗。他认为，AI只是计算神经科学的一个副产品。它是一条路径，却不是终点。在他看来，还有远比AI更重要、更根本的问题尚未解决——理解生命的本质、意识思维的本质。

这份冷静认知的背后，源于他自1995年物理学博士毕业后转向AI研究，见证过的潮起潮落、寒暑交替。吴思深知，智能的基本原理仍有待挖掘，在抽象概念、记忆、情感等领域有诸多缺失。而大脑是宇宙中一个鲜活的智能样本，用数学模型诠释大脑的工作原理，将为AI的发展奠定基础。“研究清楚大脑表征的机制之后，构建出的智能系统才会更像真正的人类。未来的人工神经网络的发展应该考虑引入这一套东西。”

比起一味追赶Geoffrey Hinton等“教父级人物”掀起的当前AI热潮，吴思更希望新一代科研人员能学习他们在无人问津的寒冬中默默耕耘、契而不舍的精神。

“大脑就是这样的网络，没有理由不工作。”他经常用Hinton 的这句话鼓励年轻人。

Q 按照《追问》的惯例，请介绍一下您的研究领域。

吴思：我的研究领域是计算神经科学和类脑计算，关注的方向是用数学方法来阐明大脑信息加工的基本原理，并在此基础上发展一些人工智能的算法和模型。

Q 您本是物理学出身，最终却成为了一名神经科学家。结合您的自身经历，请问量化学科是如何与神经科学发生交融的？如今越来越多计算机或数学背景的学者开始进入这个交叉领域，这对于神经科学的发展是一种必然趋势吗？

吴思：这是必然的趋势。我们时常将神经科学与物理学进行比较。早期的物理学以实验和观测为主，到了后期，出现了理论物理学，代表人物如牛顿和爱因斯坦等。

神经科学也有着类似的发展历程。早期，研究者主要依赖数据和实验观测去理解大脑，而后他们也像研究物理一样，试图用量化的方法归纳出大脑工作的基本原理。这样的研究方式会让我们对脑科学有更深刻的认识。

Q 您是从何时开始对神经科学领域产生兴趣的？

吴思：兴趣总是自然而然发展而成的。在我们那个年代，有这么一句话：学好数理化，走遍天下都不怕。物理学可以说是上个世纪最重要的学科。在硕士和博士阶段，我逐渐意识到物理学的发展已经趋于成熟，剩余有待开拓的理论方向已相对有限，所以我想进入一个新的且自己更感兴趣的领域。于是，我首先进入了人工智能领域，人工智能在当时仍处于发展的寒冬，因此我就思考能否从大脑的功能中获取一些灵感来发展人工智能。因此，我就很自然地转到了脑科学的方向上。

也许大家会觉得（从物理学转向神经科学的跨界之大）令人惊讶， 其实这是非常自然的。物理学利用数学方法研究大自然规律，而计算神经科学是用数学方法研究大脑的工作规律，从方法论上看其实换汤不换药。在计算神经科学领域的研究者中，早期有2/3的人都是学物理的，也许他们转变研究方向的考虑也与我类似。

Q 人工智能的寒冬大概是从什么样子的？

吴思：1995年我获得了博士学位，在硕博研究生期间我学习了理论物理和广义相对论。在学习过程中，一些偶然的机会让我接触到了关于人工智能的书，与物理相比，我感到人工智能太有意思了。由此我开始转做人工智能。那时正是人工智能的寒冬，大家都难以理解那个寒冬到底有多寒冷。在那个时候，如果我要说我是做人工智能的，大家会以为我是“骗子”。所以那时候我的简历里从不会出现“人工智能”。之所以出现寒冬主要是由于那个时代缺少大数据和算力，不像今时，人工智能在那时还无法产出一些亮眼的东西，整个社会对人工智能的发展是缺乏信任的，认为没有什么前途。

但我自己对智能是感兴趣的，也不会因人工智能的寒冬对人工智能本身失去兴趣。很自然地，我联想到大脑就是宇宙中的一个鲜活的智能样本，通过学习大脑，然后再做人工智能，不就是一个很自然的想法吗？大概在2000年，我在日本理化学研究所的时候，就开始有意识地从人工智能转到计算神经科学，目的是研究清楚大脑是如何产生智能的，为我之后在人工智能方向上的研究奠定基础。

Q 那个时候卷积神经网络已经存在了吗?

吴思:是的。在上个世纪八十年代,卷积神经网络最早的版本是由日本科学家Kunihiko Fukushima提出的Neocognitron(新认知机)1。同一时间,Yann LeCun已经在研究卷积神经网络,但当时他在整个科学界很不受待见。他的研究成果没人买账,论文也发不出来。因此,我很佩服诸如Geoffrey Hinton和Yann LeCun这些人工智能科学家,他们用超强的毅力在这条艰难道路上坚持了下来。

Q 那时您与他们有过学术层面的接触吗?

吴思:那时并不是属于他们的高光时代,那时的他们还不是学术巨星。在一些学术会议上我碰见他们,他们都是默默无闻的。那时最耀眼的是做SVM(support vector machine,支持向量机)的研究者,如Bernhard Scholkopf 和 Alex Smona等青年才俊,彼时的SVM有点像是现在的深度学习网络,在当时在整个应用层面将人工神经网络击败了。因此,那个时候的学术明星是那些做支持向量机的研究者们。

Q 人工智能未来的发展是趋于独立,还是会越来越像人类智能?

吴思:AI是个大的领域,卷积神经网络、深度学习网络和如今新的Transformer模型仅代表了AI的一种发展途径。例如卷积神经网络在早期受到脑的启发,然后慢慢脱离脑开始走向工程应用。我们在做工程应用时,总是以性能(performance)为指导,并不会因为这些应用没有遵循大脑的机理就不去使用它。

所以实现AI应用的一条路径,就如同现在的卷积神经网络、深度学习网络或 GPT,它们都是从原来的人工神经网络慢慢脱离生物,更强调以工作性能(工作得好不好)为原则。

但并不能说这些就代表了人工智能的全部。如果从人工智能“发展人造的智能”的定义出发,AI还有另一条发展路径:学习大脑。目前以深度学习网络为代表的AI还无法实现很多人类高级认知功能,那么这条道路仍值得探索,这条路可能会更复杂,实现突破可能更晚一点,但不代表这条路就不存在了。

Q 破解AI和大脑这两个黑盒,在方式上有何相同和不同之处?

吴思:这个问题比较复杂,我会慢慢说,因为我对这个问题的认识也在不断的变化。

人工神经网络的训练模式决定了它是一个黑盒:比如输入端为大数据,设定输出端是物体的标签,随后使用一个反传算法去优化中间的参数,训练好后的网络就可以工作了。但大家并不清楚为什么会这么工作(中间到底发生了什么),所以说它是个黑盒的概念。由此,人们提出了可解释性AI,一些具有数学或计算机背景的科学家会试图用数学工具去破解这个黑盒。

反观人类大脑,也是一个黑盒。计算神经科学家们不就是在研究大脑是怎么工作的吗?所以说,我们也在破解我们(大脑)的黑盒。那么这个黑盒是怎么破解的?一种做法是通过动物实验。神经科学家发展了多种多样的模式动物,比如线虫、果蝇、斑马鱼、老鼠和猴子等。我们之所以研究它,是想从中得到一些人脑的工作机理。这些实验动物各有优缺点,以斑马鱼为例,它是透明的,因此它的神经活动可以看得比较清楚。实验神经科学家会通过各种各样的实验,比如若想搞清楚生物视觉系统是怎么工作的,就会给它(实

验动物）一些视觉刺激，然后记录对应脑区的神经元活动，根据刺激和神经元活动的数据来反推信息编码的原则是什么（或者神经活动包含了哪些刺激信息）。这更像是一种以数据驱动的研究方法，通过实验观察神经元的活动，试图了解它的原理，来打开黑盒子。

我觉得，AI和计算神经科学或神经生物学之间存在一种互补关系。比方说，从神经生物学的研究中，我们理解了大脑的视觉系统原来是这么工作的，那么这些发现是否提示 ，经训练的人工神经网络也拥有类似的工作原理？有些科学家利用多层卷积网络来模仿视觉系统，通过有监督的分类任务训练好后，发现其神经元活动已经接近于我们视觉系统中的神经元行为，由此可以推断，从优化的角度要完成多物体的识别，也许就需要生成一个共同的representation（表征），这不论是在生物大脑还是人工神经网络都是适用的。这，就是一种相互借鉴的研究方法。

以上所介绍的是比较传统的观点。刚才提到，我对这个问题的认识是在不断变化的，特别是在ChatGPT问世之后，大家也不是很清楚它是怎么工作的，但它却表现出了这么高的能力。这也促使我思考，我们一味的追求什么都可解释——所谓彻底打开这个黑盒子——是否是一个不切实际的想法？因为大脑就是一个超级复杂的系统， 就像ChatGPT一样，在有了一个输入之后，它就按照一种规律，生成出一些句子或者一些话，然后你仿佛觉得它好像有智能了，对吧？

我们试图打开黑盒，要去解释它为什么能够产生这种行为的目标，是我们人类的想法。至于我们能否实现或者能否彻底实现，其实是不一定的，即不一定存在这种可解释性，或者说如果我们真要追求可解释性，也许要发展很多新的数学工具，在某些概念上有新的重大突破，才能有一天把一个这么复杂的系统给搞清楚。在这方面其实我也没想清楚，因为我自己的认识也在不断的更新。AI的最新发展对我的很多观点也造成了冲击，我要重新思考一些过去自以为是天经地义的事情，现在我发现这些（看似自然的）事情也不是那么回事了。

Q 有些学者认为，ChatGPT可能是一个由量变引起质变的产物。同理，如果我们能绘制足够精细的大脑图谱，去收集足够多的数据，是否也能涌现出一个比以往AI更强大的智能。您对此怎么看？

吴思：我觉得不是那么回事。将什么都归类为涌现，这是一种回避问题的做法。不能因为描述不了一种现象，就用一个简单的涌现来解释。但可以肯定的是，系统需要足够复杂才会产生一些涌现行为。至于它产生的原因，我们其实还缺乏很好的数学工具或概念去描述它，所以导致了今天的困境：很多东西解释不了。随着科学的逐步发展，我们发明了这样一个超级复杂的系统，但它也许尚不能够用现在已有的简单数学工具来描述。这就需要我们发展一些新的数学工具，或者新的理解方式，这可能才是问题的关键。

中国古人观察到天体运动是特别复杂的，但古人没有试图用一个共同的法则去解释它，而是想象出一个住着玉皇大帝的天宫的复杂神话故事。但是牛顿意识到，也许天上的天体运动和我们地上的物体共享一个特别简单的法则，这就是他在概念和思想上的突破。他沿着这个路子就提出和发展出了“万有引力”的概念。

为了描述力的概念，他发展了一个经典的数学工具：微积分。在那个时代，如果牛顿直接告诉我们“力”的概念而没有去预测（解释）什么东西（现象），大家会觉得他是在胡扯，没人相信他。而他不仅发展了这么一个理论框架，并能够解释很多东西，大家就会觉得牛顿说的是对的。这个时候（被牛顿理论说服的）我们

就开始强迫自己去学力的概念和微积分。最后我们就慢慢接受，并觉得这样的解释是理所当然的。我认为对生物智能和超级复杂的AI系统的理解，可能需要一些思想观念上的突破。

Q 今年，您与几位学者共同发表了“AI of Brain and Cognitive Sciences: From the Perspective of First Principles”2，文中提出了大脑与认知科学之于人工智能的六大首要原则。您为什么觉得它们是首要的原则？

吴思：首先这六个原则不是我一个人提出来的。在北京有个智源研究院，这篇文章是AI认知神经基础方向的智源学者们指导博士后共同撰写的。我负责的是其中的“吸引子网络”章节，因为吸引子网络是我自己长期研究的方向。我想谈一谈为什么说吸引子网络是一个基本法则，这也是目前的AI框架中所忽略的一点。

大脑是由大量神经元连接所组成的，是目前已知的宇宙中最复杂的动力学系统。我们所有的感知和行为都是因为我们受到了外界刺激，或者大脑内部产生了某种活动，促使我们的动力学系统（即大脑）进行演化，从而产生行为。所谓的动力学系统演化就像神经元相互作用所引起网络动态的变化。有这样一种状态，所有邻近状态都汇聚于它，那么这个状态就被称为吸引子。吸引子对应者网络能量空间中的局部最小值，所有邻近状态的能量都高于它，所以才会被“吸引”到这里。

我认为大脑是一个超级复杂的动力学系统，它在时空域做信息加工，这一点和目前的人工神经网络在计算上有着本质的不同。如果真要做一种类脑的智能系统，吸引子网络是逃不掉的。我在那篇文章里抛出了吸引子网络的重要性，认为它是大脑做计算的一个基本法则。目前越来越多的神经科学实验证明了吸引子网络在大脑中存在。

Q 吸引子网络理论对AI的发展有何启示

吴思：在AI中，吸引子网络可能会帮助解决“抽象概念”的问题。我们知道，大脑有抽象概念的能力，小时候我们可能会学习具象的知识，但一旦进入读书阶段，课堂上学习的基本上全是抽象概念了，比如说数学。所以说，抽象概念的表征是一个核心。现在的AI是数据驱动的，其实还有一块是知识驱动，我们需要一种像人一样通过抽象概念进行学习的AI，才能加快知识的获取。

另外，吸引子网络形成了大脑中的一种高效记忆系统。比如，这是我们第一次在线上见面，也许下一次我们在线下时你可能一眼就认出我来。从线上到线下，我的形象、头发或者衣服都可能会发生很大的变化，但你还能认出我来，这是因为你会把我长相中的一些抽象概念和表征给提取出来，与我的声音、名字全部联系在一起，在你的大脑中可能就形成了“吴思”的概念，这个概念由神经元活动表征。如果研究清楚大脑表征抽象概念（还有别的，比如情感）的机制之后，构建出的智能系统才会更像真正的人类。所以在未来的人工神经网络的发展中，应该考虑引入这一套东西（吸引子网络）。

不算实验的话，我所知道的目前用理论构造模型去表达抽象概念的工作还没有做出成果来，但可能有人在尝试了，我们课题组也在尝试这个研究方向。

Q 您能解释一下“吸引子网络”吗？

吴思：可以想象一下，我们大脑实际上是由很多神经元连接而成的超级复杂网络。当大脑收到一个输入之后，网络的状态就会发生变化，这个变化实际上就相当于人脑在做计算或者在进行记忆搜索。比如在一个有关噪声的探究中，吸引子就代表了去噪之后的一个稳定状态。举个例子，比方说我知道了你的名字，也听到了你的声音，那么我下次看到你大概率还能够认出你（但也可能认不出，因为现在我年龄大了容易忘事）。一旦你被我的大脑记住了，我的脑中就会存在一种可识别出你的神经网络稳定状态。只要有一个输入触发，就能够重新演化到这个稳态，这意味着我的记忆恢复了，然后我就能够识别出你。也就是说，下次我看到你的图像，大脑中的网络就会演化出对应着你的概念的状态。因此，大脑的记忆计算系统与电脑的记忆计算系统是不一样的。吸引子网络就是这么一套用数学框架来描述的网络。

当我们在做抉择时也会涉及吸引子网络，这实际上是汪小京老师所做的工作。比如，我们在两个状态之间做抉择的时候，这两个状态（选择）就相当于是两个吸引子。我们在做决策的时候，实际上是在收集证据（便于我们做出判断）。那么这个证据就会决定我最终落入哪个吸引子，最终做出抉择。

吸引子网络甚至可以用于解释我们对语言的理解。我们人类的基本语法结构实际上是先天就储存在大脑中的。它就好像是一系列的吸引子一样。当我在听你说话、或者我自己说话的时候，吸引子网络就会开始演化与展开，这样我就可以理解你在说什么话了。吸引子的概念是由物理学家提出的，因为我是物理背景，所以我自然的就接受而且喜欢这个概念，我也会用吸引子来描述我的一些科学工作。

Q 除了抽象概念之外，您觉得目前的AI和我们的人类智能还有哪些本质差异？

吴思：我觉得大家经常在批评AI或者将AI与人做对比时，可能将“智能”与“智慧”这两个概念搞混了。智能是完成特定任务（比如学习、推理）的能力，是一个解决问题的能力。而我们人类拥有更高级的能力就是智慧，也就是加入了情感的社会认知能力。在特别复杂的生活场景中，不仅需要智能更需要智慧。举个简单的例子，如何解决全球变暖？一个最简单的解决方案是解决掉全人类，彻底切断源头。但是我们知道，这是最愚蠢的行为，在完成这种综合复杂任务时，我们需要考虑智慧，这个智慧包含了我们超出智能的能力，比如需要考虑全人类共同的利益。在这种特别复杂的场景中，需要的是综合各种因素，而不是完成简单具体任务的能力。从这个意义上说，AI离人类非常远，根本还谈不上有智慧。虽然如今的ChatGPT可以做一些推理，但无法实现人类在抽象概念上的操作，所以两者之间的距离还非常非常远。

还有一点也不可忽略，ChatGPT也好，现在最新的其它AI也好，它们能够完成的大部分工作实际上是我们大脑新皮层处理的事情，比如语言上的认知推理能力。但是我们生物体有一种亿万年进化的行为叫做本能行为，从某种意义上这对目前AI来说反而更难，这就是所谓的AI具身智能*。AlphaGo再厉害，还需要找个人拿放棋子，即使是拿棋子这个简单的动作，对机器人来说也是不容易的。但这对我们每个人来说轻而易举，这实际上是我们经过亿万年进化出的结果。至于我们是怎么获得这种多样的本能行为的（比如跑步、手眼协调），这个问题目前还没搞清楚，而研究清楚它可能会加速机器人领域的发展。这样的机器人再加入具有一定思考能力的大模型（如ChatGPT），才能造出真正服务于社会的机器人。否则它就是一个没胳膊没腿的、能做一些语言加工的机器。没有具身智能也谈不上智慧，其实还有很长远的路要走。

编者注:具身智能是一种人工智能领域的概念,通常称为“Embodied Intelligence”或“Embodied AI”。它指的是拥有自身物理体验或机器实体的智能系统。这些系统不仅具备智能的认知能力,还能够与环境互动并执行任务,类似于人类或其他生物体验世界的方式。

Q 如今利用“图灵测试”来了解机器的智能已略显简单,您认为当今时代应如何通过测试来了解AI的“智慧”?

吴思:针对这个问题,我还没有系统地思考过,刚才我举了一些生活中人类认知能力的例子,比如人类的commen sense(社会常识)、与人交往的能力(比如照顾别人情绪的例子),再或是在复杂场景中,人类需要(综合)考虑各种各样的因素,我认为目前AI在这些方面是有欠缺的。认知科学家和心理学家应该把人的多种认知功能做成类似ImageNet数据集,或者做一系列标准任务,而不是简单的通过语言问答的图灵测试,才能测试AI是否具有人类的智能和智慧。但是目前这个问题几乎没人做,我觉得应该着手去做了。

Q AI在未来可能还会经历寒冬和热潮的交替,您觉得当代学者应该如何应对?

吴思:一方面,我虽然看过了潮起潮落,但我依然对现在AI取得的成果感到惊喜,甚至有点惊讶。另一方面,我也不会被外界的鼓吹给“骗到了”,因为我深知AI背后的一些基本性的科学原理还有待挖掘。因此,在课题组招生上,我主要以“是否为兴趣驱动”来选择学生,因为只有保持兴趣,才不会随着外界的潮起潮落而漂泊不定。同时,我也不会让他们追赶潮流。北大的学生是优秀的,代表着中国科学的希望,我认为赶潮流发文章这个行为“太low了”。国家给予了这么好的平台,不能因为别人做了大模型,我也做大模型(当然我也不反对其他人因为兴趣做大模型)。我会希望组里同学要坚持自己的信念,然后锲而不舍长期的工作,这才能做出原创性的成果。

在我看来,目前的AI发展只是一条路径,并不是终点,还有最fundamental(根本)的问题尚未解决:理解生命的本质、意识思维的本质。这比AI还重要,AI只是一个物质层面的应用,但是理解我们生命的本质那是更重要的东西。所以我会希望我们组里的同学抱着一种长远的理想和目标。我们不要去跟风Geoffrey Hinton等人开创这轮AI热潮 ,而是要学习他们那种低谷中锲而不舍的精神,将这种精神用在科研发展上,争取在这一交叉领域做出好的成果。

Q 您如何评价近十几年来计算神经科学在中国的发展?

吴思:(学科)还在发展中,(从事计算神经科学研究的)人还是太少,但会有越来越多的人参与进来。不光是计算神经科学,比如说现在做类脑智能的研究者,与深度学习网络不同,他们做一些偏脉冲神经网络研究。这些人也在学习计算神经科学的知识,所以说这个领域在扩展。对比来说,计算神经科学就像是理论物理,你看它的领域(范畴)是很大的。

一个月前,我课题组在线上开设了一门课程《神经计算建模》。这门课程偏学术性,非常前沿。结果一经开放,400个名额立刻就被抢空了,就好像现在买演唱会门票的感觉。受到当初办(冷泉港计算与认知神经科学)夏校的影响,做东西要么认真做,要么不要做。我们做这门课程非常认真,也控制质量,同时也很注重与学生的互动。反正至少我听到的同学们对这门课的评价是特别好的。考虑到很多人在第一期没有报上名,为了让更多的人有机会学习这门课,接下来我们会筹办第二期、甚至第三期。

Q 这十几年中，计算神经科学界有哪些突破和进展？

吴思：我本身置于这个领域，对自己所在的领域可能比较挑剔，或者说（对于学界内的突破和进展）我自己定的标准可能太高了。其实这个领域的进展也挺多的，在十几年前，我们可能做一些简单的神经元和突触建模就可以了。现在的研究者则更偏向大规模网络的建模，不再去解释一些通过实验可得的简单行为，而是试图去实现高级认知功能的解释。客观的说，这（高级认知功能的建模）其实也是我们这个领域现在所面临的一个问题：虽然绝大多数人相信我们可以从大脑中学到一些东西，但计算神经科学领域还没有出现Alpha-Go、AlphaFold这样的杀手级应用。同时也没有做出一个能够替代动物实验的神经疾病模型。后者是欧盟脑计划中的目标，但他们也没有做出来。所以，我觉着我们这个领域所朝着的大的科学目标是正确的，但目前的确还没有取得革命性的突破。

在AI处于寒冬的时候，Geoffrey Hinton 说过这样一句话：大脑就是这样的网络，没有理由不工作。这个简单朴素的信念支撑他锲而不舍地研究，最终发展好了深度学习网络。我经常用Hinton的话来鼓励学生，大脑每天都在运转，拥有各种高级认知功能，尽管它很复杂，只要我们在这个领域静下心来，共同努力，一点一点积累，我相信在十年之内会迎来大的突破。

Q LeCun说AI不是仿生学，您觉得类脑计算是如何从大脑中获得启发的，为何您认准类脑计算是突破AI瓶颈的一条路径？

吴思：首先我强调一点，沿用LeCun的观点来说，就是仿生到底要仿到什么程度？我是反对将大脑所有的精细神经结构全部构建出来，这是涉及生物智能的实现层次。类比英国心理学和神经科学家David Marr所提出的三个层次，我们首先有个行为层次，用于表现认知功能以及各种心理行为；然后中间有个实现层次，它可以描述成数学方程或者是网络模型；再下面是神经生物的实现层次，比如神经元的突触如何连接。实际上，大脑非常复杂，也不容易研究清楚。至少对我来说，全盯着实现层次是没多大意思的。

我们最终是要抽象出一个大脑工作的信息加工理论，这（理论）会是一个数学化的东西，它的下面才是神经元的实现层次。所以我们在做类脑研究时，不能只等待神经生物学的进展，去一点点从神经元、突触开始搭建模型。我们要从神经生物中得到启发，再外加一些我们人的理性思维以及数学的工具，通过这种上下夹击的方式争取把一个类脑智能研究清楚，这也是我课题组研究的路数。

Q 近年来，您认为神经科学的哪一项工作给AI带来了启发？

吴思：我认为大脑的整体框架、认知架构、高效的记忆表征系统、某个高级认知功能的建模突破以及抽象概念的表达这类东西可能会对AI的进一步发展带来启发。

Q 谈到记忆，人类的记忆与AI的记忆之间有鸿沟吗？

吴思：AI实际上没有记忆，或者说AI的记忆其实是由大数据训练好的网络的连接权重。但我们大脑记忆是一个（连续）吸引子网络系统，每个吸引子网络对应一个抽象概念的表达，网络中不同状态之间的相互作用与交流最终形成大脑高效的记忆系统。我在科研生涯中做了很多课题，但是连续吸引子网络是我20多年一直没有变过的课题，你们要是感兴趣的话，以后可以给你们讲讲具体的细节。海马是大脑的记

忆、清楚，我们在抽象概念的表达上就会有所突破，也许在AI中就用得上了。

Q AI的未来发展是会朝着追求与人类的相似性不断趋同，还是任其发展？

吴思：从工程的角度来看，面对一个专项任务，AI真的没有必要像人脑，实际上大家对AI战胜人类的事实表现得有点过于惊讶了，比如早期用珠算做加减乘除，而现在计算机在数字的计算速度和精度方面早已远超人类，我们又有什么可惊讶的呢？从我的理解，一个具体任务的数学机制一旦被破解，就没有我们人类什么事了，机器肯定超过我们。但是，强调AI类人的原因是我们不仅仅追求智能，也希望AI拥有智慧，向人类社会看齐，从而服务人类社会。我们希望它能够拥有常识、同理心、情感、道德观，能在复杂环境中智慧地处理事情。从这个意义上说，AI需要向我们人类看齐，让人类给AI很强的启示，否则的话这个东西（AI）就会完全失控。

我再补充一个我现在面临的问题：很多人认为研究计算神经科学、大脑的机制是为了AI。其实没有。AI只是计算神经科学研究的一个副产品，对于我个人来说，最重要的是去探究大脑的奥秘，我需要了解生命的本质是什么。我可以不要物质生活，我只要精神生活。而且说句实话，研究清楚大脑的工作原理对这个社会的帮助不只在AI，比如医学、教育（研究大脑在发育过程中是如何通过可塑性学习知识的）等。总之，计算神经科学还有更广阔的前景，只不过我们可以在类脑智能这方面和AI有一些交集。

答 汪小京追问：

Q 怎么定义、定量人的思维和智力？这个问题的答案会告诉我们今天的AI与人的差距。具体地说就是如何通过神经科学的方法去理解思维和智力的神经机制。

吴思：由于我自己是跨学科的，现在在从事认知神经科学研究工作，我发现学科之间是有壁垒的。做AI的人擅长通过数学化的方式去完成具体任务，而心理学家喜欢在行为层次上进行研究。比如，关于我们人的智力和思维，在心理学上有一个学科叫比较心理学，几十年前，比较心理学通过大量的实验试图回答动物到底有没有智力，比如《乌鸦喝水》中乌鸦把石子放进水瓶来升高水位的行为是智力的体现吗？研究者还会用别的动物，比如老鼠，来回答类似问题：动物到底有没有我们人的那种智力？我觉得这些研究很大层度上都被AI领域的人给忽略掉了。AI研究者应该去了解比较心理学，去看一些基于行为层次对人类认知的研究，这样对比以后也许会意识到AI在实现类人的思维和智力的道路上该怎么走。

我们做计算神经科学的人在学科交叉方面其实起到了非常重要的作用，正好把这两个领域之间的 Gap（鸿沟）给填起来。AI强调一个数学模型，而心理学研究行为层次，那么现在最需要的是建立数学模型，去真正解释人的认知行为。解释清楚之后，就可以在帮助回答思维和智力的不同之处以及人的高级认知功能本质方面起到重要作用。（记者：涵棋；编辑：韵珂）

来自@吴思的问题：

情感在我们的智能计算中起到什么样的作用？

情感这个问题很有意思，人工智能的大牛Marvin Minsky曾说过这样的话：现在的问题不是机器人需不需要情感，而是一个没有情感的机器人到底能不能产生智能。我觉得这里面涉及的一些东西就是智慧，而情感是我们智慧的核心部分。

追问专访·吴梦玥：

机器比人类更会理解声音吗？

嘉宾：吴梦玥

上海交通大学计算机科学与工程系副教授。目前主要研究丰富语音分析与处理及其医疗应用，在语音及语言处理领域顶级会议期刊发表多篇论文。

扫码查看原文

声音是人类社会重要的交流媒介，它不仅可以传情达意，还可以反映人的身体状况。在本期《追问专访》中，上海交通大学计算机科学与工程系吴梦玥老师将带我们一起进入语音世界，从多模态互动到医疗应用，共同探索声音的奥秘。

Q 请介绍一下您的研究背景。为何会对这个研究领域感兴趣？

吴梦玥：我现在的主要研究方向是丰富音频分析。当我们听某种声音的时候，如果是在听一种语言，我们不仅在乎这个人说了什么，还关心这个人说的话是如何说出来的，即他/她在说话时的情绪和情感是怎样的。再进一步想，一个人在说话的同时能够体现出他/她本身的精神状态或认知状况，这其实是把语音或者语言功能看成是大脑认知功能的外化表现。因此，从语音的角度上，我们可以做很多病理上的分析。

另一方面，我们所听到的声音中不仅包括语音，还包括自然界或我们所处环境中的一切声音。很长一段时间以来，传统做语音研究的人会认为这些自然声音都是“噪声”，但其实我们在处理所有听觉信息时，每一个细小的声音都提供了极其多的信息。现在我们把这个领域叫做“丰富音频分析”，所谓的“丰富”来源于两方面，一方面是指人声会有很多层次，可以提取到很多信息；另一方面则是指环境的丰富。我现在想做的研究就是如何把这两者很好地结合起来。

Q 丰富音频分析有哪些应用场景呢？

吴梦玥：其实从刚才我们谈到的研究内容中可以明显地找到一些相应的应用场景。比如语音上的分析，尤其是和病理方面相结合时，在医疗领域的应用场景就非常广泛。

病理上的语音研究分为几类，一类与器质性障碍有关，比如当腺样体肥大时，可能会影响整体的气流，在发音的过程中就会有阻碍，所以这些器质性的病变会引起语音信号上的不同。因此我们的研究和耳鼻喉科有很多相关部分，可以通过一个人的语音来判断他/她嗓音的变化，包括判断像腺样体肥大一类的病变，甚至也可以做喉癌的早期预测。人除了说话之外，还可以产生其他声音，有些声音与器质性改变也有关系，比如鼾声，现在有很多研究会通过检测鼾声来监测睡眠，或者查看其呼吸系统是否存在问题。

此外，在新冠全球性大流行期间也有一些研究，比如通过一个人的咳嗽声来判断他/她咳嗽的根因。这些研究不仅可以用于诊断新冠肺炎，也可以放到一个更广泛的场景中，尤其是在儿科领域。咳嗽是儿童呼吸系统非常常见的疾病，儿童出现咳嗽症状的原因非常多。我们和上海市儿童医学中心进行合作，发明了一种便于儿童携带的、可长期穿戴的设备，外型像一个麦克风或者一个纽扣，这样就可以监测儿童整个咳嗽过程的变化，从咳嗽的频率和咳嗽产生的所有语音逆向推导，比如咳嗽的性质是干咳还是湿咳，再进一步分析是由普通上呼吸道感染引起的咳嗽，还是由某一类肺炎引起的咳嗽。这些都是一些非常明确的应用场景。

除了在器质性疾病上的应用场景，神经退行性的疾病或者与情感障碍直接相关的疾病也可以进行语音研究，比如抑郁症、焦虑症、帕金森症以及老年痴呆。在对老年痴呆的患者进行语音分析和比较时，发现它和抑郁症、帕金森症有一定的相似性。一方面大部分老年痴呆患者在很长时间内会伴有抑郁的症状，另一方面这种疾病和帕金森症一样都属于神经退行性疾病，这些疾病之间的内部联系使我们的系统能在这些场景中得到应用。

从其他方面来说，还有一个非常直接的应用——婴儿啼哭的检测。比如可以在家里放一个检测器，当它收集到小孩哭叫声时可以对哭叫声进行分析，然后判断孩子的需求是什么。

此外，我们前段时间和公安机关进行了合作，在监察人口流调时如果想知道有谁从外地返乡了，就可以在返乡人员的家门口安置麦克风阵列，几户人家可以共用一个麦克风阵列，通过麦克风阵列对开门关门声音的识别来判断是否有人回来或者进出。

这项研究也可以应用到确定滴滴乘客的出行安全上，在打车出行时为了查看乘客的安全，录音是实时开着的，但即使录音实时开着，也没有人会实时查看所有的录音。所以处理录音的时候就需要检测和判定其中的异常事件，对是否有人在尖叫、吵架或者求救等情况进行检测，这些都属于丰富音频分析里我们所探讨的内容。

更进一步，可以探究如何用完整的自然语言来描述一段音频内容。比如用ASR可以直接得到一个语音的翻译，又比如在现在这个场景里，用自然语言描述的话，可以描述为“几个人在进行网络会议研讨，其中有哪些具体内容”，或者也可以直接描述一段语音为：“有人走过，同时有鸟在叫……”这些都可以很好地帮助听障人士，即使听不见声音也能通过语言文本了解此刻这个听觉世界到底在发生什么。一些手机厂商已经开始进行这方面的研究了，旨在可以更进一步地满足听障人士或者弱听人士的需求。这些是我可以想到的丰富音频分析直接对应的应用场景。

Q 在研究过程中，数据是一切的基础。您主要使用哪些类型的数据？又是如何收集和分析这些数据的呢？

吴梦玥：这是一个非常关键的问题，不管是医疗领域还是环境声音领域，相对于我们研究了很久的语音而言，这部分声音数据还是比较稀缺的。对于医疗领域的声音类型数据，我们会和医院进行合作，但是和医院的合作更多是在硬件上发明、创造或者利用现有的技术将它改造成更适用于分析应用场景的形态，然后采集音频数据，之后在实验室里进行分析。

至于环境音频的声音，首先环境声音非常多，但它最大的问题在于怎么进行标注。谈及标注时又会引起一些新的研究问题的探讨，比如是否可以用弱监督的方式描述环境音频。环境音频方面最大的数据集是Google在2017年推出的AudioSet，里面包含了527类不一样的声音事件，每一条音频里又包含多个标签，但其实没办法很精准地定位标签，比如一段音频里第一秒到第三秒有一个事件，或者第四秒到第八秒还有一个事件，这种强标签的标注方式非常耗时耗力也耗费资源。现在有一个段落级别的标注方式。怎样用弱监督的方式先进行标注，再用强监督的方式对每一帧进行标注，是我们这个研究领域里面临的较大挑战。

除此之外，我们自己在2018年首次提出了audio caption这个任务，即怎样用一段自然语言文本描述音频内容。相较于之前的标签化研究而言，这种方式则更贴近于人类的听觉感知。如果刚刚听到一声巨响，你在描述这件事时不会说“爆炸声、分号、呼救声、分号”，而是会用一个很自然的句子来描述，这就是我们希望未来机器在做听觉感知时能直接输出的结果。当然，我们创造了这样一个新任务时同样需要一个新的数据集进行支撑。

总之，我们研究的数据要么来源于真实场景，比如通过和医院合作或去自然界采集，要么就是在一些基本的数据集上发明一些新的标注方式进而解决我们当下的问题。

Q 您近期的一项研究中提到了一个叫clap的模型，用于训练这样的模型的关键数据集有哪些？以及它们是如何构建的？

吴梦玥：在前几年的时候出现了非常多结合了视觉和自然语言的大规模预训练模型，但是音频领域则非常少，很大原因在于数据集的缺乏。但在去年，包括我们在内，同期有三篇文章中提到的模型都是叫clap，因为之前clip模型是在图像（imagine）上做caption，我们把图像换成音频（audio），所以叫做了clap。

其实我们的训练方式和原来的clip非常像，关键是怎样解决音频领域里的数据集——尤其是和文本对应的数据集——从哪来的问题。

一个方法是可以基于原有audio caption数据集训练一个模型，然后用这个模型去给其他所有适用的音频打上尾标签。在打尾标签之前还有另外一个方法，可以将离散的标签加进来，把它做成一个引导，然后用这些标签去引导audio caption模型，这样生成的caption本身会更加符合原来的音频内容。以这样的方式对海量数据进行尾标签标记时，从某种程度上来说已经构造了一个音频和文本对应的数据集。

在这个基础之上，我们用对比学习（contrastive learning）的方式，比如说用两个编码器，一边输入音频一边输入文本，再加上一个对比损失（contractive loss），这样训练下来的预训练模型在很多和音频

或文本相关的下游任务中能获得较大的性能上的提升。总之，如果要做预训练，数据的来源以及数据的质和量都非常重要。一方面可以训练一个模型来标记标签，另一方面也可以利用ChatGPT为更多的音频数据生成自然语言描述。

Q 很多实验都面临“走出实验室”的问题。现实世界中，语音信号可能受到各种因素的干扰，如背景噪声，说话人的口音、语速、语调变化等，使用不同的录音设备和麦克风也可能会导致语音信号出现差异。那么，实验室训练的语音识别系统如何处理真实世界中的语音信号？

吴梦玥：与自然语言处理相比，音频分析最困难的地方确实是统筹所有不同音频的信号。我们研究中很多数据都来源于真实场景，因此在医院采集声音时，我们会规定统一的型号或采样率，从而得到一个优化较好的模型。在最后进行模型训练时，我们也会采用不一样的方法使得模型有更好的适配性或者鲁棒性，比如可能会进行不同噪声的模拟，或者额外添加一些噪声，不过这也使得原本用来训练的数据集变得更复杂。

如此一来，真实测试中可能碰见的任何情况都包含在了原始的训练数据集的分布里，但要真正让这项工作得到实地应用——无论周围有什么人、环境多么嘈杂都可以在真实世界中实现如同在实验室里一样好的性能——还是比较困难。因此，关键问题还是在于我们可接受的在真实环境中模型性能的下降范围是多少。

对于这个问题，传统的语音识别研究中同样面临真实世界的挑战——在这种非配合式的环境下，如何得到更好的研究结果，我们为此做了很多努力和尝试，但目前为止这个问题还没有被解决。

Q 您刚才提到在研究中很重要的一环是对环境声音的标注和描述。随着GPT的到来，AI模型也成为了科研当中有力的工具，包括我们知道GPT-4已能够实现对多模态数据的分析、理解、整合和输出。那么它是否能对环境声音的标注和描述有所帮助？

吴梦玥：这个问题非常有意思。如果让一个人用语言描述小提琴和大提琴声音上的差异，或者描述咖啡厅场景和餐厅场景中的声音有多少差异，人很难描述清楚。但如果向ChatGPT提出这样的请求，不管是GPT-3.5还是GPT-4，它给出的答案都非常合理，从中可以发现，ChatGPT其实是通过强大的文本能力弥补了声学编码器上的不足。所以我们认为，在对环境声音的描述上，ChatGPT可能会比人做得更好。

现在的问题关键是要给予ChatGPT怎样的提示词（prompt）才能让它既符合我们的要求和描述习惯，同时又能够精准地描述声音中具体的特性。前段时间，英国萨里大学就有一篇这样的研究，这一研究虽然只在第一步使用ChatGPT来辅助研究，但是总体上而言，我觉得这是一个很有前景的方向。

不过在语音的模型中，即使使用了ChatGPT也无法直接把图像或者语音当做素材供给它做多模态的联合训练，后续可能需要我们在自己的实验室里进行微调（fine-tune）或者做联合训练。不过这方面确实存在应用场景，ChatGPT目前拥有的对不一样模态信息的理解能力可以辅助我们做信息媒介的部分分析和处理。

Q 基于ChatGPT，您的研究团队还做了哪些尝试呢？

吴梦玥：ChatGPT的应用还是得以文本为媒介，在模型训练的过程中如果出现了样本较少的情况时可以使用ChatGPT对数据进行标注，尤其在处理非常细微的情感关系的差异时效果很好。除了对声音本身的分析外，也可以用ChatGPT做另外的研究，例如让机器人模拟医生和患者的整个以对话为基础的问诊场景——用ChatGPT做两个模拟器，一个模仿病人一个模仿医生，然后将它模拟出来的问诊情景与真实的精神科问诊过程进行对比，然后就可以探究与真实场景相比时ChatGPT在对自然语言的理解和处理上还有哪些局限。

在我们训练的所有AI模型中，ChatGPT的自然语言理解能力已经达到了极限，接下来应该怎样用模型实现和真实场景效用一样的人机问诊也是我们想结合ChatGPT进行的研究。如果自然语言理解的能力对于ChatGPT来说已经无法进一步提升，那么在自然对话与模型模拟的对话间还存在哪些因素上的差异，这些都是我们现在非常关注的。

Q 您提到由ChatGPT来充当医生与患者的模拟问诊场景，那么它所创造的模拟数据可否作为真正的研究数据使用？基于此的研究结果是否有意义？

吴梦玥：目前来看，其实不太行。它可以模拟一些比较基础的案例，但和真实的应用还是有一定的差距。具体体现在，比如说模拟医生，ChatGPT和医生的问话形式或者风格有一定的差异，ChatGPT可能会更书面化，而在平时问诊的时候，为了让患者放松，医生很有可能用的是一些更轻松的、偏口语化的问诊方式。当用ChatGPT模拟患者时，现实中患者看医生的时候，他/她不会那么坦白地告诉医生一些答案，或者很多患者并不清楚自己的症状到底是什么，但是ChatGPT作为一个这样的患者时，比如最开始我们让它加个抗拒，它可能就抗拒一次两次，你反方向再问一遍的话，它马上就说出来了，感觉就像“我有答案，但是因为你告诉我不要把这个答案说出来，我就藏两下”，它和真实患者之间的心理差距还是非常大的。

所以，我认为它可以用来做一定程度的数据增强。但是如果要把这种模拟的数据拿来做完全的训练数据，可能和实际应用场景的差距太大。

现在对于ChatGPT的应用中可以对比ChatGPT作为患者所模拟出的数据与真实病人的数据间的差异，这部分工作目前已经有了初步的结果，后续马上会发表出来。目前可以作出的较直观的结论是：如果给ChatGPT设定了较好的prompt，在患者处于配合的情况下，模拟出的场景可以十分接近真实问诊场景，而当有患者并不处于配合的状态时，对话会产生较大的困难，所以差异本身还是取决于机器人所要模拟的真实场景的复杂度。

用ChatGPT可以对简单基础的问诊场景进行模拟，但和真实的问诊应用还有一定差距。在真实问诊时为了让患者放松，医生会使用偏口语化的问诊方式，而ChatGPT模拟的医生问诊风格则偏向于书面化表达；模拟患者时也有差异，比如患者在面诊时可能不会坦白地说出答案，可能自己也不了解自己的具体症状，也可能出现一些前言不搭后语的情况，但是ChatGPT则难以完全模拟这种状况，例如在模拟患者抗拒回答时它可能仅会抗拒一两次，转换了询问方式它就不再抗拒，所以这和真实患者之间仍存在非常大的心理差距。我认为ChatGPT可以用来做一定程度的数据增强，但产生的数据与真实应用场景间的差距太大，无法用作完全的训练数据。

Q 在前段时间举行的“AI助力攻克脑疾病研讨会”上，您提到自己很长一段时间都在做基于语言功能来判断抑郁症、帕金森等疾病的研究。语音与脑疾病间有什么关联？如何利用语音检测疾病？

吴梦玥：比如帕金森疾病是一种神经性的退行性疾病，它会影响大脑中的运动功能控制（motor control），运动功能控制不仅影响对于手脚的控制，还会影响到说话前的准备阶段（speech preparation），在大脑产生“说话”的念头到控制发声器官发声这两个步骤间还存在缓冲过程，当运动功能控制的部分受到影响后，虽然脑海中已经想到了要说的词，但因为发声器官在这个时刻还没有得到控制所以没法及时发声。所以很多帕金森患者在发音时可能出现发音不清晰或一直重复某个语音的情况，也可能会在发声中出现较长时间的停顿作为发出下一个语音的准备。

因此，帕金森患者在声学表现上有一些表征，比如说话的语速会变缓，整体的词汇量会变少，话语间的停顿时长也会变得更长，对一个词的重复次数会比正常人更多。这些其实都是可以进行量化计算的特征，将这些量化内容加到最后的检测模型里，就可以通过语音去反馈很多和疾病相关的特征。

Q 目前基于语音进行疾病诊断的准确性是怎样的？是否已经一些研究已应用于医疗领域中？这其中是否会存在潜在的伦理问题？

吴梦玥：国内外的新闻中其实有对此类研究应用的准确性的报道，比如做抑郁症检测中有使用的南加州大学的数据集，用这个数据集做一个基线（benchmark），经过实验的调参后可以得到80%-90%的准确性，但将它放到真实场景或近似场景中面对不同方式采集来的数据时，它的迁移能力还是非常差的。如果不经过任何调参优化对不同的数据集进行检测，可能准确性就变成了60%-70%。面对这种情况，一方面可以结合不一样的模态进行检测，另一方面可能需要进一步寻找不受环境因素或者数据集因素所影响的特征，最后才能实现比较鲁棒或者可迁移的检测方式。

在这过程中会产生一定的伦理问题。第一个是这种模型检测能否替代医生的问题。首先，这项技术本身可以帮助医生工作，比如一个接受治疗的人可以通过心理状况筛查的小程序查看自己近期的心理状况，不需要每次复查都去医院面诊，这方面可以很大程度上增加诊断的便利性。但即使它在实验上已经达到了较好的准确性，其本身也无法替代医生面诊的检测结果。

除此之外，之所以强调使用语音来进行检测，是因为很多其他方面的信息比如脸部信息、步态等方面涉及的隐私内容可能会比语音涉及的隐私内容多，但是语音检测仍然会涉及人的隐私。比如在对于抑郁症或其他精神疾病的诊断上更多采取面诊的方式，仅仅根据患者对自己状态的描述来诊断的话客观性就会下降，所以我们在考虑是否可以使用可穿戴的设备对患者的睡眠、活动量等方面进行长期的监测，据此推断患者实际的状况，但这也会涉及另一类伦理问题：医生是否有权利获取患者日常生活中的生活轨迹来进行病情监测？因此，我认为从宏观角度来看，医疗、个人、公共卫生的管理之间都可能存在一定的冲突和矛盾。

技术本身是向前发展的，但牵涉制约技术的因素很多，技术是否能运用到实际生活中需要考虑的因素还有很多。

Q 随着AI技术的迅猛发展，您觉得未来语音领域会有怎样的突破？

吴梦玥：我们实验室之前毕业的一位博士现在在Google进行多语言语音识别的项目，这个项目就是希望做到多语言的语音识别，构建出可以对多个语言甚至100个不同语言进行识别的语音识别系统，这其中也利用了声音和文本间的对应关系，在说话过程中，音素（phoneme）和语言（character或letter）之间存在很强的对应，用音素+时长就可以实现文本和语音间的转换。

丰富音频的分析中也存在很强的对应关系，比如“鸟叫”和含有鸟叫声的一类音频间有很强的指向性，以逆向利用这种指向性来进行音频上的编码，因此，文字与语音的关系也可以帮助我们进行多模态的对声音的理解或分析。

所以我会认为未来的一个很有潜力的发展方向，就是将语言当成有更充分知识的线索来辅助研究，在与语音相关的任何研究领域里或许都会很有帮助。

Q 在ChatGPT问世后，您认为AGI相关的通用人工智能下一个阶段将向哪个方向发展？最终人工智能是否能进化得如同真正的人类一样？

吴梦玥：很早之前有一部科幻电影《她》（her），在电影中每个人都有一个视觉系统，人与人之间可以通过耳机进行对话，机器和人之间不存在信息理解的差距，这是我对未来通用人工智能功能的一种初步预想；再比如波士顿动力（Boston Dynamics）想做的陪伴型的机器狗，这也是一个研究方向。能实现这些功能的信息处理肯定是多模态的，如果机器获得的信息和人类获得的信息中间有太大的差距，就没办法帮助人进行决策。因此，在技术上来说模型还存在需要继续修正的部分，只有探究到人与机器人之间的差距再弥补这个差距，才能让机器变得与人更相像。

现在在人与机器的交互过程中，机器本身更多是以工具的形式存在，当它可以不局限于受到刺激才能回答的形式，而是可以主动进行对话时，才能使人机交互（human machine interaction）变成更接近于人与人之间的互动（human to human interaction）。

此外，当我们知道对方是机器人的时候，你会不会对机器人说“谢谢”或者“抱歉”？在我们做模拟的过程中发现，如果医生事先知道对方是由ChatGPT扮演的患者时，医生并不会产生对“患者”的共情，在诊断过程中会更多倾向于通过走完流程来确认ChatGPT是否演绎出一个合格的患者；而当ChatGPT扮演医生来应对患者时也是一样。所以，还需要了解人和人相处与人和机器相处之间存在哪些差距，探究这种差距同样是实现真正通用的人工智能的关键。

Q 您认为试图让机器和人更相像，到底是一件好事还是一件坏事？

吴梦玥：我觉得让机器和人更相像，一方面能够帮助机器拥有更好的性能，另一方面，当机器拥有了与人相似的种种能力后，人才能与机器进行更自然的沟通，否则人与机器之间仍然存在着差距。至于我们的研究中是否希望机器人更像人，这是更大范围上的伦理上的讨论。比如流浪地球中的Moss可能已经开始出现自己的意识，意识的出现对机器人来说是一件好事还是坏事，机器人存在的价值和意义到底在哪，我想这些会由哲学的老师去讨论。

如果从技术上来说，我们肯定希望通用人工智能更像人，当机器人拥有了与人相似的能力对人而言会有很大帮助，人本身将能够从很多繁复的劳动中解脱出来。至于解脱出来之后的行为能力是会上升还是

下降，这是现在谁都没有办法预计的结果。（记者：Lixia；编辑：韵珂）

来自@吴梦玥的问题：

机器理解语言和文本的过程与人类理解语言和文本的过程间有什么差别?输入机器的信息是频谱的或者直接是波形的，机器在编码这些音频信息时与人类大脑处理音频信息之间的差异是什么?

医生最想用GPT解决哪些问题?

嘉宾:毛颖

华山医院院长、天桥脑科学研究院转化中心主任

嘉宾:徐一峰

国家精神疾病医学中心脑健康研究院院长

嘉宾:王振

上海市精神卫生中心副院长

嘉宾:吴梦玥

上海交通大学计算机科学与工程系副教授

嘉宾:胡鹏伟

中国科学院新疆理化技术研究所研究员

嘉宾:陈亮

华山医院神经外科副主任

嘉宾:郁金泰

华山医院神经内科副主任

嘉宾:彭代辉

上海市精神卫生中心心境障碍科主任

扫码查看原文

Q 作为临床医生,最想用GPT这种大型AI模型解决哪些问题?

毛颖:精神疾病不是我的领域,但是我想,GPT的语言理解能力很强,而目前很多精神疾病,比如精神分裂症,其诊断大多是通过医生与患者对话进行的,所以在精神疾病的诊断方面GPT会有很好的效果。另外,对于语言的认识,英文环境下已经做得很好,但是在中文环境下我们仍然大有所为。

陈亮:有没有可能医生和患者进行一段自由交谈以后,GPT能够提取里面的信息,直接将关键字段填入病史,如果发现缺少了哪些字段,还能提醒医生再继续问一问。如此之来,病史质量提升,医生的工作量也会下降。

郁金泰:我想ChatGPT在队列建设上会起到非常重要的作用,比如阿尔茨海默症的早期识别主要是通过与患者的信息沟通。这部分患者的队列建设,尤其是社区队列,患者的数量是非常大的,如果通过人工的话,时间成本和人力成本非常高。因此,ChatGPT对于健康信息的采集显示出了巨大的优势。

另外,我们在做认知评估时会涉及各种量表,而中国早期轻度认知功能受损的患者约有4000万,如果通过人力去筛查的话很难实现早期的识别。但是通过GPT把有价值的量表提取并做成在线的数据采集,将会起到非常重要的作用。除此之外,在临床诊疗过程中,认知障碍的神经退行性疾病的诊断也依赖于病史问询,好的病史对于疾病的诊断非常重要。因此,我们可以把一些具有敏感性或特异性的信息指标,通过算法研发出有助于临床诊疗的工具,对于基层医生的诊断可以提供很好的指导价值。

Q GPT在疾病治疗的过程中又能发挥什么作用呢?

郁金泰:在认知方面,目前证据级别比较高的就是计算机辅助的认知训练,还有就是针对明确的风险因素,通过ChatGPT的相关数字疗法对此类人群进行管理。比如通过人机对话,我们可以获取患者相关的可调风险因素,针对这种风险因素制定一套对患者有效的预防模式,同时再通过进一步定期评估来观察患者的危险因素是否得到有效控制。另外,通过这种评估我们可以发现患者到底是在语言还是记忆或者空间等方面出现了问题,针对患者的问题采取相应的认知训练,这方面也非常有前景。更重要的是,认知其实和情绪一样,是一个公共健康问题,单纯靠医生无法解决,可能需要这方面的工具来解决。

Q 目前,GTP这类大型AI模型在区分不同的疾病以及对疾病严重程度的预测方面表现出了一定优势,那么如何更好地发挥它的优势,并将其应用到临床疾病的诊断中?

吴梦玥:很长一段时间我都在进行基于语言功能来判断抑郁症、帕金森、老年痴呆等疾病这方面的研究。我们发现,对比躯体症状,患者在描述自己精神上的感受时,语言表达会比较模糊,也比较主观,每个人对自己症状的理解有很大差异。通过躯体症状来判断疾病的模型是比较容易的,但是在精神疾病领域就遇到了很大的困难。从患者非常自由的口语表达提取出来的症状与医生面诊得出的结果差异性很大。而ChatGPT通过大型数据库训练后,其自然语言处理能力和理解能力,相比之前训练的模型要好很多,所以我认为在这方面GPT这种大型AI模型会有很好的表现。

另外,在声学方面我们一直存在的困扰是,抑郁症患者语音和正常人语音在声学层面有怎样的差异?我们发现单纯声学层面的表征方式迁移性效果很不好,在一个数据库上表现很好的特征组合,换成另外一个数据集,效果就大打折扣。而如果我们把传统的物理声学信息与医生对患者语言描述的信息相结合,通过这两种模态的结合很可能会得到一个比较通用的表征方式,所以我认为ChatGPT也可以帮助我们理解抽象特征。

胡鹏伟:在生命科学领域,我们过往用人工智能做蛋白质相关的研究,将蛋白质的序列转化成一些可被机器学习的数据形态。随着大型AI模型出现,这在蛋白质相关领域已经开始广泛应用。比如,AlphaFold的蛋白质结构预测,它是先用一个大模型学习了全部已知蛋白质的结构、序列等多种信息,然后再用该模型去预测未知蛋白质的结构,目前预测的蛋白质结构已经超过已知的三倍左右。在生命科学领域,不只是蛋白质,还包括RNA、DNA等等多种数据的表达,AI大模型在对它们特征的重新表达,以及基于重新表达的再预测方面都已经在发挥作用。

Q 模型的好坏很多时候取决于数据量的大小,而医疗领域通常数据量都很小,如何解决这个问题?

胡鹏伟:我们曾经做过眼底病变的预测,当时的样本量也很有限,大概仅有几万张患者的眼底图片。我们利用这几万张样本,把样本的病灶和正常组织进行区分,然后进行强化学习,最后再通过不同的组合去生成新的医疗样本。这只是一个例子,我觉得其他基于图片性质的数据,甚至是一些基于文本结构化的医疗化数据,都可以进行尝试。

吴梦玥:相对而言,医疗领域的数据确实量都比较少,用大的预训练模型,然后在小数据集上做微调,效果非常好。此外,预训练模型和迁移领域的小数据模型之间,会有一些格式或者说知识领域上的差异性,

所以我们在做数据微调时加一些合适的prompt，让它们之间的差异缩小，让微调的效果更好，也是一个比较好的研究方向。另外，我们可以引入医生的专业知识，做一个有引导的微调范式，可能也是一个比较好的方向。

Q 一个语言模态需要多大的信息量来预测抑郁症？如何将其他模态的信息整合进ChatGPT模型当中，一起做临床抑郁的诊断和预测？

吴梦玥：ChatGPT开放了API端口之后，其实是可以把很多其他模态的数据放进来的。比如说语音，只要前端把语音的整体数据提取特征之后，其实是可以匹配到ChatGPT的端口里面，最后进行定量分析。当然，现在的一大难点是，在换了一个数据集之后，效果就会很差。一是因为数据集比较小，每次的实验效果并不一致。二是它的迁移性比较难，所以我们也希望把临床的数据拿进来，可以实现一个比较鲁棒的、可迁移的检测方式。

胡鹏伟：从源头来说，首先确定哪些信号是现在能够获取的，语音是现在用得比较多的，除此之外还有患者自身的其他信息，比如基因信息、行为信息、生理信号信息等。如何把一个人的这些信息数字化到模型里，我觉得可能依赖于传递式设备，不只是录音设备、录像设备，还需要能够收集血压、血糖、心电、血氧等生理信号的设备。如果通过可穿戴式设备把这些信号收集回来，然后为每个人建立模型，再把这些信号交给ChatGPT，我觉得才可能依靠ChatGPT描述好一个人。

彭代辉：我个人做得最多的是脑影像和脑电，但是这两个维度的指标和抑郁症全病程发展过程以及内在机制之间的灵敏度和特异度非常低。从临床医生的角度来讲，对于抑郁的诊断，语言文本、面部表情这两个维度做得非常成熟，让临床医生可以辅助诊断，但是我们总希望能找到一些生物学的信息，比如说血压、心跳、呼吸。虽然它们的特异度、灵敏度和疾病机制之间可能没有直接关联，但是从医疗角度来讲，这些维度的检测是可操作的，具有可获得性和可依从性。我觉得人工智能计算机建模至少在疾病的监测维度、某些风险事件的监测维度，是有突破的，有可能实现的。从医疗角度来讲，现有的一些生物学维度的指标可能可以辅助诊断，但是我们希望看到疾病的临床现象背后，包括智能机器人问诊的背后，是不是能对接到一些生物学标志物。也就是说，ChatGPT是不是能回归到生物学其他维度，帮助我们探讨疾病的发生机制。

徐一峰：ChatGPT在精神心理领域具有广阔的前景，但是我觉得在实际应用中仍然面临着三大障碍。OpenAI非常注重数据和语料的质量，非常适合精神科应用。但它本身存在黑箱性质，或者说黑盒性质，如果把这个模型做出的决策用在人身上，不管是在诊断或者治疗当中，我都认为它是一个重大决策。那么，到底它能够用到什么程度，我觉得这是第一个要考虑的。第二点是，ChatGPT在中文环境下，面临着使用的稳定性、语言文化的适应性等问题。第三点是，在临床实践当中，精神疾病是带有病耻感的。所以在病人或者家属的强烈抗议下，我们就会做成中间性的诊断，比如说抑郁状态、焦虑状态。对于这样的数据，其基础可能本身就存在问题。

王振：如果想把人工智能用在精神心理的疾病诊疗上，结合除文本外的很多其他信息是非常重要的。比如面部微表情的识别、肢体语言的识别，再加上一些生理指标，将它们整合在一起的话，将来肯定能做很多工作。目前在精神心理领域，很多团队想借助ChatGPT来模拟医生或者专业人员的一部分前端工作，比

如预问诊、写病历、筛选出一些高风险人群等等，这样可以大幅度提高心理服务的可及性。但我觉得更加重要的是，ChatGPT做的只是人已经会做的事情，它只是提高效率。它是否能够辅助专业人员提升目前的能力，无论是在疾病的识别上还是在诊断治疗上，尤其是借助ChatGPT产生新的机制、新的知识，还有待验证。另外，ChatGPT对中文的理解，以及对中国文化特征的理解，还是不足的。希望未来我们能够建立自己的GPT平台，服务于中国的精神科临床。

Q 在医疗领域，我们是否需要建立中国的GPT模型呢？

胡鹏伟：我想这个问题的答案是肯定的。目前应用ChatGPT进行一些初步的研究是没问题的。但是涉及生命科学领域，它是受到严格监管的。我觉得未来肯定需要有一个国家级别的中间平台，来汇总这些数据，建立安全机制，然后在此之上建立不同疾病的模型。这样的话，我们在上面开展科研才会比较放心，也是比较长久的路。

吴梦玥：ChatGPT拥有非常好的建模能力，除了需要非常大的算力，还需要数据的支撑，数据本身的质量也是一个非常关键的点。其实前段时间，国内也一直在推行各个医疗平台之间数据的格式、数据之间的共享机制，如果建好的话，对于我们后续做大自己的预训练模型是一件非常有帮助的事。

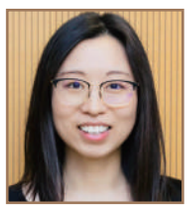
“读心术”，为什么不算准确却更受欢迎？

主持：陈光

北京邮电大学人工智能学院副教授。

嘉宾：孙静远

比利时天主教鲁汶大学博士后。



嘉宾：王少楠

中科院自动化所副研究员，纽约大学访问学者。

扫码查看原文

人类个体意识被包裹在大脑这团粉色迷人的肉球中，再经过头盖骨的厚厚包覆，形成一种坚硬的物理隔断，致使人与人之间无法直接进行意识交流。这种困境似乎在某种程度上催生了群体内的语言需求。于是，我们的人类祖先在咿咿呀呀之间与同伴交换了一部分的信息。但遗憾的是，语言有时并不能完整地反映人的全部所思所想，也由此催生了人类社会的复杂性。这也激发起了人类对准确知晓同类真实意图的兴趣，即读心。

刘慈欣在小说《三体》中曾构想出如此迷思：三体人可以通过电磁波，将真实的想法准确的传递给同类，是实实在在的透明思维。虽然这略显耿直，但历史上，人类对于能够“读心”的热情向来不减。从1920年代，首次记录到人类脑电图信号，到近些年来脑机接口技术的突飞猛进——首次破译了与手写笔迹有关的大脑信号，帮助瘫痪患者用意念写字；再到最近由chatGPT或Stable Diffusion所触发的生成式AI时代的到来，更使得人类神经信号解码步入一个蓬勃发展，尤为鼓舞人心的新阶段。那么，人工智能是否真的有可能解读人类的思维？目前关于神经解码领域又有哪些最新的进展？

陈光：在脑科学领域，如何理解“读心术”？

孙静远：我更倾向于把“读心术”称为神经解码（neural decoding）或脑机接口（brain machine interface），即通过一些技术手段，如功能性核磁共振（fMRI），脑磁图（MEG）和脑电图（EEG），来记录人类被试在看到某些刺激时的反应。人在接触到外界的刺激时，比如看一幅图画或一段视频，为了理解这

些信息，大脑内部会产生对应的神经信号活动。而这些神经信号可以被功能核磁共振、脑磁图和脑电图记录下来，接着我们可以通过一些技术手段，比如建立数学模型，在信号和刺激之间构建对应的映射关系。然后通过这些映射关系，我们就可以明确，产生对应脑信号的到底是怎样的刺激。总结而言，“读心术”或神经解码的逻辑流程是我们以比较高的信噪比*的方式去采集人的大脑活动为开始的，然后有了这些神经信号记录之后，我们再应用一些机器学习的算法，去建立模型来解析这些信号是哪些外界刺激所产生的波动。

王少楠：“读心术”（mind reading）严格来说，应该是读脑术。即我们建立某一种技术，把人的思维转换成可以直接被人类识别的文本或者图像的形式。

*信噪比（Signal-to-noise ratio，缩写为SNR或S/N）是科学和工程中所用的一种度量，用于比较所需信号的强度与背景噪声的强度。在通信领域，这个比率通常是以对数形式表达，单位是分贝（dB）。SNR的值越高，表示信号越清晰，即背景噪声越小。——来自GPT-3.5的注解

陈光：上述提到神经解码基本的逻辑流程是神经信号采集，对采集的数据进行分析，以及将信号映射并连接到所诱发它的刺激上去。那么，生成式AI在神经解码过程中应重点解决哪个环节的问题？

孙静远：人脑接收到特定刺激，进而产生对应的反应过程，这是一个非常复杂的非线性过程。另一方面，即使健康被试大脑的解剖结构有很大程度的共性，由于我们受到的教育不同，成长环境不同，这就导致不同的人类被试在看到同样的刺激的时候，所测量的大脑活动可能是非常不同的。被试个体间的差异以及人类将刺激转换成对应的神经信号过程的复杂性，导致我们想要建立信号和刺激之间的映射是极具挑战性的。在领域发展的早期，研究人员只能使用一些比较简单的数学回归模型来建立信号和刺激之间的关系。但是随着近些年，人工神经网络和机器学习的兴起，我们拥有了更好更先进的计算工具。我们能够建模比简单的线性回归要复杂很多的非线性的关系。总结而言，生成式AI填补了简单线性模型所无法涵盖的那些复杂关系。另外，生成模型也能够进一步地把刺激直接生成为人类感官更容易理解的东西。它不再是简单的两个数值之间并不直观的映射关系，而是说我们可以通过一些预先训练好的模型把我们感知到的刺激还原度较高地重建回来。

王少楠：神经解码领域目前主要有两个瓶颈，其一是神经影像信号噪声太大，随着神经降采集*机器性能的提升，这方面有望逐渐被解决；而生成式模型则主要是解决了另一个问题，在噪声这么大的情况下，相较于先前神经解码的技术只能对文本或图像的信号进行简单的二分类，现在生成式的模型则可以从信号中简单的获取稍微有信息量的东西，去大概的猜测出真实的刺激。

*神经降采样：是指将高频信号的采样率降低的过程，也就是将高频的模拟信号数字化时，将其采样频率降低，从而降低数据存储量，提高信号处理速度。——来自GPT-3.5的注解

陈光：也就是说之前所做的工作更多的是对离散信号的选择和判断；有了生成式模型之后，我们可以称之为“创新式的逼近”。生成式模型可以不是那么准确的，但是可以在一定程度上表达我们的意象。相较于之前的解码模型，我们在某种意义上讲，它变得不精确了，但是它变得更有用了。是不是可以这么理解？

王少楠：对。可以理解为，它传达出来的信息更被人接受了，或者表达的更像人了。之前输出就是一个离散的0和1，或者是或不是。这样的话，大家的使用体验没有那么好。现在的话，即使它在胡说八道，它也在说话。

陈光：从读心术要读的内容，即我们想要解读的被试的想法的目标来讲，之前的方法与现在的方法之

间最主要的区别是什么?也就是说,现在的方法能多大程度地“读心”?

孙静远:即使是现在的方法,离读心也还有很长的距离要走。如果现在,我给自己套一个便携式的脑电采集器,我想从这个机器采集的信号中实时的解析出来,我下一句要讲什么,就我目前对这个领域发展的理解,还是很大程度上不能实现的。限制主要在于所采集的信号的信噪比低。目前高质量的信号采集设备大多是侵入式的信号采集装置,但是这种设备不太容易普及。另外,尽管生成式模型让我们这个领域有了一个比较大的进步,即结果的可读性,但是其解析出来的内容的精确性是不能完全得到保证的。比如说我看到一只狗,但是我解析出来的图像虽然可能都是狗,但狗的毛色、品种,这些更为细节的特征可能和我实际看到的狗是不一样的。

王少楠:所以通俗上来说,生成式模型在神经解码上起到的作用,就是一个语言的美化器或优化器。之前可以做到的就是我输入一些脑电的信号或者是神经影像的信号,我们可以提供给被试一些不连续的离散的词汇。但有了生成式模型,就仿佛套了一层外壳,就可以把这些词汇串起来,生成人在草地上躺着,天气非常好之类的信息。这些信息不一定准确,但是人看起来会更加的舒适。

陈光:我理解神经解码可以对我们所想以及我们所感知的东西做一定程度的解码。对于目前“读心”要解码的内容而言,这两者是不是完全不一样?

孙静远:我个人理解两者是不太一样。如果我们想要重建的东西是人类实际感知到的刺激,如看到、听到的东西,那么我们所要重建出来的东西是明确的。只需要让我们的模型往目标上重建,就没有问题。但是如果说,这个东西是被试想象出来的内容,甚至连被试本身都无法准确描述他想象的是什么。这就给我们设置了一个障碍,即模型没有一个准确的重建目标;没有答案作参照就没办法评估,这个技术就很难演进。

王少楠:除了前面静远提到的实验重建目标是否明确这样一种区别以外,观察图像是一种感知(perception)行为,这是一种自发的过程。我们不需耗费太多的精力,我们生下来就会。但如果是用意念去控制一些东西的话,则是一种生产(production)的过程,这是需要被试主动地意念输出,来控制某个东西。科学研究目前还是主要集中到感知的过程,因为这样的话,实验的刺激更可控一些。

陈光:在脑科学专业领域,模型识别的结果是如何被评价的呢?

孙静远:我们目前使用的评价指标是语义分类准确率。它关注的是从内容的含义上,模型生成的东西和　被试实际看到的东西是不是同一个东西。我们会使用一个已经训练好的图像分类器,来评估模型生成的图像和被试真实看到的图像是否同属一类。另外还有一个指标是衡量模型生成的图像和被试看到的图像之间像素级别的重叠程度,即两张图像在多大程度上一样。但是,由于fMRI信号的限制,生成模型的限制,我们目前想要做到像素级别的完全复制是非常困难的。

王少楠:除了这些机器的评价指标以外,还有一种常见的评价方法,就是用人去评。选取大部分人都认同的结果作为金标准,来评价机器所产生的的文本或图像的质量。

陈光:我们前面提到了生成式模型的诸多优点,那么其在神经解码领域最主要的局限性又是什么?

孙静远：可靠性。模型生成的内容在质量上不是完全可靠的。

王少楠：大模型的词表非常大，但目前我们收入的人的神经影像的词汇量非常受限，可能最多只有几千。但是解码的时候需要在几千万个词汇上去检索，所以最后匹配到的结果并不是特别的好，所以现在的技术一般会在大模型的基础上做一些限制，解码的时候只生成与人看过的内容相关的一些词汇。所以，目前神经影像的发展可能还没有到达大模型的这样丰富的程度。

陈光：神经解码领域的瓶颈之一是神经信号采集噪声大。那么神经信号采集环节已有的可以训练系统的数据是否都是通过被试实验获取的？是否存在一些其他的方法，比如通过生成数据来提高模型的训练效率？

孙静远：举例来说，在曾经的一项研究中，我们主要采用的是给被试听中文的自然刺激，这种方式和以往的方式是有一定区别的。自然刺激是我们给被试听他们会在日常生活中听到的一些东西。自然刺激的一个优点是，相对于固定格式的刺激，采集的自然刺激数据后续所能做的实验范围更广泛。首先，我们没有在实验条件上进行非常严格的控制变量的设定，我们可以用这些数据去做其它（研究），只要是这些数据所包含的内容相关的研究。其次，在设计实验的时候，我们会在语料的筛选上面做一个设定，我们希望我们的语料能涵盖尽量宽广的语义空间，使我们的研究不局限在特定的主题。

王少楠：除了用这种真人采集的数据以外，国外一个团队曾提出这样的实验范式，他们认为既然大规模语言模型（large language model）或神经网络（neural network）几乎可以拟合一切的数据，那么我们现在有了fMRI的数据，有了模型，模型可以预测神经影像的数据。我们就可以在已有的数据上训练，来生成一些新的没有见过的神经影像数据。这是实验数据获取的另一个思路。

陈光：个体的思维习惯是否会影响信号采集的过程？

孙静远：前面提到神经解码要应对的挑战主要是被试差异和数据采集的噪声。被试差异不会影响到数据的采集过程，我们在采集过程中会要求被试全身上下都尽量不要动。而应对被试差异，我们其实更需要在解码模型上做一些突破。

王少楠：被试间的差异肯定会对我们的研究有一定的影响。但是我们现在关注的主要是共性的东西。就是想去解析人类在解码语言的时候有没有共同的规律。如果研究存在差异的话，会比较难，因为你不知道这种差异是来自被试间的，是数据质量间的，还是做实验的过程中由其他因素所引起的。进一步来说，在感知层面，不同被试之间是相同的，实际区别的是认知的过程。目前的研究，难以区分感知和认知。

陈光：用人和小鼠做信号采集的区别是什么？

王少楠：我觉得最大的区别就是，在人身上可以研究语言。如果是视觉信号的话，在动物和人身上都能开展；但是对语言研究来说，只有人具备这种能力。但是在人身上做实验大部分采用的都是非侵入式的系统。

陈光：就视觉传输而言，传输过程中的神经信号是否可以被采集，进而解码？

孙静远：我觉得对视觉传输过程中的神经信号进行解码是有意义的，但是为什么没有被实现呢？主要还是采集技术受限。目前国内外在信号采集时主要用的是fMRI，这种信号的缺点在于时间分辨率低。常见的采集间隔是1-2s。在这1-2s的过程中，视觉刺激从底层的视觉网络投射终末的视觉网络这一神经环路的传输过程已经完成，因此，在这样的时间分辨率下，我们在实验过程中采集到的不仅仅是视觉感知的信号，还有不同的视觉处理步骤信号的叠加。对这种混合的信号的解码是非常困难的。当然，如果我们有时间分辨率更高的信号，比如MEG和EEG，就有可能捕捉到视觉传输过程中某一特定的时间段。但是，MEG和EEG的空间分辨率很低。如果如提问者所言，捕捉视觉传输过程中的信号，需要一种时空分辨率皆高的采集方式。

王少楠：我觉得动态的信息肯定是非常有用的。只是现在的研究还比较初步，大家还集中在比较简单的对这种静态信息的解码上。另一方面是，fMRI相关的研究较多。虽然MEG的时间分辨率非常高，可以动态的追踪毫秒级的信号传递，但MEG设备相对来说比较昂贵。因此，从数量上看，基于MEG的研究远没有fMRI相关的研究多，引发的关注点也不多。

陈光：从几位之前的描述上看，底层的信号采集技术，似乎成为神经解码领域发展的瓶颈？近期神经解码领域内，在信号采集层面有什么突破吗？或者有什么最新的成果吗？

孙静远：便携式的EEG和MEG的采集确实有一些进展，甚至国内有一些商业化的案例。不过他们所做到的也不是说我们所谓的“读心术”，它们做到的更多是分类。比如对某些精神疾病的分类以及对某些简单操控的分类；因为这些信号本身对信噪比的要求不是特别高，所以哪怕用比较小的设备，也能够一定程度上捕捉到这部分的信号；但是对于我们目前这种较为精细的神经解码需求，会期望一种时空分辨率同步高的成像手段，能够帮我们实现高质量的神经解码。时空分辨率高，且便携的信号采集设备的发展是任重道远的。

王少楠：在fMRI神经信号的采集上，原先是3T的磁场，现在可以做到7T的磁场。磁场越强，信号采集的信噪比会更高。另外在采集实验数据的时候，有一个非常大的问题就是，我们没有办法保证被试的头部不做移动。所以目前有一些科研机构针对于此，设计出一种可以变形的头盔。被试佩戴头盔开展实验，可以在一定程度上保证数据的采集质量。

陈光：那么从整体上看，请两位老师总结一下，目前神经解码领域主要的发展瓶颈？

孙静远：我认为目前处在一种多点突破的局面，神经解码的进步离不开硬件和软件的共同发展，两者相互促进，相辅相成。硬件上的信噪比目前仍然是一个问题，信噪比的大小决定了后期模型解码结果的上限。但模型方面也仍然有许多可以继续挖掘的东西，我们希望不断逼近硬件所制约的信噪比的天花板。

王少楠：我们目前对大脑的认知实在是太少了。如果对大脑的运作机理有一些更深入的了解，对神经解码或者其他相关的应用会有一定的帮助。

陈光：生成式AI给神经解码领域带来了如此之大的发展，民众开始忧虑隐私泄露的问题。是否有一天，我们坐在公园的某张椅子上，自己全部的所思所想就会被监测到？AI读心术的伦理危机，两位老师怎么

看？

孙静远：遗憾的是，目前的研究离大家想象的这种程度相差甚远。目前非侵入式高信噪比的信号采集的硬件设施本身庞大，被试所想象的这种情况在当下几乎不可能。另外，技术的发展不能脱离监管，技术发展到哪个程度，也自然会有对应程度的监管。

王少楠：之前去过北京科技馆，有一个关于神经编码的简易设施。人们可以带上脑电帽，聚精会神地用意念控制小球左右移动。目前能实现的可能就类似这种。聚精会神的想某个东西，实现一个简单的分类的信号，这样的信号可能会驱动外界的实体做一些简单的运动。但是目前我们距离解码思维这件事情，还非常遥远。

另一方面，即使我们现在离读心术的时代还非常遥远，我们还是应该去理性思考与AI相关的伦理和隐私问题。即使读心术没有实现，现在相关的软件却已经在窃听我们的隐私信息。我觉得最好的方式是拥抱变化，首先我们应该了解我们是有这样的权利的，有这样的一种隐私保护的意识。提出要求，才有可能让监管变得更适合个体。（记者：涵棋；编辑：韵珂）

会议追问·李广晔：

脑机接口即将迎来“ChatGPT时刻”吗？

扫码查看原文

主持：陈光

北京邮电大学人工智能学院副教授。

嘉宾：李广晔

上海交通大学机械与动力工程学院副教授。

陈光：什么是脑机接口？脑机接口是如何运作的？

李广晔：在媒体上，脑机接口是近几年比较热门的话题。但从实验室层面，相关研究从几十年前就已经开始了。脑机接口（Brain Computer Interface或Brain Machine Interface）实际上就是获取大脑的神经信号，通过特定的数学模型对脑电信号进行解码来理解大脑的意图，再将这种意图转化成机器控制指令，实现大脑与外部设备的交互。例如，被试通过脑机接口打字和控制光标、机械臂运动。

陈光：这种大脑意识和信号之间的关联是有明显的对应关系，还是需要通过努力想象才能体现出一个强的脑电特征信号呢？

李广晔：脑电信号实际上是一个很微弱的信号，以控制机械臂为例，尽量努力控制自己手臂运动的动作信号可能会更强一些，但目前也有一些研究是基于运动想象，而不是尽量努力地让身体输出运动控制相关的信号来设计脑机接口范式。

陈光：这么说，意念和思想往往是更复杂、混沌的一种信号，所以脑机接口的解码不是执行一个生成式的任务，而是执行一个选择分类的任务（例如我想选择向左、向右运动）？

李广晔：是的，这种信号相对来说也是比较清晰的。从技术层面上，提取迅速变化的大脑意图目前还是比较困难的。

陈光：追问溯源，最早时期脑机接口的发展是怎样的？人们第一次想做这类尝试时大概是怎样一种情况？得到了怎样的结果？

李广晔：脑电信号最早可以追溯到1929年，德国的一位精神科医生首先在人类大脑头皮表面记录到了脑电信号。1969年，真正出现了一个比较具有代表性的脑机接口研究，美国国家卫生研究院（NIH）的Fetz在猴子大脑运动区植入电极，将电极与仪表盘连接，通过对猴子上采集到的大脑信号分析解码，如果猴子控制仪表盘指针达到目标区域，就给猴子一个奖励，这个实验相对来说是脑机接口上的第一次尝试。在这个过程中，猴子也学会了通过自己的脑电信号控制仪表盘来获取奖励，这只猴子从某种意义上也是脑机接口的第一个被试。最早在1982年，出现了第一个运动控制脑机接口的工作。研究人员在猴子大脑皮层运动区植入电极，发现当猴子做不同方向的伸手运动时，记录到的神经元放电模式是不同的。研究人员则可以基于这种不同的神经编码模式，去反向解码猴子在做什么运动。

陈光：这些早期的实验为脑机接口带来了突破性的进展，虽然实验范式比较原始，却也为后来的研究人员提供了很多启发，奠定了坚实的基础。那么，在早期的研究阶段，人们面临哪些困难与挑战？

李广晔：早期最主要的挑战是对大脑神经机制不够清楚，以及后来逐渐发现了一些对应的现象，却不清楚该如何运用这些知识。由于当时硬件条件的限制，早期在电极类型、采集通道、采样频率上都有较多的困难。并且由于早期计算机技术不太先进，当时在采集过程中信号流失或延迟也比较严重，解码算法也相对局限。

陈光：信号采集设备和技术、信号解码、排除噪声干扰，这些问题到今天依然是开展脑机接口研究的挑战。随着时间推移，新一代的技术给脑机接口研究带来了怎样的推动作用？

李广晔：首先，我认为脑机接口技术的演化非常依赖于信号采集技术的发展。早期的脑机接口采用的是非侵入式的方法（例如在头皮表面贴电极），后来出现了更好分辨率和更高灵敏度传感器，使获得的信号质量得到了改善，让前端可以得到更丰富的信息。后来也逐渐发展了侵入式电极，这些电极距离神经元更近，可以获取到更加清晰的神经信号。并且逐渐出现了更高通道的电极，甚至可以达到上千通道，可以获得更大空间范围的脑电信息。信号采集技术的发展和脑机接口技术的发展是息息相关的。

然后是信号处理和解码算法的发展。也就是我们拿到信号数据，该怎么去解释这些数据。在过去，可能采用的是一些简单的线性方法，随着计算机技术发展，滤波方法、特征提取、深度学习算法的发展有效地提高了脑机接口的性能和准确性。接着就是应用拓展方面，以前的研究更多在实验室层面，目前逐渐在临床或者人工智能领域得到了应用，例如脑卒中的康复，帕金森病、儿童多动症、睡眠障碍的监测与干预。

最后就是可穿戴性方面。早期的脑机接口设备比较笨重，现在逐渐往更便携甚至无线的可穿戴设备发展。总的来说，脑机接口就是在信号采集技术、信号处理和解码方法、应用拓展和可穿戴性这几个方面上逐渐地不断发展进步。

陈光：在技术发展领域存在一种说法叫“飞轮效应”。如李老师刚才介绍的，脑机接口在信号获取、处理、应用等方面得到了不断的发展，并且在各个技术层面形成了正反馈，即每一环都会促进其他各个环节的

大幅提升，也给其他环节创造了需求。那么，目前脑机接口领域是否进入了一个准快速发展的“飞轮时期”？

李广晔：近几年，诸如马斯克的Neuralink这类美国科技公司的发展，对于脑机接口带来了很大的宣传效果，其公司的设备在领域内也是非常具有创新性的。并且近几年，学术界和产业界合作紧密，也使得脑机接口领域得到了快速发展。

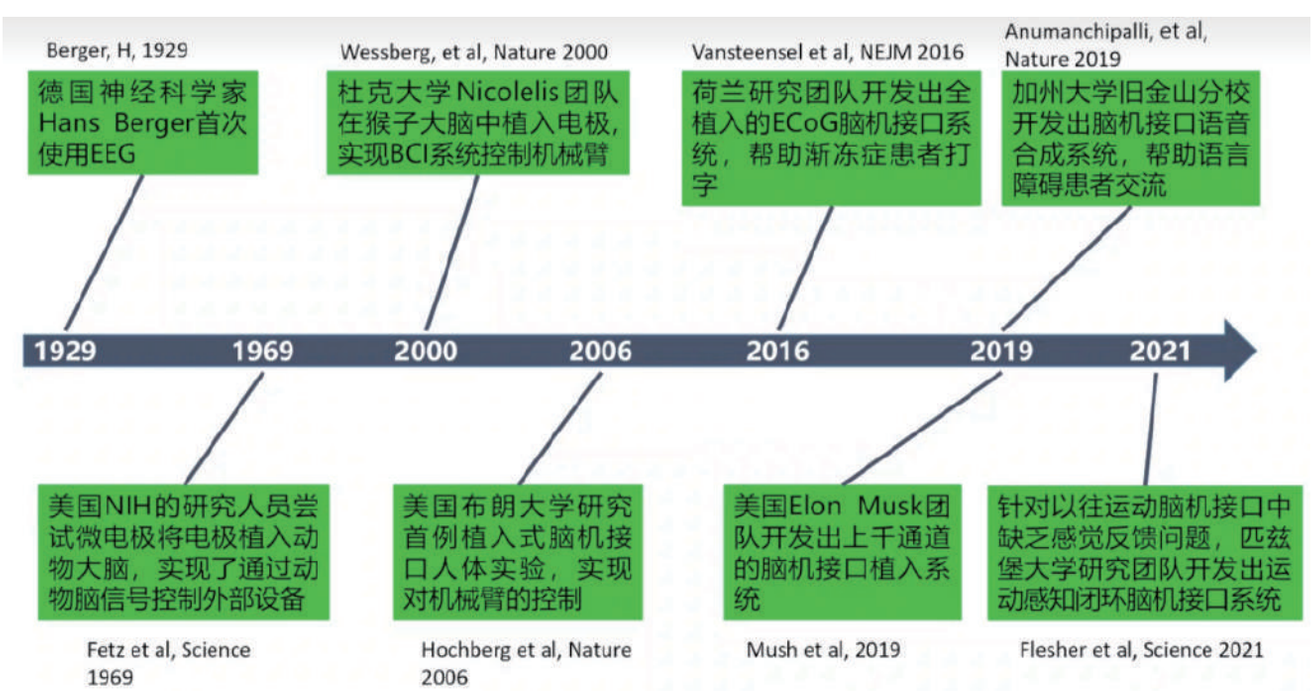
陈光：近几年，人工智能的技术在脑机接口中扮演着怎样的一个角色？为这个领域做了怎样的贡献？

李广晔：目前机器学习、深度学习之类的算法在脑机接口中主要应用在以下两个方面。

1、脑电信号处理上，脑电信号通常具有高维度性、复杂性，目前很多深度学习算法虽然是一个“黑箱子”模型，但相比之前手动提取特征，提取原始信号特征的效果很好，可以提高脑机接口的准确性、稳定性、快速性。2、改善人机交互和用户体验。比如一些自然语言处理、计算机视觉或者情感识别之类的技术融合到脑机接口系统中，这样计算机可以更好地理解用户的意图和情感，大幅度提升用户体验。

陈光：刚刚我们讨论了最早期脑机接口发展历程中的一些实验范式，那后续的发展演变中还有哪些里程碑式的节点？

李广晔：刚刚我们谈到了1929年和1969年的两个脑机接口的代表性研究成果，之后的一些里程碑式成果如下图所示。可以看到，从动物到人，再逐渐到人实验上的精细化控制，以及信号维度上的提高、多模态感觉信号的融合上，这些节点都给我们带来了不少的惊喜。并且越往后，这些节点发生的节奏是逐渐加快的。



▷图片由李广晔博士提供

陈光：脑科学和神经科学的知识是否已经融合到人工智能等其他领域上，目前对于脑科学的理解和我们对于脑机接口技术、功能的进步有怎样的联系？将来有什么趋势？

李广晔：脑机接口属于一个高度交叉的学科和技术，实际上我们对大脑信号的理解是脑机接口的基础。信号采集技术的发展促进了脑科学的进步，而脑科学的进步反过来又促进了脑机接口的发展。神经科学的探究可以帮助我们更好地了解大脑的工作原理。比如过去我们只采集运动区的信号来做运动控制系统，后来发现前额叶、顶叶等其他脑区也在运动过程中发挥作用。基于这些脑科学的知识，将来我们可以考虑构建多脑区的联合采集脑机接口系统。

陈光：近几年脑机接口技术的突破性发展有哪些？对于未来有什么影响？

李广晔：第一个方面，我认为近几年领域内的突破主要体现在更大空间范围和更高时间分辨率脑神经信号的采集。信号的质量是后续分析的基础和保障。随着这些更高维度的神经信息采集，我们可以做一些连续性、更加精细或更多自由度的运动控制脑机接口系统，将运动控制从离散转变成更连续、更自由的控制。第二个方面就是长期信号稳定采集上的进展。芯片植入系统需要考虑生物兼容性问题，如何避免长期植入引发的炎症反应对信号衰减的影响是一个关键问题。目前随着新型电极材料的发展，在长期使用上带来了很大的突破。第三个方面，自适应和闭环控制上近几年也得到了很大的突破。

陈光：这些硬件和可穿戴性设备的突破，在医学领域有什么比较大的进展，未来在哪些方面可能具有哪些进展？对于康复治疗有什么意义？

李广晔：目前脑机接口技术更多还是停留在实验室阶段，但长远来看，还是希望这些技术可以造福人类。医学领域的发展趋势目前集中在以下几个方面。1、康复治疗和运动恢复，例如由身体残疾、中风、渐冻症等导致运动障碍的患者，通过脑机接口技术帮助他们与外界交互。2、语音和沟通辅助，帮助存在语音障碍的患者实现和外界的通讯交流。3、疾病的监测和治疗，例如癫痫、帕金森病、儿童多动症与孤独症等神经系统疾病的监测，以及监测后进行干预治疗。

陈光：刚刚老师反复提到的非侵入式和侵入式脑机接口各有哪些？大概的机制是什么以及有什么优缺点？

李广晔：头皮脑电信号（EEG）是直接在头皮上佩戴脑电帽采集信号，优势是相对便携且无创；再深入的就是皮层电极（ECoG），直接在大脑皮层上采集神经信号；还有一种目前比较广泛的就是立体脑电（SEEG），主要用在癫痫的术前监测，是在大脑中植入直径0.8mm的细长针状电极，且单个电极上存在多个触点，优势在于可以同时采集到皮层和脑深部的信号；最后就是通过单细胞电位记录（single unit）采集神经脉冲信号（Spikes），可以直接采集到单个神经元的放电情况。这些方法各有优劣，ECoG和SEEG采集到的是局部场电位，是多个神经元混合的信号；Spikes则是直接采集到单个或多个神经元发放的信号，时间和空间分辨率是最高的；EEG主要是头皮上大规模的混合信号。这些信号采集方式的优缺点主要体现在相应技术是否需要开颅以及时间、空间分辨率上。

陈光：以前段时间苹果发布的Vision　Pro为例，除了脑电信号，眼动、手势的数据采集和识别也可以反映人的意图，这类多模态的数据是否可以另辟蹊径，成为未来脑机接口系统构建的另一个分支？

李广晔：这些利用人类生物传感器，如通过监测瞳孔变化识别用户感兴趣内容，包括刚刚提到的非侵入式脑机接口，未来在康复、辅助通讯、虚拟现实、娱乐等领域都有比较广阔的应用前景。例如，如果把EEG的头皮电极整合到苹果的头显内部，同时获取头皮脑电数据和其他多模态数据，可以有很多创新的应用前景。目前可以期待，但未来实现这些还是需要一些时间。

陈光：以目前的智能手机为例，存在一些私人隐私泄露的问题，对于目前脑机接口进展中，有哪些伦理上的考虑？有哪些工作已经在防止这类伦理问题的发生？

李广晔：目前世界范围对这个领域的伦理非常重视。1、个人隐私和数据安全问题，需要建立相应的准则和指南。2、用户知情权，以目前学术界为例，首先需要申请伦理许可证，实验前还需要与用户签订知情同意书，让被试知道即将做什么。3、数据按严格要求存储，未经过用户允许不可以私自传播。目前的应用暂时不存在这个风险，未来产品落地时需要在这方面制定严格的标准。

陈光：脑机接口作为一项相对神秘的技术，目前大众对脑机接口存在哪些比较普遍性的误解？

李广晔：目前的脑机接口主要还是停留在实验室阶段，一闪而过的意念识别目前还是比较困难的，媒体的过度宣传容易在这方面给大众带来恐慌。同时，大脑作为非常复杂的器官，真正理解大脑未来还需要很长的路要走。在未来的实际落地中，相应的伦理法规也会完善，因此目前大家也不要过于恐慌诸如“黑客进入我的大脑”之类的问题。

Q 在您看来，脑机接口的发展目前进入了什么阶段，距离“进入千万家”还需要多远？短期来看是否可以达到类似AlphaGo或者ChatGPT这种级别人工智能的拐点？

李广晔：从我的业内角度来说，我认为还是需要挺长一段时间的，短期内（5～10年范围内）暂时不会达到这种程度，还需要很多人的共同努力。诸如Neuralink的介入，的确会给这个领域带来一定的推动效果，但目前暂时没法给出实际的时间预测，技术还需要不断更新迭代，因此短期内应该还达不到一个新的高度。

陈光：是的，我认为具体任务（比如刚刚提到的医疗应用）的脑机接口系统还是有望在短期内实现突破，但通用型的脑机接口系统确实存在许多困难。（记者：孙广龙；编辑：Lixia）

ChatGPT与人的意识层次有何不同?

嘉宾:徐一峰

国家精神疾病医学中心脑健康研究院院长。

嘉宾:于欣

北京大学精神卫生研究所主任医师,教授。

嘉宾:林关宁

上海交通大学生物医学工程学院教授。

嘉宾:成素梅

上海社会科学院哲学研究所副所长。

扫码查看原文

2023年2月19日,天桥脑科学研究院(TCCI)、上海市精神卫生中心和上海图书馆共同举办了“柏拉图精神学园第一期对话”。本期对话以“科学与哲学视角下的精神现象反思”为题。在上海市精神卫生中心徐一峰教授的主持下,于欣教授、林关宁教授、成素梅教授,围绕意识、死亡、技术、进化等问题进行了科学追问。

Q 意识是一种认知功能,还是一种感觉?

于欣:对于意识,我们有不同的理解。心理学上,有心理学家定义的意识;哲学上,有哲学家定义的意识,精神病学也有精神病学家定义的意识。作为临床医生,我认为所谓的意识,首先是觉察,即对环境和自我状态的觉察。有觉察才有意识。最关键的是对自我的觉察,再进一步,是知道自我与环境的互动。在临床上,要判断一个人的意识状态或者意识自主性,就要评估其清醒态。人要足够清醒,就是要觉察,要保持足够的能量,注意力要保持相对的集中,对外界环境变化作出反应。

成素梅:我先讲一下人工智能。有一个经典的图灵测试叫“计算机能否进行思维活动”,因为过了奇点以后,计算机就变成了“人”。最近有讨论说,ChatGPT不仅能取代人的工作,而且可能在意识上要凌驾于人类。其实,从20世纪50年代提出图灵命测试开始,到60年代,这是哲学家们一直都在讨论的问题。传统哲学是从认识论的角度讲意识流,跟精神病学并不相关。60年代讨论人工智能的意识的时候,有一位哲学家认为无法回答“机器有没有思维活动”这个问题,因为思维活动是看不到的,人脑怎么动、机器人怎么动,都

是看不到的，只能从外在行为表现上进行判断，即使是图灵测试，依然是观察行为表现。所以说，应该问机器有没有意识。

这就涉及什么叫意识。所谓有意识，就是能跟环境进行互动，并且对环境的互动作出回应。如果一个机器能做到这两点，就可以认为它是有意识的。按这个定义，ChatGPT应该是属于有意识的行列，但和人的意识机制是不一样的。从行为上讲，或许能够跟环境互动的物，都应该是有意识的，动物也是有意识的，只是意识的层次高低不一样罢了。我们一般把意识活动当作人的一种精神活动，它有更高的要求，不仅仅是动物的要求。所以在我个人看来，从这个意义上讲，人的意识要比动物的意识高一个层次。而人工智能的意识可能还达不到，它和人的意识可能还是两个层面的。

Q 人类进化到一定阶段后，有没有可能在整体上克服对于死亡的焦虑？

林关宁：我首先从进化角度来讲，确实很多研究比较了我们的祖先，还有一些群体，比如美国一些保持传统生活的群体、亚马逊河流域的一些传统部落。在调查了他们的心理情绪状态后，可以发现这些人其实活得非常开心，他们的焦虑情绪非常低。我们的祖先也被证明，在当时，焦虑情绪是很低的。反而是随着人的现代化，我们的细化能力越来越强，焦虑、抑郁等情绪问题在逐渐上升。至于人类是否会进化到不怕死亡，可能当我们能够真正把人的意识数字化，放在人工智能里（才可能实现）。

成素梅：我可以从哲学的角度来讲一下这个问题。人什么时候不怕死？有几种方式，一种是信教。信教可以超越死亡畏惧，我认为有两个超越，一个是外在超越，就是借助某种力量去思考生死的问题，可能到一定程度后，你会超越；第二个是内在超越，就是自己修炼，去思考死亡的问题。我觉得有了一定的内在超越后，无需外在超越，你就能瞬间化解、想通，把死亡当作人生过程的终点，可能更容易接受（死亡）。第二种方法就是借助人工智能，在数字世界里永生。

于欣：死亡焦虑，其实是人类最本质的焦虑。我觉得你真的不用摆脱它。弗洛伊德一开始强调性本能，晚年的时候开始讨论死亡本能，但大家对这部分的关注不多，因为都不太愿碰这个话题。但是，死亡本能实际上是人类进步以及精神健康一个非常重要的力量源泉。好多时候我们觉得溃疡是需要治愈的，其实不是，它真的是你身上的“桃花”。太多我们认为需要抛弃的，或者认为应该割掉的东西，其实正是你身体里最珍贵或让你独一无二的东西。

所以我不觉得需要摆脱死亡焦虑，而且我觉得人类也不会摆脱。人类作为mortal（终有一死的）生物，变不成immortal（永生的），即使成为digital（数字化的）也成不了immortal。我们一直是mortal的，是独一无二的存在，我不觉得是坏事。

Q 有一个观点认为，现代科技在进步，人类社会在进步，但是人类物种在退步。如何看待这个观点？

林关宁：我们不能说科技的进步、文明的发展导致了人类物种的退化，而是说科技的发展使我们比较善于利用环境。人类修改大自然本来的状态，将之变成一个更加复杂的环境。这种情况下，人就会产生更加激进的做法，比如可能有些人会更加依赖于科技的发展，形成一种惰性，可以活得更久些。另一种说法是，按照自然界生存法则来讲，自然界非常残酷，为了繁衍可以无所不用。人类善于利用环境，使用科技让我们能够存活下去。这也付出了一些代价，就是我们保留了很多应该被淘汰的特征和形式，比如一些精神疾

病特征。

说成是人这个物种的退化，其实也不是，只是我们现在形成了一个更加复杂的物种。人类也在调整，进化成一种可能是大自然本来不希望我们走的方向。人工智能其实也是人类进化的一部分，并且进化的速度在加快。我们身体中大自然赋予的一些能力，可能跟不上进化的速度，这部分也在调整中，而人的适应性是非常强的。

于欣：我认为，其实没有所谓的“进化”“退化”，只有演化。演化相对中性，因为不知道人类是变得更好，还是变得更坏，好坏都是相对环境来讲的。我们不能够一边享受着现代科技进步带来的好处，一边又骂这科技把我们变差了，变懒了，变笨了。

从人类基因突变来讲，速度其实在加快，特别是最近100年当中，基因突变速度要更快。这种变异一定会带来正反两方面的结果。那么，为什么会有这种变化？其实是为了给人类适应环境带来更多的可能性。有推测说，人类将来会变成头大四肢小的形态，以便适应未来科技或信息爆炸的时代。有种说法是，我们要停在所谓的“幼态持续”（Neoteny）状态，像婴儿一样活着，这意味着我们要保持高度的好奇心、高度的活跃性、对外界的高度探索，同时大脑要更加发达。当然，这些推测是不是意味着未来人类一定会变成这样子呢？我不知道，因为环境会变成什么样子，我们也不知道。

成素梅：这是个科技哲学问题，关于科技与人和社会的问题。我们专业里有三种观点。一种观点叫技术决定论，认为人类社会、文明演化都是以技术为标志的，比如石器时代、铜器时代、第一次工业革命、第二次工业革命等，再到现在的人工智能。第二种观点是社会决定论，认为技术之所以能够影响人和社会，是因为人选择了它。这种观点否定了技术决定论，认为社会因素更重要，尤其是当代技术，如果没有国家的投资、资本的注入，也是做不出来的。所以，技术是社会选择的结果，是人来决定的，只是人在决定这个技术的时候没有看到它带来的负面影响。第三种观点是互动论，即技术和人是在互动过程中，共同适应、相互演化的。至于人类是退化还是进化，我不知道。

大家可能听说过卢德运动，英国当年机器大发展的时候，就有工人罢工、捣毁机器来拒绝机械化，因为他们担心机器会让人失去工作。但事实上，这个运动过后，人类不仅实现了机械化，还实现了自动化，现在还在智能化。人们在失去传统工作的同时，又产生了许多新的工作。所以，它伴随着一部分工作的消失，也伴随着一部分新工作的诞生。对于一代人来说，可能他们正好处于换代期，掌握的传统技能失效了，没有掌握新的技能，成为了牺牲者。对于年轻一代，对于在这个时代下成长起来的人而言，他们恰好就是适应的。所以这也是一个正反两方面的问题。

Q 如果人和新技术是互相适应的，那么传统的心理咨询治疗是否也能通过AI技术来实现？

徐一峰：用于心理治疗、心理咨询的技术已经有很多。我们做了很多用作认知行为治疗（Cognitive behavioral therapy, CBT）的人工智能机器人，也已经通过临床试验了。人工智能机器人、心理治疗机器人其实早就有了。在1964-1966年，MIT（麻省理工大学）人工智能实验室就做了世界上首个自然语言对话程序Eliza，它以卡尔·罗杰斯的人本主义及其学派为理论基础。

Q 精神问题严重到一定的程度，可以是一种疾病，而信息过载是一个重要的诱因。我们是不是可以更主

动些，在基因层面、外科手术层面等等，对人的基因疾病进行干预治疗？或者就像人类增强技术那样，把人变得更强，就不用担心外在的信息过载，是不是有这样的可能？

于欣：我们永远都不知道未来什么是好的，什么是不好的。精神疾病就两大类，都是非常严重的疾病。一类是发育性的，比如先天性的智力残疾，另一种是早老性的、退化性的。但有一些疾病是移行性的，是健康和不健康之间的移行过程，而且我们永远不知道这种状况在什么时候更适应这个时代。比如孤独症的孩子在社交隔离时，会更自在；而分裂症的患者可能在一个幻想空间里，会更舒服。这也涉及科研伦理，我们不知道应不应该下手，因为不知道做出来的完美基因，是不是一定是好的。我觉得精神疾病的魅力之处就在于，我们永远不知道它在当下的表现，再过10年，是好的还是不好的。

徐一峰：关于分裂症，的确有这类报道，说分裂症有其功能，有可能增加天才，或者增强知觉到别人心里的想法的能力。

林关宁：美国亚利桑那州立大学一位教授写了一本书，叫做《给坏情绪一个好理由》（Good Reasons for Bad Feelings）。书的大意是，很多精神疾病的状态，在某种情况对一些人来说是有用的，不要全都看成是有害的。从治疗方面来讲，你要把精神疾病当成有意义的东西，找到它的用处，这对于降低病感是有效的。从进化来讲，它以前是有用的，现在我们只是没有找到它的用处而已，将来可能就有用了。比如事情做得不顺利，你会不高兴，这是不是提醒你应该换一个方向？是不是应该做点其他事情来缓解一下？所以反过来想，不是所有东西都要从基因层面去治疗的。

专访张闯&张洳源：

AI和脑科学是未来十年最有前景的领域吗？

主持：陈光

北京邮电大学人工智能学院副教授。

嘉宾：张洳源

上海交通大学心理与行为科学研究院与上海市精神卫生中心双聘副研究员。

嘉宾：张闯

北京邮电大学人工智能学院教授。

扫码查看原文

陈光：正值毕业生志愿填报关键时期，首先想请两位老师分享一下当初如何一路走过来成为现在深耕在人工智能和脑科学领域的专家，当初填报的志愿与现在的专业方向是怎样一种关系？

张闯：早期选择的是模式识别与智能系统实验室，起初北邮最热门的专业是无线通信，没想到几十年过去，人工智迅速成为全社会关注的话题。风水轮流转，所以有时候是命运使然。这么多年我一直在模式识别领域做一些基础性的研究工作，走到了今天。所以，我个人认为有的时候是一种坚守和偶然。

张洳源：我本科毕业于北大心理与认知科学学院，当时称心理学系，所以我是纯心理学背景。谈到高考填志愿，我当时不太懂，志愿都是随便填的，也并没有填心理学，最后录取了心理学专业。我起初不愿意，是因为不了解，确实中国高中生接触心理学少。但进入大学后，开始了解心理学后，我就觉得非常幸运学习心理学。心理学是一个非常交叉的学科，有人文部分，如社会心理学，也有理科部分，如我从事的认知神经科学。

陈光：请两位老师简单介绍一下人工智能和脑科学两个方向各自的特点和应用领域。

张洳源：与人工智能不同，目前国内很少有大学在本科阶段开设脑科学专业（除了浙大）。包括学医的学生，也是进入临床分科室后，才接触到脑科学，这一点与国外不同。如果想选择脑科学，涉及以下相关专业（主要以前三个为主）：1.生命科学，如神经生物学；2.心理学，如认知神经科学；3.生物医学工程，如神经工程；4.其他，工科背景等专业。在应用领域，如果本科毕业，想申请硕博深造，主要分两大类：1.以动物实验

为主，是比较传统的神经生物学方式；2.以心理学、认知心理学、大脑网络为主，主要偏工科，集中在心理系。就业方面，脑机接口公司以及一些创业公司、医疗器械公司，都需要材料、算法、神经生物学知识等相关领域的人才。

张闯：人工智能是一个很宽的领域，并且是个年轻的专业。理论上讲，人工智能并非一级学科。类似于第二次工业革命之后，电、机械是“万金油”专业，现在人工智能也逐渐趋向于“万金油”专业。人工智能的发展最核心的三驾马车——算法、算力、数据，是支撑新一代人工智能技术前进的动力。就未来学科发展而言，更前沿的都在谈X+AI（即原专业+AI），用AI加持本专业。驱动新一代人工智能发展的核心动力还是算法、数理的突破。所以，基础学科依然是王道。软硬件平台、软件工程甚至是产品设计等都很重要。所以不妨把视角放宽，选择自己的爱好，再结合人工智能，更助力未来的长远发展。

陈光：脑科学和人工智能这两个领域在解决问题、推动科学进步方面有什么不同呢？

张闯：人工智能学科更侧重于解决问题，或者工程化。信息驱动的人工智能第三代的框架和体系，讲人如何构建智能的模式，甚至用计算手段实现。但今天大模型框架出现后，大家关注的是解决问题。如今人工智能已变成一种基础设施型的公众化能力，进而带动整个社会包括生产、生活方式的变革，这与前几代人工智能有本质区别，所以也会带来职业升级、知识能力的转型等方面的新需求。

张洳源：脑科学和人工智能的差别：首先是风格差异，人工智能本质还是创造一个好的机器，无论机器的功能如何，但脑科学不同，核心问题是人如何实现这一功能，它们的本质不一样。其次从某种程度来说，创造一个有感情的机器可以不同于人脑，也可以一样。反之亦然，人在解决问题的时候，不一定和现在能解决这个问题的机器一样。所以二者之间的比较才是两个学科可以互相沟通的地方。另外，人工智能过去的发展受益于算力，而脑科学受限于观测手段，观测小动物的脑如何活动比较方便，但观测人脑受限，所以在脑科学中还是存在瓶颈。

陈光：举例来讲，脑科学和人工智能目前有哪些研究方向和发展前景？

张洳源：2000年左右，人类基因组计划轰动全世界。最近十年，美国、中国、欧洲等纷纷提出“脑计划”，这是全世界继20年前就“基因是一个前沿学科”后，再次达成的共识。中国脑计划包含“一体两翼”。一体是一些基本的人脑认知问题，属于传统问题；一翼结合人工智能，统一生物脑和机器脑，催生新算法、工程应用（脑机接口等）；一翼结合脑健康，如阿尔茨海默病、抑郁症等，属于临床方向，催生大量既有基础科研又有药物研发（包括新型治疗手段）的方式。

张闯：人工智能领域的底层突破和范式创新基本都来自脑科学认知的革命。“一体两翼”或科学本身，实际上是连贯甚至是系统性的过程。从专业或发展角度来讲，如现在谈到最多的芯片，传统的CPU基本上可以国内能自主生产，但是下一个赛道是有AI能力的专用型芯片，在这方面，其实未来的发展、就业、市场都有很广阔的空间。

陈光：两位老师谈谈目前有哪些就业方向或职位，以及与未来职业规划相关的发展。

张闯：就业面很广，一方面算法工程师，主要参与基础性的算法研究，甚至在大模型到来时需要做工程

化的应用。另一方面，若可以形成对基础的框架和端到端系统的理解，可以从事规划、产品设计等前端研究。

张洳源：因为本科缺乏专门的专业，所以从事脑科学的人基本都是硕士或博士毕业。一方面从事神经生物学方向，大部分在药厂做研发。另一方面，对人的研究，比如医学影像、医疗器械公司（尤其是脑影像）、医院、高校科研机构的科研岗。任何单一的技能都不太能满足社会需求，现在需要复合技能人才。

陈光：除了从核心延展出的一些专业领域的职位外，还有在已有的职业基础上，面向应用端再去完善、应用的工作也越来越多。两位对此有什么想法？

张闯：现在人工智能是工程化、应用化强赋能的状态，人工智能未来成为基础设施之后，需要更多能驾驭基础设施的人。已经就绪的模型、框架如何实现端到端的解决方案，甚至能够提出应用的模式，在未来都有急需的场景。在这个过程中，我觉得大家可以在专业选择或求学过程中，充分利用互联网时代的优势，无论选择什么样的专业，都可以在网上找一些比较成熟的课程或专业，补充自己在不同领域的专长。

张洳源：我觉得人工智能的研究并不需要那么多人。最关键的是人工智能作为一个基础学科怎么去给其他领域赋能。AI+X是教育部最近几年新批的本科专业，以上海交大为例，本科专业AI+金融，入学后既学习金融知识，也学习AI知识，我觉得家长可以考虑这件事。更重要的是，像张老师所言，现在是信息大爆的时代，只要感兴趣就可以补充自己专业知识的短板，成为复合型人才是重要的。

陈光：脑科学和人工智能领域的发展趋势是怎样的？

张洳源：长期来看，遵循以下发展：更微观的角度，一定会聚焦到更精准、微观、细致、高效的策略和观测人脑活动。描述任何一个东西，观测是第一步，类似于望远镜研究天文，工具决定了你能看多久。更宏观的角度，有很多脑科学的知识，但缺乏如何转化到宏观大规模。如解决心理健康，除了传统药物，还有新的手段。比如EmoGPT，就是大规模利用互联网技术，用GPT陪患者聊天治病，就是更宏观尺度的应用。

张闯：从未来的发展趋势或方向来看，毋庸置疑人工智能已经进入大模型时代，甚至今天又到了一个转折点。2013年，深度学习是一个起点，10年之后，2023年，ChatGPT之后形成的各个领域的大模型是未来很重要的发展模型。大模型如何和传统领域相结合，快速形成人工智能的赋能，是很重要的。在无人驾驶等应用领域，AI的底层芯片（比如专用、通用芯片）的继续研发和制造呈现井喷式发展。从感知到认知能力的结合，类似于大模型融汇具身智能，未来有很多看点。莘莘学子如果能坚守10年，一定会成为人类社会进步发展最前沿的主力军。

陈光：就毕业生而言，哪些关键的技能和素质对将来的专业发展或学习是最重要的。请给大家一些建议。

张闯：1.良好的学习习惯。从某种意义上说，未来会存在超越人本身的智慧能力，越是这样，保持学习习惯，学会学习，是非常重要、内核的能力。高考之后才是追寻自我的起点，学会学习、终身学习，是对抗人工智能快速替代人的重要抓手。2.良好的共情能力、团队社交能力等。这些能力在未来会变得极为稀缺，越是这样越要开放一些，与同伴变成合作伙伴。3.科技向善。未来无论在哪个领域，人都会变得渺小，

尤其大模型的到来是双刃剑，科技向善其实是人更底层得精神追求和境界。无论将来做什么，走多远，都希望大家把这几点变成自己的核心竞争力。

张洳源：1.持续学习。高考不是终点，未来还有很多的坎需要不断地学习去突破。2.独立思考。自己要做什么还是要保持清醒的头脑和独立思考，不要过分的人云亦云。只有保持独立思考，才能非常明白自己能够干什么。

陈光：最后想请两位老师给大家一些建议和鼓励，在大家职业生涯里面应该考虑哪些因素，怎么样能够有一个更好的未来？

张洳源：高考确实是人生的一个节点，每个人都会有或美好或不美好的回忆，其实回头再看都不是那么重要。无论考的好坏，只是一个过程。未来的路都很长，不要争这一朝一夕。既然高考已经结束了，就好好填志愿，面对自己的大学生活、未来的人生。未来没有统一标准，虽然高考分数是一个标准，往后没有统一的标准，每个人要追求百花齐放，追求自己感兴趣的东西。

张闯：其实从某种意义上来讲，在未来寻找自己的热爱，比什么都重要。虽然热爱很短暂，学会坚持、保持长期的思维才能真正让每一位成为出彩的人生赢家。

陈光：我们相信脑科学和人工智能的迅猛发展会带来无限的新机遇，尽管这两个领域有非常多的挑战，但是对于热爱创新和探索的同学们，这些领域将成为展现自己理想的舞台。正如刚才两位老师所讲，不断学习和提升自己的技能，保持好奇心和创造力，努力成为这个时代的领军人才。让我们一起探索脑科学和人工智能的无限可能，开启属于我们更属于你们的未来！

鸣谢

这本书凝聚了2023年追问媒体部门一年的努力，更得益于那些给予我们无限支持的人们。首先，我们要特别感谢陈天桥、雒芊芊夫妇对媒体“追问Nextquestion”的大力支持。没有他们坚定支持与信任，就没有今天的我们。

此外，我们还要感谢集团每位研究员和同事们，你们的每一项研究、每一个项目、每一次讨论都为追问注入了不竭的活力。对于那些为追问媒体投稿、分享知识与见解的作者们，我们也由衷地表示感谢。你们的智慧和洞见是我们最宝贵的资产，你们的文字不仅丰富了我们的内容，也触动了无数读者的心灵。

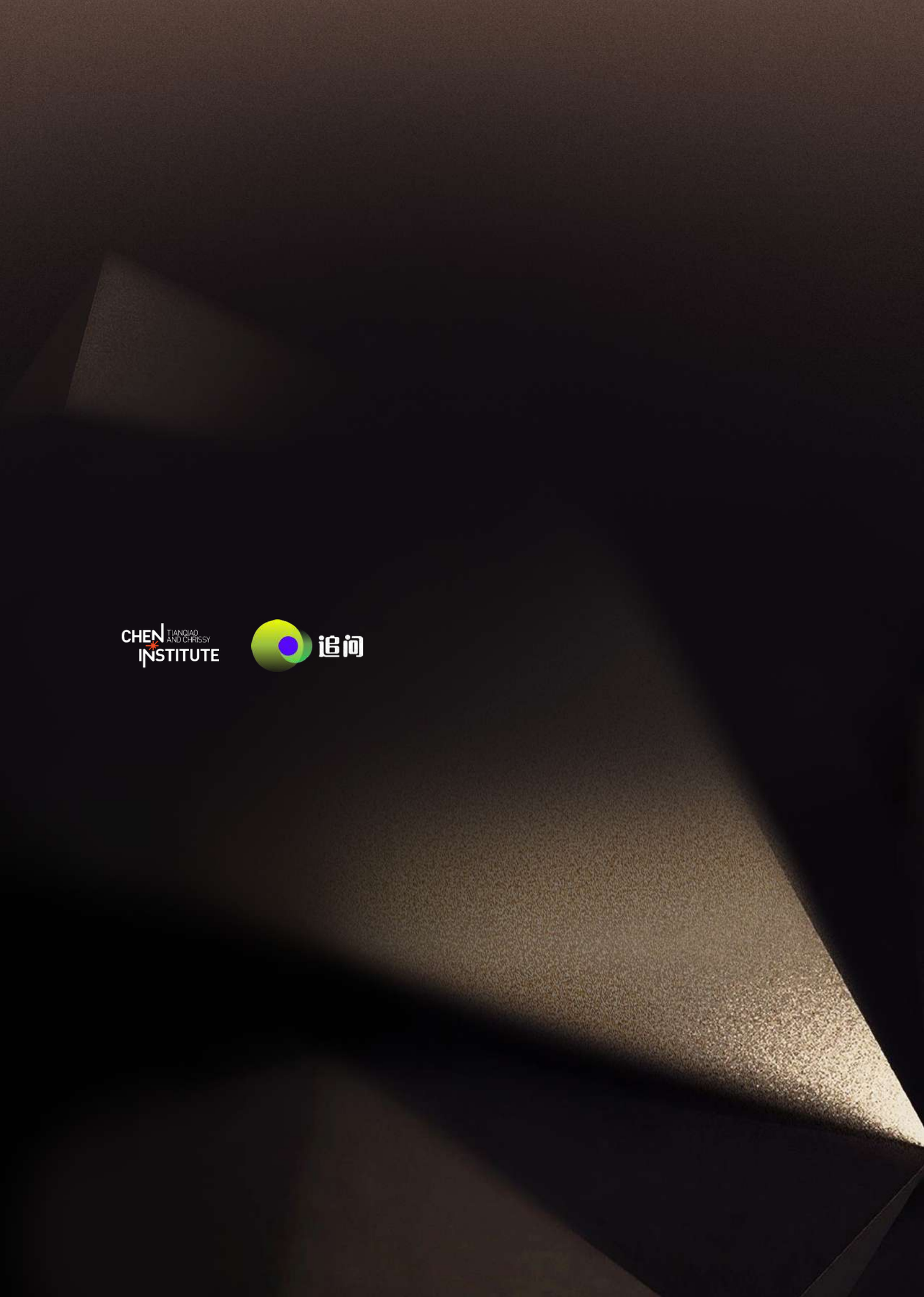
过去一年，我们还需要特别感谢参与天桥脑科学研究院（Tianqiao and Chrissy Chen Institute，TCCI）会议和接受我们采访的嘉宾们。他们是华山医院院长、TCCI转化中心主任毛颖教授、国家精神疾病医学中心脑健康研究院院长徐一峰教授、上海市精神卫生中心副院长王振教授、上海交通大学心理与行为科学研究院执行院长李卫东教授、华山医院神经外科副主任陈亮教授、华山医院神经内科副主任郁金泰教授、上海市精神卫生中心心境障碍科主任彭代辉教授、北京大学第六医院于欣教授、上海交通大学计算机科学与工程系副教授吴梦玥、上海交通大学与上海市精神卫生中心双聘研究员林关宁教授、上海社会科学院哲学研究所成素梅教授、美国旧金山音乐学院与旧金山大学Indre Viskontas博士、上海市精神卫生中心心理咨询与治疗中心主任仇剑崟主任医师、北京邮电大学人工智能学院副教授陈光、美国加州理工学院电气工程与计算机科学Yaser Abu-Mostafa教授、中科院自动化所模式识别国家重点实验室副研究员王少楠、比利时天主教鲁汶大学博士后孙静远、上海交通大学机器人研究所的助理教授李广晔（以上人名按微信文章发布顺序排序，不分先后）。

感谢纽约大学神经科学教授汪小京、中国科学院生物物理研究所的学术副所长何生教授、加州大学旧金山分校张逸真博士、北京师范大学文理学院心理学系特聘副研究员高天宇、芝加哥大学Yuan Chang Leong教授、北京邮电大学人工智能学院张闯教授、上海交通大学心理与行为科学研究院副研究员张洳源、科幻作家陈楸帆、同济大学副教授齐鹏、中央音乐学院音乐人工智能与音乐信息科技系副教授刘家丰、北京大学未来技术学院研究员段小洁、哈佛大学Alvaro Pascual-Leone教授、中科院脑科学与智能技术卓越创新中心研究员杨天明、北京大学心理与认知科学学院研究员周阳、北京大学心理与认知科学学院院长吴思教授、上海科技大学生物医学工程学院常任轨助理教授李远宁、北京大学人工智能研究院助理研究员杜凯、科幻作家顾备、马克斯·普朗克大脑科学研究所主任Moritz Helmstaedter教授等（以上人名按微信文章发布顺序排序，不分先后）。您们的精彩发言和交流探讨，极大的丰富了追问媒体的内容深度和广度。

最后，也感谢我们的读者们，是你们的持续关注和真诚反馈赋予了追问媒体以生命力和意义。没有你们的支持和鼓励，追问Nextquestion不可能持续生产高质量内容，更不可能达到今天的成就。

在探索人工智能与人类智能的未知旅程中，每一位参与者的贡献都是不可或缺的。正是因为有了你们的陪伴，我们才得以勇往直前，不断地追问问题的边界和探寻真理的本质。在此，再次感谢，并期待在未来的日子里，能有更多的朋友加入我们的行列，共同追求更多的精彩，创造无限可能。

追问编辑部


CHEN
TIANQIAO AND CHRISSY
INSTITUTE
追问