人工智能专题研究（三）

人工智能技术的未来通途刍议

徐英瑾

（复旦大学 哲学学院，上海 200433）

摘　要：现在关于人工智能的发展，社会上（甚至在行业内部）普遍存在这样的一种误解："通用人工智能"的目标本身可以通过"专用人工智能"领域内的技术累积来逐渐达成。然而，这种观点的持有者，既没有意识到将现有主流深度学习技术升级为通用人工智能技术所面临的巨大困难，也没有意识到人工智能工业的人为行业分工与人脑既有自然分工之间所存在的重大区别。在揭示这些困难的基础上，文本将给出一个消极性论点与一个积极性论点。前者是：目前的主流人工智能技术离达到"通用人工智能"的标准还很远，遑论达到"强人工智能"的标准；后者是：通向"通用人工智能"的真实道路从演化论思维"取经"，即从认知主体对于环境挑战的"适应性"与"节俭性"入手，来理解智能体运作的一般原理。

关键词：通用人工智能；深度学习；人工神经元网络；智商；全局性性质

中图分类号：TP18；B80-0　　**文献标识码**：A　　**文章编号**：1005-9245（2019）01-0093-12

DOI:10.14100/j.cnki.65-1039/g4.20180621.005

一、导论：关于人工智能技术发展现状的"乐观论""悲观论"与"泡沫论"

关于人工智能（以下依据语境或简写为"AI"）技术的发展对人类未来的影响，目前存在着三种意见。一种意见是认为人工智能的飞速发展必然会带来人类社会生活的极大丰富[①]；一种意见则认为人工智能的发展很可能会给人类的未来带来极大的危害，甚至导致人类的灭亡[②]；第三种观点则认为目前人工智能技术的发展其实并没有像媒体吹嘘得那么成熟，甚至在学科发展的基础问题上还面临着很多瓶颈，因此，过早地讨论其给人类历史带来的"颠覆性影响"，未免显得有点急躁[③]。这三种观点，可以分别被称为"乐观论""悲观论"与"泡沫论"。作为"泡沫论"的支持者，笔者近年来一直致力于对人工智能技术背后的学科基础问题进行批判性地拷问[④]。但笔者同时也注意到：除了要对以深度学习技术为代表的主流人工智能技术自身的局限性加以提示之外，我们还需要对"专用人工智能""通用人工智能"等概念的真正含义进行更为抽象的反思，以便彻底揭露包含在对于这些概念的错误运用之中的种种迷思。

而就关于人工智能的各种概念含义的理解而言，社会上（甚至在行业内部）普遍存在这样的误

收稿日期：2018-04-16

基金项目：本文系国家社科基金重大项目"基于信息技术哲学的当代认识论研究"（15ZDB020）的阶段性成果。

作者简介：徐英瑾，教育部青年长江学者，复旦大学哲学学院教授、博士生导师。

①黄欣荣：《大数据、人工智能与共产主义》，《贵州省党校学报》，2017年第5期。

②江晓原：《人工智能：威胁人类文明的科技之火》，《探索与争鸣》，2017年第10期。另外，正如这篇文章所提到的，与这篇文章类似的观点早就被斯蒂芬·霍金、伊隆·马斯克等国外名人在大量公众媒体上发布。

③李飞飞：《我曾在北京最冷的冬天感受到了人工智能的狂热》，《电脑报》，2017年11月6日。

④徐英瑾：《心智、语言和机器——维特根斯坦哲学与人工智能科学的对话》，北京：人民出版社，2013年版。

解：“通用人工智能”的目标本身可以通过“专用人工智能”领域内的技术累积来逐渐达成（很显然，正是因为受到这种误解的蛊惑，前面所提到的“乐观论”者才会将当下深度学习领域内的某些进步，视为“通用人工智能即将实现”的征兆）。然而，这种观点的持有者，既没有意识到将现有主流深度学习技术升级为通用人工智能技术所面临的巨大困难，也没有意识到人工智能工业的人为行业分工与人脑既有自然分工之间所存在的重大区别。在揭示这些困难的基础上，本文将给出一个消极性论点与一个积极性论点。前者是：目前的主流人工智能技术离达到“通用人工智能”的标准还很远，遑论达到“强人工智能”的标准；后者是：通向“通用人工智能”的真实道路借鉴了演化论思维，即认知主体对于环境挑战的“适应性”入手，来理解智能体运作的一般原理。

二、专用人工智能与通用人工智能之间的巨大距离

顾名思义，“专用人工智能”就是指专司某一个特定领域工作的人工智能系统，而所谓的“通用人工智能”（Artificial General Intelligence，简写形式：AGI），就是能够像人类那样胜任各种任务的人工智能系统。富有讽刺意味的是，符合大众对于AI之未来期待的虽然是AGI，而且西方第一代人工智能研究者——如明斯基（Marvin Minsky）、纽艾尔（Allen Newell）、司马贺（Herbert Simon）[①]，还有麦卡锡（John McCarthy）等——所试图实现的机器智能，肯定也是具有鲜明AGI意蕴的[②]，但是目前主流的AI研究所提供的产品都不属于AGI的范畴[③]。譬如，曾经因为打败李世石与柯洁而名震天下的谷歌公司的AlphaGo，其实就是一个专用的人工智能系统——除了用来下围棋之外，它甚至不能用来下中国象棋或者是日本将棋，遑论进行医疗诊断，或是为家政机器人提供软件支持。虽然驱动AlphaGo工作的“深度学习”技术本身，也可以在进行某些变通之后被沿用到其他人工智能的工作领域中去，但进行这种技术变通的毕竟是人类程序员，而不是程序本身。换言之，在概念上就不可能存在着能够自动切换工作领域的深度学习系统。由于一切真正的AGI系统都应当具备在无监督条件下自行根据任务与环境的变化切换工作知识域的能力，所以上面笔者的这个判断本身就意味着：深度学习系统无论如何发展，都不可能演变为AGI系统。

同情深度学习技术的读者或许会反驳说：为何我们不能研发出能够同时处理多个领域问题的深度学习系统呢？面对这一质疑，笔者还有一个非常重要的补充性论证。了解近年来西方认知科学哲学研究动态的读者都应当知道，美国哲学家佛笃（Jerry Fodor）曾提出过一个论证，以否定人工神经元网络能够支撑起一个完整的人类认知架构[④]。而在笔者看来，考虑到所谓的“完整的人类认知架构”与AGI之间的类似性，以及现有的深度学习机制对于传统的人工神经元网络的继承性，我们完全可以按照本文的语境要求，将佛笃的论证改造为一个对深度学习机制的“AGI化”进行质疑的论证。该论证如下：

大前提：任何一个AGI系统都需要能够处理那种“全局性性质”（Global Properties），比如在不同的理论体系之间进行抉择的能力（其根据或是“其中哪个理论更简洁”，或是“哪个理论对既有知识体系的扰动更小”，等等）。

小前提：深度学习系统所依赖的人工神经元网络，在原则上就无法处理“全局性性质”。

结论：深度学习机制自身就无法被“AGI化”。

很显然，该论证结论的可接受性，主要取决于大前提与小前提是否都是真的。笔者现在解释说明：二者都是真的。

此论证的大前提之所以是真的，是因为任何AGI系统都必须具有人类水准的常识推理能力，而常识推理的一个基本特征，就是推理过程所会涉及到的领域乃是事先无法确定的。譬如，人类投资家

---

① “司马贺”是Herbert Simon生前首肯的汉化译名（正常的译名应当是“西蒙”）。

② 譬如，“通用问题求解器”（General Problem Solver，简称GPS）的研究，就具有AGI研究先驱的意味。参见G. Ernst & A. Newell. GPS:《A Case Study in Generality and Problem Solving》, New York: Academic Press, 1969.

③ AGI组织。

④ J. Fodor：《The Mind Doesn’t Work That Way：The Scope and Limits of Computational Psychology》, Cambridge, MA：MIT Press，2000.

对于金融业务的谈论就很难规避对于国际政治军事形势的讨论（因为金融市场往往对国际军事形势的变化有非常敏感的表现），因此，我们就很难在讨论金融问题的时候预先规定“哪些领域一定不会被关涉到”。这一点甚至在做家务之类的看似琐碎的日常劳作中，也会得到体现——譬如，对于居室环境的整理在很大程度上并不仅仅关涉到“整洁”这一要求，而且还要兼顾“方便用户”这一要求，而该要求本身又指向了保洁员对于所有家庭成员的生活习惯的额外知识。换言之，跨领域的思维能力是即使连做家务的简单日常活动都需要具备的。也就是说，在涉及多样性的问题领域的时候，行为主体就必须具备对于来自不同领域的要求进行全局权衡的能力，而这就是佛笃所说的处理“全局性质”的那种能力（比如，在整理家居的时候，保洁员就必须对“视觉整洁性”与“家居物品的易取性”这两项要求进行调和或者取舍）。不难想象的是，上述要求不仅是被施加给人类的，而且也是被施加给一个理想的 AGI 系统的——如果我们希望 AGI 具有人类水准上的通用问题求解能力的话。具体而言，家政机器人、聊天机器人与军用机器人所面临的环境的开放性与复杂性，都要求支持这些机器人运作的人工智能系统具有类似于人类的处理“全局性性质”问题的能力。

上述三段论的小前提也是真的，即深度学习机制在原则上就难以处理这种具有领域开放性的全局性问题。要说清楚这一点，我们必须用最简单的语言厘清深度学习机制的基本运作原理。前面笔者已经提到，深度学习机制的前身乃是人工神经元网络，而非常粗略地说，人工神经元网络技术的实质，就是利用统计学的方法，在某个层面模拟人脑神经元网络的工作方式，设置多层彼此联结成网络的计算单位（如输入层—隐藏单元层—输出层等）。由此，全网便可以以类似于“自然神经元间电脉冲传递，导致后续神经元触发”的方式，逐层对输入材料进行信息加工，最终输出某种带有更高层面的语义属性的计算结果。至于这样的计算结果是否符合人类用户的需要，则取决于人类编程员如何用训练样本去调整既有网络各个计算单位之间的权重（见图 1）。一般而言，隐藏层计算单元只要受过适当的训练，就能够初步将输入层计算单元递送而来的“材料”归类为某个较为抽象的范畴，而所有的这些抽象范畴之间的语义关系，则可以通过某种记录隐藏层计算单元之触发模式的所谓“矢量空间”，

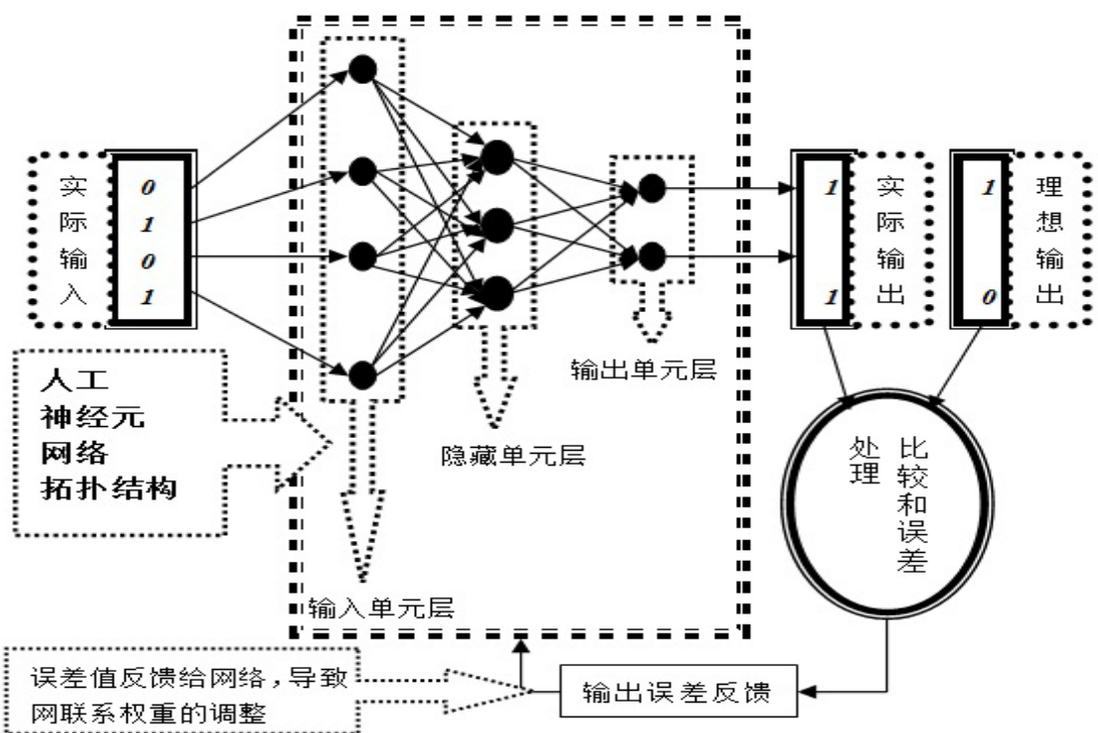

图1　一个被高度简化的人工神经元网络结构模型

而得到一种立体几何学的表征。至于作为人工神经元网络之升级版的“深度学习”，其与自己的前身之间的区别则主要在于两个方面：（1）其隐藏计算单元层的层次更多（因此可以通过更为复杂的概念层级来处理材料）；（2）系统的反馈学习算法一般会比传统人工神经元网更为复杂。这两个特征，一方面固然使得深度学习机制比传统人工神经元网络具备了更强的数据处理能力，但在另一方面也使得其运作需要更多的训练材料与更强大的计算机硬件来加以支持。

而从杰瑞·佛笃的立场上看，深度学习技术相对于传统神经元网络的这种进步，只具有有限的工程学意义，而缺乏重大的哲学意义。这是因为：深度学习机制也好，传统的神经元网络也罢，它们全都缺乏跨领域的学习能力。其背后的道理在于：任何一个深度学习系统的顺畅运作，将取决于如下关键要素：（甲）一个专门为特定任务（如人脸识别任务）打造的计算单元多层次框架（具体而言，程序员必须预先设置好每个计算单元的信息汇总函数与激发函数、计算单元的总数与层次安排方式、每一层计算单元所要提取的性质类型、整个框架的反馈算法，等等）；（乙）为系统提供大量的人脸图片识别样本。不难看出，其中的要素（甲）与要素（乙）都具有极强的领域专门性——譬如，专用于人脸识别的深度学习框架不能直接用于下围棋，而专门用于下围棋的深度学习框架不能直接用于下中国象棋，等等。同样的道理，用于人脸识别的训练样本对下围棋的训练样本也是无效的，因为二者的编码方式或许从头至尾都会有巨大的区别。

有的读者或许会感到奇怪：为何我们不能建造一个超级深度学习系统，让它能够处理不同的任务呢？说得具体一点，既然一个用来辨别物体颜色的深度学习系统能够对颜色的不同面相——亮度、饱和度、透明度，等等——进行有效的信息处理，那么，为何一个超级下棋系统就不能够对不同种类的下棋游戏进行统一有效的信息处理呢？

但不幸的是，这种超级深度学习系统其实是无法被研制出来的。相关意见如下：

第一，机器人的视觉传感器所捕捉到的一个蓝色花瓶的某种特定的蓝色色调，固然是具有亮度、饱和度与透明度等不同维度的，但这些维度都自然地依附在同样一个外部对象之上——因此，一个机器视觉系统不需要思考它自己究竟应当如何将不同的颜色属性捆绑在某个特定对象之上。与之相比较，任何一盘具体的围棋对弈与任何一盘具体的中国象棋对弈，所各自对应的都是不同的外部物理事件——除非有某种系统内部的“精神力量”将它们归为同样的范畴，否则，它们就是毫不相干的两个事件。或用哲学的行话来说，我们人类对于棋类的分类显然已经自觉地运用了某种抽象的分类概念，而深度学习运作的第一步所蕴含的视角却只能是面向“个例”（Token）而非“类型”（Type）的。退一步说，即使我们姑且承认深度学习系统是具备某种抽象概念形式的，我们也必须立即补充说：在深度学习中，那种随时可以成为人类的反思对象的概念分级系统，只能以一种前反思的形式，僵化地存在于计算单元的网状组织结构之中，而因此缺乏任何灵活性。

第二，即使前面提到的这个麻烦不存在，我们也要认识到：任何一个既有深度学习的运作，一般都会牵涉到大量的训练样本，并消耗大量的计算资源[①]。由此，一个能够跨越更多工作领域的深度学习系统，将会在相当程度上涉及到“训练样本从何处来”以及“计算资源从何处来”这两个棘手的问题——而与之相比较，人类自身却往往可以在训练样本稀缺的情况下，通过触类旁通学会新领域内的新技能。

面对以上批评，目前主流人工智能技术的支持者或许还会有如下回应意见：我们可以从认知科学家卡鲁瑟斯（Peter Carruthers）的“大规模模块理论”[②]受到启发，建造出一个具有很多模块的超级认知架构——其中每个认知模块用一种深度学习机

①美国生物统计学家里克（Jeff Leek）撰文指出，除非你具有海量的训练用数据，否则深度学习技术就会成为“屠龙之术”（Jeff Leek：《Don't use deep learning, your data isn't that big》, https://simplystatistics.org/2017/05/31/deeplearning-vs-leekasso/）。虽然深度学习专家比恩（Andrew L. Beam）亦指出，对于模型的精心训练可能使得深度学习机制能够适应小数据环境（Andrew L.Beam：《You can use deep learning even if your data isn't that big》, http://beamandrew.github.io/deeplearning/2017/06/04/deep_learning_works.html），但是比恩所给出的这些特设性技巧是否具有推广意义，则令人怀疑。

② Peter Carruthers：《The Architecture of Mind》，Oxford：Oxford University Press，2006.

制来实现，但对于它们的联合调用却使用某种传统的符号 AI 技术。这样一来，“专”与“通”的问题就可以得到某种一揽子的解决方案了。

但笔者认为，上述回应意见依然有很大问题。上述回应预设了某种比深度学习更高级的“调用系统”的存在——换言之，深度学习解决不了处理“全局性质”的难题，可以通过该调用系统来解决。但问题是：我们从哪里找到具有这种神奇功能的高级调用系统呢？上面的论证已经说明了：这样的调用系统不可能是某种更为高级的深度学习系统。那么，我们就只能期望这种调用系统乃是某种符号 AI 系统（因为只有符号 AI 系统才在某种意义上接近人类大脑的中央语义调配系统）。但既有的公理化进路的符号 AI 系统在对各个模块进行指挥与资源调配时，却只能预先将系统所可能遇到的所有问题语境予以预先规定，而这种预先规定显然会使得系统在遭遇程序员并未设想到的问题求解语境时陷入手足无措的窘境[①]。

另一个麻烦的问题是：即使上述困难可以得到某种意义上的克服，我们也无法保证对于现有深度学习模块的累积可以达到 AGI 的水平。这是因为：人类既有的大脑皮层分工是为了满足人类在采集—狩猎时代的生存需要而被缓慢演化出来的，而现有的人工智能研究内部的工程学分工方案，则主要是为了满足人类当下的商业与社会需求而被人为地制定出来的——举个例子来说，在围棋发明之前就已经出现的人类大脑架构，必定是没有一个专门用于下围棋的模块的——因此，在 AI 领域内对于诸如“围棋”模块的累积，显然无法帮助我们把握 AGI 所应当具有的认知架构的本质。或用带有中国哲学色彩的话来说，对于智能活动的“末”“流”与“用”的渐进式模拟，是无法帮助我们认识到智能活动的“本”“源”与“体”的。此外，从 AGI 的角度看，在 AI 界既有的“人为的分工方案”（见图 2）中，还存在着大量的逻辑上的分类混乱之处（比如，“常识推理能力”与“知识表征”“非确定性环境下的推理”等领域，彼此之间其实是犬牙交错、界限不清的），而这种混乱显然会进一步

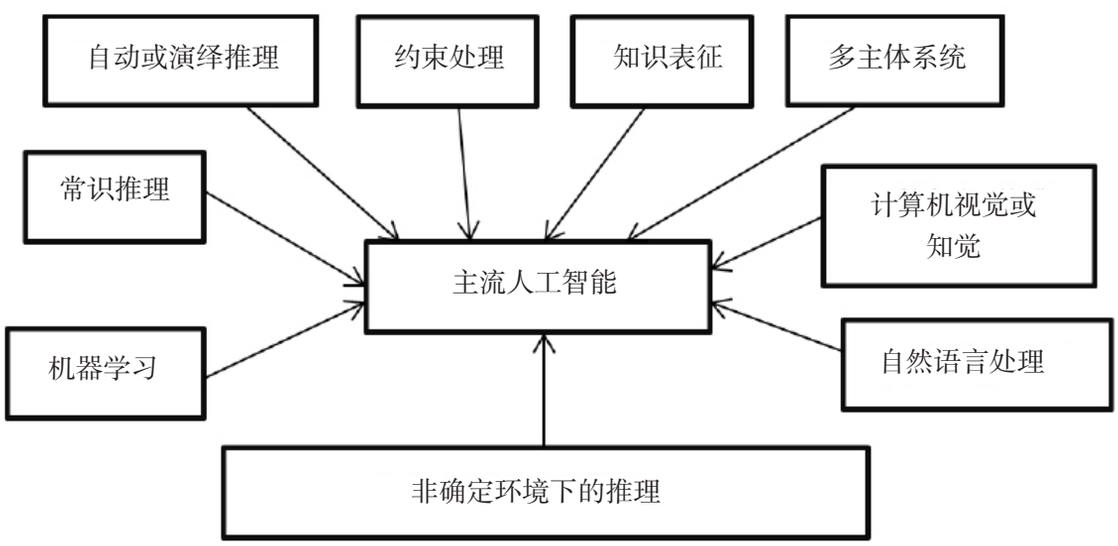

图2　目前主流AI学科内部的学术分工略图[②]

①譬如，在古哈（Ramanathan V. Guha）与麦卡锡（John McCarthy）的语境刻画工作中，研究者必须对整个系统所可能预先碰到的各种语境进行某种“未雨绸缪”式的刻画，并一一给出各个语境中的特定推理规则，以及语境之间的信息交换规则。这显然是一种非常缺乏灵活性的笨拙的做法。参见 R.Guha，J. McCarthy：《Varities of Contexts, in Modeling and Using Contexts》，edited by Patrick Blackburn et al, Berlin: Springer-Verlage, 2003: 164-177.

②该表的制定依据，乃是主流人工智能杂志《人工智能杂志》（The AI Journal）所给出的行业内部分类方案。转引自 J. Hernández-Orallo：《The Measure of All Minds: Evaluating Natural and Artificial Intelligence》，Cambridge: Cambridge University Press, 2017: 148.

使得对于这些研究分支的整合变得更加困难重重。

综上所述，目前主流 AI 技术的进展，并不能帮助我们真正制造出具有 AGI 基本特点的智能机器。那么，出路又在何方呢？一个很容易想到的方案是：机器智能的研究者必须得向业已存在的自然智能——即人类智能与动物智能——学习，由此寻找到突破的灵感。

三、如何向自然智能学习？

在通俗网络媒体中我们常常听到这样的评论：某某公司宣扬他们研制的人工智能系统已经达到了四岁或者五岁儿童的智商水准。使得这种说法具有意义的前提显然是：存在着某种横跨机器与人类智力的某种通用的“智商”概念——因此，心理测量学对于人类智商的测算方式，也可以被运用到测量机器智商之上。

虽然笔者对于用现成的人类智商标准衡量人工智能产品水准的做法持有强烈的保留态度，却对某种更抽象意义上的横跨机器与人类的“智商”概念保持开放态度。笔者的具体观点是：尽管针对人类的智商测量方式具有针对人类而言的物种特异性，但只要我们小心甄别其中的内容，我们依然可以从中寻找到某种能够沿用到 AGI 研究之上的一般性因素。

那么，心理测量学研究中的哪些因素是“特异于物种”的，而哪些因素又是具有普遍意义的呢？按照笔者的浅见，心理测量学的典型手段——如“问卷调查”——显然是带有明显的物种特异性的，因为该方法只适用于作为“语言动物”的成年人类，而不适用于动物，甚至是婴幼儿时期的人类。由此看来，在自然语言处理技术还没有完全成熟的

表1 人类基本心智能力及其检测方法和对于AI的推广意义①

能力名称	针对人类的检测方法	对于机器而言的推广意义
词汇能力	词汇填空测试（多选题）	对自然语言处理研究有意义
空间辨向能力	对于空间关系的辨别测验	对机器人的“具身化”研究有意义
归纳推理能力	残缺的单词序列接续测验	对一般的机器推理研究有意义
流畅使用词汇的能力	根据词汇规则回忆词汇的测验	对自然语言处理研究有意义
数字能力	加法测验	这是一般计算机最擅长的工作
联想能力	对于配对对象中未被明示的另一项的回忆测验	对一般的机器推理研究有意义
快速知觉能力	字母消除或是刺激比对测验	对自然语言处理研究有意义

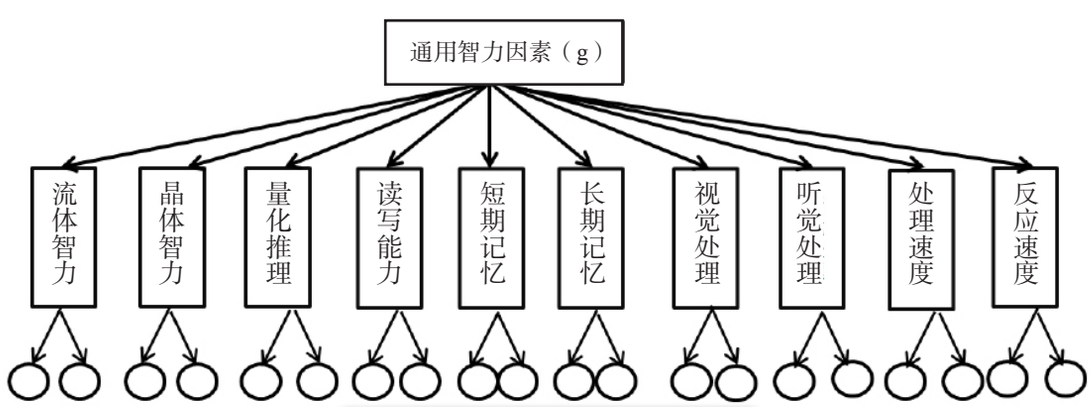

图3　卡特尔–霍恩–卡罗尔三层智力模型（底层智力项之内容在表中省略）[6]

①该表的理论依据是瑟斯顿的“基本心智能力理论”，转引自 J. Hernández-Orallo：《The Measure of All Minds: Evaluating Natural and Artificial Intelligence》，Cambridge: Cambridge University Press, 2017: 67. 笔者根据自己的理解，对原有的表格内容进行了增补（主要是添加了原表里没有的第三列内容）。

表2 卡特尔–霍恩–卡罗尔三层智力模型中的第二层与通用人工智能之间的关系

图3中第二层能力名称	与主流人工智能的关系	与通用人工智能的关系
流体智力（在缺乏前提知识的前提下创造新知识的智能）	无法再现	研究重点
晶体智力（在新语境中恰当提取并运用旧知识的能力）	无法再现（牵涉到“全局性性质”的处理能力）	研究重点
量化推理	在确定问题领域内可以精确再现	非研究重点
读写能力	在“自然语言处理”领域内可被部分再现	研究重点
短期记忆	部分实现	结合“流体/晶体智力”来研究
长期记忆	部分实现	结合“流体/晶体智力”来研究
视觉处理	在“人工视觉”领域内部分实现	结合“流体/晶体智力”来研究
听觉处理	在“人工听觉”领域内部分实现	结合“流体/晶体智力”来研究
处理速度	在确定问题领域内已被实现	非研究重点，主要与硬件配置相关
反应速度	在确定问题领域内已被实现	结合“流体/晶体智力”来研究

今天，通过问卷调查来对机器智能进行测量是没有意义的。但我们同样需要看到的是：心理测量学通过此类手段所要把握的心智能力要素，却很可能是具有横跨自然心智与机器心智普遍意义的。举例来说，心理测量学的鼻祖高尔顿（Francis Galton）就认为知觉的速度与智能的程度存在正相关关系[①]——而从计算机科学的角度看，如果“知觉速度”与传感器和相关支持软件的运作效率相关的话，那么，此类效率的提高当然就意味着系统智力的提高（当然是在其他条件保持不变的情况下）。而在笔者看来，上述从人类智能到机器智能的推理，也适用于心理学界对于各种“基本心理能力”的划分方案（见表1）。

如果说表1因为过多地对应机器的“自然语言处理能力”而依然缺乏足够的普遍性的话，那么，一种更抽象的人类智力分类方案——譬如卡特尔—霍恩—卡罗尔三层智力模型（Cattell-Horn-Carroll Three-Stratum Model）——则将覆盖更为宽泛的机

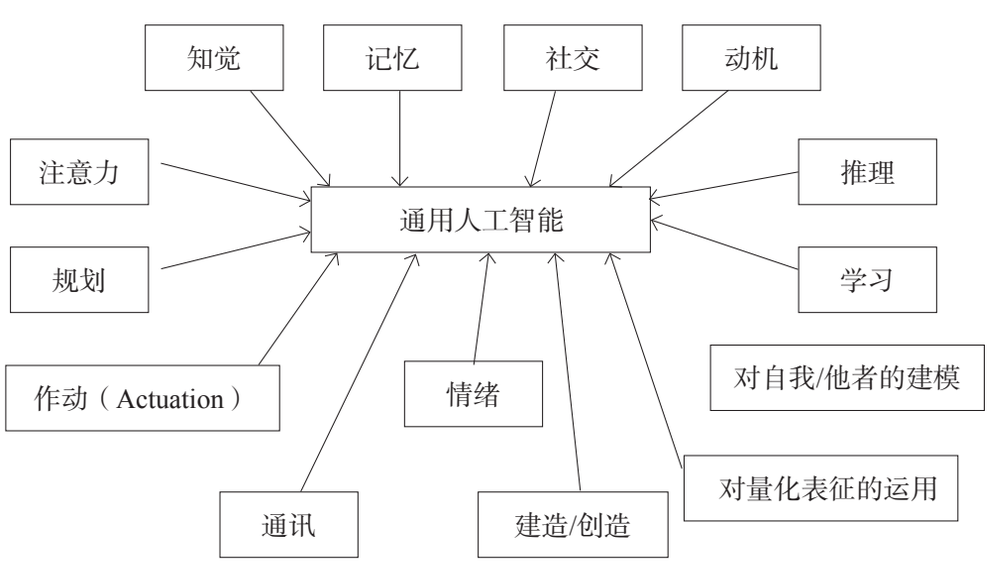

图4 AGI学界对于通用人工智能研究的子课题分布情况的勾勒[②]

① J. Hernández-Orallo：《The Measure of All Minds: Evaluating Natural and Artificial Intelligence》, Cambridge：Cambridge University Press，2017：64、68、145.

② S. Adams et al：《Mapping the landscape of human-level artificial general Intelligence》,《AI Magazine》，1993（1）.

器智能的能力范围（见图 3、表 2）。

但需要注意的是，即使我们能够依据表 2 在人类的智力能力分类与机器的智力能力分类之间找到一种对应关系，这也并不意味着我们能够立即根据这个对应表造出具备 AGI 特征的机器。其道理也是不难想象的：人类的智能是客观存在的，而心理测量学家只是试图对已经存在的事物的特性加以测量罢了。而到目前为止，AGI 仅仅是一个宏大的工作目标而已——因此，AGI 专家不可能在造出 AGI 之前就去奢谈如何测量 AGI 的程度。或说得更哲学化一点，关于人类智能构成的研究，仅仅为尚未完成的 AGI 研究提供了相应的“范导”而已（国际 AGI 学界对于这些“范导”的脉络勾勒，见图 4）——但范导本身并不意味着切实可行的工作路线图，正如一个建筑的招标书本身并不包含建筑本身的施工图纸一样。

那么，我们又当如何从自然智能那里寻找到关于建造 AGI 机制的“施工图纸”，而不仅仅是像图4 这样的“招标书”呢?

一个很容易想到的策略便是：心理学家所刻画的种种人类智力的分类形式，归根结底乃是由人类的神经系统所执行的。因此，只要我们对人类大脑的神经运作细节进行精确的描述，就可以从中抽象出一张精确的 AGI 工作图纸来。而时下方兴未艾的“类脑人工智能”（Brain-Inspired AI）研究思路，便是该思路的体现[①]。

不得不承认，与前面提到的深度学习的进路相比，类脑人工智能的研究思路的确更可取一些。虽然从字面上看，深度学习的前身——人工神经元网络——也是基于对人类大脑的仿生学模拟，但是在专业的神经科学家看来，传统的神经元网络也好，结构更为复杂的深度学习机制也罢，其对于人脑的模拟都是非常低级与局域的。与之相比较，类脑人工智能的野心则要大得多：它们要对人脑的整体运行机制作出某种切实的研究，并将其转化为某种数学形式，使计算机也能够按照“人脑蓝图”来运作。考虑到人类大脑的整体运作——而不是局域神经网的某种低端运作——能够以“神经回路”的方式向我们提示出更多的关于人类智力整体运作的信息，类脑人工智能的研究显然能够比主流的深度学习研究减少类似“盲人摸象”的错误几率。

不过，基于如下理由，笔者依然认为类脑人工智能研究还是隐藏了不少的风险。

理由之一：人脑的运作机制非常复杂，譬如，关于人类大脑的海马区是如何处理记忆信息的，现在的神经科学家也无法打包票说我们目前得到的认识是基本准确的。换言之，脑科学研究投入大，研究前景却不确定。在这种情况下，如果我们将 AI 研究的“鸡蛋”全部放在脑科学研究的“篮子”里，那么，AI 研究自身的发展节奏也将完全“受制于人”，而无法有效地分摊研究风险。

理由之二：目前对于神经回路的研究，占据了类脑人工智能研究者的主要注意力，因为对于神经回路的模拟似乎是相对容易着手的。但是我们很难保证某些神经细胞内部的活动不会对智能的产生具有关键性作用。而这一点也就使得类脑人工智能研究陷入了两难：如果不涉及这些亚神经细胞活动的话，人工智能研究或许就会错过某些关键性的大脑运作信息；但如果这些活动也都成为模拟对象的话，由此带来的计算建模成本将会变得完成不可接受（因为单个神经细胞内部的生化活动所对应的数学复杂性，就可以与整个大脑的神经网络所对应的数学复杂性相提并论[②]）。

理由之三：从已有研究来看，我们尚且不清楚“意识”（Consciousness）的存在是否是使得智能活动得以展开的一个必要条件。但假设其存在的确构成这样的一个必要条件，那么由此引发的问题便是：我们依然缺乏一个关于“意识如何产生”的成熟的脑科学理论，因此，我们无法预测何时我们能够从关于人脑的意识学说中得到关于“机器意识”的建模思路。更麻烦的是，如果彭罗斯（Roger Penrose）的“量子大脑假设”[③]是对的，那么我们就必须从量子层面上重新思考意识的本质。对于类

①该研究规划的代表项目是由瑞士牵头的“蓝脑计划”（Blue Brain Project），其目的是将人类整个大脑的神经联接信息全部用一个完整的数据模型予以记录。

②[英]玛格丽特·博登：《人工智能的本质与未来》，孙诗惠译，北京：中国人民大学出版社，2017 年版，第 106-107 页。

③ Roger Penrose：《Shadows of the Mind: A Search for the Missing Science of Consciousness》，New York：Oxford University Press，1994.

脑人工智能研究来说，这一方面会使得研究者放弃建立在经典物理学构架上的传统图灵机计算模型，另一方面又会迫使他们去思考“如何在与传统计算机不同的量子计算机的基础上去构建认知模型”这一艰难的课题[①]。这一切无疑会使得类脑人工智能的工作进度表越拉越长。

理由之四：即使人类目前已经掌握了大脑运作的基本概况，我们也无法保证由此得到的一张大脑运作蓝图可以被机器所实现。其背后的道理是：使神经活动得以可能的底层生物化学活动具有一种与电脑运作所依赖的底层物理活动非常不同的物理学特征，而正是基于这种不同，科学界才将前者称为“湿件”（Wetware），以便与后者所对应的“硬件”（Hardware）相互区别。但麻烦也出在这里：我们都知道，高性能航空发动机的运作蓝图，一般都需要非常特殊的航空材料来加以“落实”，因为这些蓝图本身已经在某种意义上透露了关于相关运作材料性质的信息——与之相对比，我们又怎么能够期望关于大脑的运作蓝图，可以不包含对于特定生化信息的指涉，而可以被完全运用到硅基器材之上呢？

分析到这一步，读者可能会问：既然简单地模拟我们的人脑并不是“向自然智能学习”的方便法门，我们又有什么别的出路呢？

该问题的答案便是：我们必须学习胡塞尔的“想象力自由变更”的办法，对“智能”的本质进行直观剖析。套用到本文的语境中去，该办法的具体操作步骤便是：对各种可能的智能类型进行展列，并由此为出发点对各种可能的智能形式进行想象，最终剔除关于智能的偶然性成分，找到智能的本质性要素。

提到“人类之外的自然智能”，很多人或许会马上联想到灵长类动物的智能。但考虑到人类与其他灵长类动物之间的类似性，出于“剔除人类智能实现方式中的偶然性因素”这一目的，我们最好还是在非灵长类动物中寻找智能活动的标记。譬如，我们在作为软体动物的章鱼那里找到了复杂的行为模式，尽管作为软体动物的章鱼具有与灵长类不同的神经系统（这种另类的神经系统使得章鱼可以在大脑与吸盘处分置记忆系统）[②]；甚而言之，有些专家还认为植物也可以具有“短期记忆”与“长期记忆”等心智功能，尽管植物是没有严格意义上的神经系统的[③]；甚而言之，有人还认为细菌也可以体现出某种“群体智能”——譬如，通过信号传导蛋白质（而不是神经组织）的帮助，大量聚集的细菌可以解决一些复杂的计算机程序才能够解决的优化问题[④]。

从以上列举的这些例子中，我们不难发现使得智能活动得以存在的偶然因素与本质性因素。神经系统的存在，恐怕就是可以被“约分”掉的偶然性因素，因为植物与细菌的智能都不依赖于其存在。特定智能行为对于遗传代码的依赖性，看来也必须被“约分”掉，因为对于章鱼的研究表明，章鱼大量复杂的捕猎与逃逸行为都是后天习得的（换言之，遗传基因只能为章鱼获得复杂行为的潜力进行编码，而不能对具体的行为本身进行编码）。与之相比较，不能被“约分”掉的本质性因素则包括：（1）如何面临环境的挑战并给出应战的模式；（2）如何在给出这种应战的同时最有效、经济地利用智能体所具有的资源。这也就是说，尽管自然智能的具体表现形式非常丰富多彩，但是其所具有的一般功能结构却是相对一致的。

不过，即使是这样相对简单的对于“智能”的

①笔者本人曾于2017年6月在美国加州圣迭戈召开的世界意识科学大会上与彭罗斯爵士交谈，向其讨教“量子计算机是否能够实现量子意识”这一问题。他对这一问题给出了否定的回答，因为他认为量子计算机的运作依然需要经典计算机的运作提供某种基础。虽然笔者不敢肯定他的这个回答一定是正确的（因为据笔者所知，像“D-WAVE”这样的“退火量子计算机”在硬件构成上就与传统计算机非常不同），但笔者至少可以肯定的是：即使关于大脑的量子意识理论是对的，也并不是说任何意义上的量子物理学现象都可以引发意识（否则“意识”就本该是无处不在的）。因此，在量子计算机研究与对于量子意识的机器实现之间，应当还是存在着大量的理论空白需要填补。

②对于章鱼的行为与心智的研究，当下已经成为西方学界的一个新热点。参见Peter Godfrey-Smith：《Other Minds: The Octopus and the Evolution of Intelligent Life》，London: William Collins, 2017.

③对于植物“心智”的研究，参见Daniel Chamovitz：《What Plants Knows》，London: Oneworld.

④E. Ben-Jacob：《Learning from bacteria about natural information processing》，《Annals of the New York Academy of Sciences》，2009（1）.

功能性界定，也足以对当下主流的 AI 研究构成某种严肃的批评：

批评之一：自然智能虽然是为了面对环境的挑战应运而生的，但对于这些环境挑战的种类与范围，则往往没有非常清楚的界定。举例来说，乌鸦肯定是在人类建立起城市的环境之前就已经演化出了自己的神经系统，因此，其所面临的原始环境肯定是不包含城市的——但这并不妨碍日本东京的乌鸦成为了一种高度适应城市环境的生物，并因此成为困扰东京市民的一项公害。这也就是说，即使是鸟类的自然智能，可能也都具有佛笃所说的那种处理“全局性性质”问题的能力，尽管我们尚且不知道其是如何获得这种能力的。与之相比较，传统的 AI 系统（无论是传统符号 AI 还是深度学习，还是所谓的“遗传算法”[①]）却需要对系统所面对的环境或者是其所要处理的任务类型作出非常清楚的规定，因此是不具备那种针对开放式环境的适应性的。

批评之二：自然智能往往采用相对经济的方式来对环境作出回应——譬如，我们很难设想一只猴子为了能够辨认出其母亲，需要像基于深度学习的人脸识别系统那样先经受海量的“猴脸”信息的轰炸，就像我们很难设想柯洁在获得能与 AlphaGo 一决高下的能力之前，需要像 AlphaGo 那样自我对弈几百万次棋局一样。需要注意的是，尽管德国心理学家吉仁泽（Gerd Gigerenzer）曾在“节俭性理性”的名目前系统地研究过自然智能思维的这种经济性[②]，但至少可以肯定的是，“节俭性”并不是目前主流人工智能所具有的特性。相反，对于信息的过分榨取，已经使得当下的人工智能陷入了所谓的“探索—榨取两难”（The Exploration-Exploitation Dilemma）：换言之，不去海量地剥削人类既有的知识，机器便无法表现出哪怕出于特定领域内的智能——一旦机器剥削既有人类知识“上了瘾”，机器又无法在任何一个领域内进行新的探索[③]。与之相比较，相对高级的自然智能却都具备在不过分剥削既有知识的前提下进行创新的能力（比如少年司马光在“司马光砸缸”这一案例中所体现出的创新能力）。

综合本节的讨论来看，向自然智能借脑固然是未来 AGI 发展的必经之路，但如何确定我们所需要模拟的自然智能的“知识层次”，则是在相关研究展开前率先要被回答的问题。与类脑人工智能研究者对于大脑具体神经细节的聚焦不同，笔者主张通过“想象力自由变更”的方法，对智能活动的本质——有机体对于开放式环境的节俭式应答——进行抽象，由此获得对于 AGI 的一些更富操作性的指导意见。至于如何将这些意见落实到具体的 AGI 研究工作中去，笔者曾在别处给出更富有技术细节的介绍，在此不再赘述[④]。

四、引申性讨论：再谈“强人工智能”与“超级人工智能”

本文既有的讨论，其实已经足以澄清这样一个论点：既有的专用人工智能之路，并不能真正通向 AGI，因为后者对于智能活动本质的涉及，并不是前者的题中应有之义。而对于有些读者来说，这样的澄清似乎还漏掉了在目前的媒体中被广泛炒作的两个概念：一个是“强人工智能”（Strong AI），一个是“超级人工智能”（Artificial Superintelligence）。那么，AGI 与它们之间的关系又是什么呢?

首先可以肯定的是，强 AI 既不是 AGI 的对立面，也不是其同义词。与强 AI 对应的概念乃是“弱 AI”（Weak AI），而弱 AI 也既非专用 AI 的对立面，亦非其同义词（尽管二者在外延上有高度重合，详后）。说得更清楚一点，弱 AI 指的是计算机对于自然智能的模拟，而强 AI 指的是计算机在上述模拟的基础上对于真实心智的获得，二者之间的区分，牵涉到的乃是“虚拟心

① Gigerenzer，G，Todd，P.，the ABC Research Group：《Simple Heuristics that Make Us Smart》，Oxford：Oxford University Press，1999.

②在所谓的“遗传算法”中，程序员必须对程序所面对的问题求解环境预先进行刻画，并确定程序的所有“基因型”与“表现型”。这种做法依然是比较僵化的，无法满足 AGI 的要求。关于“遗传算法”的详细评论，参见《心智、语言和机器——维特根斯坦哲学与人工智能科学的对话》，北京：人民出版社，2013 年版。

③ J. Hernández-Orallo：《The Measure of All Minds: Evaluating Natural and Artificial Intelligence》，Cambridge：Cambridge University Press，2017：64、68、145.

④徐英瑾：《心智、语言和机器——维特根斯坦哲学与人工智能科学的对话》，北京：人民出版社，2013 年版。

灵”与“真实心灵”之间的区别[①]。与之相比较，专用 AI 与 AGI 之间的区分则是 AI 系统自身运用范围宽窄之间的区别。因此，从概念的外延角度上看，一个 AGI 系统或许可能是弱 AI，也可以是强 AI（因为一个达到 AGI 标准的系统是否能够配得“真实心灵”，依然是一个有待争议的心灵哲学话题）——而专用 AI 系统则只可能是弱 AI（因为真实的心灵肯定是具有跨领域的问题处理能力的）。这几个概念的关系，可以通过图 5 得到概括。

再来看“超级 AI”这个概念。笔者个人认为这是一个非常含糊的字眼，因为“超级”本身的含义就非常含糊。如果就 AI 系统在单项能力上对于人类的超越的话，那么现在的 AlphaGo 就已经是这样的超级 AI 了。但如果“超级 AI”指的是某种能够比人类更为灵活地统调各种能力与知识领域的 AI 系统而言，很显然这样的系统还没有出现。但即使存在着这样的系统，如何界定“超级”二字的真正含义，依然会成为一个值得商榷的问题。其背后的道理是：正如前面笔者已经指出，任何智能体都无法不在尽量节俭地使用资源的前提下，对开放环境中存在的挑战进行“无所顾忌”的回应，因此，即使是所谓的“超级 AI”，也不可能在其运作中无限地挥霍其运算资源，并要求无限的前设知识作为其推理前提。换言之，这样的系统依然是与我们人类一样的“有限的存在者”，并与我们人类一样面临着某种终极的脆弱性。

不过，如果我们将“超级”的门槛降低，并在“比人类稍微更灵活、更具创造性一点”这一意义上使用“超级 AI”这个字眼的话，那么，制造出这样的系统，在概念上是可能的。说得更具体一点，我们当然可以由此设想：某种 AGI 系统能够以一种比人类更具效率的方式进行联想与类比，找到问题的求解方略，而这种系统所接受的某些外围设备的强大物理功能，显然也能够使得整个人造系统获得比人类整体更强大的决策与行动能力。

那么，这样的一种超级 AGI 系统，是否能够对人类的文明构成威胁呢？对于该问题，笔者暂且保持开放的态度。不过，正如本文所反复说明的，即使有一天这样的 AGI 系统问世了，其技术路径也会与主流的人工智能技术非常不同。因此，那种凭借主流人工智能技术的进展就大喊“奇点时刻即将到来”[②]的论调，依然是站不住脚的。换言之，虽然“AI 威胁论”并非永远会显得不合时宜，至少就目前的情况而言，高唱此论调的确显得有些杞人忧天。

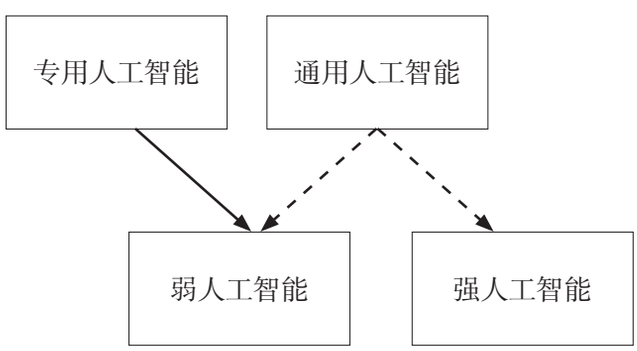

图5 关于专用人工智能、通用人工智能、弱人工智能与强人工智能四者关系的概括

注：图中实线表示“从属关系”，虚线表示“可能的从属关系”。

①强、弱 AI 区分的根据在此文献：John Searle：《Minds，Brains，and Programs》，《Behavioral and Brain Sciences》，1980(3)：417-424.

②“奇点”在库兹威尔（Ray Kurzweil）这样的未来学家那里，特指人工智能技术能够颠覆性改变整个人类文明的那个历史时刻点。但在笔者已知的范围内，在国际主流的科学哲学界与心灵哲学界，很少有人认真看待这种“奇点理论”。

On the Future of Artificial Intelligence

XU Ying-jin

（School of Philosophy，Fudan University，Shanghai 200433）

Abstract：It is widely believed that the development of mainstream approaches in Artificial Intelligence（AI）will soon lead to great achievements in Artificial General Intelligence（AGI）. But what is ignored in this view is the huge obstacle that AI researchers have to confront when they attempt to update their AI systems into AGI systems，as well as the huge difference between the division of labor within the industry of AI and that within the genuine human cognitive architecture. Hence，it is by far a "cake work" to pave a highway to connect the current technologies in AI to AGI，needless to say Strong AI. A more promising approach to AGI，instead，has to be based on the observation of how natural intelligence evolves in order to respond to environmental challenges, and AGI researchers have to accordingly view "adaptivity" and "frugality" as the key words constituting general principles guiding the behaviors of any intelligent system，no matter whether it is natural or artificial.

Key words：Artificial General Intelligence（AGI）；Deep Learning；Artificial Neural Network；Intelligence Quotient；Global Property

[责任编辑：马瑞雪]
[责任校对：李　蕾]