对李德毅院士关于“新一代人工智能十问”的哲学思考

Philosophical reflections on academician LI Deyi’s “Ten questions on the new generation of artificial intelligence”

魏屹东[1,2]

（1. 山西大学 哲学学院, 山西 太原 030006; 2. 教育部人文社会科学重点研究基地, 山西 太原 030006）

关于智能的基本共识，李德毅院士做了很好的总结：“智能是学习的能力以及解释、解决问题的能力；人工智能是脱离生命体的智能，是人类智能的体外延伸。”在人工智能和认知科学领域，这意味着智能是一种包括学习、解释和解决问题的认知能力，人工智能是人类智能的衍生智能或离身智能。而所谓的“通用人工智能”就是在不同情境中能够完成各种认知任务的普遍智能，在专门领域（如计算、下棋）可超过人类智能。

李院士针对新一代人工智能提出的“十问”和相应的十个基础问题非常具有挑战性，不仅激发了对人工智能的科学技术研究，也激发了对人工智能的哲学思考。20 多年来，笔者致力于自己提出的“认知哲学”的研究，其认识论和方法论的核心是“适应性表征”（自组织系统特别是认知系统，具有在特定环境或语境中自主地表征目标对象的能力，且这种能力能够随着环境或语境的变化而自主调整和提升）。这里笔者运用适应性表征概念以哲学的方式尝试回应李院士的“十问”。

“一问”：意识、情感、智慧和智能，它们是包含关系还是关联关系？是智能里面含有意识和情感，还是意识里面含有智能？

答：从认知哲学来看，意识、情感、智慧和智能等是具身的心理或精神属性，也就是基于生命体的高级认知属性，区别于低级的感觉或感知属性。意识是生命体将自身与外部环境区分开来（如人们从睡眠中醒来，儿童区分生命体与非生命体）的基本认知功能或属性。情感（在抑制和激活意义上）是有意识的状态，如喜怒哀乐。智慧是有意识的巧妙策略或计谋，与直觉相关，往往无章可循，类似于哲学上的“洞见”、科学上的“奇思妙想”或“啊哈”现象。人类智能是基于知识的认知能力，如符号操作、知识表征和问题解决，而人工智能则是剥离了心理属性的纯粹符号操作和知识表征。这些心理属性（除了人工智能）都属于认知范畴，既有包含关系（如情感中有意识），也有关联关系（如智慧是基于意识的）。这样看来，意识是高级认知属性所依托的基质，或者说意识是其他高级心理属性展现的一个平台。而意识又是基于感觉（feelings）和感知（perception）的。所以，感觉和感知是生命体更基本、更普遍，也就是更通用（所有生命体都有）的属性，而认知能力只有高级物种特别是人类才有的。人工智能是人类智能的衍生物，其智能是抽象符号表征意义上的，这种符号表征能力是比意识、情感更高级的认知能力（意识、情感可以不依赖语言）。所以，这些描述心理属性的概念，是相互包含的，也就是混合的。当然，这些概念涉及大众心理学、认知学科和人工智能等学科，需要做进一步的澄清。从适应性表征的角度看，认知行为，无论是低级还是高级属性，都是适应性表征的不同表现方面。因此，适应性表征是一切自组织系统，尤其是认知系统的“共通性”或“通用性”，在概念上是统摄自然智能和人工智能的统一范畴。

“二问”：如何理解通用智能？通用智能一定是强智能吗？通用和强是什么关系？

答：“通用”有两层意思：一是“普遍的”（universal）或“广义的”，即在时空上是遍历的，如宇宙（universe）；二是“一般的”或“概括的”（general），即哲学上的高度提炼，如世界统一于物质。根据这两层意思，通用智能就是在所有领域能够解决问题或完成任务的认知能力。按照笔者的理解，人工智能是相对于自然智能（人和动物）而言的，在计算主义语境中特指“机器智能”，如图灵机，在“计算”意义上这两种智能是相通的（否则人就造不出计算机）。人工智能的强弱之分，据笔者的考察，来源于哲学家塞尔在 20 世纪 80 年代对试图制造人类水平或超人类水平人工智能的称呼。“强”的含义是指让机器智能达到甚至超越人

类智能，包括有意识、有情感甚至有道德感；“弱”的含义是让机器能够在一定程度上模拟人类的智能行为，如功能模拟（如物理符号假设）、结构模拟（如人工神经网络）和行为模拟（如感知行动），不要求有意识。目前，弱人工智能（计算意义上）已经实现，强人工智能（有意识意义上）还遥遥无期。简单来说，强人工智能=理性能力（计算、推理）+心理能力（意识、情感），弱人工智能=理性能力（计算、推理）。通用人工智能（general artificial intelligence, GAI 或 universal artificial intelligence, UAI）和人工通用智能（artificial general intelligence, AGI）在人工智能文献中都出现过，通常不加区别，它是相对狭义（窄）人工智能而言的。这里的“通用”与“广义”或“宽”意思相近。粗略来说，通用人工智能对应于强人工智能，狭义人工智能对应于弱人工智能。不过，在笔者看来，通用人工智能与强人工智能还是有所区别的，这主要表现在两方面：一是在时间上，强人工智能概念早于通用人工智能，前者是塞尔提出的，后者是美国通用智能研究所的本•格策尔（Ben Goertzel）和佩纳钦（Pennachin）2007 年提出的；二是在含义上，强人工智能强调机器智能的类人心理属性，目标是实现人类水平甚至超人类水平的“具身智能”，而通用人工智能强调在感知层次实现通用性和自主性的“离身智能”，不强求有意识和情感。在这个意义上，通用人工智能稍弱于强人工智能，或者说，通用人工智能介于强人工智能和弱人工智能之间。相同之处是它们本质上都具有跨学科的特点。

“三问”：目前几乎所有人工智能的成就都是在计算机上表现出来的“计算机智能”，存不存在更类似脑组织、能够物理上实现的新一代人工智能？

答：目前的计算机和人工智能的实质是模拟自然智能。在方法论上是功能主义或行为主义。从哲学上来看，人工智能是“好像”（as-if）智能，或“假装”（make-believe）智能，也就是模拟出的“虚拟智能”，不是基于肉体的“真正的”人类智能。也就是说，计算机只能执行设计者预先设定的程序（算法）和目标，不具有“真正的”人类智能。在这里认知科学中的“硬件要紧”是一个关键，即构成认知主体或智能主体的组成成分很重要（碳基的还是硅基的）。这意味着人类智能和人工智能借以实现的物理载体完全不同。于是问题来了，组成成分完全不同的人脑和电脑是否一定不能实现相同的认知任务？答案是否定的，人脑和电脑都能执行计算任务，就如同飞机与鸟飞行的原理完全不同，但都能执行“飞行”任务。然而，目前的人工神经网络不论多么复杂，都无法与人脑相提并论。如果有一天人工神经网络也达到人脑神经元的量级（10^{11}），情形会怎样呢？笔者无法预测结果。但可以肯定，即使机器智能、机器意识等人工制品实现了类似人类意识的东西（如感受性），我们也不知道，就如同我们不知道他人的感受一样。这意味着人工神经网络生成的东西，本质上不同于生命体拥有的东西（生命、意识、心灵），尽管功能上甚至结构上相似或相同。

“四问”：机器人不会有七情六欲，还会有学习的原动力吗？如果没有接受教育的自发性，还会有学习的目标吗？

答：这个问题涉及认知的具身性（embodiment）和本能（instincts）。生命体是具身的，其功能大多是本能（遗传决定的或先天的），我们人类也不例外。机器人是离身的（disembodiment），其智能是人类智能的体外延展，这是当代认知科学的具身认知和延展认知观。这里的问题的逻辑是：情感（兴趣）是学习的原动力，机器人没有情感（兴趣），自然不会有学习的动力。情感是否是学习的原动力学术上是有争议的。在笔者看来，除了兴趣外，目标追求也是学习的动力之一。只要人工智能是目标导向系统，或者是我主张的适应性表征系统，它就是自我学习系统，如机器学习中的无监督学习。人类的学习（主动的或被动的）在很大程度上是后天教育的结果（语境化），机器人虽然缺乏人类的先天自发性和接受后天教育的主动性（非语境化），但目标导向的系统使其具有了学习的目标（人为机器设置语境，如知识库）。这是笔者主张的人工智能的自语境化能力。人类是自语境化的（自我融入新情境），如果让机器人也拥有了自语境化能力，它就会自主融入新情境（从一个情境进入另一个情境）。这种自语境化能力就是适应性表征能力。

“五问”：人的偏好和注意力选择是如何产生的？新一代人工智能如何体现这一点？

答：人的偏好和注意力是有意识的认知行为。偏好是兴趣引导或目标导向的，注意力是意识的聚焦问题。根据巴尔斯的认知理论，注意力就是意识的“聚光灯”，舞台背景包括剧务人员就是意识的“语境”，这种语境支撑着作为意识的注意力。由此给笔者的启发是，未来的人工智能的认知架构可能是：语境（知识库）+感受器（输入）+控制器（中枢系统）+执行器（输出）。强化学

习就是运用“奖励”(偏好)的方法。如果能让强化学习与语境化方法相结合,也就是将强化学习“语境化”,新一代的人工智能就会更接近甚至达到人类智能。

“六问”:如果说计算机语言的元语言是数学语言,数学语言的元语言是自然语言,前一个比后一个常常更严格、更狭义。那么,人工智能怎么可以反过来要用数学语言或者计算机语言去形式化人类的自然语言呢?

答:这是一个关乎人工智能如何表征的问题。在笔者看来,人工智能总体上是理性的事业,理性原则上是排除非理性的,如意识、情感、自我、心灵、自由意志等。在这个意义上,智能机器无需意识、心灵这些带有神秘色彩的非理性东西。理性的呈现或表征方法主要是数学和逻辑语言,即形式语言。而形式语言是高度抽象的符号表征,这是人类理性的最高认知境界。人工智能就是这种抽象认知的结晶。从人类语言形成与发展的历史看,先有口语、再有书面语(自然语言),然后才有了数学和逻辑,才有了科学理论(数学语言书写的)。语言的这种生成过程是不可逆的,这与人类社会的发展是同步的,即从低级到高级。人工智能能否反过来形式化自然语言呢?我认为可以,虽然语言的自然形成是不可逆的(进化选择和文化选择),但高级认知能够表征低级认知,比如自然语句的命题逻辑表达、自然现象的数学刻画。这似乎回到了古希腊哲学家毕达哥拉斯关于宇宙是“数的和谐”的思想。如果“数的和谐”假设是对的,那么用数学语言或编程语言表征自然语言就不是问题。

“七问”:如何体现新一代人工智能与时俱进的学习能力?

答:笔者的看法是通过适应性表征。与时俱进的学习能力就是适应性表征能力,因为适应性意味着进化适应,表征意味着目标导向的语言表达。我们只要让认知机器人或社交机器人具有适应性表征能力,它们就能够完成人类为其设置的目标,至于它们有无意识和情感,则无关紧要。这个问题也是人工心理学、人工认知神经科学、具身人工智能以及认知哲学要研究的。

“八问”:在新一代人工智能架构的机器人中,基本组成最少有哪几种?各部分中的信息产生机制与存在形式是什么?它们之间的信息传递是什么样的?

答:根据人类智能的构成,笔者认为人工智能的认知架构至少要有一个智能体(人工脑)、身体(物理载体)和模拟对象(环境)3 个部分,而且它们之间的交互是必须的,因为极有可能正是交互(大脑、身体和环境的耦合)才涌现出了智能。由于交互是一个动力学过程,这个过程必然包含时间。若加上时间因素,人工智能的认知架构就有 4 个基本组成成分。虽然交互中的信息产生的机制和智能生成之间有密切的关联,但二者是不同的,也就是说,信息产生不等于智能生成,毕竟大多物理系统都有信息产生但没有智能(如温度计、恒温器)。笔者不是人工智能研究者,具体的技术细节并不清楚,但笔者直觉地认为,人工智能架构不论是什么,“适应性表征”可能是其运作的内在机制以及不同部分之间的交互环节,因为适应性表征在同层次和不同层次之间(物理的、生物的、认知的)都存在。

“九问”:新一代人工智能如何具有通用智能?不同领域的专用智能之间是如何触类旁通、举一反三、融会贯通的?如何体现自身的创造力,如能不能形成自己软件的编程能力?

答:通用人工智能或人工通用智能是一个跨学科领域,其最终目标是实现人类水平或超人类水平(某些方面)的智能。要实现这个目标,“通用”应该是人类和机器人共享的属性或特征。问题是:什么是人机的共享属性呢?笔者认为最基本的才是通用的,这种最基本属性就是感知(人是感官系统,机器是感受器);什么是两种智能都共有的呢?笔者认为是适应性表征。即使是各类专家系统或专用智能(下棋、扫地、家庭服务等机器人),其智能体(agent)必须是适应性表征系统,即完全围绕要完成的目标来行动,如扫地机器人要能够执行保证房间干净的目标。这意味着有了适应性表征能力,就有了某种创造力,因为适应是与新奇(意外)相关的,新奇(意外)是创造性的源头,而表征本身就是再创造(呈现新奇现象的方式),如量子力学的不同表征(狄拉克标记法和薛定谔方程)。至于智能机器人能否自己拥有创造力、自己编程,这与机器人的自主性问题(哲学上是主体性问题)有关。笔者看来,只要机器人以某种程度的独立性或自主性进行一些操作,如抓取东西,我们就应该认为它们具有一定的自主性,就像计算机病毒自复制一样。有了自主性才会有创造性(必要条件)。这就要求人工智能研究者想方设法让机器人有自主性(一种适应性表征能力)。从协同涌现的视角看,单一智能体恐怕没有创造性,创造性应该是多智能体交互的一个整体特征。如果是这样,我的想法是:人工智

能的创造力要通过多智能体的协同来实现（产生“交互”意识），单个智能体很难拥有创造力（这和一个人有创造力不同）。

“十问”：基于新一代人工智能机器人，存不存在停机问题？机器人的“发育”，即软硬件的维修管理和扩充升级，如何解决？

答：在逻辑上，停机问题本质是高阶逻辑的不自洽和不完备问题，在人工智能中是判断一个程序是否会在有限时间之内结束运行的问题。如果哥德尔不完备定律是正确的，人工智能中的停机问题原则上是存在的，就像人的认知能力有至上性（有限性）一样。至于机器人的“发育”（自生长、自修复、自提升），我认为，只要机器人有了适应性表征能力，就可通过具身性和适应性表征来解决。在这个问题上，笔者持乐观主义的态度。

概言之，新一代人工智能要实现“通用”，也就是在不同层次系统（物理的、生物的、认知的）、不同学科和社会领域都是适用的，笔者认为适应性表征是其“通用的”实现机制和统一解释的概念框架。“通用”意味着最基本和最普遍，也就是适用于所有系统，如能量守恒定律适用于物理系统、生物系统和智能系统。因此，一个人工智能体，无论它最终是否有意识、有情感、有道德感，这都不重要，重要的是，它必须拥有适应性表征能力。

作者简介：

魏屹东，教授，博士生导师，山西大学科学技术哲学研究中心学科带头人，《国家哲学社会科学成果文库》入选者，国家哲学社会科学重大项目首席专家，长期从事科学史、科学哲学、认知哲学、人工智能哲学以及语境论的认识论和方法论研究，提出“认知哲学”（philosophy of cognition）概念并举办了首届国际认知哲学会议，建立了认知哲学的培养体系和研究团队；提出“适应性表征”概念，进一步将其提升为认识论和方法论，以此作为统摄自然认知和人工认知的统一范畴，并用于解释两种系统的意识或智能生成问题。

中文引用格式：魏屹东. 对李德毅院士关于“新一代人工智能十问”的哲学思考 [J]. 智能系统学报, 2023, 18(6): 1352–1355.

英文引用格式：WEI Yidong. Philosophical reflections on academician LI Deyi’s “Ten questions on the new generation of artificial intelligence”[J]. CAAI transactions on intelligent systems, 2023, 18(6): 1352–1355.