全国政协常委，中国科协荣誉委员，国际核能院院士，中国人工智能学会会士、不确定性人工智能专委会主任、智慧医疗专委会顾问，中国知识产权研究会副理事长兼学术顾问委员会主任；国家核电重大专项战略咨询专家组组长；清华大学博士后校友会会长；清华大学核能与新能源技术研究院和计算机系双聘教授、博导。中国科协原党组副书记、副主席。

卷首语 Foreword

# 对新一代人工智能的思考

张勤

如果将达特茅斯会议作为人工智能（AI）诞生的标志的话，我正好与 AI 同岁。但是，迄今为止，国际上对 AI 是什么仍众说纷纭。比较有共识的似乎是对 AI 的代际划分，即第一代 AI 是基于知识的，第二代 AI 是基于数据的。而现在，大家正在开启对第三代或新一代 AI 的研究。但要弄清楚新一代 AI 的特征，首先必须搞清楚第一代和第二代 AI 存在的问题。

我认为，第一代 AI 之所以不够成功，主要问题有三个：第一，其知识是以规则的形式表达的，因而是碎片化的，尽管有一套格式，但很难说是一个严谨的科学体系，更像是一种实用技术。然而，没有严谨的科学理论支撑的技术是走不远的。第二，缺乏严谨的不确定性表达和推理算法。现实世界绝大多数都有不确定性，只有证明数学定理等少数情况无不确定性。显然，没有严谨的处理不确定性的算法限制了第一代 AI 的应用。第三，追求通用 AI（AGI），将目标定得过于宽泛，要求其像人一样能够学习、识别、分析、归纳、推理、感知等，是不现实的。

第二代 AI 基于大数据机器学习，能够有效应对不确定性，其代表是深度学习 (DL) 模型。主要问题也有三个：第一，由于 DL 的本质是数据拟合，模型缺乏可解释性。对很多应用而言，没有可解释性就难以可信和应用。例如疾病诊断，要综合患者的症状、体征、实验室检查、影像学检查，以及性别、年龄、病史等各种信息，根据彼此的关系才能下诊断结论，而且结论只能由作为用户的医生下，由医生承担责任，否则会出现很多法律问题。这就需要 AI 可解释，包括计算结果可解释（怎么算出来的）、计算模型可解释（能理解）、计算方法可解释（物理意义清晰）。然而数据拟合是一种黑箱方法，从原理上就不具备可解释性。第二，依赖数据独立同分布假设。这里的同分布不仅指训练集与测试集之间，而且指训练和测试集与真实应用场景之间，都要同分布。但很多情况不符合这一假设。例如基层医疗机构诊病的数据样本空间与用于 DL 模型训练和测试的数据集（三甲医院病历）的样本空间就不同，导致泛化问题。第三，与第一代 AI 一样，追求 AGI，想用一个模型解决所有问题，但结果往往不佳。事实上，由于开放式问题的模式不可穷尽，总有拟合不到的。这就为其应用悬了一把达摩克利斯之剑。这也是当前自动驾驶所面临的困境。至于数据获取难、清洗加工难、保护隐私难、数据产权难、训练能耗大等非技术问题，就不在这里讨论了。

事实上，计算机只能执行人事先设定的程序（包括算法和数据），不具有真正意义上的智能，至少目前如此。就某个具体问题而言，用人设定的计算机程序来代替人解决问题是完全可能的，并且其表现往往超过人（例如 AlphaGo 战胜李世石）。一旦应用超越了事先设定的计算机程序所要解决的问题的边界（例如用 AlphaGo 下以前没有见过的半个棋盘或两倍棋盘的围棋），就很难保证 AI 仍有上佳表现，但人却可以举一反三。在当前人们对生物脑知之甚少的情况下，在尚未解决自我意识是什么和怎样产生的情况下，用计算机模拟人这样的生物脑很难，因为要模拟的对象是什么不清楚。一个显著的区别是：人能够通过自我意识理解事物，而计算机没有自我意识，也理解不了事物（缺少理解主体）。从这个意义上讲，学习（Learning）这个词用在计算机上是不恰当的。拟合（Fitting）更准确，但不够吸睛。

我的看法是：当前 AI 应着力研究两个领域：第一，研究生物脑的工作机理，这主要是医学和生物学的事情，以及相关学科的事情（例如生物电镜）。第二，研究能解决具体问题的 AI 模型，无论其基于知识还是基于数据，不一味追求 AGI。不同领域有不同需求，从而适用不同模型。例如人脸识别，无所谓可解释还是黑箱，即使有一定错误率也问题不大，这时用深度学习模型就很好。当然还要在现有基础上精进。对于诊病（不只是看片或疾病筛查）和工业系统故障诊断来说，没有可解释性的模型是不可信，因而也不宜用的；基于知识的模型（当然要系统化和具备处理不确定性的能力）才是可信和可用的，大数据学习并非必由之路。事实上，核电站几乎没有可供学习的故障数据，但要求 AI 能够诊断从来没有出现过的故障。对核电站操作员的要求也相同，所以这一要求并不过分。

归纳一下：新一代 AI 要有可解释性，不依赖或少依赖数据独立同分布，知识应系统化，能够处理不确定性，基于数据或知识或两者均可，但未必通用。

我对 AI 的定义是：AI 是一门科学技术，将由人类智能解决的问题转化为由人造机器来解决。这里首先要明确所解决的问题是不是智能问题。如果是，且由机器来解决，就是 AI。适合于解决本领域智能问题的模型就是强 AI 模型，与其是否通用无关。