·科学论坛·

健壮人工智能模型与自主智能系统*

吴　飞[1]　段书凯[2]　何　斌[3]　兰旭光[4]　黄　如[5]　吴国政[6**]

(1. 浙江大学 人工智能研究所，杭州 310027；
2. 西南大学 电子信息工程学院，重庆 400715；
3. 同济大学 控制科学与工程系，上海 200092；
4. 西安交通大学 人工智能与机器人研究所，西安 710049；
5. 北京大学 信息科学技术学院，北京 100871；
6. 国家自然科学基金委员会 信息科学部，北京 10085)

[摘　要] 基于第 216 期“双清论坛”主题报告和分组讨论，本文从开放、动态环境下人工智能的前沿理论与方法推动场景人工智能进展这一角度，提出并分析了健壮人工智能模型的本质特点，给出了发展具有自主与感知、决策与控制、群智与协同特点的自主智能系统面临的科学挑战问题，探讨了未来 5 年人工智能前沿研究领域和科学基金基础研究重点资助方向。

[关键词] 健壮人工智能；贝叶斯模型；自主智能；神经形态计算

人工智能是研究模拟、延伸与扩展人类的感知、认知、决策和执行的理论、方法、技术及系统的科学，经过 60 多年的演进，新一代人工智能正呈现深度学习、跨界融合、人机协同、群智开放、自主操控等新特征。作为引领这一轮科技革命和产业变革的战略性技术，人工智能具有辐射效应、放大效应和溢出效应，正在引发链式突破，加速新一轮科技革命和产业变革进程，成为新一轮产业变革的核心驱动力，是经济社会发展的新引擎。

世界主要发达国家正在把人工智能作为提升国家竞争力、维护国家安全的战略利器，加紧出台了一系列规划和政策。2017 年 7 月，国务院印发《新一代人工智能发展规划》。2017 年 11 月科技部启动了“新一代人工智能”科技创新 2030—重大项目，第一批项目指南于 2018 年 10 月向社会公布。2018 年 10 月，中共中央政治局就人工智能发展现状和趋势进行了集体学习，习近平总书记强调加强基础理论研究，支持科学家勇闯人工智能科技前沿的“无人区”，努力在人工智能发展方向和理论、方法、工具、系统等方面取得变革性与颠覆性突破，确保我国在人工智能这个重要领域理论研究走在世界前面，关键核心技术占领制高点。

近年来，我国人工智能基础研究和应用在计算机视觉、自然语言处理、语音识别和生物特征识别等

吴飞　2002 年获浙江大学博士学位。浙江大学求是特聘教授，博士生导师，浙江大学人工智能研究所所长。主要研究领域为人工智能、跨媒体计算、多媒体分析与检索和统计学习理论。国家杰出青年科学基金获得者、教育部新世纪优秀人才支持计划入选者、2018 年入选“高校计算机专业优秀教师奖励计划”、教育部人工智能科技创新专家组工作组组长(2018.8—2020.12)。

吴国政　信息与通信工程工学博士。副研究员，现担任国家自然科学基金委员会信息科学部三处副处长兼人工智能与智能系统项目主任。自 1997 年在国家自然科学基金委员会工作，先后任《自然科学进展》杂志编辑、信息科学部综合处副处长，自 2016 年起在信息科学部三处工作。

收稿日期：2019-06-04；修回日期：2019-07-11

* 本文根据第 216 期“双清论坛”的研讨整理。

** 通信作者，Email：wugz@nsfc.gov.cn

任务场景明确领域取得了显著进展。当前，以深度学习为代表的人工智能方法的兴起，在数据、模型、算力和明确场景结合下，提高了机器学习算法的层次和性能，很多新的应用和产品已经惊艳亮相。然而，当前人工智能发展还存在着感知智能适应性差、认知智能“天花板”低、强人工智能发展瓶颈突破乏力等挑战。

在此背景下，国家自然科学基金委员会第 216 期“双清论坛”于 2018 年 12 月 12～13 日在浙江杭州召开。本次论坛围绕制约人工智能发展的科学问题和未来能推动人工智能实现突破性发展的颠覆性技术展开深入探讨，凝练和提出我国在该研究领域急需关注和解决的重要基础科学问题以及相应的资助模式。来自北京大学、清华大学、浙江大学、中国科学院和阿里巴巴等 40 余所高校、研究机构和大型企业的 50 余位专家以及国家自然科学基金委员会政策局、信息科学部、数理科学部和管理科学部的相关工作人员参加了此次论坛。

与会专家从“人工智能理论基础”、“人工智能验证平台”和“人工智能芯片与器件”三个主题方向，提炼急需关注和解决的重要基础科学问题，确定技术爆发点，为今后 5～10 年的重点资助方向提供导向性。

在会上，与会专家就人工智能在基础理论研究、自主智能系统和神经形态芯片等方面存在的热点问题进行了分析。

1 健壮人工智能及其面临挑战

合理的判断和决策能力是人类生存及发展的基本能力。在这一认知过程中，人类面临的许多问题具有不确定性、脆弱性和开放性的特征，因此智能算法应该能够面对这些挑战，具备“健壮”能力。健壮人工智能指模型算法“对可能发生的错误具有稳健性、对未建模问题具有稳健性”。在开放世界中，无法为所有问题及其解决方法进行建模，对构建健壮的人工智能系统提出了巨大挑战。

经典人工智能理论框架建立在以递归可枚举为核心的演绎逻辑和语义描述基础方法之上，由于先决条件问题（Qualification Problem，即枚举描述促发某一行为发生的所有前提条件）和隐性分支问（Ramification Problem，即枚举刻画某个行为可能导致的所有后续潜在结果）的存在，难以事先拟好智能算法能够处理的所有情况，因此智能算法在处理不确定性、开放性和动态性等问题时难以发挥作用，需要借鉴神经进化（Neuroevolution）[8] 机理来推动这一方面的研究。

与计算机执行常规程序中遇到错误会崩溃不同，人脑能够应对不确定性、对错误具有一定的容错机制（Fault Tolerance）。如研究人员在小鼠大脑中发现一种容错机制，这有助于进一步理解大脑是如何工作的[3]。在认知学领域有一个具有争议的理论，认为感知、运动控制、记忆和推理优化等大脑功能，都通过大脑将“现有经验”和“未来期望”进行比较而进行。也就是说，大脑在所接受外界现有信息基础上主动构建模型，计算所构建模型与所期望“假设空间（Hypothesis Space）”之间的概率，并加以调整优化，即比较“现有经验”和“未来期望”进行决策[4,5]。Yang Tianming 在清醒猴的侧内顶叶（Lateral Intra-Parietal，LIP）区域神经元开展的关于决策神经机制的研究表明[10]，该区域神经元是基于对数似然比（Log Likelihood Ratio，LogLR）的贝叶斯决策模型，完成概率推理（Probabilistic Reasoning）等高级脑功能；科学家发现完成大脑奖励功能的多巴胺神经元对期望回报和真实回报的差异进行编码，基于预测误差来帮助动物更新对未来的期望，并做出决策[7]。

为了设计更加健壮的人工智能，需要提升模型错误的稳健性（如通过稳健优化、模型正则化对风险敏感的目标进行优化以及采用稳健推理的算法）、提升对于未建模问题的稳健性（如通过模型扩展、利用因果模型和采用组合模型来监测模型表现以检测异常）。

健壮人工智能研究需要重点考虑学习算法的泛化能力（Generalization）、在模型复杂化和简单化之间平衡的正则化（Regularization）、探究某一任务是否可以被学习的概率近似正确学习（Probably Approximately Correct Learning，PAC Learning）理论等[9]。

为了应对这一挑战，各国政府也开始资助若干项目。2018 年 9 月，美国国防高级研究计划局（DARPA）启动了被称为“加速第三波”的人工智能探索（Artificial Intelligence Exploration，AIE）项目，探索类人水平的交流和推理能力，以增强对新环境的自适应能力。DAPRA 认为：第一波人工智能是以符号主义为手段，主要处理语言和可描述信息；

第二波人工智能在数据建模基础上、从数据中学习模式，以模型假设的机器学习为手段；第三波人工智能以自适应和推理为核心目标。

美国国家科学基金会（National Science Foundation，NSF）2018 年 12 月启动了“健壮智能（Robust Intelligence）”的项目，旨在加强在复杂和真实环境下的人工智能理解能力。

2　自主智能系统面临挑战

自主智能系统一直是人工智能领域中最受关注的研究和应用方向之一，自主驾驶系统又是自主智能系统研究中的热点。从 1966 年的 DARPA 资助的斯坦福大学 Shakey 项目，到 1995 年卡内基梅隆大学的 NavLab 项目，再到 2007 年的 DARPA Urban Grand Challenge，自主车技术的研究已经进行了半个世纪。从 1925 年纽约街头的 Linrrican Wonder，到 1950 年通用汽车的 Firebirds，再到今天的 Tesla、Waymo 和 Otto，我们看到大部分的 IT 巨头都参与了自主车的产业应用。

由于自主智能系统面临的场景态势瞬间变化，其对通用性要求高，即自主智能系统应该具有对“未知的未知建模”特点。当前自主智能系统在实现中暴露出许多问题，如决策混乱、控制不力和故障频发等。为此，需要发展健壮人工智能的理论与方法来提升自主智能系统性能。

自主智能系统面临如下挑战：由于任务环境不确定、决策信息不完全、通信交互受限而导致无人系统协同决策智能化程度低和协同任务不能有效完成的不足；系统呈现非线性的复杂性以及通信存在时延而导致多个自主智能系统交互融合鲁棒性差；多个自主智能系统组成更大规模的复杂结构，使得单个节点计算和通信能力无法支撑巨大计算开销；如何将人为干预控制与系统的自动化算法相结合，从而实现人—机、机—机协作。

3　神经形态芯片面临挑战

神经形态计算是借鉴大脑神经网络架构和信息处理原理，构造适合处理动态复杂信息的计算范式，具有计算存储融合（存算一体）、时空整合编码、异步、高容错、复杂互连等特点，更适合非结构化数据与智能任务处理，在能效、硬件代价等方面优势显著，同时具有高度并行、自适应性、高鲁棒性等优势，可与传统计算互补共生[6]。

与专用智能加速芯片不同，神经形态芯片更加强调器件、架构和算法等原理上的仿生，而非面向特定智能任务。现有神经形态计算芯片的技术路线主要包括基于 CMOS 的神经形态芯片（如 TrueNorth、BrainScaleS、NeuroGrid 和 Loihi 等）和基于新型器件的神经形态芯片（如忆阻器、相变单元、自旋器件和 flash 等），但该领域研究仍面临多方面的挑战。基于 CMOS 的神经形态芯片采用的神经元和突触模型通常较为单一，难以反映生物系统中神经元和突触功能的多样性，单个神经元和突触单元的实现需要较多 CMOS 器件，芯片规模提升存在天花板。在基于新型器件的神经形态芯片方面，人工突触器件的性能、可塑性、一致性有待进一步优化，人工神经元器件研究仍较为有限，缺乏理想的人工突触器件和神经元器件。此外，现有神经形态器件支持的算法仍然比较有限，一方面现有器件对包含时序信息的动态复杂信息处理算法等缺乏有效支持，另一方面面向神经形态器件和芯片的算法研究不足，已有算法尚未充分发挥神经形态器件的信息处理动力学优势。神经形态器件的大规模集成技术仍是难点，缺乏基于新型神经形态器件的芯片级验证。

为了解决以上问题，需要寻找更适于仿生智能任务处理的底层器件、芯片架构和应用算法，并开展器件—电路—架构—算法的协同优化研究。探索如何减少硬件开销和功耗，如何用较少的硬件代价“有效”实现多样性神经元行为、多样性突触行为、智能高效动态信息处理，如何“有效”实现自主学习、自主演化、高容错性，以及如何从硬件上实现从“自动”到“半自主”、最终走向“全自主”的智能系统。

4　未来 5 年拟重点资助方向的建议

(1) 认知行为的信息处理机制

从脑认知机理和神经科学中获得灵感和启发，发展新的人工智能计算模型与架构，让机器具备对物理世界最基本的感知与反应，使机器具有“常识”推理的能力，能快速思考、推理和学习，能凭直觉了解真实世界。

(2) 新的机器学习(计算)方法与架构

人工智能算法能对“环境”(图像、视频、语言、语音、情感、行为等)进行“自然”理解,具备强泛化能力、知识与规则引导的小样本学习、结构自优化等能力;研究新型、非神经网络的深度模型;研究弱监督学习,缓解标注数据获取的困难;研究开放环境下的机器学习,突破封闭静态环境的束缚,鲁棒适应任务环境变化。

(3) 机器智能的评价体系与评价方法

图灵测试已无法适应对机器智能进行量化分析,无法用来评估机器对“任务”理解和执行的水平;发展新型智能评价体系与评价方法来驱动和引导人工智能深入研究。

(4) 面向人工智能的开源平台

为使用者提供多种灵活的解决方案,支持多种数据源和数据格式,丰富的 API 接口为多种编程环境提供支持,为各种大数据提供便捷、可拓展的数据建模和分析;建设人工智能领域的基础算法库、知识库和数据库,建设开放共享的人工智能计算平台,建立国家级大规模语言知识库。

(5) 面向神经形态计算的智能芯片与器件

研究支撑神经形态器件的新机理与新器件结构,实现高效、多功能、可集成神经形态器件,包括神经元器件和突触器件;研究基于神经形态器件的集成技术、基于 CMOS 和新器件的存算一体新架构,研制基于神经形态器件的智能芯片,并实现器件、算法和应用开发环境的协同优化。

(6) 新一代智能的数学模型

研究网络化学习博弈与决策、在线算法、量子算法、分布式群体智能计算、变分推理、复杂网络动态演化和涌现机制、不确定性量化和适应性设计、算法的鲁棒分析和设计等内容。

(7) 自主智能系统基础理论与关键技术

自主智能无人系统作为最高级别人工智能的体现形式是未来人工智能的主要发展方向,其基础理论研究的创新将是未来人工智能发展的源动力。研究视、听、触、气、激光等新型智能传感器关键技术和智能自主控制器。研究具有高精度自主定位与导航、自主规划与智能控制、环境自适应与进化等特点的无人终端控制系统、智能协同控制技术以及感知芯片与计算系统。

(8) 人机共融工业智能系统

研究知识驱动的制造过程决策自动化、基于模式认知的智能自主调控、安环指标的智能预警与溯源等技术。针对全球化的市场需求,基于互联网和信息物理系统,通过自主学习和主动响应来重塑供应链并优化商业行为,实现企业经营管理和决策的智能化。面向工业生产过程和系统,基于智能感知和人机交互技术,构建智能化和绿色化的柔性制造模式,重塑产业链和价值链,实现工艺优化和生产全流程整体优化。

上述研究主题可归纳为“健壮人工智能前沿理论与方法”和“自主智能系统”两个重点方向,具体如下:

1) 健壮人工智能前沿理论与方法。内容包括:对未知问题建模与环境自适应能力;健壮人工智能的可解释性和可验证性;复杂动态和开放环境下的博弈对抗;人机混合增强智能。这四大方向支撑健壮人工智能的前沿理论与方法。

2) 自主智能系统。内容包括自主与感知、决策与控制以及群智与协同,这三个方向包含了自主智能系统研究的三个层面的工作,三个层面分别为异构多智能体系统、单智能体以及器件与控制。

上述两个重点方向具有“理论支撑场景,场景带动理论”的关系,即:开放、动态环境下人工智能的前沿理论与方法能够推动场景人工智能的进展,自主智能系统能够带动脑与认知、机器学习、健壮人工智能、人机混合智能等相关理论与方法的发展。

致谢 感谢孙家广、孙优贤、柴天佑、郑南宁、郝跃、钱锋、桂卫华、陈杰、樊邦奎、王成红、周志华、陈云霁、张建伟、吴枫、陈恩红、陈俊龙、洪弈光、胡斌、焦李成、王伟、王国胤、于剑、庄越挺、段振华、华先胜、李少远、罗钟铉、乔俊飞、孙长银、陶建华、王龙、郑庆华、朱文武、陈启军、窦勇、黄铁军、李远清、潘纲、王耀南、肖侬、俞成浦、曾志刚等专家学者对本文的贡献。

参 考 文 献

[1] 中国人工智能 2.0 发展战略研究项目组. 中国人工智能 2.0 发展战略研究, 杭州: 浙江大学出版社, 2019.

[2] Hassabis D, Kumaran D, Summerfield C, et al. Neuroscience-inspired artificial intelligence. Neuron, 2017, 95(2): 245—258.

[3] Li N, Daie K, Svoboda K, et al. Robust neuronal dynamics in premotor cortex during motor planning. Nature, 2016, 532(7600): 459—464.

[4] Ma WJ, Jazayeri M. Neural coding of uncertainty and probability. Annual Review of Neuroscience, 2014, 37: 205—220.

[5] Pouget A, Beck JM, Ma WJ, et al. Probabilistic brains: knowns and unknowns. Nature Neuroscience, 2013, 16(9): 1170—1178.

[6] Schuman CD, Potok TE, Patton RM, et al. A survey of neuromorphic computing and neural networks in hardware. arXiv preprint arXiv, 2017, 1705.06963.

[7] Starkweather CK, Babayan BM, Uchida N, et al. Dopamine reward prediction errors reflect hidden-state inference across time. Nature Neuroscience, 2017, 20(4): 581—589.

[8] Stanley KO, Clune J, Lehman J, et al. Designing neural networks through neuro-evolution. Nature Machine Antelligence, 2019, 1: 24—35.

[9] Ben-David S, Hrubeš P, Moran S, et al. Learnability can be undecidable. Nature Machine Antelligence, 2019, 1: 44—48.

[10] Yang TM, Shadlen MN. Probabilistic reasoning by neurons. Nature, 2007, 447(7148): 1075—1080.

Robust artificial intelligence and autonomous intelligent system

Wu Fei[1]　Duan Shukai[2]　He Bin[3]　Lan Xuguang[4]　Huang Ru[5]　Wu Guozheng[6]

(1. *Institute of Artificial Intelligence, Zhejiang University, Hangzhou* 310027;
2. *College of Electronics and Information Engineering, Southwest University, Chongqing* 400715;
3. *Department of Control Science and Engineering, Tongji University, Shanghai* 200092;
4. *Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an* 710049;
5. *School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871;
6. *Department of Information Sciences, National Natural Science Foundation of China, Beijing* 10085)

Abstract　Focusing on the outputs of the 216th Shuangqing Forum of National Natural Science Foundation of China, this paper analyses the recent advances and main scientific challenge in terms of robust artificial intelligence and autonomous intelligent system. At the same time, this paper gives out the key research directions funded in the coming 3～5 years.

Key words　robust artificial intelligence; bayesian model; autonomous intelligence; neuromorphic computing