

# 人工智能的知识本源

王飞跃<sup>1,2</sup>

(1. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室, 北京 100190;

2. 澳门科技大学创新工程学院, 澳门 999078)

**摘要:** 人工智能的兴起并非单纯源于计算技术的突破, 而是哲学思辨、数学逻辑与认知科学等学科长期交汇演化的结果。从知识本源的角度, 系统梳理人工智能的思想根基与学科脉络。哲学层面, 从东方“赋、比、兴”与西方“文法、逻辑、修辞”出发, 结合本体论、平行智能以及“存在-生成-信念”的三元框架, 阐释智能的层次结构、虚实共生机制与平行演化理念。数学层面, 追溯3次数学基础危机对形式化、可计算性和复杂性理论的推动, 展示从希尔伯特纲领、哥德尔不完备性定理、图灵机、控制论、人工神经元模型到范畴论与朗兰兹纲领的逻辑演进, 揭示人工智能的数理根基。认知科学层面, 围绕莱特希尔报告、人工智能寒冬、隆科-希金斯对认知科学的命名、米勒认知革命及认知科学六角形框架, 说明认知科学如何为人工智能突破符号主义局限, 走向深度学习、大模型与具身智能提供理论支撑。人工智能是人文思想、形式理性与认知机制共同作用的产物, 对其知识本源的追溯有助于重新理解智能的边界、局限、演化路径与未来发展方向。

**关键词:** 人工智能; 知识本源; 平行智能; 智能哲学; 认知科学

**中图分类号:** TP39

**文献标志码:** A

**doi:** 10.11959/j.issn.2096-6652.ZNKX26067

## The knowledge origin of artificial intelligence

Wang Fei-Yue<sup>1,2</sup>

1. The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

2. Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China

**Abstract:** The rise of artificial intelligence is not merely a consequence of advances in computing technology, but the result of a long-term convergence of philosophical reflection, mathematical logic, and cognitive science. From the perspective of knowledge origins, this paper systematically examines the intellectual foundations and disciplinary evolution of artificial intelligence. At the philosophical level, it traces early forms of intelligent activities to the Eastern tradition of “Fu, Bi, Xing” and the Western trivium of “Grammar, Logic, Rhetoric”, and further interprets intelligence through ontology, parallel intelligence, and the triadic framework of Being, Becoming, and Believing. At the mathematical level, the paper reviews the role of the three foundational crises of mathematics in promoting formalization, computability, and complexity theory, and reconstructs a logical trajectory from Hilbert’s Program, Gödel’s Incompleteness Theorem, the Turing Machine, Cybernetics, and McCulloch-Pitts’ Model of Artificial Neurons to Category Theory and the Langlands Program. At the level of cognitive science, it discusses the Lighthill Report, the AI Winter, Longuet-Higgins’s naming of cognitive science, Miller’s cognitive revolution, as well as the hexagonal framework of cognitive science, showing how cognitive science helped artificial intelligence move beyond symbolic limitations toward deep learning, large models, and embodied intelligence. The paper argues that artificial intelligence is a product of the sustained interaction among humanistic thought, formal rationality, and cognitive mechanisms. Reconstructing its knowledge origins provides a fundamental framework for understanding the boundaries, limitations, evolution, and future directions of intelligence.

**Key words:** artificial intelligence, knowledge origin, parallel intelligence, philosophy of intelligence, cognitive science

## 0 引言

文献[1]讨论了智能的本质与意义。本文系统梳理人工智能的知识本源,从哲学、数学、认知科学(cognitive sciences)三大维度,揭示智能科学的思想根基与学科脉络。在哲学层面,以东方“赋、比、兴”与西方“文法、逻辑、修辞”为源头,立足本体论、平行智能与“存在-生成-信念”框架,结合哥德尔定理与《道德经》思想,阐释智能的层次性与虚实共生的平行理念;在数学层面,追溯3次数学危机对形式化、可计算理论的推动,从希尔伯特纲领、罗素《数学原理》到哥德尔不完备性定理,再经图灵机、控制论、人工神经元模型直至范畴论与朗兰兹纲领,完整呈现了人工智能的数学逻辑演化链条;在认知科学层面,围绕莱特希尔报告引发的人工智能寒冬、隆科-希金斯定名认知科学的历史契机,结合米勒认知革命与六角形学科架构、控制论“金三角”的兴衰,展现认知科学如何为人工智能突破困境、走向深度学习与大模型时代提供理论支撑。本文贯通人文思辨、数理基础与交叉学科演进,为理解人工智能的本质、局限与未来方向建立起系统的本源认知框架。

## 1 人工智能与哲学

智能与哲学同根同源,一体纠缠,无法分离。至少从第一轴心时代开始,关于智能的研究一直是哲学的“专利”。本源都是生命对美好的想象与追求,其原始形态同人类的文学艺术技巧密切相关。例如,古代中国《诗经》“三法”的“赋、比、兴”之文学表述范式<sup>[2]</sup>,古希腊柏拉图《对话录》“三艺(trivium)”的“文法(grammar)、逻辑(logic/dialectic)、修辞(rhetoric)”之论辩表现方式<sup>[3]</sup>,都是人类智能“描述、演绎、生成”过程知识化的早期形态。用今天的人工智能术语来说,就是“赋”主表示,包括感知、表征、陈述、数据,即知识表示工程;“比”主推演,包括类比、逻辑、关联、推断,即推理与学习;“兴”主创造,包括想象、灵感、涌现、生成,即决策与行动。这正是当代人工智能从感知、推理到创造完整的基本框架与流程。

实际上,中文的“哲学”本意就是“智慧之学”或“智能之学”。东汉许慎编著的《说文解字》中写道“哲、知也”,更早的先秦《尔雅》说得更直接——“哲,智也”。古文“哲”字在形、音、

义3个方面都体现了人类智慧是穿透曲折之后的洞明,反映了智能与复杂的内在关联。英文更直截了当,其“哲学(philosophy)”一词就是“爱”与“智慧”的结合,是古希腊哲人的创造。

### 1.1 本体之道

人类经历了漫长的“赋、比、兴”之后,形成了关于存在(being)和变化(becoming)之哲学本体论,成为现代人工智能本体方法的思想基础。本体的核心任务就是人类通识的表征,主要包括常识(common sense)和共识(consensus),分属自然哲学和社会哲学。

常识与共识的有效表述和利用是形成标准、提高效率的前提。人工智能自成为一个单独的学科之始,常识问题一直被认为是最复杂和最困难的问题。实际上,在人工智能大模型技术涌现和普及的今天,许多专家仍然认为常识依然是最复杂和困难的问题。

哲学本体论的演化启发了许多人工智能早期“常识愚公”式的项目,如“赛克”(Cyc)和“词网”Wordnet等<sup>[4-5]</sup>,它们使数据库技术发展为常识建模的基础,从自然语言处理、知识图谱和深度学习等一跃成为今日的大语言模型(LLM)等人工智能技术。然而,面对目前LLM的“幻觉”“腐烂”“漂移”等严重问题,下一步如何发展?为此,我们不妨回顾一下中国哲学本体论之初的最核心理念,即“道”的原始哲理。

如图1所示,老子《道德经》以“道可道非常道”6个字为开篇。古文没有标点符号,传统上这6个字被解释成两句话——“道可道,非常道”,即说出来的道理,应该有个界定或使用范围,而有了界定就不再是放之四海而皆准的永恒道理。在数学上,与之对应的就是“狭义哥德尔定理”,即任何一个数学体系,不可能既是完备的,又是一致的:如果一种理论能够从其体系中推导出其所有的知识(完备性),那就无法保证其知识中没有自我矛盾的东西(一致性)。换言之,“既死又活”的薛定谔之量子猫是逃脱不掉的命运。

然而,近代长沙马王堆等考古发现,这6个字应被断成3句话——“道,可道,非常道”,其中“非常道”是西汉时为回避汉文帝刘恒之名讳,由“非恒道”演化来的。这一新的断句不但让《道德经》与“广义哥德尔定律”联系起来,更因由“恒”到“常”的一字之变,为人工智能的新本体

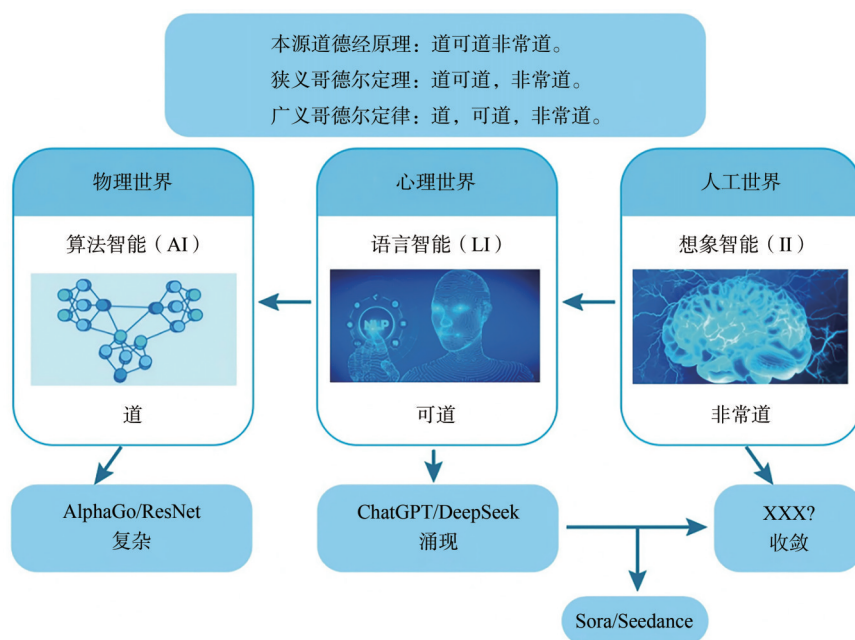


图1 我们时代的智能：广义哥德尔定理

和新智能发展开拓了思路。

王浩是最早从事人工智能研究的华人哲学家和数学家，其博士导师奎因（Willard Van Orman Quine）是逻辑实用主义宗师，号称“分析哲学的定海神针”，被公认为“20世纪最有影响力哲学家之一”。王浩20世纪五六十年代就开始利用原始的计算机进行数学定理的机械化证明研究<sup>[6]</sup>。他于1958年用IBM704计算机3分钟就证明了名著《数学原理》中的数百条逻辑定理，其方法被誉为“一击七蝇”（seven flies in one blow）<sup>[7]</sup>。这些定理的证明曾花费其作者罗素和怀特海近十年甚至是毕生的精力，影响巨大，直接导致哥德尔发现其不完备定理。罗素得知王浩的机器证明后，感叹且幽默：早知今日，何必当初。王浩于1983年获得人工智能定理机械证明的“里程碑奖”，是人工智能在数理逻辑领域的标杆成就。晚年王浩回归哲学，主要研究哥德尔的数学哲学思想。

据王浩总结，哥德尔“认为机器或计算机不可能超越人脑，除非数学不是人类发明的。而且就算数学不是人类发明的，机器计算机还是无法超越人脑”，即“广义哥德尔定律”<sup>[8]</sup>。从智能技术的角度去理解《道德经》开篇的新解释，可与广义哥德尔定律呼应，为其“非常之思”提供“非常之道”的思路：人类的智能分为3个层次，首先是以物理第一世界的运动规律及其感知规则为主的自然

之“道”及“算法智能”（algorithmic intelligence, AI），即自然中可运行的道<sup>[9]</sup>；其次是以心理第二世界的思维规律及其表示规则为主的交流之“道”及“语言智能”（linguistic intelligence, LI），即可以说出来的道<sup>[10]</sup>；最后是以第三人工世界的认知规律及其知识体系为主的创新之“非常道”及“想象智能”（imaginative intelligence, II），即只能用想象表现的道<sup>[11]</sup>。

这3个层次之间的智能存在下列界定：

**广义哥德尔定律**  
 $AI \ll LI \ll II$   
**算法智能远小于语言智能，语言智能远小于想象智能**

换言之，人类的算法智能（可道之道）远远小于其语言智能（可名之名），人类的语言智能又远远小于其想象智能（非常道，非常名，即可想之想）。西方谚语的“一幅画胜过千言万语”（A picture is worth a thousand words）、“想象是心灵之眼”（Imagination is the eye of the soul），还有爱因斯坦的名言“想象比知识更重要”，都是广义哥德尔定律的不同反映。

目前，AlphaGo/ResNet等已是算法智能“AI”方法的新经典，ChatGPT/DeepSeek等已是语言智能“LI”技术的新时尚，给我们的时代带来巨大的冲击和变化。我们期待想象智能“II”的突破，或许，Sora/Seedance等视频流媒体生成系统已经是

这一突破的开端，这也是世界模型的目标与希望。当然，我们更希望并相信新的想象智能技术不但能使人类想象力从传统意义上的精神和艺术能力升华为驱动社会进步的核心生产力与竞争力，而且还能够管理和控制好技术创新与社会治理之间的平衡，在赋能社会发展的同时，有效防范风险，保障人类安全，维持人类的根本福祉与主体地位。

## 1.2 平行之智

随着人类经验与知识的扩展，经历了“先行后知”“先知后行”“知行相资”“知行合一”等认识阶段之后，哲学本体终于从第一轴心时代的以唯物或唯心为主体的“一元”本体，发展到第二轴心时代的“心灵-心身”（mind-body）兼顾的“二元”本体。在第三轴心时代，波普尔的三个世界之平行本体<sup>[12]</sup>，更为人工智能从“新AI”的自主智能到“新AT”的自主技术，最终走向想象智能“II”和想象技术“IT”（imaginative technology）提供的一条“三元”本体道路。

实际上，不但《道德经》，中国先秦儒家哲学和宋代新儒理学的哲学思想中都有三元本体的影子。例如《中庸》中的“教、性、道”，《孟子》中的“行、心、志”，王阳明的“物、心、意”等理念，都体现了与波普尔相似的思想。特别是《中庸》的“天命之谓性，率性之谓道，修道之谓教”，以及王阳明的“心者身之主，意者心之发，物者意之用”，而且“格物者，格其心之物也，格其意之物也，格其知之物也”，突破二元的“物者，心之物也；心者，物之心也”，在一定程度上阐明了三元之间的关联与纠缠。此外，更重要的是，孔子在《论语》中表述的核心思想就是试图构造一个理想的“人工社会”，通过教育把活生生的自然人教化为“人工社会”中的“人工人”，其间大家都讲、信、实践“仁、礼、中庸”等人伦“人工”理念，以确保人类社会的和谐与和平。利用想象智能，今天我们可以通过知识机器人或智能体技术重新设计孔子的“人工人”构想，构建小社会模型（Small Social Models, SSMs）让人类在新的层面重归自然，让人有人用、机有机用、数有数用。

西方哲学也走过类似的道路。从柏拉图向虚而生的“理型”到其洞穴寓言中火的影子的比喻，再加上亚里士多德的形而上学（后物理学）及其逻辑推理的三段论，差不多为三元本体之“后”形而上学或“后后”物理学的发展铺平道路。在柏罗丁

（Plotinus，也称普罗提诺）的新柏拉图主义里，本体已经成为“太一，理智，灵魂”（The ONE, nous, soul）三位一体的三元本体，而且都是“太一之影”（The shadow of the ONE）。新柏拉图主义认为，太一“不可说，不可名，不可思”，差不多就是“名可名，非常名”，其认知与东方哲学“道生一，一生二，二生三，三生万物”几乎有相同的表述，似乎在哲学上，DAO is the ONE。虽然道之自然的生成、太一之影的超越构成了东西方哲学各自的特色，但体现了共同的人类精神，是人类命运共同体的保证。

波普尔的三个世界平行本体不但可以有机地融入现象学和过程哲学等现代思维方式，更为哲学思想与智能技术的“虚实平行相交”提供了实施的平台。图2给出一种实现“小知识→大数据/大模型→深智能”过程一体化的哲学与技术框架，其中DAO是技术之道，由分布式智体构成的组织实现，由自然人、机器人、数字人组成，受区块链制约，通过智能合约联系，并且具有“真”（TRUE）的性质：可信（trustworthy）、可靠（reliable）、可用（useful）、正效/高效（effective/efficient），即以正确的方式，做正确的事。在此，所谓的元宇宙（metaverse）实质就是社会物理赛博系统（CPSS），就是平行智能发展与应用的新时空。在人类的新人工智能技术之下，老子古老的“鸡犬之声相闻，老死不相往来”之原始的至德至治太平世界，将以“平行世界相通，虚实天涯比邻”的安全、健康、可持续的发展方式实现。

## 1.3 智能哲学

新时代的设计与建设离不开新的科技，但仅凭科技远远不够，新时代还需要新的思维和新的哲学。首先，从“赋、比、兴”或“文法、逻辑、修辞”起步，传统哲学基本上以存在（being）与生成（becoming）为核心，特别是西方哲学，主要讨论这两个“B”；智业时代的智能哲学需要走向“三个B”，即加入“believing”，必须研究以being、becoming、believing为新主体且三者平行交互的新哲学。三个世界、三种意识，使第三个“B”信念（believing）与另外两个“B”平行独立，并通过新人工智能技术使之工程化，成为“信用力”和“注意力”作为新时代关键“资产”或“商品”的哲学与技术保障。为此，我们还需要建立与三个世界对应的3种主体知识和哲学，以满足人类在物质和精神方面的需求，具体如下。

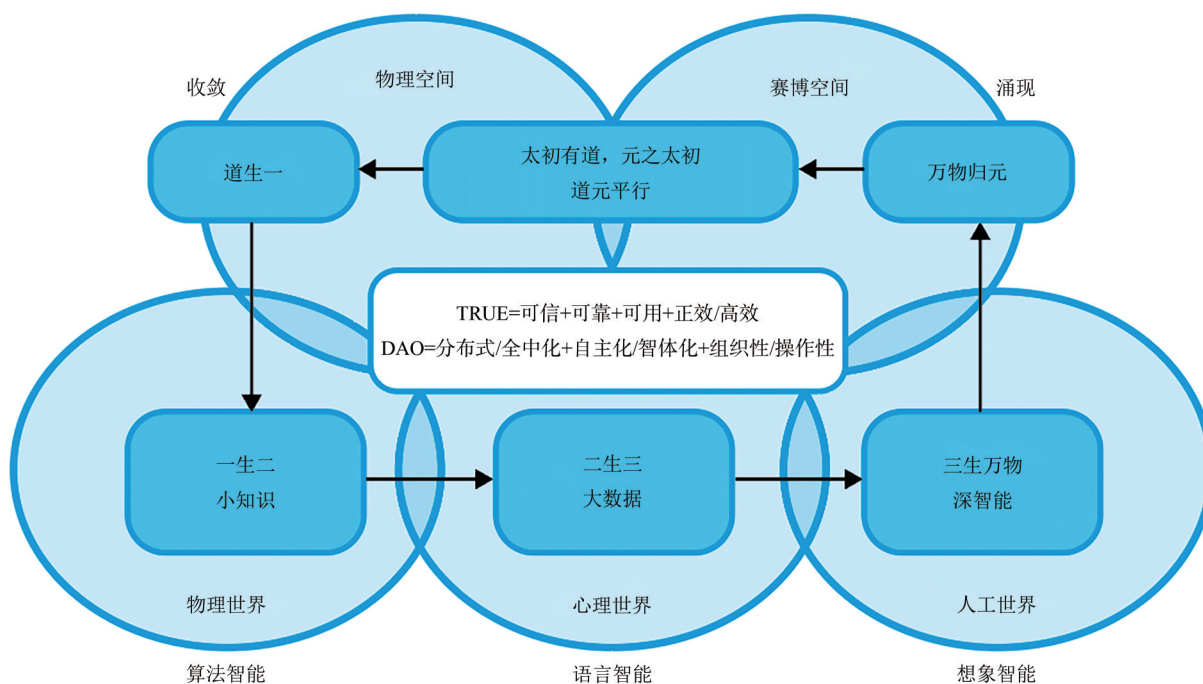


图2 基于社会-物理-赛博系统的平行智能演化架构

### 3种主体科学和哲学

|                    |                                                    |
|--------------------|----------------------------------------------------|
| (1) 存在 (being)     | 与之对应的关于物理自然世界的描述性知识 (descriptive knowledge) 及显象哲学  |
| (2) 生成 (becoming)  | 与之对应的关于心理认知世界的预测性知识 (predictive knowledge) 及过程哲学   |
| (3) 信念 (believing) | 与之对应的关于人工生成世界的引导性知识 (prescriptive knowledge) 及平行哲学 |

面向智能时代的新本体架构，由“三个B”构成的智能哲学的核心结构：

- 创世存在 (being-in-the-world)，源自哲学家海德格尔<sup>[13]</sup>；

- 变世生成 (becoming-of-the-world)；

- 为世相信 (believing-for-the-world)。

这三者不仅共在，而且更各自居于其间 (in-the-between，源自海德格尔<sup>[14]</sup>)、“与之平行” (among-the-parallel)，彼此依存、互为呼应。

这一框架将笛卡儿“我思故我在”拓展为循环贯通3个世界的自觉与循环，即：

我思故我在，立足创世存在，指向实在之基；

我在故我信，依托变世生成，展开演化之行；

我信故我思，经由为世相信，引发导向之力。

三者并不分属不同世界，而是在物理自然世界、心理认知世界与人工生成世界之间平行互动、循环相生，形成智能时代的循环因果论：我思确立

我在，我在生发我信，我信引导我思。这一思想上溯至古希腊目的论 (Teleology)；1943年维纳在《控制论》中重构现代目的论，以循环因果与反馈机制将其转化为可工程化的科学原理。受此启发，麦卡洛克与皮茨提出 McCulloch-Pitts (MP) 人工神经元模型，奠定人工神经网络基础<sup>[15]</sup>，并逐步发展为深度学习、Transformer 架构与今天的大模型体系<sup>[16-17]</sup>。

从更深层的视角看，这一循环共生、不可分割的本体结构，与量子力学中的量子纠缠具有深刻的内在同构性：三个B之间并非彼此独立、线性作用的，而是相互关联、互为条件、不可约简的，一方状态的变化会整体联动，恰如纠缠粒子之间非定域、不可分的关联。传统世界观强调分立、线性、单向因果，而新本体架构与量子纠缠共同指向一种整体性、平行性、循环性的思维图景：存在、生成与信念不再是孤立环节，而是深度纠缠、协同演化

的统一整体，共同构成智能时代最底层的本体论与科学基础。

当然，哲学并不是总从正面促进人工智能的发展。历史上，哲学家德赖弗斯（Hubert Dreyfus）在其于1965年为兰德公司（RAND Corporation）撰写的报告《炼金术与人工智能》<sup>[18]</sup>、1972年的著作《计算机不能做什么》<sup>[19]</sup>以及后续的《计算机依然不能做什么》<sup>[20]</sup>中，基于海德格尔与梅洛-庞蒂的现象学思想，对早期人工智能，特别是符号主义人工智能展开了激烈批判。他的核心观点直指传统人工智能的根本缺陷，认为其将智能简化为符号运算与逻辑推理，严重忽视了智能的具身性、情境性与默会（隐性）知识。德赖弗斯的批判早于1973年英国科学研究理事会（SRC）委托数学家莱特希尔完成的《莱特希尔报告》<sup>[21]</sup>。这份权威评估判定人工智能研究承诺过高、成果严重不足，与德赖弗斯的报告相互呼应、叠加影响，直接被英国政府采纳为削减经费的依据，导致全球人工智能领域遭遇近10年的发展寒冬。

从历史视角回望，由德赖弗斯发起的这场哲学批判实则具有重要的正面意义，使符号主义之外的人工智能方法得到更多重视，客观上有益于具身思想在人工智能研发上的影响，实际上帮助了人工智能向今天的方向发展。当年被他讥讽为“炼金术”的微弱火苗，如今早已成长为改变世界的熊熊大火；当下蓬勃兴起的具身智能，更是对其思想的直接呼应与深化。这一历程深刻启示我们，有时哲学批判并非否定技术，而是帮助人工智能回归智能本质，在反思与校准中实现更扎实、更深远的发展。

未来，哲学将为人工智能提供价值方向与存在根基，引导人工智能从单纯算力、算法走向可信、

可控、可理解的智能新形态，确保技术服务于人。人工智能则将倒逼哲学突破传统思辨框架，直面具身智能、平行世界、数字主体等全新议题，推动本体论、认识论与价值论的整体革新。二者深度融合，必将催生面向智业时代的智能哲学新体系：以存在、生成、信念为核心，在物理、认知、人工3个世界之间平行互动、循环共生，最终实现科技与人文的统一，构建人机协同、虚实相生的未来文明新范式。

## 2 人工智能与数学

人工智能与数学的深层关联在于二者共同遵循着一条“向虚而生”的演进之路。数学最初以整数与可公度为基础，构建起自洽而直观的朴素体系，却因第一次数学危机打破了这种圆满，不得不从有理数走向无理数，从实数拓展到虚数，再将哲学意义上的范畴转化为严格的数学范畴，从此踏上不断超越直观实在、走向抽象建构的历程。正如利奥波德·克罗内克（Leopold Kronecker）所言：上帝创造了整数，其余都是人的工作。数学“人的工作”之人工与后来人工智能之“人工”异曲同工，形成跨越时代的呼应。此后数学的每一次突破都在向着更形式、更抽象、更超越经验的方向延伸，让计算越来越复杂。而“智能是复杂化的计算”，由此，数学的抽象化进程与人工智能的逻辑建构便成为同一条理性脉络下的彼此呼应：数学向虚而生，人工智能循虚而智，生成智能，共同构筑人类理性的抽象与技术进化之路。

### 2.1 人工智能与数学危机

数学的发展与人类文明阶段相伴而行，不同时代有不同的核心数学形态，也对应着人类文明的三大轴心时代与数学史上3次基础危机。

#### 三大轴心时代与数学史3次基础危机

|                  |                                          |
|------------------|------------------------------------------|
| 第一轴心时代<br>(农业时代) | 数学以直观的算术与几何为核心，与之相伴的是不可公度危机              |
| 第二轴心时代<br>(工业时代) | 数学以微积分、动态变化为核心，与之相伴的是无穷小危机               |
| 第三轴心时代<br>(智业时代) | 数学以计算、逻辑、形式化与复杂性为核心，与之相伴并为之奠定基础的，则是集合论危机 |

3次危机分别对应着人类文明从农业时代、工业时代走向智业时代的历史进程，也对应着第一轴心时代、第二轴心时代、第三轴心时代的思想跃迁，共同构成了人工智能得以诞生的数学根基。

#### (1) 第一次危机：不可公度危机

古希腊人认为一切量都可以用整数或整数之比表示，甚至“万物皆数”。但边长为1的正方形对角线长度 $\sqrt{2}$ 无法表示为两个整数之比，它不是一

个有理数，说明存在不可公度的量，动摇了当时数学的基础。这揭示出：有限的理性框架无法覆盖真实世界的复杂性。人们拓展数域、承认无理数，但花了上千年的时间才为无理数建立了严格的数学基础，数学实现“向虚而生”。这一教训启示我们：世界无法被有限、离散、精确的规则完全穷尽，必须从有理走向无理，人工智能也不得不从基于符号传递逻辑推理走向计算演化，以处理连续、复杂、开放现实世界的各项任务。

### （2）第二次危机：无穷小危机

微积分中的无穷小量在运算中时而当作0、时而又当作非0，出现明显逻辑矛盾，引发数学基础动摇。后以 $\delta$ - $\varepsilon$ 严格极限语言重建严谨性，让我们可以有顺序的过程从有限走向无限。显然，这一突破为人工智能的优化算法、梯度下降、连续学习等各种计算方法提供了严密的数学支撑。

### （3）第三次危机：集合论危机

朴素集合论中允许定义“包含所有不包含自身的集合”，由此产生自指矛盾，这就是罗素悖论。通俗地讲就是：一位理发师说“我只给所有不给自己理发的人理发”，那么他该不该给自己理发？无论怎样都自相矛盾。这一悖论撼动了数学基础，哥德尔不完备性定理进一步表明形式系统无法绝对完备。数学走向公理化、形式化与可计算理论，以有限理性从有序走向无序，直接催生图灵机、计算机、范畴理论与计算复杂性科学研究，为人工智能的诞生提供理论源头。

3次数学危机恰好与哲学上的“三个B”理念对应：涉及对存在（being）的认知，从有理到无理，从实数到虚数；改变对生成（becoming）的理解，从有限到无限，从静态到动态；开创对信念（believing）的技术化研究与应用，从有序到无序，从单一到多样，从确定到不确定，从简单到复杂，从涌现到收敛，最后化哲学信念为数学概念，为对立统一创立数学基础。

3次数学危机的相继突破，共同构筑了人工智能诞生与发展的思想基石和数学基础。

## 2.2 人工智能与近代数学

人工智能的核心思想并非现代产物，而是一批数学家跨越千年对推理可机械化、逻辑可计算的持续探索。

13世纪西班牙学者鲁尔（Llull），在著作《科学之树》中提出知识系统分类与逻辑组合思想，成

为人类尝试将理性思维系统化、机械化的最早先驱之一。

17世纪的莱布尼茨（Leibniz）深受鲁尔影响，他的博士论文《论组合的艺术》直接继承并改造了鲁尔的组合艺术思想，发展出自己的组合理论。莱布尼茨真正追求的是能够对逻辑推理进行计算的“脑积分”，却在这一过程中与牛顿共同创立了微积分；他同时系统建立二进制，并将其与中国八卦相互关联，为现代计算机奠定了数制基础。

17世纪，帕斯卡（Pascal）设计出早期的机械计算器（Pascaline，1642年），开启了人类机械计算的实践探索，其与费马共同奠定了概率论的数学基础，而概率论现已成为人工智能与机器学习的核心数学工具之一。19世纪，巴比奇（Babbage）设计了分析机（analytical engine，1837年），成为现代计算机理念的重要先驱。19世纪，布尔（Boole）在著作《思维的定律》中创立布尔代数，首次将人类的逻辑思维转化为严格的数学运算，让逻辑真正实现符号化、计算化。布尔代数也成为现代数字设计、数字电路与计算机硬件的核心数学基础。

## 2.3 人工智能与现代数学

1900年，在第三届国际数学家大会上，希尔伯特提出了震撼整个数学界的23个问题，为20世纪的数学划定了方向。比这23个问题更深远，却常常被人遗忘的，是他同时提出的一套数学公理化与形式化的思想规划，后世称之为希尔伯特纲领——他梦想将整个数学大厦彻底公理化、形式化、完备化，并且可以进行机械化推理与证明。如果这个梦想真的实现，那本身就是终极的超级人工智能。

这场运动背后，还藏着一段学术史上的著名交锋——法国数学学派与德国数学学派之争，具体表现为庞加莱与希尔伯特的对立。庞加莱更偏向直觉、直观与几何洞察，而希尔伯特坚持严格、形式、逻辑与公理系统。这场争论既是今天人工智能领域通用人工智能与具身智能路线之争的最早源头，也是人工智能早期以推理为主的“纯净派”与以计算为主的“邋遢派”之争的思想雏形。

希尔伯特纲领一经提出，点燃了全球数学青年的热情，他们中间最时髦的口号就是：“打起你的背包，到哥廷根去！”年轻人奔向希尔伯特所在的哥廷根大学，希望为其心中那座完美的数学大厦添砖加瓦。

罗素和他的老师怀特海正是这场浪潮的核心推

动者。1900年第三届国际数学家大会结束后，两人便下定决心携手撰写一部专著，为希尔伯特的形式化数学大厦打下最坚实的逻辑基础。

为了这份使命，他们特意将书名定为《数学原理》，并沿用拉丁文标题 *Principia Mathematica*——这正是他们对牛顿的直接致敬。机械运动的物理世界通过牛顿的《自然哲学的数学原理》(*The Mathematical Principles of Natural Philosophy*)一书及其三大定律，变成可以精准计算的数学世界。而罗素与怀特海则想用自己的《数学原理》完成一次反向的伟大循环：把数学世界重新变回可以机械化推演与证明的机械世界。他们的目标是：只要给出公理与前提，就能在有限的机械步骤里，推出所有定理，并且保证系统既一致又完备，永远不出错。原本计划一年完成的工作，两人最终投入近10年时间，三卷本《数学原理》于1910—1913年陆续出版<sup>[22]</sup>。为了这份极致严谨，《数学原理》中仅仅证明 $1+1=2$ 都要用几十页正文，草稿更是长达几百页。《数学原理》出版后引起世界性轰动，但罗素本人认为，全世界只有5个人真正通读过这部巨著。

1931年，哥德尔发表里程碑论文，标题正是“论《数学原理》及有关系统的形式不可判定命题”。哥德尔不完备性定理横空出世，从数学上严格证明了数学系统不可能既一致又完备，彻底击碎了希尔伯特纲领的完美设想。

哥德尔打破了希尔伯特美好的数学大厦之梦，也打破了他“我们必须知道，我们必将知道”的坚定信念，却没有给出出路。

不久之后，年轻的图灵登场。1936年，他给出了计算的数学理论，即图灵机<sup>[23]</sup>，让计算机化之路正式开启。他这篇论文引用的第一篇参考文献，正是哥德尔的工作。这一成果与1935年邱奇的递归函数理论相互印证<sup>[24]</sup>，共同构成了邱奇-图灵论题。该论题成为现代计算机科学的理论源头，并催生了人工智能史上第一个也是最有名的计算机语言——Lisp。

维纳受《数学原理》影响，研究兴趣从动物学转向数学研究，最终创立控制论(cybernetics)<sup>[25]</sup>。“cybernetics”源自古希腊语“掌舵术”，安培曾将这一术语定义为治理国家的科学，即社会管理；维纳则用它研究机器与生命系统中的控制和通信，聚焦于机器与大脑的行为规律，即机械大脑。

麦克洛奇(McCulloch)与皮茨(Pitts)同样

受《数学原理》形式逻辑思想的启发，试图用逻辑方法构建大脑模型。1943年，两人在维纳控制论研究成果的启发下，发表经典论文《神经活动中内在思想的逻辑演算》，提出首个人工神经元模型，成为今天基于深度神经网络的深度学习和Transformer的思想源头。

冯·诺伊曼是最早深刻认识到哥德尔不完备性定理重大意义的学者，也是最早发现图灵研究价值，并推荐图灵到普林斯顿大学在邱奇指导下完成博士学位的人。在此基础上，他参与早期电子计算机的研制，1945年为EDVAC计算机项目撰写了著名的《EDVAC报告草案》，在其中正式提出至今仍在使用的冯·诺伊曼体系结构<sup>[26]</sup>。

值得注意的是，在这份奠定现代计算机体系的里程碑报告里，唯一引用的学术论文正是McCulloch和Pitts于1943年发表的神经元模型文章。此外，冯·诺伊曼创立的博弈论成为当今博弈智能(game intelligence)与决策智能的重要理论基础，是人工生命和人工智能的核心方向之一。

冯·诺伊曼也深刻影响了麦卡锡与明斯基这两位人工智能领域的重要学者。两人均为普林斯顿大学数学博士。麦卡锡正是在聆听了冯·诺伊曼关于控制论的相关报告后，走上了自动机与智能系统的研究道路；冯·诺伊曼也曾在明斯基的博士答辩委员会中，助力其完成博士学业。麦卡锡与明斯基后来共同成为1956年达特茅斯会议最核心的两位发起者，人工智能学科由此正式诞生。

希尔伯特、罗素、怀特海、哥德尔、图灵、维纳、McCulloch-Pitts、冯·诺伊曼、麦卡锡、明斯基，一条从现代数学出发，贯穿逻辑、计算、神经模型、计算机体系，最终走向人工智能的完整的数学思想与方法链条，就此成型。

## 2.4 人工智能与当代数学

当代数学正在为人工智能提供一种全新的底层框架，这就是范畴论(category theory)。它不再以实体、元素、集合为中心，而是以关系、结构、映射、过程为核心，被称为“数学的数学”，也正在成为下一代人工智能最具潜力的理论根基。这一趋势的革命之处就是：从对象本体走向关系本体。

1945年，数学家塞缪尔·艾伦贝格(Samuel Eilenberg)与桑德斯·麦克莱恩(Saunders MacLane)在论文《自然等价的一般理论》中正式创立范畴论<sup>[27]</sup>，标志着这一抽象数学体系的诞生。范畴论

的基本结构由对象与态射构成，它不关心对象内部的构成，而专注于对象之间如何连接、如何变换、如何在变换中保持结构不变。这种以“关系”为核心的思维方式跳出了传统数学的静态视角，成为能够统一代数、拓扑、逻辑、计算等多个领域的通用语言。

范畴论的思想并非凭空产生，其“范畴”一词直接继承了西方哲学的核心认识论传统。亚里士多德在《范畴篇》中提出十大范畴，用于界定存在的基本形式与语言世界的方式；伊曼纽尔·康德（Immanuel Kant）则将12组先天范畴作为人类理性认知世界、组织经验的基本框架。无论是古典哲学还是近代认识论，范畴都是人类用来结构化、系统化理解世界的根本工具。数学中的范畴论正是将这一哲学追求推向严格公理化的结果，它与智能系统对世界进行表征、推理、泛化的内在需求高度一致。

在范畴论中，单子（monad）是极具思想深度的核心结构，它在数学上呈现典型的三位一体特征，由函子与两组自然变换共同构成，满足结合律与单位律，是一种具备自洽性、封闭性与组合性的高阶结构。这一概念直接继承自莱布尼茨的单子论，莱布尼茨将世界视作由不可分割、自身封闭、能够映射整体的基本单元——单子所构成。范畴论中的单子正是对这一哲学思想的严格数学实现，凭借组合、封装、状态传递与上下文关联的特性，适宜作为智能系统的最小认知单元，也有望成为未来智能科学与技术中的基本单元——智子。

近年来，范畴论为人工智能的理论构建提供了重要支撑。相关研究将范畴作为知识表示的基本形式，用函子描述不同领域之间的结构保持映射，用自然变换刻画学习模型之间的一致性关系，为神经网络、表征学习、因果推理、规划系统提供了严格的形式化语义。范畴论也被用于分析深度学习模型的组合性、泛化能力与可解释性，为构建更通用、更稳健、更接近人类认知模式的人工智能系统提供数学工具<sup>[28]</sup>。

除范畴论之外，当代数学还存在一项旨在统一全领域的宏大构想——朗兰兹纲领（Langlands Program）<sup>[29]</sup>。它与范畴论在思想上高度相通，均强调结构、对称性与统一性，共同代表了当代数学走向整合与系统化的核心方向。从人工智能的长远发展来看，朗兰兹纲领提供的跨领域对应关系、对

偶性与对称性原理同样具备重要潜力，有望在知识统一表示、推理架构、几何智能与符号计算融合方向发挥关键作用，或许是未来人工智能重归符号主义路线的重要工具。

传统数学擅长处理静态、确定、封闭的系统，而新一代人工智能需要处理动态、开放、依赖关系与结构的复杂场景。范畴论凭其强大的抽象性、组合性与结构性，能够统一表示知识、逻辑、感知、行动与世界模型，为人工智能从数据驱动走向认知驱动、推理驱动提供真正意义上的统一数学框架。

范畴论立足当代数学的重要前沿，也为人工智能通向成熟与通用提供了一条重要的可能之路。在未来的智能时代，范畴论或许将成为支撑整个人工智能体系的核心数学基础。

相当意义下，代数的本源是存在，第一个B，Being，不是度量；几何的本意是变化，第二个B，Becoming，不是形态；范畴的本质是信念，第三个B，Beliving，不是关系。数学、哲学、智能的纠缠，是人工智能发展的源泉。

综上所述，数学不仅为人工智能提供了逻辑基础、计算工具与形式化语言，更以自身在3次危机推动下的学科演进为人工智能的发展提供了深刻启示。然而，数学界对人工智能的发展并非只有正向理论支撑，有时也产生严重的反向作用。最典型的例子，莫过于早期人工神经网络遭遇的理论“滑铁卢”。在那个狂热的探索期，学者正是依靠严密的数学推演，给当时的技术幻梦浇下了一盆冷水，无情揭露了其架构潜藏的理论死局：风行一时的Rosenblatt Perceptron神经网络，被明斯基数学上证明居然连简单的“异或”（XOR）问题都解决不了，直接判了一段神经网络的“死刑”。这种由严谨数学论证“盖棺定论”的悲观论调，宛如推倒了多米诺骨牌，随后在宏观战略层面引发了极具破坏性的连锁反应。此外，身为剑桥大学卢卡斯数学教授（这一席位自牛顿出任始，便承载着巨大的学术影响），数学与力学领域权威学者莱特希尔发布的《莱特希尔报告》，直接引发了人工智能发展史上最严重的一次“寒冬”。莱特希尔并非因预判AI会危害人类而反对研究，而是认为彼时人工智能研究已有25年，仍未取得值得持续资助的实质性进展，其质疑直接导致全球AI科研经费大幅收缩。而接任其卢卡斯教授席位的霍金，与他持截然相反的观点，霍金明确警示不加管控的人工智能或将成为人

类文明的终结。也正是这份由数学家主导的报告带来的行业困境，倒逼人工智能领域展开底层反思与学科重构，并直接催生了认知科学这一独立学科的诞生。

### 3 人工智能与认知科学

一般人绕不过“认知”一词，如果你相信弗洛伊德精神分析学说，那你就连梦里也躲不开“认知”的纠缠。然而，“认知”到底是指什么，特别是科学意义下的“认知”是什么，关心的人并不多，真正明确其内涵的更是少数，尽管这个词现在几乎每时每刻都会出现在人们的词汇和交谈之中。

其实，就连1983年商务印书馆出版的《现代汉语词典》里都没有“认知”的词条，那时“认知”差不多就是“认识”，在中文语境里更是如此。从汉字本源来看，“认”从言，唯有能用语言清晰表达、加以界定的，方可称为认；“知”从矢，如箭矢一般锐利精准、直抵本质，这样的理解才称得上知。认与知合一，便是既能言说表述又能深刻洞见的完整心智过程。

#### 3.1 什么是认知科学

目前关于“认知”的定义可归纳为：获取知识并通过思考、体验和感觉对其理解与加工的心理活动及过程。认知的核心内容包括注意力、记忆、判断、评估、推理、计算、领悟、决策、知识形成和语言生成等，其使命就是已有知识之应用和新知识之产生。因此，“认知”不但要像孔子那样“温故而知新”，还要有老子那般的“非常道”之创新能力。

要实现认知的使命，我们需要“认知科学”——一门“年轻”而古老的学科。相当程度上，认知科学与人工智能同根同源，有时差不多就是同义语，是同一功能和目标的不同说法或不同侧面。实际上，“认知科学”的诞生是人工智能开拓者面对重大科研生存危机时的一次“新瓶装旧药”式的绝地反击，出人意料地成了其陷入“严冬”过程中催生出的“一株春芽”。如今，这株“春芽”不但已经茁壮成长成为大树，而且已成林，构成形形色色的认知科学研究领域，正成为新时代人工智能与智能产业的核心科技。

认知科学可用一句话定义：心智（mind）及其过程的科学研究。再具体一点，就是对一般性认知的性质、任务和功能的探索。认知科学家的主要

工作就是研究智能与行为，核心是弄清楚中枢神经系统（CNS）是如何表示、处理、转化信息的。换言之，脑与神经是认知科学的支撑。话虽不多，但“内涵越小外延越大”。外延有多大？无人知晓——已有上万部专著，论文更是无数。

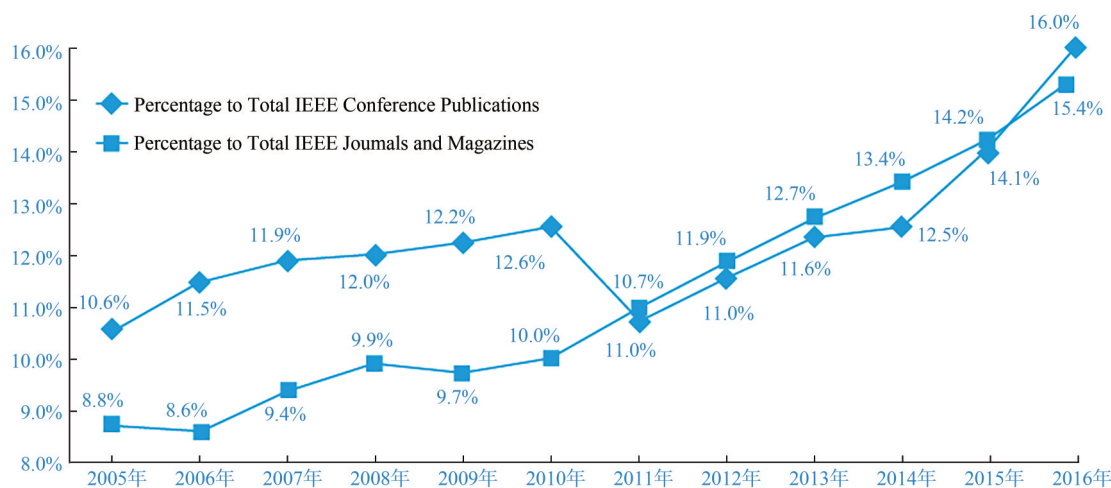
图3所示是对IEEE发表的认知科学相关论文的统计<sup>[30]</sup>。由图3可见，至2016年，认知科学的核心内容——脑及神经相关论文已占IEEE出版总量的16%，而且增速更快。比起脑与神经，认知科学论文总量至少是其2倍。IEEE又是世界上最大的相关专业组织，其会议和期刊是工程信息各类论文的最大集散地。据此，按保守估计，认知科学的论文也要占技术类文章总量的20%。实际上，专业认知科研人员在总科研人员中的比例不会超过2%，因此，这是一个令人难以置信的比例。这种迅猛的发展势头或许得益于认知科学独特的基础体系构成，使其成为当代的热点学科之一。

认知科学被认为是人工智能的“物理学”。表面上，莱特希尔1973年的报告曾让人工智能陷入发展危机，隆科-希金斯为留存其火种，以“新瓶装旧酒”的方式提出“认知科学”之名，希望延续人工智能的研究工作，也因此被公认为是这一术语的提出者。实际上，早在1956年，米勒就已掀起一场堪称“反革命”的认知革命，他提出的六角形体系搭建起认知科学的核心框架，是这门学科真正且主要的奠基者之一。

#### 3.2 人工智能冬天里的春芽：认知科学

隆科-希金斯之所以提出“认知科学”这一术语，直接导火索正是那份颇具争议的莱特希尔报告。这份报告将人工智能打入冷宫，使之陷入了近十年的AI严冬。莱特希尔报告是应英国科学研究理事会的委托撰写的，起因是以爱丁堡大学为代表的人工智能研究团队申请大额经费购置大型计算机（DECsystem-10，即PDP-10）用于AI研究，引发资助方对投入产出与研究前景的激烈争议，故委托学术权威莱特希尔独立调查评估。从正式委托到最终提交，整个过程费时约3个月，而莱特希尔本人坦言，实际调研与撰写仅花费了两个月时间。

莱特希尔在报告中质疑，人工智能是否有必要作为一门独立学科存在。他甚至带有贬义地认为，当时所谓的人工智能研究不过是些入门级、浅尝辄止的基础工作，用英文里带有轻视意味的说法就是“ABC”——指最浅显的基础知识，暗含“小菜一



2005: 9% IEEE's IPA related to brain and neurotechnology, 2016: 16%

图3 2005—2016年IEEE脑科学与神经技术知识产权资产占总出版物比例的增长趋势

碟、不值一提”的嘲讽意味。其中，“A”代表先进自动化（Advanced Automation），“C”代表基于计算机的中枢神经系统（computer-based CNS）研究，“B”则代表试图把“A”和“C”连接起来、制造机器人的架桥活动（bridging activity of building robots）。

然而，如图4所示，四分之一世纪过去了，“A”与“C”之间依然存在着巨大鸿沟，依旧是两个隔离、几乎没有交流、互不打交道的世界。作为架桥活动的“B”，特别是那些要架桥的机器人，更是不见踪影。难道“B”这座桥还沉在水下吗？人工智能没有多少值得进一步支持的进展，不如把经费分配到其他相关的学科。当时英国学者普遍认为，人工智能的开始应从图灵1947年关于智能机械的技术报告算起，而非美国人主张的1956年达特茅斯会议，已经研究了25年。

伴随此报告发表的还有4份评论，第一和第四都是直接反击，认为莱特希尔外行，甚至有些职业

操守问题，没有正确理解和弄清人工智能的内涵与意义就“乱下结论”。“B”应当被看成智能的基础科学（basic science）或智能理论（intelligence theory），而不是制造机器人去建桥的活动。第二份评论差不多是附和莱特希尔，只有第三份隆科-希金斯的评论耐人寻味，提出“认知科学”作为人工智能新的研究方向，结果更是出人意料。

隆科-希金斯认为，认知科学大体分为数学、语言、心理、生理4个主要方面，就其研究与理解认知而言，不可能将这些方面孤立开来。他还特别指出，单数形式的认知科学要比复数形式的认知科学更合适。

当时美国相关学者正处在“认知（反）革命”的高潮之中，学术界有些迷失方向，各种观点分裂巨大，特别缺少一面可以形成共识的鲜艳旗帜。隆科-希金斯的这一建议就像一场及时雨，立即被多数学者接受。就这样，在人工智能即将进入寒冬之际，隆科-希金斯“天时、地利、人和”一般地催

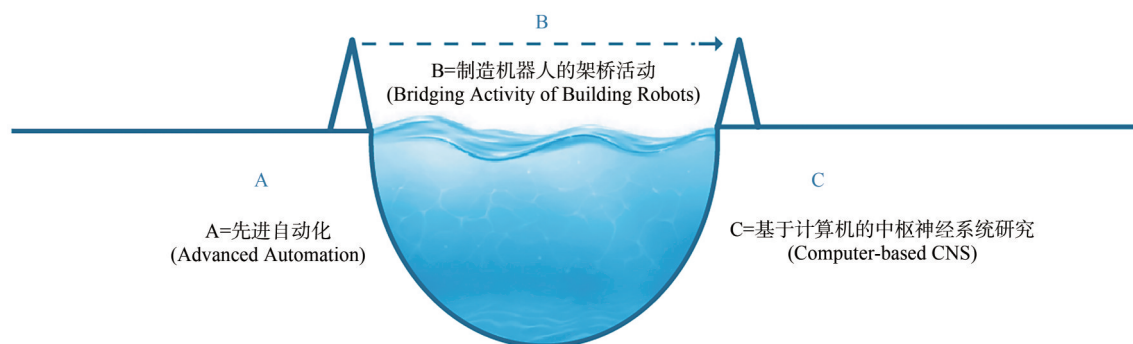


图4 莱特希尔的AI认知：ABC

生了认知科学，使其成为一门独立的正式学科。

尽管隆科-希金斯为认知科学确立了学科名称，让这一领域从AI寒冬中为人工智能的研究又树起一面新旗帜，但彼时英国学术界对认知科学并未给予充分支持，导致其博士毕业生、团队核心研究成员欣顿，于1978年远赴美国加州大学圣地亚哥分校担任认知科学相关项目的研究助理，这成为其学术生涯的重要转折点。

1977年，认知科学领域核心期刊 *Cognitive Science* 创刊于加州大学圣地亚哥分校；1979年，国际认知科学学会也在此校正式成立。学术期刊与专业学会的双重奠基，让认知科学从此真正成为国际认可的独立学科。1986年，该校更是成立了全球首个认知科学系，为学科发展搭建起正式的学术建制。1981年，欣顿参与该校认知科学研究学者鲁梅尔哈特、麦克莱兰主持的认知研究核心项目“并行分布式处理”（parallel distributed processing, PDP）项目研究。

彼时数学界已经证明，多层神经网络只要3层就足以逼近任意给定的函数功能，因此学术界普遍认为神经网络层数不应该超过3层。而PDP项目却打破了这一传统认识，大胆尝试超越3层的深层网络结构。尽管当时人们还不能完全理解其意义，但这一探索为后来的深度学习、大模型等研究方向做出了最早、最关键的开创性贡献。该项目最核心的贡献，便是让多层神经网络成为认知科学新的研究方法与研究工具。

相关研究成果最终集结为两卷本《并行分布式处理：认知的微观结构探索》<sup>[31]</sup>，于1986年正式出版，在当时人工智能领域引起广泛关注，并产生了巨大而深远的影响。书中提出的广义德尔塔学习规则（generalized delta learning rule）成为多层神经网络训练的核心算法支撑。但学术界很快发现，该算法的核心思路早已由沃博斯在其1974年的博士论文中提出<sup>[32]</sup>，由此引发了一场有关学术不端的争议。

实际上，欣顿从未声称广义德尔塔学习规则是自己的发现，也并不事先知道沃博斯的相关工作。但受此争论与当时研究环境的多重影响，欣顿随后离开美国，前往加拿大继续从事神经网络研究。

此后发生的一切，早已成为人们熟知的历史。在欣顿等人工作的基础上发展起来的深度学习与Transformer研究，催生了AlphaGo、ChatGPT等新

一代人工智能工具。沿着这一路线展开的认知科学研究也已然成为当今人工智能研究的主流，认知科学与人工智能再度合二为一。如今具身智能领域的视觉-语言-动作（vision-language-action, VLA）体系，正是这一融合方向最前沿的代表工作。

隆科-希金斯是人工智能史上一位十分具有传奇色彩的学者。他早年在剑桥攻读理论化学，是计算化学的开拓者，大学二年级发表重要论文，正确揭示乙硼烷（ $B_2H_6$ ）的分子结构，推翻了当时教科书与诺贝尔奖获得者的权威结论。许多人认为，他理应由此获得诺贝尔化学奖。

他29岁受聘为伦敦国王学院理论物理学教授，培养出希格斯（希格斯后来因“上帝粒子”研究获得2013年诺贝尔物理学奖）等学者。31岁时，隆科-希金斯任剑桥大学理论化学教授，此后转向脑科学与人工智能领域，并培养出欣顿（欣顿后来获得图灵奖与诺贝尔物理学奖）等学者。

作为开创英国人工智能研究的三剑客之一，隆科-希金斯于1967年在爱丁堡大学参与创办世界上首个专注人工智能的机器智能感知系（后更名为人工智能学院）。他提出计算机视觉基本矩阵法与八点估算法，为现代视觉计算奠定基础。他去世之后，他的学生和学术界在国际重要会议CVPR上设立了隆科-希金斯奖，以纪念他的杰出贡献。音乐是他一生的爱好，他指导了多位博士生利用人工智能进行音乐创作，并开创了音乐认知这一研究领域。

一些英国学者认为，隆科-希金斯的学术生涯并不顺利，成就没有得到相应的认可，并为此不平。但在“认知科学”一词的命名权上，他又是十分幸运的。这一命名把智能和行为的探索同生命的计算研究凝聚成了一个整体，为人工智能在低谷时指明了一条上山的突围之路，同时也为已然成规模但方向不清的“认知革命”运动树立了一面鲜明的旗帜，并很快得到全世界相关学者的响应，快速推动了认知科学的蓬勃发展。

可这一切对学者乔治·米勒（George Miller）又意味着什么？作为“认知革命”的真正发动者和认知研究早期及长期以来最主要的开拓者之一，米勒从哈佛到普林斯顿，再到圣地亚哥，攻城略地，开疆拓土。可一夜之间，自认为是棋手的米勒，醒来却发现自己肩上扛着别人的旗帜。这位心理学大师又作何感想？

### 3.3 一场认知“反”革命：米勒数与认知六角形

米勒是20世纪杰出的心理学家，也被众多学者视为认知科学真正的创始人。1956年，他发表经典论文《神奇的数字 $7\pm 2$ ：我们信息加工能力的某些局限》<sup>[33]</sup>。这篇文章不仅开启了撼动整个心理学领域的认知革命，更成为心理学、认知心理学与认知科学史上引用率最高的经典文献之一。

在心理学早期，主流是依靠内省法研究意识的学派，主张通过自我反省分析心理结构。后来兴起的行为主义推翻了内省学派，主张只研究可观察的外部行为，完全拒绝讨论意识、注意、记忆等内部心理过程，把人脑当作一个不可研究的“黑箱”。而米勒开启的认知革命又推翻了行为主义，重新将内部认知过程与信息加工带回心理学的核心，实现了范式的再次翻转。也正因为如此，这场由米勒推动的认知革命，实质上也可以看作一场认知反革命。

米勒提出的神奇数字 $7\pm 2$ （米勒数），揭示了人类短时记忆的基本规律：人在同一时间内，只能对少至5个、多至9个、平均7个独立信息单元进行稳定有效的处理；超出这一范围，就会记不住、处理不过来，产生问题。

这一发现的革命性在于，米勒直接将香农1948年创立的信息论引入心理学，把人脑看作有带宽上限、信息通量限制的“通信通道”，用信息熵、信道容量等通信理论解释人类的认知极限。更重要的是，从历史视角回看，米勒是运用大数据思想的前驱之一，开创了数据驱动揭示人类认知规律的先河。

这不仅终结了行为主义主导的时代，更开创了计算心理学，并深刻影响了生物信息学、生物计算、社会计算、计算社会科学等一大批交叉学科的诞生与发展。

与米勒数形成清晰对照的是邓巴数（Dunbar's number,  $150\pm 50$ ）。

- 米勒数 $7\pm 2$ ：个体认知层面——同一时间可稳定处理的信息单元上限，人类短期内只能与 $7\pm 2$ 个人维持深度密切的互动关系。
- 邓巴数 $150\pm 50$ ：人际社会层面——能够长期稳定维持密切关系的人数上限。

邓巴数由英国牛津大学进化人类学家罗宾·邓巴（Robin Dunbar）提出<sup>[34]</sup>。他基于灵长类大脑新皮层容量与群体规模的关联研究，推算出人类受大脑认知能力限制，能够稳定维持密切关系的上限约

为150人。

米勒数与邓巴数以个体的认知能力为中心，一内一外，一个管即时认知处理，一个管长期社会关系，共同说明人类无论在认知还是社交上，都存在由大脑结构决定的硬约束。人工智能创始人之一的司马贺提出的有限理性原理指出，人类在现实决策中不存在绝对最优解，只有满意解，这一思想也正是建立在这一信息加工能力有限的基础之上的。

米勒提出的短时记忆容量限制思想与现代人工智能中的长短期记忆网络（LSTM）<sup>[35]</sup>在设计理念上具有内在关联。LSTM通过记忆单元与门控机制，对信息进行选择性保持、更新与遗忘，在机理上与人脑有限认知容量与信息处理瓶颈的思想高度契合，是时序信息处理领域的重要基础模型。

认知科学六角形是米勒的杰作，用他自己的话说，是其“反革命”的杰作。如图5所示，此“六角形”形似人头，神似人脑：从顶顺时针向下为哲学、语言学、人类学，以感性知识为主，归人文社会科学，差不多就是右脑的功能；从底顺时针向上为神经科学、人工智能、心理学，以理性知识为主，属理工自然科学，就像左脑的功能一样。看到此图，“脑神经”3个字往往影像般地涌现脑海，十分形象地代表了认知科学的核心。

20世纪70年代末，闻知斯隆基金会将资助MIT的明斯基等以人工智能为主开展认知科学研究后，米勒“横插一杠”向基金及项目负责人提出自己的想法：钱给MIT就是多买些计算机而已，况且人工智能只是大趋势的一小部分，认知科学要以心理学、语言学、计算机科学为主，辅以神经科学、人类学、哲学，并在给基金会的报告中加上了图6所示的唯一的一张图<sup>[36]</sup>。当时，Sloan基金会正担心被世人视为MIT的专门基金，结果本应MIT独享的经费，最后被分散至许多按六角形选的研究机构，其中之一就是支持康奈尔大学的迈克尔·加扎尼加（Michael Gazzaniga）开始了认知神经科学研究，他最后成为此领域的大家，由此开启了认知科学史上的一段重要的发展历程。作为这一历史插曲的副产品，认知科学“六角形”问世，并演化为今天的图5。令米勒“遗憾”的是，最终，在很大程度上，认知科学还是如明斯基所愿被人工智能主导，而不是他所希望的心理科学。

实际上，米勒的“六角形”并不全是针对以明斯基为代表的以人工智能为主的认知科学观点，更

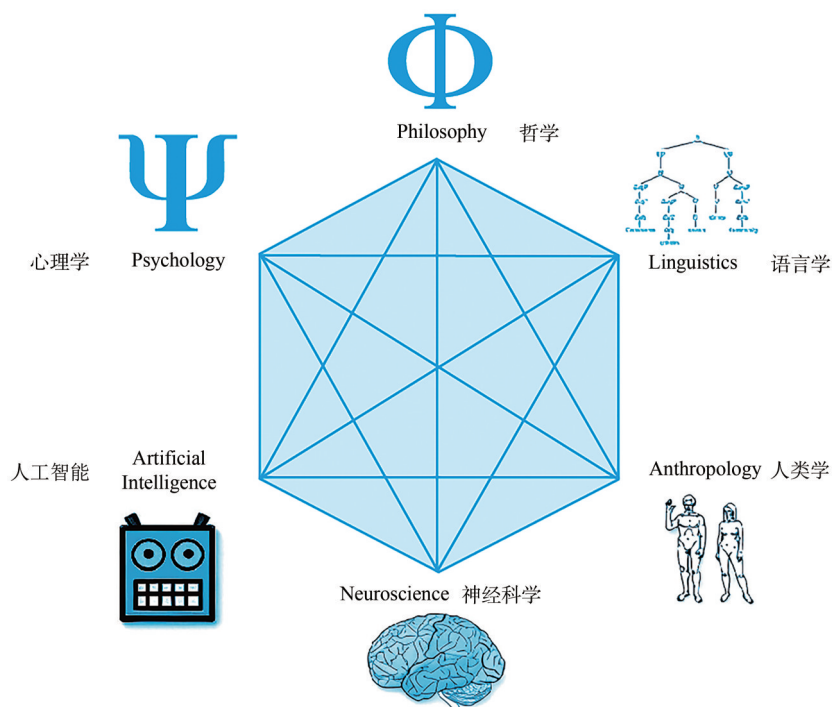


图5 认知科学的“六角形”

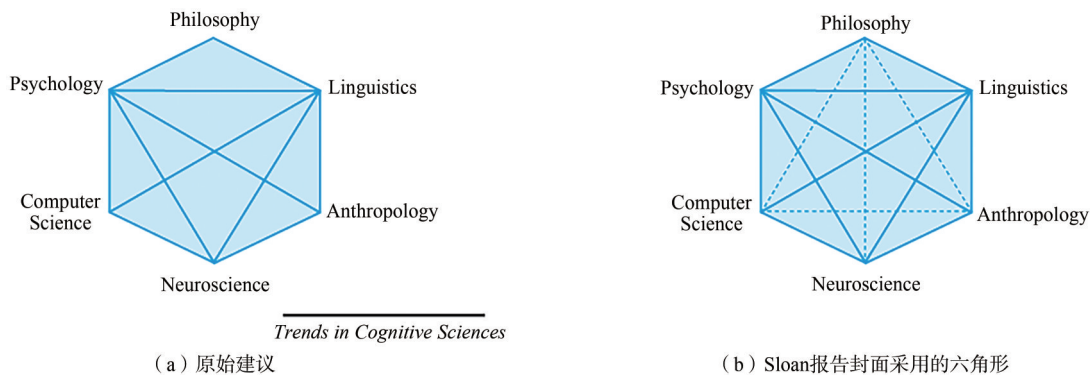


图6 米勒的认知六角形

主要的是基于自己长期的体验与认识：研究认知的专家由于各种各样的原因，有些纯粹是因为历史的偶然，往往有自己特定的角度和方法，对其他的工作和成果略知一二，尽管很难掌握全局情况，但都意识到自己面临的一些问题的解决方案关键性地取决于传统上其他学科对其的求解，就是自己好多“门”的钥匙都在别人手中！据此，米勒提出认知六角形，强调认知科学必须多学科、跨学科交融发展。在图5中，两个学科之间的连线代表它们之间已有联系，或必须联系。当时只有11根连线，如今，图5已是完全连通的图，15根连线全有了。一张图的演化，象征着认知科学的成熟与崛起。

米勒晚年更因组织领导了词网“Wordnet”项

目而成为如今文本分析、自然语言处理和人工智能从业人员无法绕过的一座丰碑，有力地推动了大型智能变革。从其相关文章中可以看出，米勒当仁不让地视自己为现代认知科学的创始人，至少是最主要的领袖之一。今天，学术界已公认米勒为认知革命的发起者、认知心理学与认知科学的主要奠基人，是当之无愧的学科创始人。在智能研究史上，英国人开创了人工智能，命名权却归了美国人；美国人开创了认知科学，命名权却归了英国人。在这之后，还有更加让人感叹的故事。

### 3.4 认知科学与脑机交互：控制论“金三角”的缺憾

其实，在米勒开启认知革命、欣顿从认知科学

角度推动神经网络与深度学习兴起之前，早在1943年就已正式出现一条远比它们更早，也更接近生物脑本质的研究脉络。这条脉络直接连通控制论、神经科学与智能研究，不仅孕育了最早的人工神经元模型，更开启了脑机交互与具身智能的源头探索。它比认知革命、深度学习的正式兴起都早了数十年，却因一个学术共同体的突然解体，成为学科史上一段被长期低估，甚至被忽略的重要序幕。这便是控制论“金三角”的故事，也是认知科学与脑机交互真正的起点。

20世纪40年代，控制论的诞生为理解生命的智能本质与机器的认知模拟，以及后来的认知科学与脑机接口开辟了重要的思想路径。作为这一交叉领域的奠基人，维纳提出的信息反馈与神经计算理念<sup>[37]</sup>，从一开始就深刻影响了以麦卡洛克、皮茨为核心的研究团队。后来皮茨和麦卡洛克先后转入MIT，三人共同构成了基于控制论与神经科学交叉领域的认知研究“金三角”，成为智能与认知科学最早的奠基力量。

事实上，这3位学者虽背景各异，却都在《数学原理》所代表的数理逻辑思想深刻影响下，最终汇聚到同一研究领域。维纳早年曾关注动物研究，博士阶段最初研习哲学，受罗素和怀特海《数学原理》<sup>[22]</sup>的影响转向数学领域；毕业后前往剑桥追随罗素深造，又先后与哈代、希尔伯特学习与合作，并在此期间结识了当时希尔伯特的助手冯·诺伊曼、邱奇等青年学者。麦卡洛克在大学时期便深入钻研数理逻辑，他认为大脑皮层顶层神经网络连接方式，与数理逻辑刻画的形式结构具有内在一致性，并由此从心理学转向神经科学，专门研究大脑机制。他曾在耶鲁大学绘制出世界上第一张大脑解剖功能图，之后前往芝加哥创建了大脑研究实验室。皮茨的经历更具传奇色彩，这位中学未毕业、出身底层的少年，因数理逻辑与麦卡洛克结缘，最终在维纳的帮助下，在无正式中学学位的情况下，破格成为维纳在MIT的博士生。三人因共同的思想源头与学术追求走到一起，创立控制论和人工神经元模型，在人工智能与认知科学的早期发展史上，写下了一段令人感叹的传奇，也留下了一段令人唏嘘的悲剧。

维纳从1925年开始与布什合作研究计算机，并在布什的帮助下招收李郁荣为他的第一个博士生，可能也是他唯一的工程博士生。二人一起开展

模拟网络的研究。然而与布什不同，维纳主张计算机应当基于数字信号而非模拟信号，当时绝大多数计算机仍采用模拟信号实现，并且他坚持计算机应采用二进制，而不是十进制。布什作为模拟计算机领域的权威，对此并不十分认同，两人之间因此产生分歧。在李郁荣的帮助下，维纳对于反馈，特别是负反馈电路、电路网络中的循环因果现象，以及相应的振荡产生了极大兴趣。二人不仅共同发明了“李-维纳网络”<sup>[38]</sup>，并由此衍生出现代通信的技术基础，还萌生了重要思想：受负反馈和循环因果逻辑支配的目的性行为，是实现智能的本质。这也成为后来创立控制论的关键因素。维纳于1935年到1936年在清华大学访学与合作，在此期间他希望继续在清华大学开展计算机研究，并建议向MIT采购相关器件，尽管这一建议被布什否决，他仍将这段在清华的时期视为自己学术思想发展的关键阶段。返回美国后，维纳曾在冯·诺伊曼家中停留数日，向他介绍同布什关于计算机的争论以及在清华大学的工作，并推荐冯·诺伊曼前往清华大学访问，但因卢沟桥事变未能成行。维纳后来回忆：“在我的学术生涯中，如果说有一条分界线，标志着从专门学术走向真正独立开创，那么我认为是1935年的中国之行。”

麦卡洛克在1942年5月的梅西基金会交叉学科研讨会上，从维纳的好友罗森布鲁斯那里了解到维纳团队关于循环因果的思想与研究。他立刻意识到，这种新型因果关系（即目的论）不仅支配着生物与机器的目的性行为，更为他长期困惑的神经网络模型中的时序表示问题提供了关键思路。对于他而言，循环因果本就十分自然——大脑神经网络本就是首尾相连的结构，在他看来必然形成因果循环。会后麦卡洛克返回芝加哥，立刻与皮茨展开合作。在维纳、罗森布鲁斯与毕格罗在《科学哲学》发表那篇著名的目的论论文后不久<sup>[37]</sup>，1943年，麦卡洛克与皮茨也共同发表了包含人工神经网络的McCulloch-Pitts模型（MP模型）<sup>[15]</sup>。1943年的这两篇论文分别开启了认知科学与人工神经网络研究，既是计算智能的奠基工作，也是计算机智能的起点，更是今天深度学习、AlphaGo等技术的深层思想源头。

二战期间，弹道计算与曼哈顿原子弹项目，尤其是冯·诺伊曼推动的蒙特卡罗方法，让他深刻认识到算力与计算机的重要性。二战结束前夕，他又

了解到军方研制现代计算机的进展,因此将兴趣与精力转向计算机领域,并重新与维纳恢复密切交往。1944年12月,由维纳起草、冯·诺伊曼与哈佛大学负责Mark计算机的数学家艾肯共同署名的邀请信,发给了由维纳亲自挑选的一批学者。三人认为,一场伟大的技术革命即将来临,计算机将在其中发挥重大作用,提议为即将到来的宏大科学探索界定核心主题,即维纳提出的“目的论新原则”,并将未来研究领域的专业学会定名为“目的论学会”。学会的研究围绕四大领域展开:通信工程、计算机工程、控制工程,以及神经系统的通信与控制。该学会致力于研究目的如何在人和动物的行为中得以实现,以及研究如何通过机械与电子手段模拟目的性行为。

1945年1月6日至7日,“目的论学会”会议在普林斯顿召开,由维纳与冯·诺伊曼共同主持。这场学术革命的构想,完全是在维纳科学思想的启发与引领下展开的。从某种意义上说,这是1956年达特茅斯会议之前,关于智能与计算最严肃、最具奠基意义的学术研讨。但这批学者只开了这一次会,之后就再也没有活动,所谓的“目的论学会”也无疾而终。这一失败很大程度上与“目的论”这一略显古怪的哲学名称直接相关。倘若当时选用一个更中性、更易被学术界接受的名称,或许人工智能这一学科便会由此正式开端。对于智能科学史而言,这无疑极具研究价值的一段插曲。

维纳对“目的论学会”及其研讨会的失败并不十分介意。在麦卡洛克的大力支持下,从1946年起,维纳的主要兴趣已转移到梅西基金会专为他资助的系列会议,并把主要精力放在认知科学,尤其是基于动物研究的生理认知科学。他的好友罗森布鲁斯从哈佛大学回到墨西哥,成立专门实验室开展相关研究,并与维纳密切合作。皮茨也成为维纳的博士生,来到MIT跟随他开展研究。此后,麦卡洛克放弃了芝加哥的实验室,加入维纳在MIT的团队,合作达到顶峰。他们在MIT完成并于1959年发表了影响深远的经典论文——《青蛙的眼睛告诉青蛙的大脑什么》<sup>[39]</sup>,成为认知神经科学的里程碑工作。

这一期间,团队获得了大量资助,购买了大批新的仪器设备,研究势头正盛,在国际上形成了十分重要的影响力。但由于一场与学术无关的纠纷,维纳突然、决绝地中断了与皮茨、麦卡洛克的全部

合作,在没有任何公开预兆的情况下,宣布永远断绝与两人的一切联系。这一变故让麦卡洛克与皮茨陷入绝境,一蹶不振。皮茨彻底沉沦,拒绝博士学位,烧毁全部手稿,自我封闭,最终(1969年)英年早逝,年仅46岁。备受打击的麦卡洛克也于同年去世。智能与认知研究的“金三角”就此崩塌,这也成为当时许多学者不解并远离维纳及控制论的重要原因。

尽管“金三角”就此破裂,但仍有许多维纳的学生继续从事认知科学与智能相关研究,奥利弗·塞尔弗里奇(Oliver Selfridge)便是其中的代表人物。他参加了1956年夏季的达特茅斯人工智能会议,同年9月又在MIT组织召开了认知科学专题研讨会,延续了控制论与认知研究的脉络。塞尔弗里奇在人工智能与计算机视觉领域贡献卓著,他开创了光学字符识别(OCR)方向,研制出早期实用化的字符识别系统。1959年,他提出经典的鬼域(Pandemonium)模型<sup>[40]</sup>,采用分层特征提取与并行处理机制,成为现代卷积神经网络(CNN)与深度学习的重要思想源头,也为模式识别、机器感知和计算机视觉奠定了早期理论基础。

维纳的控制论思想不仅奠定了智能科学的根基,更跨越学科、跨越时代,深刻影响了后世无数顶尖学者。他的信息与反馈理念直接启发了诺贝尔奖得主克里克对生命本质的研究。克里克随后正式提出“信息是生命系统的一项基本属性”这一重要观点,清晰揭示了生命与信息的新奥秘。克里克自己也坦言,生命中信息运行的模式与“10年前维纳提出的模式惊人地相似”。

MIT人工智能先驱明斯基在1986年提出的智能体理论与心智社会思想<sup>[41]</sup>,以及MIT机器人专家罗德尼·布鲁克斯(Rodney Brooks)于1986年提出的分层包容(subsumption architecture)<sup>[42]</sup>、1991年完善的具身智能与行为机器人思想<sup>[43]</sup>,同样延续着维纳认知科学的思想脉络。从脑机接口、脑机交互到现代机器人与智能系统,诸多前沿领域能发展到今天,都离不开维纳当年的杰出贡献。

值得一提的是,国际系统、人与控制论(Systems, Man, and Cybernetics, SMC)学会在维纳思想的深刻影响下创立。图3正是该学会前任会长的总结,其中标注的IoM(Internet of minds)就是智联网。在互联网、物联网之后,人类社会必将进入智联网时代,其核心是连接各类AI智能体、数字

人与大模型。未来人类将通过数字人，借助互联网与AI智能体、大模型及他人进行交流，这一模式远比当下的脑机接口更强大、更高效，应用范围也更广泛。当前出现的MCP、ACP以及A2A（agent-to-agent）等智能体交互协议还只是互联网的早期原始形态，而这正是维纳当年开创的控制论与认知科学思想在智能时代的延伸与未来发展。

面对人工智能的快速发展，不少人担忧机器会取代人类，甚至威胁人类生存。当年面对自动化浪潮，社会同样出现过机器会替代人、冲击人类地位的担忧，而维纳正是针对这类关切作出回应，提出“人有人用，机有机用”，并以此为核心思想，写下名著《人有人的用处：控制论与社会》<sup>[44]</sup>。今天重温维纳的观点，依然具有重要启示。

## 4 结束语

人工智能的知识本源根植于哲学、数学与认知科学，三者共同构成智能科学的理论基石。本文梳理了人工智能的思想源头，哲学层面从本体、存在与信念的三元结构出发，揭示智能的虚实共生与平行理念；数学层面以3次数学危机为线索，展现从形式化、可计算理论到图灵机、控制论的逻辑链条，为智能计算奠定数学基础；认知科学层面则在人工智能寒冬中崛起，以六角形学科架构为框架，形成跨学科融合的智能研究范式。本文指出，人工智能并非单一学科的产物，而是哲学思辨、数学逻辑与认知机制长期交融的结果，为理解智能的边界、演化与未来方向提供了本源框架。

## 参考文献：

- [1] 王飞跃. 智能的本质与意义[J]. 智能科学与技术学报, 2026, 8(1): 1-11.
- [2] Wang F Y. On intelligence: the origin and the goal[J]. Chinese Journal of Intelligent Science and Technology, 2026, 8(1): 1-11.
- [3] Anonymous. The book of songs (Shi Jing)[M]. Trans. Waley A. London: George Allen & Unwin, 1937.
- [4] Plato. Complete works[M]. Cooper J M, Hutchinson D S, eds. Indianapolis: Hackett Publishing, 1997.
- [5] Lenat D B. CYC: a large-scale investment in knowledge infrastructure[J]. Communications of the ACM, 1995, 38(11): 33-38.
- [6] Miller G A, Beckwith R, Fellbaum C, et al. Introduction to WordNet: an on-line lexical database\*[J]. International Journal of Lexicography, 1990, 3(4): 235-244.
- [7] Wang H. Proving theorems by pattern recognition I[J]. Communications of the ACM, 1960, 3(4): 220-234.
- [8] Wang H. Toward mechanical mathematics[J]. IBM Journal of Research and Development, 1960, 4(1): 2-22.
- [9] Wang H. A logical journey: from gödel to philosophy[M]. Cambridge: MIT Press, 1996.
- [10] Wang F Y, Zhang J J, Zheng X H, et al. Where does AlphaGo go: from church-turing thesis to AlphaGo thesis and beyond[J]. IEEE/CAA Journal of Automatica Sinica, 2016, 3(2): 113-120.
- [11] Wang F Y, Miao Q H, Li X, et al. What does ChatGPT say: the DAO from algorithmic intelligence to linguistic intelligence[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(3): 575-579.
- [12] Wang F Y, Miao Q H, Li L X, et al. When does sora show: the beginning of TAO to imaginative intelligence and scenarios engineering[J]. IEEE/CAA Journal of Automatica Sinica, 2024, 11(4): 809-815.
- [13] Popper K R. Three worlds[M]//McMurrin S M. The tanner lectures on human values. Salt Lake City: University of Utah Press, 1980: 141-167.
- [14] Heidegger M. Being and time[M]. Translated by Macquarrie J, Robinson E. Oxford: Blackwell Publishers, 1962.
- [15] Heidegger M. Contributions to philosophy (of the event)[M]. Translated by Rojcewicz R, Vallega-Neu D. Bloomington: Indiana University Press, 2012.
- [16] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The Bulletin of Mathematical Biophysics, 1943, 5(4): 115-133.
- [17] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
- [18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30. [S.l.: s. n.], 2017: 5998-6008.
- [19] Dreyfus H L. Alchemy and artificial intelligence[R]. Santa Monica: RAND Corporation, 1965.
- [20] Dreyfus H L. What computers can't do: a critique of artificial reason[M]. New York, NY: Harper & Row, 1972.
- [21] Dreyfus H L. What computers still can't do: a critique of artificial reason[M]. Cambridge: MIT Press, 1992.
- [22] Lighthill M J. Artificial intelligence: a general survey[C]//Artificial Intelligence: A Paper Symposium. London: Science Research Council, 1973.
- [23] Whitehead A N, Russell B. Principia mathematica: Vol. 1-3[M]. Cambridge: Cambridge University Press, 1910-1913.
- [24] Turing A M. On computable numbers, with an application to the Entscheidungsproblem[J]. Proceedings of the London Mathematical Society, 1937, s2-42(1): 230-265.
- [25] Church A. An unsolvable problem of elementary number theory[J]. American Journal of Mathematics, 1936, 58(2): 345.
- [26] Wiener N. Cybernetics: or control and communication in the animal and the machine[M]. Cambridge: Technology Press, 1948.
- [27] Von Neumann J. First draft of a report on the EDVAC[M]. Philadelphia: Moore School of Electrical Engineering, University of Pennsylvania, 1945.
- [28] Eilenberg S, MacLane S. General theory of natural equivalences[J]. Transactions of the American Mathematical Society, 1945, 58(2): 231-294.
- [29] Fong B, Spivak D I. An invitation to applied category theory: seven sketches in compositionality[M]. Cambridge: Cambridge University Press, 2019.
- [30] Langlands R P. Problems in the theory of automorphic forms to Salomon Bochner in gratitude[C]//Lectures in Modern Analysis and Applications III. Heidelberg: Springer, 1970: 18-61.
- [31] Wang F Y. A winter's spring bud: the story of cognitive science (I)[J]. Intelligence and Complexity, 2018, 12(4): 2-7.
- [32] Rumelhart D E, McClelland J L, PDP Research Group. Parallel distrib-

- uted processing, Volume 1: explorations in the microstructure of cognition: foundations[M]. Cambridge: MIT Press, 1986.
- [32] Werbos P J. Beyond regression: new tools for prediction and analysis in the behavioral sciences[D]. Cambridge: Harvard University, 1974.
- [33] Miller G A. The magical number seven, plus or minus two: some limits on our capacity for processing information[J]. Psychological Review, 1956, 63(2): 81-97.
- [34] Dunbar R I M. Neocortex size as a constraint on group size in Primates[J]. Journal of Human Evolution, 1992, 22(6): 469-493.
- [35] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [36] Sloan Foundation Committee. Sloan cognitive science: report by a sloan foundation Committee on the state of the art of cognitive science[R]. Massachusetts Institute of Technology Libraries, Department of Distinctive Collections, 1978.
- [37] Rosenblueth A, Wiener N, Bigelow J. Behavior, purpose and teleology[J]. Philosophy of Science, 1943, 10(1): 18-24.
- [38] Lee Y W, Wiener N. Electric network system: U.S. 2024900[P]. 1935-12-17.
- [39] Lettvin J Y, Maturana H R, McCulloch W S, et al. What the frog's eye tells the frog's brain[J]. Proceedings of the IRE, 1959, 47(11): 1940-1951.
- [40] Selfridge O G. Pandemonium: a paradigm for learning[C]//Proceedings of the Symposium on Mechanisation of Thought Processes. London: HMSO, 1959: 511-529.
- [41] Minsky M L. The society of mind[M]. New York: Simon and Schuster, 1986.
- [42] Brooks R. A robust layered control system for a mobile robot[J]. IEEE Journal on Robotics and Automation, 1986, 2(1): 14-23.
- [43] Brooks R A. Intelligence without representation[J]. Artificial Intelligence, 1991, 47(1/2/3): 139-159.
- [44] Wiener N. The human use of human beings[M]. Boston: Houghton Mifflin, 1950.

#### [作者简介]



王飞跃 (1961- ), 男, 博士, 中国科学院自动化研究所复杂系统管理与控制国家重点实验室主任, 澳门科技大学特聘教授, 主要研究方向为平行系统的方法与应用、社会计算、平行智能、知识自动化。