中国人工智能系列白皮书

——心智计算：构建脑与心智启发的人工智能

中国人工智能学会

二〇二三年九月

《中国人工智能系列白皮书》编委会

主　　任：戴琼海

执行主任：王国胤

副 主 任：陈　杰　何　友 刘成林　刘　宏　孙富春　王恩东
王文博　赵春江　周志华

委　　员：班晓娟　曹　鹏　陈　纯　陈松灿　邓伟文　董振江
杜军平　付宜利　古天龙　桂卫华　何　清　胡国平
黄河燕　季向阳　贾英民　焦李成　李　斌　刘　民
刘庆峰　刘增良　鲁华祥　马华东　苗夺谦　潘　纲
朴松昊　钱　锋　乔俊飞　孙长银　孙茂松　陶建华
王卫宁　王熙照　王　轩　王蕴红　吾守尔·斯拉木
吴晓蓓　杨放春　于　剑　岳　东　张小川　张学工
张　毅　章　毅　周国栋　周鸿祎　周建设　周　杰
祝烈煌　庄越挺

《中国人工智能系列白皮书----心智计算：构建脑与心智启发的人工智能》编写组

曾　毅　史忠植　施路平　蔡恒进　张丽清　曹存根
库逸轩　黄铁军　丁世飞　李清勇　候　彪　赵菲菲
张　倩　赵宇轩　鲁恩萌　赵卓雅　冯　慧　丘铱可

目　录

第 1 章 引言.....4
第 2 章 心智计算研究概述.....6
2.1 心智计算的发展历程.....6
2.2 心智计算的科学问题.....7
2.3 心智计算的哲学展望.....10
2.4 本章小结.....11
第 3 章 心智计算的理论模型.....12
3.1 图灵机.....12
3.2 物理符号系统.....13
3.3 ACT-R.....14
3.4 SOAR.....15
3.5 CAM.....17
3.6 BrainCog.....19
3.7 本章小结.....22
第 4 章 心智计算中的心理揣测.....23
4.1 心理揣测概述.....23
4.2 心理揣测的实验范式.....24
4.2.1 以动物为被试的实验范式.....25
4.2.2 以人为被试的实验范式.....26
4.3 心理揣测的神经基础.....28
4.4 心理揣测的计算模型.....29
4.4.1 基于贝叶斯的心理揣测模型.....29
4.4.2 基于深度学习的心理揣测模型.....32
4.4.3 基于脑启发的心理揣测模型.....32
4.4.4 基于其他方法的心理揣测模型.....34

4.5 本章小结 35
第 5 章 心智计算中的情感共情 36
5.1 情感共情概述 36
5.2 情感共情的实验范式 37
5.3 情感共情的神经机制 37
5.4 情感共情的计算模型 40
5.5 本章小结 43
第 6 章 心智计算中的意识理论 44
6.1 意识理论概述 44
6.2 意识理论的实验范式 45
6.2.1 裂脑人实验 45
6.2.2 遗忘症与情节记忆 45
6.2.3 最小神经关联物 46
6.3 意识理论模型 46
6.3.1 高阶理论 47
6.3.2 全局工作空间理论 47
6.3.3 整合信息理论 48
6.3.4 再入/预测处理理论 50
6.3.5 因果链重构理论 51
6.4 本章小结 52
第 7 章 总结与展望 54
第 8 章 参考文献 56

第 1 章 引言

心智计算（Mind Computation）以多学科交叉的方式融合来自人工智能、认知科学、脑与神经科学、演化生物学、人类学等学科的研究方法与计算范式，对生物智能与心智活动的计算机制机理进行多视角、多尺度系统性的探索，在研究动物与人类心智计算理论与模型的基础上，发展受脑与心智启发的通用人工智能。重点研究生物与人工心智的计算理论体系、心智建模、学习与记忆机制、常识构建与理解、生物与人工意识、社会认知等的科学原理和计算理论与技术。

神经科学、脑科学、认知科学从多视角、多尺度系统地揭示生物脑的结构、功能和机制，理解生物心智活动的自然智能本质；人工智能通过计算建模来模拟、延伸和拓展动物与人类智能的方方面面，达到类生物水平的理解、思考、学习、决策及社会认知能力。尽管人工智能经过近七十年的发展已经在某些领域取得了显著进展，但动物与人类心智涉及到的复杂认知和意识、情感、想象力和创造力等，在当前的人工智能系统中仍难以模拟和重现。为此，心智计算旨在充分实现多学科交叉融合，深度借鉴脑与心智的工作机理，通过计算建模模拟生物心智活动的多尺度结构和功能可塑性，在计算系统中重现动物与人类的心智。

动物与人类的心智是思维和认知能力的总体，包括感知、学习、记忆、决策、推理、情感、心理揣测、意识等。当前对于心智的计算建模更多关注感知、学习、记忆、决策等方面，对与自我认知相关的情感、意识、心理揣测的研究相对较少。为此，本白皮书聚焦于心智计算的理论模型、心理揣测、情感共情以及意识理论展开详细的介绍。

本白皮书首先回顾心智计算的研究历史与发展历程，汇总主要的科学问题，从哲学视角介绍心智计算的愿景。紧接着，详细介绍六种心智计算的理论模型，心智计算的理论体系模型旨在同时集成感知、记忆、决策、运动，以及意识、共情、心理揣测等社会认知能力到一

套通用、系统的框架中。进一步地，本白皮书以心智活动中与自我认知紧密结合的心理揣测、情感共情、意识理论为切入点，深入介绍经典的实验范式、神经机理以及计算理论模型。最后，本白皮书简要的总结与展望心智计算的研究。

第 2 章 心智计算研究概述

2.1 心智计算的发展历程

心智计算的发展历程可以追溯到 20 世纪 60 年代，由最初的对心智的具体问题求解，逐渐演变为系统的心智计算理论形态，包括了以表征—计算为核心的第一代心智计算理论[1]，以及以具身性为理论特征的第二代心智计算理论[2]。

20 世纪早期，Warren McCulloch 和 Walter Pitts 最先提出了神经活动具有计算性的观点，并认为认知可以由计算来解释。20 世纪 60 年代起，经过 Hilary Putnam、Jerry Fodor、David Marr 等人的发展，心智的计算理论（Computational Theory of Mind，CTM）正式被提出并成为第一代心智计算理论的核心，在人工智能和认知科学领域逐渐占据了主流地位。心智的计算理论认为，心智是一个通过大脑神经活动物理实现的计算系统，认知和意识都是一种计算形式。1975 年，Jerry Fodor 提出思维语言假设后，心智计算理论逐渐演变为包含符号计算和连接计算等不同范式。

20 世纪 80 年代之后，在 Hubert Dreyfus 和 John Searle 等人对强人工智能激烈批判刺激下，以具身性（Embodiment）观念为其理论特征的第二代心智理论逐渐登上历史舞台。Shaun Gallagher 曾以包含具身认知（Embodied Cognition）、嵌入认知（Embedded Cognition）、延展认知（Extended Cognition）和生成认知（Enactive Cognition）的“4E 认知”概括了具身心智理论的核心理念[2]。具身心智理论认为身体和心智是相互依存的，动物与人类的智能和认知过程不仅依赖于大脑，也依赖于身体和环境。

无论是以表征—计算为核心的第一代心智计算理论，还是以具身性为理论特征的第二代心智计算理论，都是过去几十年来人工智能领域围绕“心智是如何进行计算”这一问题的不断思考。随着近年来受生物神经机制启发的、以深度学习为代表的连接主义的重要应用突破

和对融合符号主义和连接主义的混合路径的不断探索，传统的心智计算理论有望在这些新突破的基础上焕发新生，为实现结构和机制受脑与心智启发、认知行为达到乃至超越动物与人类水平的智能系统提供更多启示。

2.2 心智计算的科学问题

心智计算作为发展面向人工通用智能的核心思路之一，其研究范围尤其广泛，从感知、学习、记忆、常识知识构建与理解、因果推理到情感、意识、创造、心理揣测、伦理道德等。尽管当前人工智能在某些领域取得了应用体验方面的显著进展，如图像分析、自然语言处理等，但对于重现动物与人类心智仍面临诸多挑战，特别是在复杂的情感、思维、意识等认知方面仍需要持续的探索和努力。从应用意义而言，心智计算研究中最关键的科学问题是如何集成人脑多方面的智慧到一个通用的架构当中，实现真正意义上的通用人工智能。为达到这一目标，包含但不限于如下几个方面的科学问题：

1. 多感觉融合

当代深度学习仍面临着关键挑战，如复杂场景理解、认知、推理能力不足，深度学习易被欺骗，维数灾难等瓶颈问题。理解动物与人脑学习机理，特别是不同感知系统的编解码机制、多感觉信息融合与处理机制等，并将其转化为计算模型，提高感知认知与理解等能力是心智计算中的一大科学问题。

文献[3]探索了类脑张量分解-高维数据结构挖掘问题，采用类脑结构，可以在数据缺失的情况下训练模型，利用正则部分补全缺失数据。文献[4]探索了传统信息瓶颈算法在压缩信息的同时会严重影响模型预测能力，提出了基于有监督解耦的信息瓶颈算法，解决了传统信息瓶颈算法中的压缩-预测项权衡问题。

2. 知识表征与推理

为构建准确、全面的常识知识库，需要大量的来源广泛的常识知

识，包括人类社会层、人类个体层、物体层、以及抽象层等多个层次的信息融合，以确保学习到准确的常识知识。通过对文本的分析，挖掘和抽取出常识规则、常识知识，对知识进行有效的表示和推理，以及在复杂的推理问答任务中学习到常识、逻辑、演绎推理能力亦是研究难点。文献[5,6]等对传统常识知识库的构建方法进行了系统地分析和比较。当前随着大语言模型等新形式的知识推理工具的不断发展，更需要探索如何汲取传统知识表征和推理方法的准确性优势，改进大语言模型自身存在的模型幻觉等问题。

3. 记忆

当前人工智能大模型的训练和推理所需能耗巨大，而人类仅仅需要 20 瓦的能量就可以协同多项认知功能完成复杂的认知任务。工作记忆作为人脑的一项重要认知功能，每时每刻都伴随着人类的学习与决策。而工作记忆仍存在容量限制等问题，即能够同时存储或处理的信息量是有限的。人类可以通过与长期记忆协同等方式来灵活地减轻负荷，高效地调度工作记忆来帮助解决不同任务，类人脑的存算一体机制也是当前人工智能需要向人类学习的。

4. 创造力

人类所具备的创造力仍是现有的人工智能系统所欠缺的，机器大多是基于预先编写好的程序执行指令，尚不能像人类一样进行创造性的思考和行为，不具备创新能力。从本质上理解人类创造性的神经机制，进而让机器自主地探索以习得创造能力仍是一个研究难点。尽管当前生成式人工智能的发展已经让机器展现出前所未有的“创造性”，但这样的创造性仍被认为缺乏人类创造力所源自的对于真实世界的经验、情感和体验，因而尚难于达到人类的创造力水平。

5. 社会认知

在具备了感知、学习、记忆、决策等认知能力的基础上，社会认知是人类及其他动物在社会交互中表现出来的对自我、对他人的理解

和认知能力，在提升社会技能和行为方面起到重要的作用。意识、情感共情、心理揣测等社会认知能力在人工通用智能中也是极其关键的。而在复杂的社会决策环境中探究意识的本质、情感的产生与加工机理、共情及心理揣测的神经基础，赋能人工智能以自我意识等社会认知能力仍是一项亟待攻克的科学问题。目前已有一些工作对社会认知的计算实现进行了初步探索，例如，武汉大学蔡恒进教授提出认知坎陷（意识片段）的“附着”与“隧道”，从新的视角探讨心智的工作模式，探索机器自我意识、情感机制、记忆机制等问题，形成具有“自我”认知的、基于理解的人工智能技术[7]；中国科学院自动化研究所曾毅团队提出了一系列脑启发的心理揣测和共情模型，从计算角度揭示心理揣测的神经机制[8]的同时实现智能体帮助他人避免安全风险[9]、提升与他人的合作性能和效率[10]、以及实现情感共情和利他救援[11]等。

6. 认知功能的自主协同

人类心智能够应对复杂场景、复杂任务的核心机制之一是认知功能的自组织协同。在不同尺度可以将人类认知功能划分为数百种，而这些认知功能并不像工作流一样被预先组织在一起按照既定的模式工作，而是通过多感觉的输入刺激，以自组织的方式自主协同并应用于解决复杂问题。不仅能够举一反三，还能够通过自组织的分解与组合解决没有见过的问题。其背后在心智层面的自组织、自主协同机理是心智计算理论的核心。

7. 软硬件协同构建脑与心智启发的智件

受脑与心智启发的人工智能是实现双脑融合的有效途径，是发展人工通用智能的基石。如何利用多学科交叉融合，实现理论、芯片、软件、系统和应用协同发展的脑与心智启发的通用人工智能体系架构，从面向人工智能的硬件、软件的融合发展为智件（AIware），是亟待突破的关键问题。清华大学施路平团队提出了异构融合的类脑计算架构—“天机”类脑计算芯片[12]，能够模拟大脑中神经元之间信号传递

的方式，融合人工神经网络及脉冲神经网络两条技术路线，在多学科交叉融合的软硬件协同设计上进行了前沿探索。中科院自动化所曾毅团队研制的软硬件协同类脑脉冲神经网络体系结构智脉·萤火(BrainCog FireFly)系列研究融合了类脑认知智能引擎“智脉 BrainCog”的认知体系结构，并以 FPGA 为平台实现软硬件协同创新，打造脑与心智启发的智件体系。然而智件体系不仅是脑与心智启发的软件模型与计算体系结构的协同设计，还包括具身本体与软件及计算体系结构的三元融合，是具身心智理论的体现、实践与发展。

2.3 心智计算的哲学展望

心智计算的发展需要从哲学视角树立愿景。将心智收敛至 Mind，Shimon Edelman 在《Computing the Mind》中提出“I am my mind, I live in my brain”，指出了心智与脑及自我之间的关系。而严格意义上讲英文中的 Mind 并不能囊括中文表达的心智，德文中的 Gemüt 比 Mind 要更达意，如莱布尼茨将 Gemüt 解读为“思想的能力、感觉和意志活动的统一”，如 Justus Georg Schottelius 和 August Friedrich Müller 将 Gemüt 定义为：“知性和意志的合体”。这似乎与中文的“心灵”又更为接近，而从计算视角而言，“心智”则比“心灵”更为广泛。王阳明心学的讨论可以被认为是心灵与心智的哲学指引。即使以侠义的视角而言，心智计算的目的也应当是为人与动物的心智构建计算理论基础，启发人工心智（Artificial Mind）的研究与应用。

心智与心灵的核心从“自我”出发。从演化和计算的视角而言，有了自我感受与体验，能够区分自我和他人，为心理揣测/认知共情提供了基础，在此基础上情感共情和利他才成为可能，才具备了产生道德直觉的前提。人类与人工智能据此拥有道德，合乎伦理。

当代人工智能无“我”无“心”，心智计算研究的目的是在揭示人与动物心智的计算本质基础上，为未来人工智能“立心”。目前数据驱动的人工智能，模型算法在数据输入之前可谓“无善无恶”，是

本心之体。通过数据编码与训练之后可谓“有善有恶”为意之动。通过与人交互达“知善知恶”。再通过价值观校准实现“为善去恶”。从“无”到“为”是心智形成、升华的过程，是“知行合一”的实践。

2.4 本章小结

心智是人类思维和认知能力的总体，包括但不限于感知、学习、记忆、决策、推理、思维、情感、心理揣测、意识、伦理道德等，这些能力帮助个体感知外在世界、学习和记忆事物、对环境做出判断来采取不同的行为。心智计算就是希望通过计算建模的方式去理解、模拟、实现人类和动物的心智。本章首先回顾了心智计算的历史和发展历程，紧接着总结了心智计算七个视角的科学问题及哲学视角的愿景。心智计算的学术贡献首先是自然与人工心智的计算理论体系，在此基础上，构建并实现脑与心智启发的通用人工智能，打造从软件、硬件协同创新到智件的跃迁。

第 3 章 心智计算的理论模型

心智计算的理论模型能够协调和集成感知、学习、推理、规划、决策、意识等多项类脑认知功能，旨在理解心智的工作机理，并在计算建模中重现心智是怎样工作的[13]。文献[14]讨论了心智建模的标准，即灵活的行为、实时性、自适应的行为、大规模的知识库、动态行为、知识集成、自然语言、意识、学习、发育、演化及生物脑认知的神经实现。本章分六节介绍心智计算的理论模型并进行总结。

3.1 图灵机

艾伦·图灵（Alan Turing）于 1936 年提出图灵机的概念[15]，图灵机是一种无限记忆自动机，如图 3-1 所示。它由一条无限长的纸带、一个读写头、一个状态寄存器和一套控制规则组成。纸带上的格子可以记录“0”或“1”。在带子上方移动一个读写磁头，它是由有限记忆自动机 L 来控制的。自动机 L 按周期工作，关于符号(0 或 1)的信息，由磁头从带子上读出，而反馈给 L 的输入。磁头根据在每个周期中从自动机 L 得到的指令而工作，它可以停留不动或向左、向右移动一小格。与此同时，磁头从自动机 L 接收指令，执行收到的指令，它就可以更换记录在磁头下面方格中的符号。

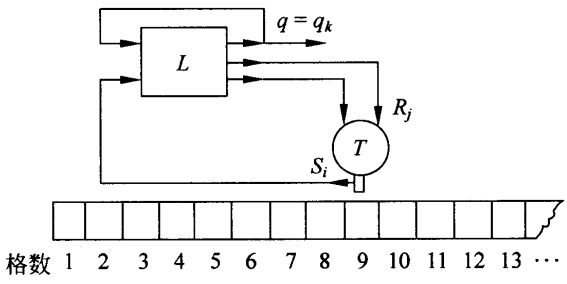

图 3-1 图灵机

图灵机的工作唯一地决定于带子方格的初始存储和控制自动机的变换算子，这个算子可以表示为转移表的形式。我们用S_i (S_0＝0, S_1＝1)表示磁头读出的符号；用R_j（R_0=停止，R_1=左移，R_2=右移）表示

移动磁头的指令；用q_k（k＝1，2，...，n）表示控制自动机的状态，则表 3-1 给出了图灵机状态转移表。

表 3-1 图灵机状态转移表

输　入	状　态	
	$S_0 = 0$	$S_1 = 1$
q_1	S_0, R_2, q_k	S_1, R_1, q_m
q_2	S_1, R_0, q_s	S_0, R_2, q_1
q_3	S_1, R_1, q_p	S_0, R_2, q_2

从表 2-1 中看出，自动机 L 的动作依赖于输入 q 和它的状态 S。对于给定值 q 和 S，将有 q，R，S，这三个量的某一组值与之对应。这三个量分别指明，磁头应在磁带上记录什么符号 q，移动磁头的指令 R 是什么，自动机 L 将变到什么新状态 S。在自动机 L 的状态 S 中至少应当有这样一个状态 S* ，对于这个状态来说，磁头不改变符号 q，指令 R=R_0（停止），而自动机 L 仍处于停止位置 S*。

图灵机看似简单的结构，却可以在理论上模拟数字计算机的一切运算，成为了计算机信息加工的理论基础。1950 年，图灵设计了图灵测验，通过问答来测试计算机是否具有同人类相当的智力。

3.2 物理符号系统

20 世纪 70 年代，Allen Newell 和 Herbert A. Simon 提出了物理符号系统假设（physical symbol system hypothesis，PSSH）[16]，他们认为物理符号系统具有充分且必要的条件进行通用智能行为。由此假设得到了三个推论：人具有智能，因此人脑一定是一个物理符号系统；计算机是一个物理符号系统，它就一定能够表现出智能；既然人脑和计算机都是物理符号系统，那么我们就可以用计算机来模拟人的心智活动[16]。人脑和计算机一样都是物理符号系统，因此都可以简化为具有 6 种功能：输入符号、输出符号、存储符号、复制符号、建立符号

结构以及条件性迁移[16]。

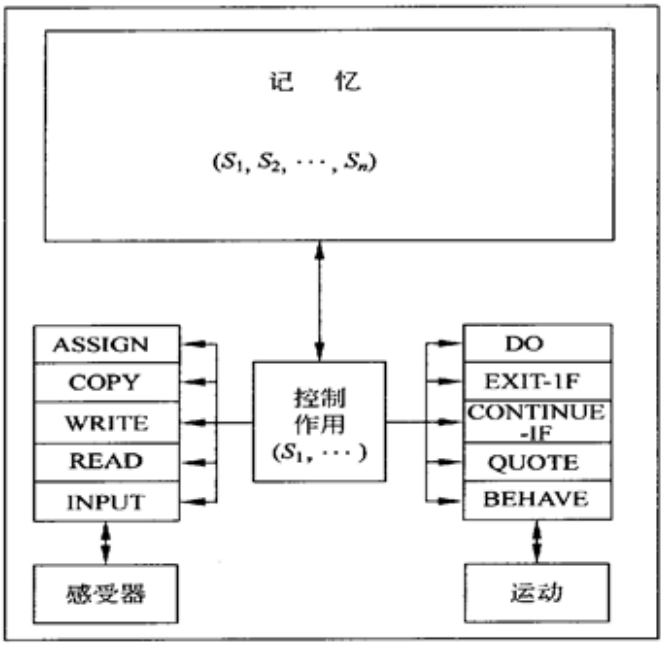

图 3-2 物理符号系统[17]

图 3-2 给出了物理符号系统的一种框架[17],它由记忆、一组操作、控制、输入和输出构成。它通过感受器接受输入，输出是特定的行为带来的外部运动。外部行为的输出和执行也会影响后面接受到的感觉输入。物理符号系统中的记忆和控制交互协同,得到不同的内部状态。基于符号结构组成的记忆不断地更新、组合、表达,发挥不同的作用,进而根据输入来产生一系列的活动。

3.3 ACT-R

John R. Anderson 融合人类联想记忆模型与产生式系统结构，提出思维的自适应控制 ACT （Adaptive Control of Thought）模型[18]。ACT 的系统结构由工作记忆、产生式记忆和陈述性记忆组成(见图 3-3)[19]。工作记忆将当前编码的外部世界知识存储至陈述性记忆中，在其中以组块为单元建立起语义网络，并根据需求提取至工作记忆。在与外部世界的交互中，陈述性记忆不断地被提取来解决问题，并试图通过组合弱方法来产生许多子目标及对应的陈述性知识。逐渐地，在应用过程中新的产生式规则就会生成，并转化陈述性知识为程序性的知识，这一过程也叫程序化。产生式规则通过与工作记忆的匹配来执

行相为匹配的行动。

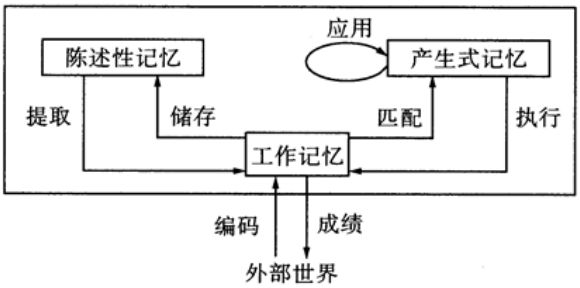

图 3-3 ACT 的系统结构[19]

在 ACT 基础上发展的 ACT-R[19]采用产生式规则系统来实现类人的认知功能，并集成有感知、记忆、语言、决策等多个认知模块，突出了对多个脑区功能细节、层次化结构的借鉴，以达到更准确地模拟人类的认知过程。

3.4 SOAR

1987 年, Allen Newell 和 John Laird、Paul Rosenbloom 提出了一个通用解题结构 SOAR[20]：即状态 State，算子 Operator 和结果 Result, 表示弱方法的基本原理是不断地将算子作用于状态，以得到新的结果。如图 3-4 所示，产生式记忆器和决策过程形成处理结构。产生式记忆器中存放产生式规则,它进行记忆搜索及控制决策：首先，所有规则被并行地用于工作记忆器，判断优先权，决定哪部分语境进行改变以及如何改变；进一步地，决策阶段决定语境栈中要改变的部分和对象。

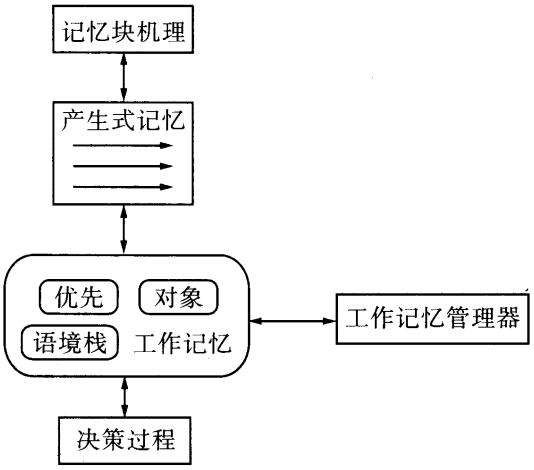

图 3-4 SOAR 的框图[20]

SOAR 中的所有成分统称为对象，这些成分包括状态、状态空间、

算子和目标。在 SOAR 问题求解过程中，大体上是一个分析-决策-行动的三部曲。

（1） 分析阶段

输入：库中的对象；

任务：从库中选出对象加入当前环境；

增加有关当前环境中对象的信息角色；

控制: 反复执行，直至完成。

（2） 决策阶段

输入：库中的对象；

任务：赞成，或反对，或否决库中的对象。选择一个新的对象，用它取代当前环境中的同类对象。

控制：赞成和反对同时进行。

（3） 执行阶段

输入：当前状态和当前算子；

任务：把当前算子应用于当前状态。

如果因此而产生一个新状态，则把新状态加入库中，并用它取代原来的状态。

控制：这是一个基本动作，不可再分。

SOAR 系统运行过程中，在分析阶段，任务是尽量扩大有关当前对象的知识，以便在决策阶段使用。决策阶段主要是进行投票，投票由规则来做，它可以看成是同时进行的，各投票者之间不传递信息，不互相影响。在执行阶段，如果当前环境的每个部分都有定义，则用当前算子作用于当前状态。若作用成功，则用新状态代替旧状态，算子部分成为无定义，重新执行分析阶段。

每当问题求解器不能顺利求解时，系统就进入劝告问题空间请求专家指导。专家以两种方式给以指导。一种是直接指令方式，这时系统展开所有的算子以及当时的状态。由专家根据情况指定一个算子。

指定的算子要经过评估，即由系统建立一个子目标，用专家指定的算子求解。如果有解，则评估确认该算子是可行的，系统便接受该指令，并返回去求证用此算子求解的过程为何是正确的。总结求证过程，从而学到使用专家劝告的一般条件，即组块。

另一种是间接的简单直观形式，这时系统先把原问题按语法分解成树结构的内部表示，并附上初始状态，然后请求专家劝告。专家通过外部指令给出一个直观的简单问题，它应该与原问题近似，系统建立一个子目标来求解这个简单问题。求解完后就得到算子序列，学习机制通过每个子目标求解过程学到组块。用组块直接求解原问题，不再需要请求指导。

SOAR 系统中的组块学习机制是学习的关键。它使用工作记忆单元来收集条件并构造组块。当系统为评估专家的劝告，或为求解简单问题而建立一个子目标时，首先将当时的状态存入工作记忆单元。当子目标得到解以后，系统从工作记忆单元中取出子目标的初始状态，删去与算子或求解简单问题所得出的解算子作为结论动作。由此生成产生式规则，这就是组块。如果子目标与原问题的子目标充分类似，组块就会被直接应用到原问题上，学习策略就把在一个问题上学到的经验用到另一个问题上。

3.5 CAM

人的心智中记忆和意识是最为重要的两个部分。其中记忆存储各种重要的信息和知识，意识让人有了自我的概念，能根据自我需求、偏好设定目标，并根据记忆中的信息进行各项认知活动。为此史忠植等人主要基于记忆和意识创建了 CAM（Consciousness And Memory）心智模型[21]。下面重点介绍 CAM 的系统结构和认知周期。

心智模型 CAM 的系统结构如图 3-5 所示，包括 10 个主要功能模块：视觉、听觉、感知缓存、工作记忆、短时记忆、长时记忆、高级认知功能、动作选择及响应输出[21]。人的感觉器官包括视觉、听觉、

触觉、嗅觉、味觉。CAM 模型中重点考虑视觉和听觉的感觉输入。感知缓存是最直接、最原始的记忆，只能保存感觉信息在很短的时间，约几十到几百毫秒。工作记忆由中枢执行系统、视觉空间画板、语音回路和情景缓存构成。短时记忆存储信念、目标和意图等内容。长时记忆包括语义记忆、情景记忆、程序性记忆等。在心智模型 CAM 中，意识关注系统的觉知、全局工作空间理论、动机、元认知、注意、内省学习等自动控制问题[21]。除意识外，CAM 模型还实现脑的学习、记忆、语言、思维、决策、情感等高级认知功能，并针对特定任务做出动作选择及响应输出。

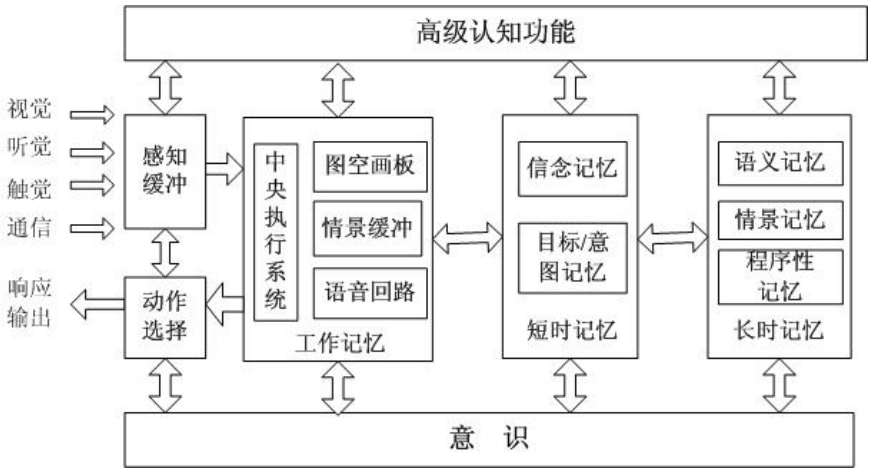

图 3-5 CAM 的系统结构[21]

认知周期是认知水平心理活动的基本步骤。人类的认知是由反复出现的脑事件的级联周期。在心智模型 CAM 中，每个认知周期感知当前的境况，通过动机阶段参照需要达到的目标，然后构成内部或外部的动作流，响应到达的目标[22]。CAM 认知周期分为感知、动机、动作规划三个阶段。感知阶段是通过感觉输入，实现对环境的觉知过程。使用传入的知觉和工作记忆的信息作为线索，本地联想，自动地检索情景记忆和陈述性记忆。动机阶段侧重于学习者的信念、期望、排序和理解的需要。根据动机的影响因素，如激活比例、机会、动作的连续性、持续性、中断和优惠组合，构建动机系统。动作规划将通过动作选择、规划以达到最终目标。

1. 感知阶段

感知阶段认识或理解环境，组织和解释感觉信息的处理。感官接收到的外部或内部的刺激，是感知阶段产生意义的开端。觉知是事件感觉、感知、意识的状态或能力。在生物心理学中，觉知被定义为人类或者动物对外界条件或者事件的感知和认知反应。

2. 动机阶段

在心智模型 CAM 的动机阶段，根据需要确定显式目标。一个目标列表中包含多个子目标，可以形式地描述为：

$$G_t = \{G_1^t, G_2^t, \dots, G_n^t\} \qquad at\ \ time\ \ t$$

在心智模型 CAM 中，动机系统通过短时记忆系统完成。信念记忆存储智能体当前的信念，包含了动机知识。愿望是目标或者说是期望的最终状态。意图是智能体选择的需要现在执行的目标。目标/意图记忆模块存储当前的目标和意图信息。在 CAM 中，目标是由子目标组成的有向无环图，执行时分步处理。一个个子目标按照有向无环图所表示的路径完成，当所有的子目标都完成之后，总目标完成。

3. 动作规划阶段

动作规划是由原子操作构建复杂动作以实现特定任务的过程。动作规划可以分为两个步骤：首先是动作选择，即从动作库选择相关的动作；然后使用规划策略使被选的操作组装一起。动作选择是实例化动作流，或可能从以前的动作流中选择一个动作。有很多的选择方法，它们中的大多数基于相似性的标准匹配目标和行为。规划对动作组合提供了一个可扩展的和有效的方法。它允许一个动作组合请求被表示为目标的条件，规定一组约束和偏好。

3.6 BrainCog

类脑认知智能引擎(Brain-inspired Cognitive Intelligence Engine，BrainCog) “智脉”[23]是一个基于全脉冲神经网络（Spiking Neural Network, SNN）的类脑人工智能与脑模拟计算平台，用于在多个尺度

上建模、模拟不同物种的认知大脑，并受此启发实现类脑与心智的人工智能。智脉以多尺度神经可塑性为基础，同时支持脑启发的人工智能及脑多尺度的结构功能模拟，为受脑与心智启发的人工智能、计算神经科学等多个学科提出一套通用、完备的、系统的基本组件。

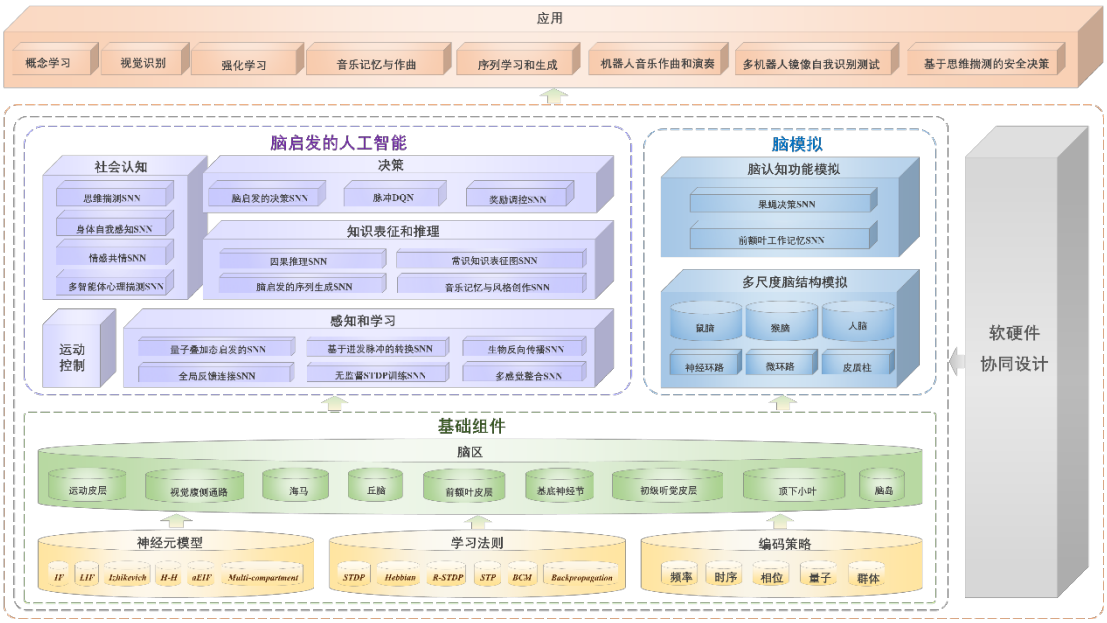

图 3-6 类脑认知智能引擎“智脉”的基本组件与应用[23]

如图 3-6 所示，智脉的基础组件包括丰富的生物神经元模型、多种类脑突触可塑性法则、不同脉冲编码方式、脉冲神经网络的连接模式以及多个功能性脑区模型。基于以上基本组件，智脉提供五类认知功能组件：感知与学习、知识表征与推理、决策、运动控制、社会认知。目前发布约 40 个脑启发的人工智能计算模型映射到 28 个关键功能性脑区。智脉还支持软硬件协同设计，以及机器人为载体的类脑认知智能应用。

1. 感知与学习

智脉支持多种有监督和无监督的脉冲神经网络学习算法，包括短时突触可塑性、脉冲时序依赖突触可塑性，基于代理梯度的反向传播算法，和基于 ANN（Artificial Neural Network）到 SNN 的转换算法，在图像识别、分类、检测任务上得到充分的验证，并展现出小样本学习、抗噪性等能力。智脉还实现了类人概念学习的多感觉融合框架，

以及量子启发的脉冲神经网络，结合多房室神经元在噪声环境下取得稳健的性能。

2. 决策

智脉提供了多脑区协同的决策脉冲神经网络以及深度强化学习脉冲神经网络。前者在 Flappy bird 游戏上实现了类人的学习能力，并具备支持无人机在线决策的能力，能够实现类果蝇的线性和非线性决策以及反转学习。后者实现了深度脉冲神经网路和强化学习的结合，在 Atari 游戏上的得分超过传统深度强化学习模型。

3. 运动控制

智脉借鉴人脑运动控制的神经机理，构建了多脑区协同的机器人运动控制脉冲神经网络，实现了人形机器人的钢琴弹奏。

4. 知识表征与推理

智脉集成了符号序列记忆与生成、常识知识表征、因果推理脉冲神经网络，实现了初步的概念知识生成及推理认知。类脑的音乐记忆与乐曲创作脉冲神经网络实现了对音符序列的表征与记忆，并能创作不同风格的乐曲。

5. 社会认知

智脉涵盖的类脑社会认知脉冲神经网络模型赋予智能体以理解自我和他人的能力，实现机器人通过镜像测试、橡皮手错觉、错误信念实验，使得智能体能够帮助他人规避安全风险。

人类的心智活动是集成有多项认知功能的复杂的思考和行动过程。为此，智脉还提供了一个多认知功能协同的人形机器人应用，即情感驱动的机器人乐曲创作与演奏。在该任务中，机器人需要调用智脉的感知与学习功能来识别图片中的情感，并调用知识表示与推理功能依据情感生成乐曲，最终由运动控制模块实现机器人的乐曲演奏。

在脑模拟方面，智脉支持不同尺度的脑结构与认知功能模拟，其中脑功能模拟实现了前额叶皮层工作记忆及果蝇线性、非线性决策功

能的模拟。脑结构模拟实现了对鼠脑、猴脑、人脑的微环路、皮质柱、全脑等多尺度的全脉冲神经网络计算建模。

3.7 本章小结

本章介绍了六个心智计算的理论模型，包括图灵机，物理符号系统假设，ACT-R，SOAR，CAM 和 BrainCog。这些代表性的心智计算理论体系探讨了感知、记忆、决策、运动等认知功能，以及意识、共情、心理揣测等社会认知相关的能力。对心智计算的体系架构建模旨在受脑与心智的神经机制启发，同时集成人类心智的方方面面到一套通用、系统的框架中，朝向通用人工智能的方向迈进。

第 4 章 心智计算中的心理揣测

4.1 心理揣测概述

通过对比人类和其他动物的大脑新皮质的面积可以发现，人类的新皮质最为发达，同时相比于其他动物，人类也形成了最稳定的社会群体。社会脑假说认为人类进化出比其他物种更复杂的大脑功能是为了处理更为复杂的社会关系，做出适应性的社会行为。进化心理学的观点认为基于社会规范、互惠形成的大规模合作可能是人类在自然界中占据主导地位的原因之一。在探索复杂社会关系与社会决策的过程中，心理揣测（Theory of Mind, ToM）这一社会认知功能对人类的社会认知能力的形成有着巨大的作用。

心理揣测又被翻译为心理理论，最早由大卫普瑞马克（David Premark）创造。它可以被简单地理解为智能体在观察到他人行为之后，可以揣测他人产生行为的原因（归因）。这些原因代表着他人的心理状态（mental state），心理状态的内容一般涵盖感觉（如“疼痛”）、情绪（如“愤怒”、“难过”）、信念（如“吃冰激凌可以降体温”）、需要/愿望（如“想要冰激凌”）或者目标（如“去超市买冰激凌”）等。与心理状态紧密关联的内容比较多，包括一般性的世界知识、价值取向、优先级策略等。这些内容在对他人进行心理揣测时也起到重要作用。借助这种心理状态，智能体可以区别于自己的内心想法来预测他人的行为。

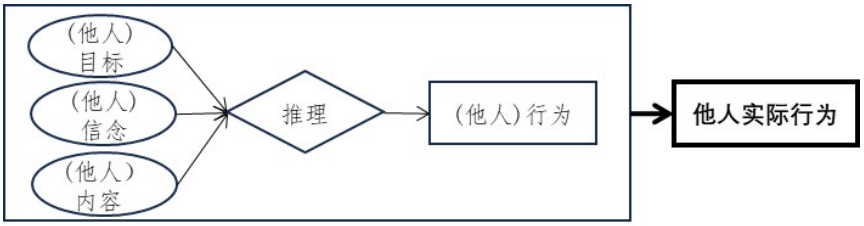

图 4-1 心理揣测示意图

在对他人的心理揣测过程中，模拟者需要把自己的目标、信念以及其他心智内容归因到他人脑中，如图 4-1 中方框内容所示，然后根

据该模拟，预测或理解他人的实际行为。

如果我们可以预测他人的行为，我们就可以预见并避开麻烦，或者利用所预见的机遇。在预测复杂行为时，与仅仅将某个特定的身体运动与特定结果相关联的方法相比，根据他人的心智状态，例如意图或知觉，解释他们的行为更能增强预测的有效性。其中一个原因是：相同的行为可能有完全不同的意图。例如，竖起大拇指在中国是好的意思，而在泰国是走开的意思。除此之外，心理揣测更吸引人的地方在于，让智能体的交互更加独立且灵活。举个例子来说，两个人一起捡苹果，如果我距离两个苹果一样近，而我的朋友距离其中一个更近，我会推测他想要近一点的苹果，所以我会打算要距离朋友远的苹果；而我的朋友推测我会把距离他近的苹果让给他，因此他会去捡距离他近的苹果。由此，心理揣测会衍生出一些有助于社会决策的行为。而事实上，神经科学的研究也确实表明心理揣测对社会决策有影响，社会决策任务也会激活与心理揣测相关的区域（例如，颞上沟、颞顶交界处和内侧前额叶皮层），这与处理自己和其他玩家的行动和意图一致。因此构建一个类脑的心理揣测模型可以提高对他人行为预测的准确度，同时可以在现有的智能决策方法基础上辅助社会决策。在人工智能领域，目前现有的心理揣测建模工作普遍从概念中抽取关键词用深度学习方法建模；也有从贝叶斯的角度建模简单任务的心理揣测过程。除此之外，随着认知神经科学大量关于心理揣测在婴幼儿及儿童时期的研究文章的涌现，借鉴认知神经科学的研究结果启发构建心理揣测网络也在蓬勃发展。

4.2 心理揣测的实验范式

心理揣测发展过程中的里程碑之一是获得了错误信念归因的能力，也就是说，认识到其他人可能对世界有不同的信念[24]。基于此，研究人员设计了很多实验范式来研究动物和人类的心理揣测能力，此处介绍几种比较有代表性的实验范式[25]。

4.2.1 以动物为被试的实验范式

知情者-猜测者实验[26]：被试动物和两个人在一个房间中。一个人作为“猜测者”，首先离开房间；另一个人作为“知情者”，需要从房间中的四个盒子里选择一个并把食物放到其中。所有盒子被挡板挡住，被试动物能看到哪个人放了食物，但不知道放到哪个盒子里。等“猜测者”返回房间，挡板被移走，两个人指向盒子。“知情者”指向装有食物的盒子，“猜测者”则随机指向另外三个盒子。被试动物需要通过上述线索，选择其中一个盒子进行检查来寻找食物。

竞争性喂食实验[27]：社会层级较低的下级动物和社会层级较高的首领动物在测试场地的两侧，场地中有两块挡板 A 和 B。所有测试中，实验人员进入场地并把食物放到挡板 A 的下级动物侧（即下级动物能看到食物，首领动物看不到），在一些测试中，实验人员几秒后返回场地并把食物移到挡板 B 的下级动物侧。下级动物的笼子在实验人员放食物时开着。控制条件是在放食物过程中，首领动物的笼子是打开的还是关闭的，因此下级动物可以看到或看不到首领动物。放食物结束后，两个动物都被放入测试场地，其中下级动物比首领动物早几秒释放。若下级动物能通过心理揣测测试，则在如下三种条件下下级动物会更倾向于去拿食物。（1）单次放食物时，首领动物的笼子是关闭的；（2）第一次放食物时首领动物的笼子是打开的，但在转移食物时，首领动物的笼子是关闭的；（3）单次放食物且首领动物笼子打开，当下级动物在实验最后阶段认为首领动物没有看到食物，下级动物更可能去拿食物。

眼镜实验[28]：黑猩猩亲身经历戴两种眼镜的经验，一个眼镜是透明的，另一个是不透明的，两个眼镜的颜色和形状不同。测试实验中，黑猩猩向两个人乞讨食物，一个人戴透明的眼镜，一个人戴不透明的眼镜。如果黑猩猩具备心理揣测的能力，它将更频繁的向那个戴透明眼镜的人乞食。

错误信念实验[29]：基于意外地点转移任务修改，测试的基础是追踪动物的凝视点。有物体位于位置 A，当第二个人离开时，第一个人把物体藏到位置 B，然后第二个人回来后寻找物体。若动物首先并较长时间注视位置 A，则说明动物认为第二个人仍然认为物体在位置 A。

4.2.2 以人为被试的实验范式

Sally-Anne 测试[30]：Sally-Anne 测试（如图 4-2 所示）是一个经典的意外地点转移任务，是认知心理学家验证被试是否具备心理揣测能力的经典实验。给被试描述如下场景：Sally 和 Anne 在一个房间中，Sally 有一个篮子，Anne 有一个盒子，Sally 把球藏到篮子里，然后离开房间，Anne 把球藏到盒子里，等 Sally 返回房间后，询问儿童被试“Sally 会去哪里寻找球？”。只有四岁以上的儿童被试能够正确回答“Sally 会去篮子里寻找球”，四岁以下的儿童被试则回答“Sally 会去盒子里寻找球”，即只有四岁以上的儿童被试可以正确推断他人信念，并理解他人具有错误信念。

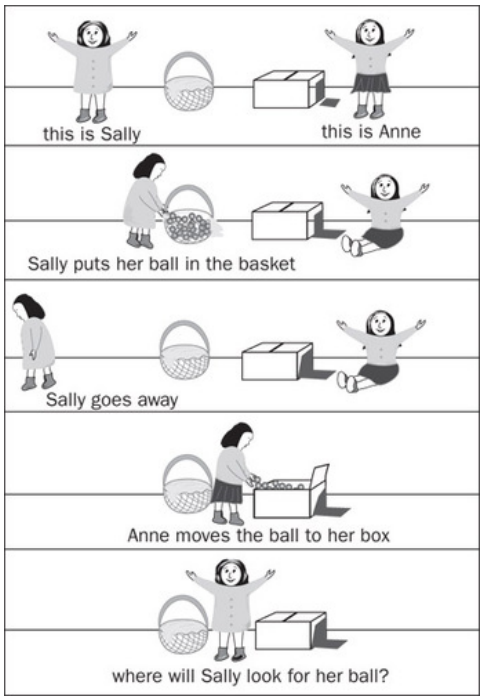

图 4-2 Sally-Anne 测试[30]

预期注视实验：预期注视实验为婴儿提供一个表演者即将会做出的某一种行为的提示，代替行为本身，然后记录婴儿注视的位置，通

过婴儿的注视位置是否同表演者信念的位置一致来检测婴儿是否具有他人信念推理的能力。Senju 等人[31]调查了 18 个月大的婴儿是否会利用自己过去的视觉经验，将感知和随后的信念归因于他人。婴儿被分成两组，一组戴着不透明的眼罩，另一组戴着看起来不透明但实际上是透明的眼罩。不透明眼罩和透明眼罩看起来一模一样。对戴眼罩的婴儿展示不同的玩具和图片，并询问物体的位置。在测试过程的熟悉阶段，木偶将玩具放在或藏在左侧或右侧的盒子里，表演者去寻找玩具，使婴儿知道表演者寻找玩具的意图。在测试过程中，木偶将玩具藏到一侧的盒子里，在表演者带上眼罩后，木偶把物体从场景中移开。表演者将眼罩摘下，表明即将寻找玩具，此时记录婴儿的注视位置。研究结果显示，不透明眼罩组的婴儿会看向物体最后一次出现的位置，表明理解了表演者的错误信念，而透明眼罩组没有表现出对特定位置的偏好。

期望冲突实验：期望冲突实验通过引入令婴儿感到意外或不符合他们预期的情境来测试他们对于世界的认知和理解。这种任务通常涉及到在一个特定情境中引入不寻常的事件，观察婴儿是否会表现出惊奇的反应，如注视时间更长。Soughtgate 等人[32]设计了一种新的期望冲突实验，旨在检验 2 岁婴儿是否能理解错误信念。实验分为两个熟悉试次和一个错误信念试验，其中熟悉试次 1 和 2 交替进行。在熟悉试次中，木偶交替地将球放置在左边或右边的盒子中，随后消失，伴随两个小窗户同时亮起和“滴”的声音刺激。随后，表演者的手从相应的小窗中伸出，指向球所在的盒子。这一过程会在每次试次中反复出现，形成条件反射，提示被试表演者将寻找球。在错误信念试验中，木偶首先将球放入左侧盒子，然后移动到右侧盒子，接着表演者离开场景。在此时，木偶将球从右侧盒子移回左侧盒子。然后，表演者重新进入场景，伴随着“滴”的声音，两侧小窗户同时亮起，提示被试反应。使用眼动仪记录数据，婴儿会首先注视右边的盒子；若表演者

从左边的盒子拿球，婴儿的注视时间更长。

偏好注视实验[33]：偏好注视实验旨在研究儿童在听到故事后，是否会观看与其所听内容相匹配的图片。实验过程中，婴儿在听故事的同时会呈现两张图片：一张与故事内容相符，另一张则不符。随后，在故事末尾，会对表演者的行为目标进行叙述。研究人员将观察婴儿是否更倾向于注视与表演者信念相符的图片。

预期指示实验：预期指示实验探究了婴儿是否能根据表演者在行动前是否需要帮助，来做出与表演者信念相符的行为。婴儿在认为表演者需要帮助时，会通过手势提示或告知表演者，以防止不符合表演者预期的结果发生。如 Knudsen 和 Liszkowski[34]的研究表明，18 个月的婴儿在表演者持有错误信念且期望找到物体时，会积极地使用手指指向包含物体的容器，以提示物体的当前位置。而在表演者知晓物体位置或不打算找物体的情况下，婴儿很少采取这种指示行为。

意外内容实验[35]：在意外内容实验中，实验者给被试呈现外观同内容物不相符合的物品，判断儿童是否具备心理揣测的能力。实验中，实验者向被试展示一个外表看起来像糖果盒的容器，通常人们会认为里面装有糖果。然后，实验者询问被试："容器里面是什么？"被试回答"糖果"。接着，实验者打开容器，发现里面装的是铅笔而不是糖果，然后问被试："其他儿童在打开容器前，认为容器里面是什么？"结果显示，4 岁以下的儿童回答"铅笔"，而 4 岁以上的儿童回答"糖果"。

4.3 心理揣测的神经基础

心理揣测的神经基础，即所涉及的脑区、脑区环路及神经机制尚未研究清楚，但一些脑区在心理揣测实验中发现被重复激活。在健康被试的各种心理揣测任务中，腹内侧前额叶(ventromedial prefrontal cortex, vmPFC)、双侧颞顶联合区(Temporo-parietal Junction, TPJ)和楔前叶(Precuneus)等脑区一直发现被激活。

Schurz 等人[36]元分析了来自 73 个心理揣测影像学研究的 757 个激活点，涉及 1241 名参与者，他们的元分析包含六个不同的任务组：错误信念与照片、特质判断、策略游戏、社交动画、眼睛中的思维和理性行动。他们发现在所有任务组中 mPFC 和双侧后 TPJ 都被激活。在错误信念与照片故事任务组中，他们发现 TPJ、顶下小叶（Inferior Parietal Lobule，IPL）、楔前叶、后扣带回、mPFC 连接簇 3 和 4、mPFC 腹侧部分、前扣带回（Anterior Cingulate Cortex，ACC）、右前颞叶和岛叶相邻部分被激活。

Molenberghs 等人[37]对 144 个数据集（涉及 3150 名参与者）进行了一系列的激活似然估计（ALE）荟萃分析，以探讨与特定类型的 ToM 任务有关的大脑区域。就共性而言，在内侧前额叶皮质和双侧颞顶交界处发现了一致的激活。

Schurz 和 Perner[38]回顾了 9 个关于心理揣测在大脑中如何实现的最新神经认知理论，并根据最近由 RN111 进行的 meta 分析的结果对它们进行了评估。根据认知过程与大脑某些区域相关的理论，他们推断出不同类型的 ToM 任务应该涉及哪些区域。这些脑区包括 mPFC、后颞上沟 (posterior Superior Temporal Sulcus, pSTS)、TPJ 和 IPL。

4.4 心理揣测的计算模型

在心理揣测计算建模方面，大致可以分为基于贝叶斯的心理揣测模型、基于深度学习的心理揣测模型、包含连接主义建模及认知架构设计等其他方法的心理揣测模型、以及基于脑启发的心理揣测模型。

4.4.1 基于贝叶斯的心理揣测模型

基于贝叶斯的心理揣测模型是最具代表性的一类心理揣测模型，麻省理工学院的 Goodman 等人[39]建立了两个贝叶斯模型，这两个模型都支持预测和解释。简单模型中 Sally 的信念只与玩具的位置有关，而复杂模型中 Sally 的信念不仅与玩具位置有关还与她对玩具的视觉感知相关，即 Sally 是否能够看到玩具移动。这一区别使得简单模型

无法通过错误信念任务，而复杂模型会成功。Bayesian 推理通常是通过逆强化学习（Inverse Reinforcement Learning，IRL）进行的。正如 Jara-Ettinger[40]所描述的那样，“通过模拟具有假设信念和欲望的 RL 模型来预测其他人的行动，而通过反演该模型来实现心理状态推断”。

类似于这种想法，麻省理工学院的 Baker 等人[41]提出了一个贝叶斯心理揣测模型（Bayesian Theory of Mind，BToM），这个工作将 belief 建模为智能体在一时刻为某一状态的概率，以此为基础构建的动态贝叶斯网络（dynamic Bayes net，DBN）可以预测环境中智能体的目标。Baker 的工作将心理揣测中抽象的名词，例如信念（belief）、想法（desire）进行符号化，使得模型的可解释性更强。

除此之外，Baker 将 IRL 的思想与部分可观察马尔可夫决策过程（POMDPs）结合用于建模心理揣测模型，被揣测对象在环境中的行动是可观察的，并以此为后验来对被揣测对象的信念和目标进行逆向推断。该模型可以根据智能体在空间中的移动方式，来推断它的信念、期望和知觉。在两个心理学实验中，该模型获得了和人类被试相似的实验结果。实验结果表明，贝叶斯心理揣测模型可以根据他人的行为揣测他人的信念、期望和知觉，以及用他人的想法和行为揣测环境的状态。在这种基于概率（贝叶斯）方法建模心理揣测的过程中，逐渐衍生对心理揣测进行递归建模的思路[42-44]。

以两个人的场景为例，递归推理可以通过下图描述：拥有零阶心理揣测（图 4-3）能力的智能体可以根据对另一个人的行为观测生成一个概率分布以作为它对另一个智能体信念推断的依据。拥有一阶心理揣测（图 4-4）能力的智能体同时具有推测他人零阶信念和一阶信念的能力。一阶信念是指智能体认为另一个人如何推断自己的概率分布。然后拥有一阶心理揣测的智能体会将其一阶预测与零阶信念进行集成，并将该集成信念用于最终决策。预测对代理行为的影响程度由它的一阶置信度决定，如果预测正确则增加置信度，反之则降低。这

样的贝叶斯心理揣测模型普遍需要很高的计算成本来形成和维持信念。因此，模型通常针对特定的场景进行优化，比如石头剪刀布、或者一些假设性较强的特定任务中。由于心理揣测是智能体思考的过程，因此心理揣测的实验会伴随着智能体的决策，而要单一研究心理揣测模型就需要保证智能体的决策是完全正确的。以上的这些研究确实在心理揣测建模方面取得了一些进展，但是对于解决复杂的或者实际应用问题仍过于理想化。

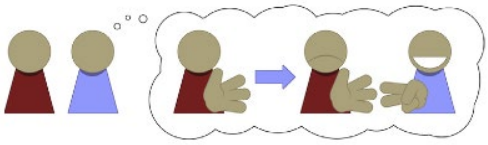

图 4-3 一个零阶心理揣测的例子[44]

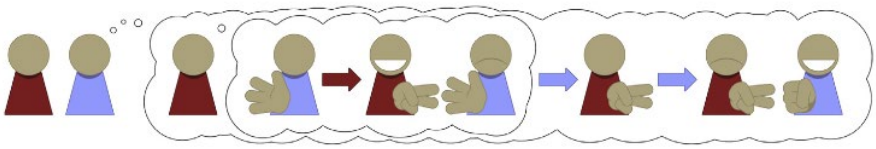

图 4-4 一个一阶心理揣测的例子[44]

麻省理工学院的 Lee 等人[45]定义了一个用于人机交互的非语言交流的双重计算框架。他们使用贝叶斯心理揣测方法来模拟讲故事时的交互作用。讲述者利用声音线索来影响和推断听者的注意状态，将其作为一个部分可观测马尔可夫决策规划问题进行计算。听者通过自己的反应传达注意力，将其作为一个动态贝叶斯网络计算。通过人机交互实验证明模型在注意力识别和传达的有效性。爱丁堡大学的 Patacchiola 和 Cangelosi[46]提出了一种基于信任和心理揣测的发展认知架构，该架构受心理和生物学的启发，由演员-评论家框架和贝叶斯网络组成，这些模块分别对应于大脑中用于心理揣测的脑区。最后，他们用 iCub 仿人机器人进行了两个心理学实验，结果与儿童的实验数据一致，有助于揭示儿童和机器人基于信任的学习机制。

4.4.2 基于深度学习的心理揣测模型

受益于深度学习的飞速发展，基于深度学习的心理揣测模型也取得了很大进展。Google DeepMInd 团队的 Rabinowitz 等人[47]设计了一个 ToM-net 神经网络模型实现通过元学习对其他智能体的建模，他们的网络包含了建模被观测者特点、内心状态的模块，并通过结合这两部分的输出以及被观测者当前的状态来对被观测者进行揣测。他们构建了一个能够收集智能体行为轨迹的观察者，其目标是预测其他智能体的未来行为。他们将提出的 ToM-net 模型应用于简单的网格环境中，结果表明观察者可以有效地为智能体建模并通过 Sally-Anne 测试。而观察者自身不需要执行任何动作。

加利福尼亚大学洛杉矶分校的 Akula 等人[48]基于心理揣测的思想提出了一个可解释人工智能框架 CX-ToM，用于解释深度卷积神经网络做出的决策。该模型可以显式的建模人类用户的意图、人类用户对机器的理解，以及机器对人类用户的理解，通过人类用户和机器之间的多轮交互，提高模型的可解释性，并增加人类对模型的信任。

除此之外，心理揣测的思想也正在影响着多智能体强化学习。为了辅助智能体决策，心理揣测通过观测对手的历史信息（比如，位置、行为、是否结束游戏等信息）来推断对手的目标、行为趋势等。Yang 等人[49]提出了 Bayes-ToMoP 的新方法，可以有效地检测对手使用的固定或更高级的推理策略。Bayes-ToMoP 还支持检测以前从未见过的策略，并相应地学习最佳反应策略。除此之外，深度版本的 Bayes-ToMoP，通过使用深度强化学习技术将 Bayes-ToMoP 扩展到足球游戏这种复杂的多智能体强化学习任务中。

4.4.3 基于脑启发的心理揣测模型

与认知心理学和脑科学关系更紧密的是脑启发的心理揣测模型。在这个方向上，中科院自动化所曾毅团队取得了一系列研究进展。Zeng 等人[8]借鉴心理揣测的多尺度神经可塑性机理，即相关脑区、脑

区功能及神经环路，提出类脑心理揣测脉冲神经网络模型。该模型实现了机器人的自我经验学习，并能够利用自我经验实现对他人信念及行为的揣测，使机器人可以通过错误信念任务，获得初步的心理揣测能力。该模型探索了自我经验、相关脑区和脑区间连接的成熟度，特别是抑制控制机制对心理揣测能力的影响，有助于从计算角度揭示心理揣测的神经机制。

Zhao 等人[9]在此研究的基础上，提出了多脑区协同的心理揣测脉冲神经网络模型，该模型由四个部分组成：视角采集模块（模拟 TPJ 和额下回脑区功能）、策略推断模块（模拟 vmPFC 脑区功能）、动作预测模块（模拟 dlPFC 脑区功能）和状态评估模块（模拟 ACC 脑区功能），模型采用了模拟生物神经元的 LIF 神经元、网络的学习过程采用了与突触可塑性相关的 R-STDP 方法、网络的连接是参考心理揣测各个脑区之间连接建立的。因此每个子模块的输出都是可解释的，训练过程也是受脑启发的。该模型可以区分并对不同类型的智能体进行揣测，并且基于揣测来预判他人未来的安全状态。最后将该模型应用到安全风险任务中，实验证明，具备心理揣测能力的智能体可以帮助他人避免安全风险。

心理揣测的心理状态往往是抽象的，难以直接观测和表征。因此，Zhao 等人[10]不显示地构建对他人的心理状态而是采用网络隐层表征他人的心理状态，进而预测他人行为。该方法中每一个智能体都有自己的决策网络以及心理揣测网络。心理揣测网络的输入是对环境的观测以及对他人行为的观测，隐藏层编码了智能体对他人内心状态的归因，输出层表征了对他人行为的预测。心理揣测的结果可以丰富智能体对当前状态的表征从而帮助提升多智能体合作的性能和效率，并提高智能体在竞争中的竞争力。同时模型还借鉴大脑中 TPJ 模块可以区分自己和他人的功能，包含了存储自身经验和对他人观测的模块，以便模拟智能体使用不同信息进行决策。实验结果表明，在自身经验的

帮助下智能体更容易对陌生的智能体产生准确的判断；而随着智能体之间交互变多，智能体通过对方的历史信息来推测对方会更有效率。

格拉斯哥大学的 Roth 等人[50]认为 Zeng 等人[8]提出的脑启发的心理揣测模型同认知双重过程方法一致，区分了更自动、快速、更少受控制的过程和更刻意、更缓慢和有意识的过程，并与区分内隐和外显的心理揣测模型一致。因此，借鉴类脑心理揣测模型[8]，特别是该模型提出的四条通路：自我经验学习通路、动机理解通路、自我信念推理通路和他人信念推理通路，Roth 等人提出了一项新的心理学实验范式，并在 60 名人类被试上进行了该实验，实验结果进一步证明了脑启发心理揣测模型的有效性和合理性，有助于进一步揭示心理揣测的神经机制。两项工作将人工智能和心理学在心理揣测方面的研究紧密结合在了一起，在计算建模和心理学实验间形成了良好互动。

4.4.4 基于其他方法的心理揣测模型

另外还有一些从连接主义建模、认知架构设计等方面构建心理揣测模型的研究。麦吉尔大学的 Berthiaume 等人[51]提出了一个连接主义模型来模拟错误信念任务。通过增加隐藏层神经元来提高模型的计算能力，该模型可以成功模拟错误信念任务由失败到成功这一转变，他们认为，这种转变的根源不在于对信念的理解，而是由抑制自身信念处理资源的增加导致的。图卢兹大学的 Milliez 等人[52]提出了一个时空推理系统 SPARK，借助该系统，机器人可以以更加自然的方式实现有效地沟通和互动。该系统可以使机器人通过 Sally-Anne 测试，并在对话消歧方面表现良好。

西英格兰大学的 Winfield[53]基于内部模拟模型提出了一个心理揣测模型，并部署在 NAO 机器人上，该模型可以在内部模拟机器人下一个可能的动作，从而预测这些动作对自己和其他个体可能产生的后果，对增强机器人的社会交互能力十分重要。西英格兰大学的 Bremner 等人[54]提出了一个信念-期望-意图模型，其逻辑结构通过记

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/677042160146006036