第十九章　超智能体系统理论

作为科学研究的四大跨世纪难题之一，人类心智的产生（起源）至今还令人迷惑不解，而对人类未来智能发展的憧憬，更是令人神往。当今信息革命所改变的，不仅是人们的生活，由此而引发的知识革命和智能革命，所改变的更将是人类发展的未来。今天，我们之所以可以笑傲古今，是因为我们是站在了前人所构筑的知识和智能大厦的平台之上。一个人的生命和能力总是有限的，而集成了人类智慧的集成智能系统，其威力将是无穷的；借助这一系统的“神力”，我们每个人未来或许都可以成为“超人”。我们坚信，构建人-机集成的超智能体系统，既是人们对智能和智能系统进行研究和探索的初衷，也将是智能和智能系统研究的最终归宿。本章，我们将给出一个人-机综合集成的超智能体系统的理论框架，更深入的研究和探讨，我们将在后继文章中给出。

19.1 综合集成智能系统研究与开发--未来的发展

我们知道，人工智能是研究如何构建出一个人造的智能机器或智能系统来模拟人类智能活动的能力，以延伸和拓展人类智能的学科。我们可以说人工智能（学科）是计算机科学中涉及研究、设计和应用智能机器或智能系统的一个分支；也可以说人工智能（能力）是智能机器或智能系统所固有的或可执行的通常与人类智能有关的智能行为的能力，如判断、推理、证明、识别、感知、理解、交流、设计、思考、规划、学习和问题求解等思维活动。智能的工程化，即智能工程，从本质上讲，就是要创建更多各种形式的智能系统。它既是为了适应当代社会经济和科学技术发展的需要，也是智能科学和复杂系统理论研究成果的实际应用。

目前，对智能的工程模拟，主要有三条途径：一是，**对人类心理机制及功能的模拟，特别是对人类思维机制与功能的模拟**。该方法认为，智能产生的基础是信息、知识和思维。其中，思维论者强调，智能的核心是思维，人的一切智能都来自于大脑的思维活动，人类的一切知识和理智行为都是人类思维的产物，因而通过对思维机制与方法的研究和模拟即有望揭示出智能的本质。而**知识阈值理论**则认为，智能行为主要取决于知识的数量及其精准化的程度；一个系统之所以具有智能，是因为它具有可运用的知识；智能就是在知识的引导下在一个巨大的搜索空间中迅速找到一个满意解的能力。我们可以认为，**智能就是知识与思维能力的综合**；其中，知识是一切智能行为的基础，而思维能力则是获取知识并运用知识求解问题的核心，是人脑智力活动的最主要体现。更进一步，我们还可以认为，人的脑力劳动本质上就是进行基于符号的信息处理；其中，抽象思维，特别是逻辑思维，是完全的符号处理过程，而形象思维或直觉思维，也常常是伴随有“亚符号处理(信号处理)”的符号处理过程。这一理论和方法，在智能工程模拟的研究中，至今仍有着最重要的影响，如人工智能、知识工程和专家系统等，都是在这一理论的指导下发展起来的。二是，**对人脑生理结构及功能的模拟，特别是对人脑神经系统结构与功能的模拟**。持这种观点的人认为，人类智能或认知产生的“基础”不是“符号”而是“神经细胞（神经元）”，**智能行为或认知过程是大脑中相互联接的神经元的集成（涌现）功能**。因而，只有从大脑神经元及其连接机制出发进行研究，搞清楚大脑的结构以及它们进行信息处理的过程和机理，才能揭示出人类智能产生的奥秘，从而真正实现人类智能在（智能）机器上的模拟。其中，联接主义者正尝试建立各种分布式并行计算系统（人工神经网络）对大脑的神经结构进行结构和功

能模拟，并希望所构建的系统能具有自学习、自组织、自适应、联想、模糊推理等各种能力和特性。三是，**对属人世界生态进化过程与功能的模拟**。该方法认为，**人类智能的本质，应是在（真实世界）动态环境中的决策和行为能力，对外界事物的感知-反应能力，以及在特定环境中生存和发展的能力；能产生智能行为的主体，应是具有某种（生存与发展）意愿和特定（智能）行为能力的可学习或进化的主体；其智能，应是（在特定的先天基础条件下）在后天（不断地）“实践”中（逐步）形成的，它源于主体对外界刺激所做出的恰当反应；而智能只能由智能体在现实世界中与周围环境的交互作用中才能表现出来**。因此，智能系统的智能应该属于具体的意愿与环境，应该有与环境的交互作用；**智能，只能是在（智能）系统各个部分的交互作用下，在系统与环境的交互作用中，所涌现出来的一种总的行为特性**。其中，基于行为主义的研究者甚至认为，知识的形式化表达和模型化，已成为人工智能进一步发展的重要障碍；其实，智能行为的产生，可以不需要复杂的知识，也不一定需要复杂的知识表示和推理，生物智能就主要取决于其感知和理智行动；应该直接利用（智能）机器对环境作用，以环境对作用的响应为原型；**智能行为应体现在现实世界中，应通过与环境的交互而表现出来**。而进化论者则强调智能系统智能进化的可能性与必要性，认为人工智能可以像人类智能一样通过逐步进化实现，分阶段地发展和增强。他们认为，智能模拟应该以复杂的现实世界为背景，先研究简单动物（如昆虫）的信号处理能力并模拟和复制，沿着进化的阶梯逐步向上进行。而多智能体系统理论则强调智能的**社会性和集群效应**，认为，**生物智能，有很多就是群体协作的结果，具有一定意愿和简单规则的多个个体的群体行为，也可“涌现”出复杂的智能。**

应该说，无论是符号主义、联接主义还是生态进化主义的方法，都是智能模拟的有效途径。不过，它们在各自取得众多令人鼓舞的成果的同时，也都面临着各自不同的困境。因为现实已经证明，用符号主义、联接主义或生态进化主义方法，建立在现有计算机基础上的人工智能系统，只能模拟人的部分智能行为。以单纯的还原论为指导思想的符号主义研究方法或以单纯的整体论为指导思想的联接主义研究方法等，都不可避免地存在着某种局限性和片面性。如何改进现有的理论和研究方法，提高一个人工智能系统的效率和性能，使其能更快、更好地发展，从而更接近于人类智能，将是智能系统研究的最重要课题和方向。

应当认识到，人工智能系统目前还只能是人类改造世界和拓展自身智能的“工具”，我们不应奢望在目前情况下就能制造出可自主达到甚至超越人类总体智慧的人工智能系统。在“智能工具”的框架内，在“人-机结合”技术路线的指引下，运用整体论和还原论相结合的“**系统主义**”和“**机制主义**”研究方法，采用生物计算机、神经计算机与传统的冯•诺依曼计算机相结合的多种可模拟人脑功能的新（计算）系统，才有可能使人工智能的研究突破目前的困境。未来的智能系统，（在可以预见的未来），一定是人-机共生共存的集成智慧系统。

对于智能系统未来的发展，我们坚定地认为：

1.未来智能系统构建的基本模式，一定是人-机结合的综合集成系统

我们强调未来的智能系统，一定是人-机结合的综合集成智慧系统，不仅是因为，在当前阶段，所有成功的智能系统，都是人-机结合的综合集成系统；还是因为，我们坚定地认为，以人为主、人-机结合的综合集成方法，在可以预见的未来，也一定是智能系统研究和开发的主流；集信息、知识、人类智能和机器智能之大成的**大成智慧**，是现阶段解决现实生活中复杂问题的一种最可行的方法，也是未来人和机器、知识和智慧集成发展的根本方向。

我们知道，对于人工智能的发展，一些乐观主义者曾认为，人工智能的研究是前途无量的，人工智能不仅能完全替代人的智能，而且还将超过人的智能；而一些悲观主义者则把人工智能贬为现代的“炼金术”，认为它们不过是在某些局域的表面上对人类智能的某种粗略模仿，还远未达到对人类智慧的实质性了解。而人工智能的发展也似乎并没有单独支持哪一派的立场：它既朝着深入认识、反映人类智慧的方向在一步步前进；但又远远未能实现当初乐观者的许多既定目标。随着智能技术的不断发展，人们对人类认知和思维进行模拟的范围正在不断扩大，人工智能系统在功能上也许会不断向人类智能接近；但是，从本质上看，它们之间的“界限”也许是“永远”都不会（完全）消除的。因为功能模拟只是近似而不是等同，更不可能是总体的（智慧）超越。从已有的研究来看，目前的人工智能还只是人类智能的“局域功能拓展”，与人类智能还有着本质的不同。比如，人类智能是在人脑生理和心理活动基础上产生的，是基于人的主观世界的主观能动的过程；而人工智能目前则主要是基于机械的、物理的运动，它并不具备由世界观、人生观、情感、意志、兴趣、爱好等心理活动所构成的主观世界，而只是“机器”“无意识”的运动过程。人类智能行为是自觉、能动的，有目的性和可控性；而目前的人工智能系统则还没有（人们理性中的）自觉性、目的性和主观能动性；在解决问题时，它并不探求任务本身的含义，它只是机械地在执行指令。人能在认知的基础上，提出新概念，作出新判断，创造新事物，具有丰富的想象力和创造性；而目前的人工智能则至今还无法表现出想象力和创造性；它还不能自主地提出问题，创造性地解决问题，在遇到没有列入程序的“意外”情况时，也会束手无策。人类智能具有社会性，作为社会存在物的人，其大脑的很多功能是适应社会生活的需要而逐步形成和发展的；人们的社会需要远远超出其直接生理需要的有限目的，是由社会的物质文明与精神文明的发展规律所决定的；作为人脑功能的思维能力，是通过社会的教育和训练，通过对历史上积累下来的文化的吸收而逐渐形成的；人的内心世界所以丰富多彩，是由于人的社会联系是丰富的和多方面的；而目前的人工智能系统则没有人的社会性感受，要让它把人脑的功能全面“模拟”下来，也许需要在一定程度上再现人的思想发展的某种历史逻辑，而这是无论多么“聪明”的人工智能系统目前都无法做到的。因此，从总体上看，目前的人工智能还仅是人类意识自我认识的产物，是人类意识器官的延长，是人类在认识和改造世界的实践过程中智慧积累的结晶。人工智能系统的功能也许在某些局部功能领域已大大超越了人脑，但在本质上，在技术上，它目前都还不具备可以完全达到或超越人类（总体）智能水平的客观条件。

而将人和机器相结合，将人的智慧与人工智能（机器智能）相结合，发挥其各自的优势，却是可以产生出他们各自无法达到的效果的。人的意识活动是丰富而能动的，为了认识世界和改造世界，人始终发挥着主导作用；但是，人又是有局限性的，人的感受、思维、知识和精力，都是有局限的。尽量开拓（智能）计算机和其它各类自动机的功能，让它们尽可能多地帮助人来工作，才是人类社会发展的正确途径。人-机结合的综合集成智能系统之所以更能取得成功，就是因为它是一个“人-机结合、优势互补”的系统。人和人之间、人和智能机器之间，可以通过精心设计的集成环境，进行信息和知识的共享以及思维和智慧的交流，实现“人”、“机”智能的互补和有效利用，进而实现“人助机、机助人”的智能决策支持。在增强了“人”解决问题能力的同时，也大大地增强了“机器”的功能。比如，在可按形式化方法进行逻辑推理或需要进行精确数值计算的场合，“逻辑计算系统”显然能够显示出巨大的威力；在某些智能型的信息处理领域，如模式识别、模糊处理、自适应控制和组合优化等方面，“神经计算系统”也将能发挥出其独特的优越性。为了充分发挥人的心智的优势和智能（计算）机器的高性能，把人和智能机器相结合，应是最恰当的途径。目前，在人-

机结合的综合集成智能系统中，综合集成已从算法、模型的综合集成扩展到了感知、认知、思维和智慧等多个方面的综合集成。其综合集成的意义已不仅仅表示系统是由基于多种方法的多种模块所组成的，而是根据问题在某时刻的特定需要，在系统集成的意义下动态地构成了包括“知识系统”和“社会智能组织”在内的若干个子集，在动态的信息和思维交互的过程中去求解问题。人-机结合的综合集成智能系统，就是使计算机（智能机器）与人之间形成一种合作关系；人-机智能系统，就是“人脑”与“电脑”构成的优势互补的统一体，就是让人与机器各自去完成自己所擅长的任务；其智能，将是人-机合作的产物，肯定会超越单一的智能。“机器智能”，作为人类智能的模拟、延伸和拓展，在人类“智慧”的“引导”下，会（始终）作为“人类”智能的“一部分”而存在。“硅基”智能系统未来会不会“失控”？也许有可能。但以人类现有的“智慧”，会让这种情况发生吗？

2 **未来智能系统构建的基本方法，一定是基于复杂系统理论和系统工程方法的**

我们知道，在对智能本质揭示和模拟的进程中，一直存在着还原论和整体论两种最基本的范式。近代物理科学以还原论为指导思想，充分发展了其基于还原主义的研究和分析方法——通过解析把高级运动形式分解或还原为低级运动形式，也取得巨大的成就。它告诉我们，茫茫世界中所有的巨大的和较大的实体，如星云、星球、山川、万物等，都是由一些较小的物理实体构成的，而这些较小的物理实体，本身又是由更小的物理实体（量子）构成的；我们可以对它们进行层层分解，直至最终到达分子、原子和基本粒子（量子）层次。面对智能问题，一些人也希望能沿用这种基于还原主义的研究范式。他们认为，现实生活中的每一种现象，都可看成是更低级、更基本的现象的集合体或组成物，因而，通过解析的方法，我们也可以用低级运动形式的规律来组合出高级运动形式的规律来。由还原论所派生出来的方法论和研究手段，就是对研究对象不断地进行分析（解析），化复杂为简单，直至其最原始的形态。但是，在面对诸如生命、意识、智能和人类社会运动等这样一些复杂且“活生生”的对象时，单纯基于还原论的方法却有些失灵了。智能科学的研究表明，单纯的还原主义方法目前还无法完全揭示智能的本质，在对智能和智能系统的研究中，任何忽视了自主性和意向性等这些人类行为重要特质的研究方法，都将是行不通的。还原论者看到了事物不同层次间的联系，想从低级水平的运动入手来探索高级水平运动的规律，这种努力是很可贵的。但是，低级水平的运动形态和高级水平的运动形态之间，毕竟有着本质的区别；如果不考虑所研究对象的特点，简单地想用低级运动形式的规律来描述高级运动形式的规律，显然是行不通的。

不错，在经典物理学的世界中，（客观）世界完全是由各种粒子、力和场等，最终是弦和能量（物能），构成的；宇宙中存在的一切，都不过是运动中的物质（物能）和物质（物能）的运动。在这一物理世界中，由更小事物构成的较大事物的许多特性，主要是由更小事物的特性和行为决定的，并且可以根据更小事物（的特性和运动）来解释。在这一单纯的物理世界中，**“经典物理学并没有为意识留下一个合理的位置，因为它在逻辑上就已是完全的了**。”而在现实世界中，人的行为是包含着主观性和意向性的东西的，它是有意向的行为，而不只是一种身体的运动。在心理学中，人们常以**意向性**作为区分心理现象和物理现象的标志；认为，物理现象只涉及物理对象本身，而心理现象则包括了心理主体和它指向的外部对象的方式。**意向性最鲜明的特点是“指向性”和“关于性”。任何意识活动都是指向一定对象的，统摄对象是意识活动的前提。意识活动的特点还表现在既意识到对象的存在，也意识到意识自身的存在，而且是通过对象的存在而意识到意识自身存在的。意向性结构的一端是代表主体的自我，另一端是与(内部和外部)世界发生着广泛联系的意向对象。意向性与生命现象所表现出的目的性有着内在的联系**。可以说，包含在意向性中的目的性是对认知

过程中盲目性、散漫性的克服，也使意向性“带领”心灵走出了原始的混沌状态和低水平的组织层次。由此可见，要想达到对人类智能的揭示，解析意向性应是极其关键的一环。

在科学研究中，若面临的研究对象十分复杂，而研究方法又很不成熟，在某种程度上，对研究对象进行科学分解，在更适合的研究水平上进行研究，对揭示科学规律来说，不仅是可取的，而且是必需的。但这种分解必须考虑不同学科的特殊性。在智能科学研究中，任何方法都不能在分解中丢掉心理现象的特殊意义，将生动丰富的心理现象变为毫无生气的“物理元素”的集合体。单纯还原主义的方法，对事物间关系比较简单，因果关系明了，能够用精确的数学方式表述的事物，会是十分有效的。但面对复杂性问题，简单性不存在了，确定性也不存在了，单纯的还原主义方法，也就陷入“困境”了。

随着单纯还原主义的研究范式在对复杂性问题进行研究过程中走入困境，基于整体论的研究范式逐渐兴盛起来。整体论与现代有机论、系统论、突变论、完形论、理性直观说等相互渗透，使当代整体论的反还原主义特征和实用化的方法论主张得以彰显，曾极大地改变了数百年来西方科学文化的根基——基于牛顿-笛卡儿的自然观、知识论和科学范式，使我们对复杂事物的研究进入到了一个“**系统**”研究的时代。

系统论尤其是随后陆续出现的以复杂性研究为核心内容的系统科学群，开始了向整体论的转移。在以社会科学与生命科学为背景的系统研究中，都无一例外地引入了整体论和系统论的思想，它们与还原论的有机结合，已经提出了不少可用于对复杂系统进行研究的新理念和新方法。

贝塔朗菲曾把一般系统论界定为关于整体性的科学，把整体性界定为一种“**涌现**”的性质。在这里，系统科学首先是提供了一种从整体出发思考和解决问题的观点，引导人们将事物视为一个有机的整体，充分考虑它所有的因素之间那种相互关联、不断变化的复杂关系。当然，这种整体论和系统论的思想，在深入到现代科学的诸多经典理论中时，也会产生出多种不同的具体形式。

在基于**系统分析**的所有理论中，整体论的思想发挥着重要作用。但是，在系统论的发展过程中，单纯整体论观点的局限性也日渐凸现。整体论的局限性即在于其过于强调整体，对其部分与个体的“**自由**”却重视不够甚至忽略。整体论对事物的处理，大方向是对的，但是因为不注重深入了解细节，这样的处理常常会带有主观主义和经验主义的色彩。系统科学的产生和发展历程表明：“**不要还原论不行，只要还原论也不行；不要整体论不行，只要整体论也不行。不用还原论去考虑整体也不行。不还原到元素层次，不了解局部的精细结构，我们对系统整体的认识也只能是直观的、猜测性的、笼统的，缺乏科学性。没有整体观点，我们对事物的认识只能是零碎的，只见树木，不见森林，不能从整体上把握事物，解决问题。科学的态度应是把还原论与整体论结合起来。”系统科学的使命即在于将还原论和整体论取长补短，进而实现还原论与整体论的统一**。任何事物的发展都是不断地自我完善与发展的过程，同时，这一过程是由一系列存在着内在联系的阶段性过程构成的。由于事物发展的复杂性、曲折性，**对还原论与整体论实行在“辩证统一”基础上的有机结合，将是科学发展的必然趋势。**

鉴于人脑的功能是成千上万具有不同专门功能的子系统协作的结果，人类智能的本质不可能归结为几个像波函数或运动学三定律那样规整、简洁、漂亮的基本原理，人工智能也很难以单一的理论向前发展，在“**系统主义**”的基础上同时融合还原论和整体论的研究策略，将不失为智能和智能系统研究的一种有效的方法。

我们坚信，**未来的智能系统研究，将是还原论与整体论相统一的；未来研究对智能本质的揭示**

和模拟，将是多层面的；未来智能系统的构建，也将是自上而下的解析与自下而上的集成“涌现”相结合的系统方法。

对于未来智能系统的构建，我们看好系统工程方法的应用，更看好**复杂适应性系统理论**的发展。不过，为复杂适应性系统（简称CAS）建立一个完整的理论体系，目前还是十分困难的。因为复杂系统的整体行为，并不是其各个部分行为的简单叠加，系统中充满了非线性。复杂适应性系统，一是系统参量多；二是结构层次多，聚集体结构呈多元化；三是其结构法则难以把握。因此，虽经近半个世纪的努力，我们至今还难以构建出复杂适应性系统完美的功能模型，也难以构造出其完美的行为模型。既然我们很难把我们所不熟悉的复杂情形解析为我们所熟知的基元，就更不用说将其构建成为基于经验的完整模型了。但是，复杂适应性系统理论，确实会是我们未来研究（复杂）智能系统的重要基础和方向。将系统主义与机制主义相结合，既考虑其系统运行的宏观整体动力学规律，又深入解析其微观产生机制，我们就一定能构建（或模拟）出一个完美的类人的复杂智能系统。

3 **未来智能系统研究的核心内容，必定是包括着思维和智能的集成理论的**

我们认为，未来，对人-机综合集成智能系统的研究，会包括着对智能产生机理的揭示，也会包括着对人工智能系统（体系）的构建，更会包括着对多个智能系统智慧集成的研究。

研究生物智能系统——特别是人类---智能产生的机理，研究使人类个体智能得以提高的方式和方法，无疑是智能科学研究的根本；而人工智能系统的构建，人-机综合集成智能系统的构建，更是人们对智能科学研究的期待。未来的智能系统研究，必须从研究“人类智能与机器智能的统一”出发，创造出智能系统研究的新的途径和方法来。我们知道，思维科学着重研究的是思维的规律，认知科学着重研究的是人的认知，目标旨在揭示（人类）智能产生的基础和机理。与思维科学和认知科学不同，人-机结合的智能系统研究，其着重点还在于将人的智能与智能计算机器的高性能进行有机结合，构建出人-机结合的“使两者取长补短的人-机共生”的综合智能系统来，让人同计算机一起去完成目前谁也无法单独完成的事情。这些研究，其核心无疑会是思维和智慧的集成。人类社会的进步和智能的提高，一直是在社会组织的主导下个体与社会共同努力的结果；人工智能系统也会如此。未来，不可能会是一个智能系统“独霸”“江湖”，多智能体（系统）的合作与集成是必然的趋势。研究多智能体之间思维和智慧的集成，也许比研究如何构建一个单一的智能系统更重要。

4 **未来智能系统的发展，必将是建构、进化与自组织相统一的**

未来的智能系统，首先应是建构的。在系统的建构中，会有信息和知识系统的构建，也会有人工智能系统的构建，更会有集成系统的构建。在人工智能系统的构建中，行为模拟是必要的，但心理的解析和思维的动态表达更是重要的。所构建的系统，将是外在行为与内在心理统一描述的系统。而将符号主义、联接主义和行为主义（生态进化主义）的方法等综合运用，进行优势互补，实现有机统一，应是现阶段最可行的一种做法。

未来的智能系统，是建构的，也会是进化的；系统不仅具有集成的特征，还必须具有进化的特征。在这里，进化（evolution）是指广义的进化，是指系统的历史演化及发展。当然，未来集成智能系统的进化首先是“人”的进化，是人认知能力和问题解决能力的提高。再就是信息和知识资源的不断扩充和信息自动化处理系统的不断完善。但更少不了的，则是各类人工智能系统的发展。

目前，人工智能系统的发展还离不开人的指导和帮助，因此，人工智能系统所谓的“进化”，确切地讲应该称作“升级”。人们通过对人类智能本质的揭示，得到相应的机理，再通过构建人工智能系统来进行模拟实现。它既是人类智能的产物，也将是人类知识和智能水平得以提高的有力工具。

人工智能系统和信息自动化处理系统的进化，包括着软件的进化，当然也包括着硬件系统的进化。现有的信息处理技术，主要是以冯•诺依曼结构的计算机系统为技术支持的。冯•诺依曼结构的计算机系统的优势，在于其高速的数值计算能力。人工智能系统的进一步发展，或许更需要新一代的在很大程度上类似于人脑的“计算机”系统。目前，世界上处于研究热点的，被众多科学家寄予厚望的，代表着未来发展方向的“计算机”，主要有“光子计算机”、“生物计算机”和“量子计算机”等。光子计算机以光子作为传递信息的载体，以光互连代替（电）导线互连，以光元件代替电子元件，以光运算代替电运算，或将使运算速度在目前基础上呈指数上升。量子计算机是根据原子或原子核所具有的量子特性来工作，运用量子信息学，基于量子效应构建的一类完全以量子位（量子比特）为基础的计算机。目前，它是利用一种链状分子聚合物的特性来表示开与关的状态，利用激光脉冲来改变分子的状态，使信息沿着聚合物移动，从而进行运算。生物计算机则是以生物元件为核心部件构建的计算机。它利用蛋白质有开关特性，用生物工程技术产生蛋白质分子，并以它们做元件制成集成电路----也就是生物芯片。生物芯片本身具有天然独特的立体化结构，其密度要比平面型的硅基集成电路高5个数量级，因而会比硅基集成芯片更小。更为期待的是，由于生物具有自我修复的功能，生物芯片一旦出现故障，也许不需要人工修理就可以进行自我修复。所以，生物计算机将具有“半永久性”和更高的可靠性。再者，生物计算机的元件是由有机分子组成的生物化学元件，它们是利用化学反应工作的；所以，其工作只需要很少的能量就可以了。基于生物芯片的神经计算机也许是最值得期待的。**生物神经计算机**具有仿人的判断能力和适应能力的特性。它可以判断对象的性质与状态，可同时并行处理实时变化的大量数据，引出结论并能采取相应的行动。若用众多微型生物神经处理器来模仿人脑的神经元结构，采用大量的并行分布式网络来构成“（生物）神经电脑”，其前途或许是不可限量的。

未来的集成智能系统，其主流或许是建构的和进化的，但也不排除会有自组织系统的存在。最起码，目前，有不少网络论坛、QQ群和微博等，就是自组织的。未来，随着信息技术的进一步发展，自组织的“沙龙式”的集成智能系统只会越来越多，并在提高人的认知和解决问题能力方面会发挥越来越重要的作用。特别是当智能体系统具有了“自主特性”之后，主动或被动的“自组织现象”将成为常态。

19.2　超智能体系统理论的基本概念

1．主体

在超智能体系统理论中，**主体**是指一个可运行于动态环境中的具有较高自治能力的系统或实体。该系统或实体可以是一个生物体、一台自动机，或是一个可控计算机软件系统；它具有特定的信息加工能力，可接受特定的任务并能对外提供特定的服务，它与其它主体或环境之间具有较为松散或相对独立的关系，通常具有**交互/协作性、目标/任务驱动性、自主/可控性**等特征。

自主性(Autonomy)：自主性或自治性是主体的最根本特性。这也就是说，其活动是可以“自主”的；其内部运行是不由人或其它系统直接控制的；它对自己的行为和内部状态拥有一定的控制权。在没有人或其它系统直接干预的情况下，它可根据外部条件和内在（认知）规则自主地控制其行为，自主地运行和执行其操作。对遇到的事件（问题或任务），在不受外界监控指导的情况下，可由自己的“决策机制”决定应采取何种活动；在需要时，还能够主动与服务对象、系统资源或其它的主体进行通讯和协作。

社会性（Socialability）：社会性或协作性是说，主体具有相互协作的能力，它可通过与其他主体的合作来共同完成某项任务。这是主体可作为系统（整体）的一部分而顺利工作的关键。这种协作可以从简单到复杂，从单一的服务提供到基于智能方式的多个主体的协同和协商以合作完成复杂任务。当有冲突发生时，主体还具有通过协商解决这种冲突的能力。因此，各主体之间通常存在着相互依赖、相互制约的关系。它们的行为既有局部效应，又有全局效应。

通信能力（Communicability）：通信或交流（交互）能力是说，当需要合作以共同完成某项任务时，主体具有通过某种通信语言与其他主体进行交互或进行信息交换的能力。这种通信既要保证主体之间的相互交流，又要不至于影响主体之间的相互独立性。在系统中，主体间的通信是相互协作、协商的基础，是思维和智能得以集成的基础。

反应能力（Reactivity）：反应或响应能力是说，主体能对特定任务或其所在的环境的改变及时地做出反应。主体能接受一定的任务请求，并及时响应；能感知其所处环境的变化，并通过行为作用于环境。

2．智能（主）体

智能（主）体即具有智能行为能力或智能化信息处理能力的主体。我们可以把“智能体”看作是能够通过其感知机制感知其生存环境的信息，并借助于其执行机构作用于环境的任何智能系统。它能够在目标或任务的驱动下主动采取包括学习、通讯、社交等各种手段，感知其外在环境的动态变化并（根据认知）做出恰当地响应。

在超智能体系统理论中，一个智能（主）体通常具有如下典型特征：

主体性：一个智能体应是一个相对独立的主体，具有主体的所有典型特性。

智能性：毫无疑问，智能性是智能体的本质特性。智能体所有的行动，都具有或基于一个既定的目标（任务）或信念；它能够感知环境的变化或接受指派任务，并依据其智能判断或根据以前的经验和认知，做出（智能化）行为决策，以及时对这些任务或变化做出反应（响应）。

智能性也意味着智能体具有知识和能力。一般认为，智能是指一个智能体信息处理或问题解决活动中所表现出来的一种能力，即其可对客观事物进行合理分析、正确判断以及有目的的行动和有效的处理周围环境事宜的一种综合能力。而这也要求，第一，它必须具有感知能力（也包括主动探索能力）。即通过视觉、听觉、味觉、触觉等感知外部世界的能力（感知智能）。当然，智能也体现在其可感知其自身的存在。第二，它必须具有记忆和思维能力。记忆用于存储所感知到的外部信息以及由其思维所产生的（内在）经验和知识；思维用于对信息进行处理，即利用已有知识对信息进行分析、计算、比较、判断、联想及想出对策等。思维是一个动态过程，是获取知识及运用知识求解问题的根本。当然，思维有逻辑思维（抽象思维）、形象思维（直感思维）以及顿悟思维（灵感思维）之分，智能体必须具备其中的部分功能。第三，它必须具有学习能力、适应能力及实践行为能力。学习能力是指可通过接受“教育”或通过实践等过程来丰富自身的知识和技巧（经验）的能力；适应能力是指在各种不同环境下都能做出恰当决策和应对的能力；行为能力是指可以把思维决策转化为行动的能力（行为智能）。

合作性(Collaboration)：即，一个智能体可以与其它智能体分工合作，共同完成单个智能体无法完成的任务。

主动性(Proactivity)：即，一个智能体应能够遵循其意愿或承诺采取主动行为，表现出某种主观能动性的特征。

由于上述特性，因此，一个基于智能体的系统，应是一个集灵活性、智能性、可扩展性、鲁棒性、组织性等诸多优点于一身的系统。

智能体的构成要素包括其结构状态、心理意识状态和行为能力状态等。人类心理状态的要素主要有认知（信念、知识、学习等）、情感（愿望、兴趣、爱好等）和意向（意图、目标、规划和承诺等）等。智能体的行为应是由其“意识”控制的，而智能体的意识则由其**“自主意识”**控制。我们也把具有完全的或部分自主意识的智能体称为**智能主体**。认为，智能主体的行动主要受其心理状态驱动。在一个自然或人工的环境中，一个智能主体必须能利用其在与环境的交互中所获得的经验和知识自主修改其内部状态（心智状态），以适应环境变化和问题求解的需要。智能主体的内部结构将直接影响到系统的性能和智能。智能主体的体系结构使得其感知系统可感知其系统内外环境的状况和发展变化，其控制机制则可把系统的意图有选择地传递给其执行机构去影响其它智能体或环境。**人们常用信念、愿望、意图来刻画一个智能主体的心理意识状态。这里，信念描述智能主体对世界及环境的（总体）认识，表达对可能发生的状态的坚定态度；愿望可从信念直接得到，描述智能主体对可能发生情景的判断；意图代表着智能主体的思考状态，是其在某时刻要选择执行的计划（规划）；意图来自愿望，制约着智能主体，是目标的组成部分。信念、愿望、意图与行为具有某种因果关系。**智能主体的核心部分应是其决策机制或问题求解功能，它起着智能主体的主控作用。它可接收全局状态、任务和时序等信息，指挥（系统）相应的功能操作，使智能主体实现从感知到实体动作的各项功能，包括感知、决策、动作、模式重建、目标规划和通信等。

我们也可以认为，**智能（主）体就是具有社会和领域知识，能依据其心理状态（信念、期望、意向）自主工作，并具有语义交互操作和合作行为协调能力的系统或实体。**作为可参与协调和合作的“软构件”或智能主体，它不仅可为实施紧凑一致的协同工作提供有力的支持，也为建立面向分布计算的开放性、可重构和可伸缩的新型计算环境建立了基础，更为模拟人类的社会组织智能开辟了新的途径。

3．人-机集成智能体

在超智能体系统理论中，我们称一个为了完成某一特定任务而由人和由人操控的工具设备依一定方式共同构成的智能化信息处理及行为系统为人-机集成智能体（人-机集成智能系统）。它是一个由“人”和由人操控的“（智能）机器”共同构成的智能系统。

“人”，无疑是一个智能水平最高，系统功能最全，体系结构最复杂的生物控制与信息处理系统。我们可以认为，**人类智能系统，包括着 4 个功能子系统：意志系统、感知系统、思维系统和行为系统，分别执行人的信念、感知、（认知）思维和行为功能**。在上述四个功能子系统中，**意志系统是最高级的中枢系统，其主要功能是进行感知、思维和行为等的控制，产生有意识的控制行为以协调系统有意识的整体活动。思维系统是主管逻辑（抽象）思维和形象思维等的系统，具有联想、记忆、学习、分析、判断、推理、决策等思维功能。感知系统是视觉、听觉、触觉、嗅觉、味觉等各种感觉的信息处理中心，负责对外周神经系统传入的各种感觉信号进行时空整合与信息融合，具有感知、（初步）认知、识别和（直观）理解等感知功能。行为系统的主要功能是协调人体的运动和行为，控制人体的动作和姿态，保持运动的稳定和平衡，对人体全身的运动和姿态进行协调控制，具有对其运动和行为进行协调、计划、优选、调度和管理等功能。**人类智能系统，还具有**多中枢自协调机制**。其中，**意志系统具有全局协调机制，可进行系统的全局运行协调，系统的总的目的和行为的协调；也包括辨识和注意等。感知系统具有感知协调机制，可对外界并行或串行传入的感觉信

息进行时空整合、信息融合与内外协调等。行为系统具有行为协调机制，可根据意志系统关于运动或目的行为的指令，以及感知系统关于系统本身和外界环境的感知信息，对系统运动和姿态进行协调控制。

在人-机集成智能系统中，“机器”是指所有由人操作、控制或使用的工具和设备（包括智能设备和系统）。这些工具和设备包括：能协助或提高人的感知能力的工具与设备（可具有感知智能）、能协助或提高人的行为能力的工具与设备（可具有行为智能）、能协助或提高人的信息处理能力的工具与设备（可具有认知智能）等。

需要指出的是，一个人-机集成智能体，是一个由人和机器构成的智能系统，但这一系统必须是“集成”的，是由统一意愿和思维意识指导的。它可被看作是一个由单一思维主体所主导的系统，或人-机合一的智能体。

4．超智能体

在超智能体系统理论中，一个超智能体是指由多个思维主体或智能系统以一定方式有机集成的，对外有完全一致行为的智能体；其特点是，他们可以被看作是一个超出一般智能体的统一的智能体系统。

超智能体可以全部由“人”组成，即各类**“智能组织”**；也可由多个人工智能体组成，即**“多智能体系统”**；更可由人和（智能）机器共同组成，此时，我们也称其为**人-机综合集成智能体**。

5．人-机综合集成的超智能体系统

在超智能体系统理论中，一个人-机综合集成的超智能体系统，是指由多个“加盟”**智能主体**（包括人工智能体、人-机集成智能体和超智能体等），服务设施和公共交流、协作及信息自动化处理工具，公共信息/知识资源，依一定规则共同组建的，可在一定的环境下对特定信息或特定领域问题进行共同处理的，**具有某种“主体”特征的（智能）集合体或智能化信息处理系统**。

我们称一个系统为人-机综合集成的超智能体系统（简称超智能体系统），若

超智能体系统==｛加盟智能（主）体，服务设施和自动化信息处理工具，信息/知识资源；规则[系统构建规则，系统运行规则，系统交互与行为规则，系统进化规则]；系统领域及问题空间；系统背景空间[环境，历史]｝

其中，超级智能体系统的**加盟智能（主）体**可以是“人”、人-机（集成）智能体、人工智能体或超智能体。系统的加盟智能（主）体可分为**“主管智能体”**和**“功能智能体”**。**主管智能体**可依据上层意愿或内在群体意愿从事特定管理工作：作为系统“发起人”或“法定代理人”，对内，协调加盟的各个智能体的工作，可有任务支派和系统行为最终决策权；对外，有选择愿意加盟的智能（主）体，与外界环境或其他智能体进行协调和协作的权力。**功能智能体**来自系统环境中的（领域）智能体世界，可依据兴趣和系统请求加盟系统；作为系统的一分子，它具有一定的意愿、兴趣和特定的知识和能力，在加盟系统中，可从事某一特定的或专门化的工作或参与决策等。

一个超级智能体系统，可以是一个完全自治的适应性系统，也可以是一个受外部控制的受控智能体系统。当其主管智能体的意愿是受外部控制时，它即是一个受控智能体系统。而在一个自治智能体系统内，任何一个智能体都有可成为主管智能体。

超级智能体系统中的**服务设施和自动化信息处理工具**包括所有与系统运行有关的、可提高人的认知能力和行为能力的设备和系统，如计算机和网络等，它们既是系统中进行信息处理的工具，也是系统进行信息交流或加强自身功能的工具。

超级智能体系统中的**信息/知识资源**包括所有与系统运行有关的信息资源和知识资源，如，各类专业数据库、各类知识（库）系统和网络即时信息等。

超级智能体系统中系统的**构建规则**包括智能（主）体加盟的条件和系统管理的规则、规程等；加盟条件可以根据系统需要而定；其管理包括**强势管理**[主管智能体有任务指派权和基于民主集中制的最终决策权]和**弱势管理**[如：松散联盟、自由沙龙；主管智能体对加盟智能体无任何强制权力]。故智能系统也就包括了从“人-机合一”的“超人”到“自由沙龙”的所有系统。未来，他们将共同构成一个**超级智能体社会**。

超级智能体系统的**运行规则**包括：议事及行事规则；研讨及决策规则；任务指派规则等。

超级智能体系统的**行为规则**将规定系统的可允许或不允许的行为。系统的行为包括对内改变自己认知的学习和对外给出问题解决方案或动作规划的决策行为等。

超级智能体系统的**领域**及**问题空间**规定了系统可接收或可处理的问题范围。系统可处理的问题可以是被（系统）上层指派的任务或系统本身感兴趣的问题。问题主要来自实践，也可由某一智能主体提出。

超级智能体系统的**环境**包括**相关智能体世界、问题背景空间、自然-物质约束条件、社会-人文约束条件等**。**智能体世界的智能体可包括信息数据智能体、知识智能体和（认知）思维智能体**等；它们或是单一功能的，或是复合功能的。其中，**信息数据智能体**主要从事公共或专业数据的采集、评价、整理，数据库构建(可信数据汇集、入库，存疑数据备注)以及信息数据服务（合法请求识别；对合法请求进行数据推送）等工作。**领域知识智能体**主要从事**相关专业知识的采集[获取]、评价、整理**，知识库构建（可信知识表达、汇集，疑义知识备注）以及**专业知识服务**（合法请求识别；对合法请求进行知识推送）等工作。

智能体世界的智能体主要是**认知（思维）智能体**。它们是一些具有特定知识和（认知）能力的智能体。认知（思维）智能体的能力包括**感知能力、思维能力**[认知能力、分析问题解决问题能力]、**学习能力**[提高自己知识和智能水平的能力]、**工具使用能力**[使用工具来拓展自己能力的能力]和**交互协作能力**[社交能力]等；也可认为其具有感知智能、认知智能、行为智能和社会交互智能。认知（思维）智能体有自己的私人知识集合、行为规则和私人信息集合，具有解决特定问题的能力。在认知与问题解决过程中，其思维过程包括：相关问题识别；相关问题（解决）空间构建；相关知识激活为问题求解知识；问题解决方案提出；问题处理后反馈学习；相关专业知识的主动学习；认知、兴趣、意愿的修正；等等。认知（思维）智能体具有主体意识。**主体意识包括本我、自我和超我。它以本我为基础，以自我为中心，又受超我控制。主体意识可产生智能体自己的意愿、信念和兴趣。信念、意愿和兴趣会使智能体有自己独特的感知与理解能力，有自己的行为规则；认知和学习可改变其信念、知识、行为规则和能力。**

智能体世界的**智能体**可凭自己的意愿或管理智能体的请求加入或离开一个超智能体系统。

超智能体系统的**关联（交互）**包括超智能体系统外部的关联和超智能体系统内部的关联（交互）。超智能体系统内部的关联是指在系统的统一或非统一的意愿下，各智能体依据自己的知识和行为规则，研讨性或协作性的处理问题时的交互与联系；超智能体系统外部的关联是指系统作为一个统一体，与环境或其他智能体的交互与关联。

超智能体系统的**运行环境**包括：**单一的物理世界。**这是一个完全由物质、能量、运动和信息构成的世界。在这一单一的物理世界中，现存的丰富多样的有形或无形的物质，在由各种物质能所生

发的各种力的作用下，依客观的物理学原理，进行着多样的变化和运动；并以客观信息的形式进行着自我的展示。其运动包括机械的运动、声光的运动、电磁的运动、分子与核子（量子）的运动等。**有生命的自然世界**。即有生命体存在的客观世界。生命体的本质在于，它是一个有自主能力，可进行自组织、自我控制、自我繁殖，依据感知信息进行自主反应的有机体；是一个具有等级层次结构，能够进行新陈代谢运动，可通过DNA-RNA-蛋白质秩序进行自我复制的自组织系统。**有高级意识存在的属人世界。**即有高级智能体—“人”--存在的现实世界。高级智能体是一类具有主观能动性和智慧的主体。他们可以主动地去认识世界并设法改造自己的环境，可恰当地处理自己生存和发展中所遇到的各种问题，创造出各种“人造物”，为自己的生存和发展积极创造有利条件。**有“高级人造智慧生命体”共存的智慧世界**。这是人类社会发展的未来世界。在未来世界，除了碳基智慧生命，预测也会有（人造）“硅基”智慧生命等的存在。由此，它将会是一个由具有高度智慧的人与具有高度智能的“人造智慧生命体”和谐共存共进的理想世界。

超智能体系统的**智能行为**包括：**智能化信息感知与处理；智能系统或其功能单元的（被动或主动）学习；智能系统或其功能单元的（功能和结构）进化（演进）；智能系统或其功能单元的行为规划与协作等。**当然，其所有行为，包括问题求解、学习与理解、经验总结、分析与综合、判断与推理、规划与决策等，都应是基于（理性）思维和智能的。

我们可以认为，**超智能体系统，是一种人-机结合的、多层次的、包含多种“智能化信息处理单元”的集成智能系统。**从某种意义上讲，超智能体系统的构建，既是智能系统研究的**“最初目标”，**也将是智能系统研究的**“终极目标”。**

19.3 关于超智能体系统中信息/知识资源构建与共享的理论研究

19.3.1 关于公共信息/知识资源构建与共享的本体理论研究

1．公共信息/知识资源系统构建的基本原则

公共信息/知识资源，是人-机综合集成超智能体系统中极其重要的组成部分。它要求我们去构建各类信息和知识资源系统，并可供集成系统在运行中随时使用。由于这些公共信息和知识资源必须满足“人-机共享”的要求，因此，它也就需要解决信息和知识的**“表达（表示）”**问题。信息和知识表达问题的本质，就是要提供一个让所有“人”和“（智能）机器”都可以对这些知识和信息进行“理解”和“处理”的统一框架；人类对知识进行表达的信息框架主要是**语言系统**；而要让机器理解人类语言，对人类的认知和语言进行某种形式的“规范化”和“形式化”将是必要的；但更为理想的方式是，让“（智能）机器”可以“理解”各类“人类自然语言”和各种“表情”。

就内容而言，客观信息所表现（展现）的是“（客观）事物运动的状态和运动状态变化的方式”；认知信息是对人所关心的某种事物的某种不确定状态的某种肯定（认定）；知识所表现的则是“某类事物运动的状态和状态变化的规律”在人头脑中的反映，是人对客观世界和社会事务的认知，是对客观事物存在状态及发展变化规律的认知，是对如何正确处理各种（社会）事务的经验，是对这些认知和经验的系统化总结。它包括对事物是什么或怎么样或会怎么样的认知，也包括对（各类）事务应该如何（恰当）处理及为什么要如此（处理）的认知。信息和知识尽管有本质的区别，但信息资源系统和知识资源系统的构建与共享，却有着诸多一致之处。

就信息和知识资源系统构建时的**设计准则**而言，系统设计的基本准则应是：系统应便于实现；系统应便于应用和共享。为了便于应用和共享，系统中信息和知识的清晰、明确和一致性的表达将

是必要的。这就要求对信息和知识要“**对齐**”。系统应清楚地向外界表明其所要传递的意义，不能含混不清；系统所要采用的语言应是在特定领域中人-机“共用”的，是由“共许语言”构成的。另外，它还应该是以“通用资源”的视角来描述、管理和共享这些信息和知识的。而为了便于实现，一致性、可扩展性、最小承诺和简洁高效的编码等，将是需要认真考虑的。其中，可扩展性要求，系统应该为可预料到的知识和信息的增加提供可扩充的基础，并可支持在已有基础上系统的更新或特殊的应用需求，而无需修改已有的（系统）结构。而为了共享，对信息和知识进行某种规范化的说明也是必须的。

2．基于本体的信息和知识资源系统的构建方法

我们知道，信息/知识资源系统构建的关键是信息/知识的表示、获取和运用(包括各种形式的加工)。现在，在人类社会的各个领域，都存在着大量的信息/知识资源系统，这些信息/知识资源系统由于应用目标或构建技术的不同，它们之间常常是千差万别的；即使是同一领域中基于相似应用目标而构建的应用系统，也往往会存在着很大的差异。导致这种差异出现的原因很多，例如，系统构建者本身的知识构成，系统所基于的技术平台，等等。这种情况直接导致了在新的信息/知识资源系统构建时，工程人员必须对同一领域中的信息/知识资源系统不断地进行“重新设计”；这既造成了人、物、财的大量浪费，也为信息和知识的共享带来很多麻烦。为了有效地解决这些问题，迫切需要有一种全新的信息/知识表示方法，这种方法要能够统一人们对各个领域中领域知识的认识，也要使计算机能尽其可能去“理解”。基于本体论(Ontology)的系统构建方法，非常适合这些需求。

一般认为，本体(Ontology)**是共享概念模型的明确的形式化规范说明**(An ontology is a formal, explicit specification of a shared conceptualization)，是对特定领域中的概念及相互间的关系的抽象描述。这里，**概念化**(conceptualization)是说，它是为了某种目的而对要表示的世界的一种抽象，是客观世界现象（认知）的抽象模型。每一种信息/知识库或表达信息/知识的系统，都应显式的或隐式的遵从于某种概念化。概念化不关心实际含义，只关心其形式化结构，并且与用来描述它的语言无关；或者说，其表示的含义是独立于具体的环境状态的。**明确**(explicit)是指，概念及它们之间的联系都需要被“**显式**”的精确定义；**我们当然可以认为ontology就是一个“词汇表”，但是，仅仅由词汇表组成的ontology是没有太大用途的，ontology必须给出这些词汇的（精准）含义；亦即，对其所使用的概念及使用这些概念的约束都要有明确的说明**。**形式化**(formal)是指，所定义的ontology应该是机器也可读的。而**共享**(share)则是指，本体中体现的应是领域共同认可的知识，是相关领域公认的概念集。它所针对的，应该是团体而不是个体。从根本上说，Ontology的作用就是为了构建领域模型，或通俗地讲，**本体就是用来描述某个领域甚至更大范围内的概念以及概念之间的关系，使得这些概念以及概念间的关系在共享的范围内具有大家共同认可的、明确的和无歧义的意义**。这样，人-机之间以及机器之间，就可以更方便地进行交流和数据共享了。

“本体”作为一种（认知）知识共享模式，为特定领域的人和应用系统的交流提供了极大的便利，也正因为如此，本体的研究和应用已迅速延伸到知识工程、自然语言处理、信息检索、智能化信息集成和知识管理、信息交换和软件工程等领域。

本体的基本建模元语包括：

• 类(Classes)，也即是概念(Concepts)。它可以用来描述任何“具体”事物；例如，“水果”，它是一个概念，描述了现实生活中的苹果、橙子等我们所熟知且常常被认为有关联的一类事物。概念也可以描述“抽象”事物；如功能、行为特性、策略和推理过程等。从语义上讲，它表示的是对

象的集合；而其定义，我们当然也可采用诸如框架(frame)结构等来表达之。其内容可包括：概念的名称，与其他概念之间的关系，以及用自然语言对概念的描述等。

·**关系(Relations)**，即领域中概念之间的交互关联和作用。它在形式上可定义为n维笛卡儿积的子集：

$$R：C_1\times C_2\times\cdots\times C_n$$

在语义上，关系对应的是对象元组的集合。本体中存在着四类基本关系：

part-of，kind-of，instance-of，attribute-of

这四类基本关系是概念之间最基本的关系，几乎每个领域中的概念之间，都会存在这些关系。其中，part-of是整体与部分的关系；kind-of是概念之间的继承关系；instance-of是概念与所其所描述的某一具体事物之间的关系；而attribute-of是概念与概念具有的某一方面特征之间的关系。

当然，除了上述基本关系之外，本体论中概念之间也存在着更多其它的关系。根据领域的不同和本体应用目标的不同，我们可以明确定义本体中其他的（概念间）关系。这些（概念间）关系的确定对不同的应用目标可能会有着不同的意义。

在所有的关系中，**函数和公理**是两类较为特殊的关系。函数(functions)表达了概念之间的某种确定（性）关系，该关系的前n-1个元素可以（唯一）决定第n个元素，其形式化的定义为

$$F：C_1\times C_2\times\cdots\times C_{n-1}\rightarrow C_n$$

公理(axioms)即永真断言。函数和公理在利用本体进行逻辑推理的应用中将非常重要。

·**实例**(instances)是概念所涵盖的具体事物（或称对象），可以将实例理解为概念的具体化；例如，苹果是水果的一个实例。

上面所描述的几种本体建模元语，或已能满足本体的一般建模过程；根据实际应用情况，我们也可以引进更多的建模元语，以构造可满足实际应用的各种本体。

目前，人们所提出的本体模型结构已有很多。其本体模型：

或可以**3元组**

$$O=\{C，P，R\}$$

表示，其中，C是概念集；P是属性集；R是概念间的语义关系集。

或可以**5元组**

$$O=\{C，R，H，rel，A\}$$

定义，它是将**现实世界中的事物集映射为了概念集C，将现实世界中的关系映射为了关系集R，将概念分类树映射为了概念层次H**，将**概念间的关系映射了为概念关系集rel，将现实世界的内在规律映射为了公理集A**。该结构是把现实世界完整的“映射”到了一个本体模型结构中，或能较好地反映现实世界。

我们则主张，用于公共信息/知识资源系统构建的“本体”模型结构，最好用一个6元组来描述：

$$\Omega_C=\{C，A^C，R，A^R，H，X)$$

其中，　$C=\{c_1，c_2，\cdots\}$ 为领域的**概念集**；

$A^C=\{A^C(c_1)，A^C(c_2)，\cdots\}$ 为**概念的属性集**；

$R=\{r_1(c_1，c_2)，r_2(c_2，c_3)，\cdots\}$ 为**概念间的关联集**；

$A^R=\{A^R(r_1)，A^R(r_2)，\cdots\}$ 为**概念间关联的属性集**；

$H=\{(c_1，c_2)，(c_2，c_3)，\cdots\}$ 为**概念集C的层级关系**；

$X=\{x_1, x_2, \cdots\}$ 为**（公理）规则集**，是定义在概念、关系和属性上的一组（集体约定）规则，用于对模型构建进行约束等。

其实，将本体引入信息系统的构建中，就是要从语义层次上来考察事物的运动状态及状态的变化方式，把本体意义上的信息赋予更具体的内涵。本体对客观现实抽象本质的描述，使得它能够在**通信交流、互操作和系统化**等方面发挥重要作用，并日益成为了知识工程中各类系统构建和使用时的核心内容之一。其中，本体在交流或通信方面的作用，可以理解为是为了共享的交流（交互）。

我们知道，超智能体系统中的交流（交互）有三种方式：人与人之间的交流，人与机器之间的交流，以及机器与机器之间的交流。本体模型通过建立**信息结构**的理解共享机制，可为上述三种交流提供很好的桥梁。这里所说的**信息结构，**可以具体化为本体模型中的各种建模元语。在一个特定的领域中，理解的共享往往是通过领域术语(Domain Term)来实现的。

在互操作方面的作用是说，系统的集成往往要涉及到互操作的问题，包括不同系统的相互作用和系统内部不同模块之间的相互作用等。由于本体模型可使领域知识独立于系统的操作知识，基于领域知识系统之间或系统内部的不同模块之间进行概念映射，将有可能实现其互操作的无缝连接。

而对系统化的作用是说，系统化的工程方法如今已是一种科学化的工程管理方法，它提倡从整体出发，去合理开发、设计、实施和运行一个系统。引入本体模型，一方面可以对领域知识进行系统分析，并对领域知识进行系统地显式说明；另一方面，本体模型的形式化表示方法，也使得系统的一致性检查成为可能。

从现实来看，各类信息/知识资源系统的构建往往是基于某个领域，而且绝大多数知识是通过文本信息进行描述的。因此，在知识工程中进行本体构建，可以定位为基于某一特定领域及该领域中存在的领域文献的本体构建。一种快速有效且具有相当领域通用性的领域本体构建方法，对于发挥领域本体在知识工程中的巨大作用，从而推动知识工程在各个应用领域中的快速发展，将是具有重要意义的。

在知识工程领域，**本体模型有时也被称为领域模型(Domain Model)或概念模型(Conceptual Model)**，它是关于特定知识领域内各种对象、对象特性以及对象之间可能存在的关系的内容理论。通过对应用领域的概念和术语进行抽象，领域本体形成了应用领域中共享和公共的领域概念，可以描述应用领域的知识或建立一种关于知识的描述。本体的抽象，可以是高层次的抽象，也可以是针对特定领域的概念抽象。

就一般而言，一个领域本体系统应由以下几个方面构成：**该领域概念的层次体系、概念的属性及属性的取值范围、概念之间其他的语义关系、一定的构建规则等**。领域本体系统通过对特定领域内概念及概念间关系的明确描述，将成为人-人之间、人-机之间、机器与机器之间互相理解的语义基础。**基于本体，我们将可构建出一个领域概念网络(Concept Net)。它将是基于某一领域本体所形成的一个复杂的（抽象语义背景）网络。在该网络中，是由概念来充当结点，连接不同概念的关系将形成网络的边，领域内的公理系统将成为维持网络结构相对稳定的必要条件，而此网络也就成为了对领域信息和知识进行规范化描述的基础。**

我们也可将基于本体构建的领域概念网络看作是对信息和知识资源进行规范化描述和共享的一个框架；如此，则“**构建、悬挂或关联**”在此框架上的信息和知识，将结成一个可供大家随时共享的“**知识藤网**”。该知识藤网所对应的，就是在本体概念框架的基础上所表达的关于事物及事物状态之间的有关联关系的知识。**一个基于本体的知识藤网，其结构模型或可表达为：**

$\Phi_{\Omega}=\{O, R^{CO}, T[R^{C/O}], S \mid \Omega_C (C, A^C, E^A; R; AX)\}$

其中，Ω_C 为基于本体的概念背景框架，

$\Omega_C=\{C, A^C, E^A; R; AX\}$；

C为领域内的概念集合；其中的概念是某一对象集合的抽象描述，与该对象集合形成对应关系；该集合又可表述为：

$C=\{c_i \mid c_i \longleftrightarrow \Phi(Obj_i), i \geqslant 0\}$，即$c_i$是从对象集合$Obj_i$中抽象出来的概念，

Φ为抽象函数；

$A^C=\{A^C_i\}$为定义在C上的属性的集合，其中的A^C_i 将与c_i相对应；而A^C_i可表达为

$A^C_i=\{slot_{ij} \mid slot_{ij} \in c_i, j \geqslant 0\}$；

E^A为定义在概念C上的属性约束的集合，$E^A=\{E^A_i\}$；E^A_i又可表达为

$E^A_i=\{res_{it} \mid res_{it} \longleftrightarrow f(s_{i1}, s_{i2}, \cdots, s_{it}, \cdots, s_{in}), s_{it} \in A^C_i, t \geqslant 0\}$。

关系R是概念之间的关系集合，表达概念之间某种相互作用、相互影响的状态，包括基本关系和函数等。

$R=\{NR, FN\}$，

其中，NR为基本关系的集合，包括part-of，kind-of，instance—of，attribute—of等，它们构成了概念网络的基础层次结构；

FN是函数的集合，是定义在概念类上的映射关系，

$FN=\{f \mid f(c_1, c_2, \cdots, c_i, \cdots c_{n-1}) \rightarrow c_n, ci \in CS, n \geqslant 0\}$，

f是$c_1, c_2, \cdots, c_i, \cdots c_{n-1}$ 到c_n的n维映射函数，NR也可以理解为FN的2维特殊函数。

AX为领域内公理的集合，

$AX=\{ax \mid ax \longleftrightarrow A(c_1, c_2, \cdots, c_i, \cdots c_m), c_i \in CS, m \geqslant o\}$

O是领域内的事物集合，也即与概念C相关的**实例或对象集合**；

$O=\{Obj_i \mid Obj_i \longleftrightarrow I(c_i), n \geqslant 0\}$，

I为实例函数。

R^{CO}是**实例描述集合，**

$R^{CO}=\{r(Obj_i, c_i)\}$

$r(Obj_i, c_i)$表达事物Obj_i归属概念c_i的**关系。**

$T[R^{C/B}]$是人们所认知的C或B中事物[状态]之间的关联关系集合体。

$S[T[R^{C/B}]]$为对以$T[R^{C/B}]$表达的知识或信息的**信度**。

基于本体进行信息和知识描述有规范化和清晰化的一面，但也存在着对许多知识难以准确表达的不足。因为在现实世界中，有很多信息和知识是含有不确定性的。我们或可以将“本体”扩展为**模糊本体**或**粗糙本体**。

比如，可定义粗糙[或模糊]本体是对应于不确定信息的一类本体，是由粗糙[或模糊]概念和粗糙[或模糊]关系组成的，并认为一个粗糙[或模糊]本体也可用一个三元组：

$D=<Cz, Pz, Rz>$

表示，此时，Cz为粗糙[或模糊]概念集，Pz为属性集，Rz为粗糙[或模糊]概念之间的粗糙[或模糊]关系。

而一个粗糙概念更可进一步表示为

cz={y，l，u}

其中，y称为cz的内涵(Intent)，表示该粗糙概念所具有的属性；

u称为cz的上近似外延(upper Approximation Extent)，表示可能被cz的内涵所覆盖的对象集；

l称为cz的下近似外延(Lower Approximation Extent)，表示满足内涵y中全体属性的对象集。

粗糙关系Rz则是粗糙元组的有穷集合，它是集合叉集 P(D1)×P(D2)×…×P(Dm)的一个子集。其中，Di是属性域，P(Di)为Di的幂集。

但是，即使如此，只用本体，我们还是难以对各类信息进行更进一步地处理。因此，我们在提倡基于本体对信息和知识资源进行构建和共享的同时，更欣赏基于**因素空间理论**的信息和知识的表示和系统建构方法。

19.3.2 基于因素空间理论的信息和知识系统的构建

1. 基于因素空间理论进行信息和知识系统构建的一些考虑

因素空间理论，是由汪培庄教授首先提出和深入研究的。它为知识表示提供了一个数学工具和一个可用于认知与知识描述的形式化框架。

因素空间理论的研究者认为，对事物、信息和知识的（精准）描述首先需要有一个框架。例如大家所熟悉的物理系统，要准确描述一个物理对象的运动，就需要有一个（时空）坐标系，即建立描述其运动的“架子”；有了坐标架以后，现实世界中的物理对象及其运动，便可以“挂”在这个架子上，变成了坐标系中的一个点，一条曲线，一个曲面或一个区域。这也就是说，**描述工作是先有了“坐标系”，我们才可以应用这个“坐标系”去描述实际的对象**。对于信息和知识资源的描述，什么是其“坐标系”？怎么建立这个“坐标系”？在这个“坐标系”中如何描述实际的对象？描述的充分程度又如何？这些都是信息和知识的表示理论需要考虑的问题。因素空间理论引入了因素、因素空间、描述架和知识藤网等概念，并以它们作为对信息和知识进行统一描述的“操作平台”。不仅如此，因素空间理论还进一步研究了如何在此描述框架下对信息和知识进行处理的问题，给出了它们的“逻辑运算”、充分性测度和数学拓扑特性等。

我们一直主张以因素空间理论作为对信息和知识进行描述、表达和处理的理论基础，是因为我们坚定地认为，**人类对信息和知识进行表达的主要方式是共许的语言，关联用语言表达的知识和现有感知信息的主要是人的知识背景以及思维和理解能力**。人类语言[**自然语言**]的机器理解，也需要有一个人-机**共许的语言**和**背景知识系统**；而**要让机器可对语言进行“理解”和“处理”，这一“人-机”共许的语言必须是形式化的和可计算的**。

符号模式可实现对规范语言及确定性知识的形式化描述；逻辑系统可实现对确定性思维系统的模拟以及基于规则的“计算处理”。但对于现实中众多具有各种不确定性的信息、知识和知识系统进行表达和处理，采用单一的符号模式或逻辑方法将是相当困难的。

我们或许可以把确定性系统作为相应非确定性系统的背景框架；把非确定性系统作为在确定性系统骨架上可随机变化或认识不清的可模模糊糊的准确定性系统（若这种不确定性也可被形式化描述和计算，那就更好了）。如此，则可**在一定构架和一定变动范围内对不确定性系统进行（确定性）描述并作进一步处理了**。在这里，构架应是基本确定的，而关联在此框架上的信息和知识的状态，则应是在一定范围内（模糊）可变的。

知识的背景框架我们或可以用基于本体的基本概念框架来架构，如此，则领域知识的基本架构则最好用基于因素空间理论的“**知识藤网**”来编织。**基本概念框架是人类共许或公认的认知表达体**

系；而知识滕网则是对领域知识进行表示和处理的特定知识系统。但是，要在此认知基本网络的基础上编织一个在特定领域内大家可以共用的、“可与活生生的现实世界相匹配”的知识网络和作业（比如思维）网络，其概念与知识的状态也必须是大家认可的，且是带有某种不确定性的，是需要用**模糊集或随机集**来表达的。并且，由于在背景框架上“编织”的领域知识系统需要人-机共用，因此，其表达必须满足特定的形式化和可计算要求。

本体模型所给出的，主要是一个共许的“**认知表达基础框架**”，是一个“**共许的语义网络**”，是一个“**人-机信息交流的语义背景空间**”。**因素空间**，则是一个**可对事物状态及其关系进行认知的基础框架**，也是一个**可用于对不确定性知识进行描述的空间**；而基于因素空间理论所构建的信息和知识系统，也将会是一个可对各类信息和知识进行充分刻画和处理的“人-机共许”的知识系统。

2 因素空间理论的一些基本概念简介

在因素空间理论中，因素是一个元词汇，其研究者认为，它的含义可从归因性、解析性、描述性等角度进行刻画。从归因性的角度理解，因素这一概念主要有两层含义，其一是由结果寻找原因，这时的因素可理解为引起某种结果的事物；其二是由状态或特征选择名称，此时的因素便作为一类状态或一组特征的标号。从解析性的角度考虑，一个事物可用不同的方式从不同的侧面加以描述，得到事物的不同状态或者特征。因此，因素可以理解为解析或识别现实世界的一种方式，是一类状态或一组特征的公共标志。从描述性的角度讲，任何事物都是诸因素的交叉。这种交叉意味着可以建立一种描述事物的广义坐标框架，不同的事物将对应为这种广义坐标系中的不同的点。因此，因素可以理解为这种广义坐标**系维**的名称。

在因素空间理论中，所谓事物u与因素f**相关**，是指从f谈论u，有一个状态f(u)与之对应。我们可称(U，V]为一个左配对，如果U与V分别是由一些对象和一些因素组成的集合，且对任意u∈U，一切与u有关的因素都在V中。并进而称一个左配对(U，V]为一个配对，记为[U，V]，如果对任意f∈V，一切与f有关的事物也都在U中。若给定一个配对[U，V]，我们可以在U与V之间规定一个二元关系R：

R(u，f)=1　<=> u与f有关

称R为相关关系。若记

D(f)= {u∈U，R(u，f)=1}

V(u)= {f∈V，R(u，f)=1}

则因素f∈V可视为一个映射，它作用在一定的对象u∈U上可获得一定的状态f(u)：

f：D(f)=X(f)

u→f(u)

这里，$X(f)=\{f(u)_{u\in U}\}$叫做f的状态空间，X(f)中任何一个元素都叫做f的一个状态。

因素空间理论中的因素，还是具有一定逻辑关系的因素。因此，在各因素和因素状态之间，我们可以定义并实施各种逻辑运算。由此，因素空间的提出者所给出的因素空间的公理化定义是：

给定左配对(U，V]，取因素族$F\subset V$。一个因素空间是以一个完全的布尔代数Ψ=Ψ(∨，∧，c，1，0)为指标集的集合族$\{X(f)\}_{(f\in F)}$，它们满足公理：

(F1) X(0)={θ}，其中θ表示空状态；

(F2) $\forall T\subset F$，若（$\forall s,t\in T$)(s≠t，=>s∧t=0)，则

$x(\vee_{f\in T} f)=\prod_{f\in T}X(f)$；

(F3) 对任意f,g∈F，若f∧g=0，则∀u∈U，有

$$(f \vee g)(u)=(f(u),\ g(u))$$

这里**F叫做因素集**，**f∈F叫做因素**，X(f)叫做因素f的状态空间。

在数学上，因素空间或许只是一种抽象的数学结构，但是，它在现实中又是有着广泛的实际应用背景的理论框架。比如，如现代控制论中的状态空间，模式识别中的特征空间和参数空间，现代物理中的相空间，医疗诊断中的症候空间等，都是因素空间的特殊情形。还需要指出的是，因素空间并不是一个单一的因素状态空间，而是一簇因素状态空间，是一个联合的因素和因素状态空间。更为可贵的是，它还可被视为是一个状态可变的状态空间或维度可变的状态空间。“**变维**”，是因素空间理论的核心思想之一。

3.基于因素空间理论的知识描述方法

为了利用因素空间来对知识进行描述，这里，我们先给出一个关于**描述架**的概念：

假定要讨论一组概念C=｛α，β，γ，…｝，它们的论域记为U。取因素族V，使U与V组成一个左配对(U，V]，再取因素集F⊂V，使F是充足的，即满足条件

$$(\forall u1,u2\in U),\ (\exists f\in F)\quad (f(u1)\neq f(u2))$$

此时，我们称三元组(U，**C**，F]或(U，**C**，$\{X(f)\}_{(f\in F)}$] 为**C**的一个描述架，其中X(f)为因素f的状态空间。

我们知道，概念是人脑思维活动的基础，是知识形成的基本要素。概念主要有两种基本的描述方式：一种是指明概念所具有的本质属性的内涵方式；一种是界定符合某概念的全体对象所形成的范围的外延方式。从因素空间的观点看，概念内涵的描述及量化实质上就是用与概念相关的因素和因素状态的表现外延来量化。在人工智能系统中，要让机器来学习概念，通常就是通过输入属于某一概念的实例的不同属性或相关因素的表现外延来进行的。

设(U，**C**，$\{X(f)\}_{(f\in F)}$] 为**C**的一个描述架，若

（1） $\{X(f)\}_{f\in F}$为一个因素空间；

（2） $\forall \alpha\in$**C**，$\exists G\subset F$，使α的内涵I(α)在表现外延的意义下满足

$I(\alpha)\in$**F**$(\prod_{f\in G}X(f))$，其中G为F的独立因素集；

（3） $\forall A\in$**F**$(\prod_{f\in G}X(f))$，存在$\alpha\in$**C**，使α的内涵满足I(α)=A

则我们称(U，**C**，$\{X(f)\}_{(f\in F)}$]为闭合描述架。若（2)与（3)中的G为有限集，则称(U，**C**，$\{X(f)\}_{(f\in F)}$]为有限生成的闭合描述架。

设(U，**C**，$\{X(f)\}_{(f\in F)}$]为闭合描述架，对∀f,g∈F，f与g独立；对∀A∈**F**(X(f))，∀B∈**F**(X(g))，定义

$$(A\vee B)(x,y)=A(x)\vee B(y),\ \forall x\in X(f),y\in X(g)$$

$$(A\wedge B)(x,y)=A(x)\wedge B(y),\ \forall x\in X(f),y\in X(g)$$

$$(A--B)(x,y)=A(x)-B(y),\ \forall x\in X(f),y\in X(g)$$

$$A^{c}(x)=1-A(x),\qquad \forall x\in X(f)$$

则称A∨B以内涵析取方式定义了**C**中一个概念，记为C(A∨B)；A∧B以内涵合取方式定义了**C**中一个概念，记为C(A∧B)；A-B以内涵取差方式定义了**C**中一个概念，记为C(A-B)；A^{c} 以内涵取余方式定义了**C**中一个概念，记为C(A^{c})。

以上几种以不同的逻辑形式生成概念的方式，可统称为概念的逻辑生成方式。

设(U,**C**,$\{X(f)\}_{(f\in F)}$]为闭合描述架，f_i(i=1,…,n)为一组独立因素集，$\forall A_i\in$**F**$(X(f_i))$，i = 1,

2,…, n，定义

$$(\odot^{n}_{i=1}A_i)(x_1,x_2,\cdots,x_n) = Mn(A_1(x_1),\cdots,A_n(x_n))$$

$$\forall x_i \in X(f_i),\ i = 1,2,3,\cdots,\ n.$$

其中，Mn为n维标准综合函数，则称$(\odot^{n}_{i=1}A_i)$以内涵的标准综合方式给出了C中一个概念，记为$C((\odot^{n}_{i=1}Ai))$。

经典集合论可以描述清晰概念的外延，对于模糊概念，用Fuzzy集合进行描述则更为贴切。

给定一个描述架(U，C，F]，任取一个概念c∈C，它在U中的外延是U上的一个模糊子集$A^{\sim}\in F(U)$，$A^{\sim}$是一个映射

$$A^{\sim}：U\rightarrow[0,\ 1]$$

$$u\rightarrow A(u)$$

A(u)叫做u对c或A的隶属度。当A(u)=1时，称u绝对地符合c或完全属于$A^{\sim}$；当A(u)=0时，称u绝对地不符合c或完全不属于$A^{\sim}$。

给定一个描述架(U，C，$\{X(f)\}_{(f\in F)}$]，取一个概念α∈C，其外延为$A^{\sim}\in F(U)$，∀f∈F，记

$$f(A^{\sim})：X(f)\rightarrow[0,\ 1]$$

$$x\rightarrow f(A^{\sim})(x)==\vee_{f(u)=x}A(u)$$

$f(A^{\sim})$是表现论域X(f)的模糊子集，即$f(A^{\sim})\in F(X(f))$，称为概念α在表现论域X(f)中的表现外延。

因素空间理论认为，任何事物、关于事物的信息和知识，都是由各种要素所组成的，各种不同的因素和因素状态的联合，将构成一个广义的（联合）因素空间。任何事物所表现出的信息，都可以在此因素空间中找到一个相应的描述。此因素空间[坐标系]可以对事物的全貌进行全面地描述，每个因素[或坐标轴]将可描述事物在某个方面的信息。由此，基于因素空间理论构建信息和知识系统的建模方法的总体思路将是：**首先，基于因素空间理论，引入“描述架”作为信息和知识（资源）统一描述的“操作平台”。根据需要和实际情况，确定好事物及其相关信息和知识描述的概念、因素和因素状态空间，确定好各个概念和因素之间的关系并规定好其间的各种逻辑运算，以此为基础为各类信息和知识资源建立一个统一的描述和处理框架，并最终完成资源的划分和抽象描述。其目标则是实现各类资源在特定环境下的共享和协同服务等**。

19.4 超智能体系统中人工智能体的构建理论

19.4.1 人工智能体的构建--需要研究和考虑的主要问题

在人-机综合集成的超智能体系统中，人工智能体是指一类具有自主“意识”和“思维”能力的知识系统，它可不断地与环境发生交互作用并可在集成系统中承担一定的（智能化）信息处理功能。一个人工智能体的状态，可认为主要由其心智部件，如信念、能力、选择、意图等组成；其本身也具有自治性、对环境的交互性、协作性、可通讯性，以及自适应性和实时性等特性。而对于一个具有问题解决能力的智能体，我们所关心的主要是其所具有的知识和思维能力。由于该智能体所要面对的，是一个不断变化的环境；在这样的环境中，该智能体不仅要具备对紧急情况的及时反应能力，还要具备在一定的策略下对中短期的行为做出规划的能力，通过对外部世界和其它（智能）主体状态的分析来预测未来的能力，以及通过通讯语言实现和其他（智能）主体进行协作或协商的能力等。因此，该智能体的构建，需要考虑多个方面的问题。

人工智能体的构建，首先需要考虑的是其知识的获取、表达和运用问题。整个知识工程所研究

的内容，其实都是围绕着这些问题展开的。知识的主要作用，是可以指导人的认知和行为，让人知道什么是什么，或什么怎么样[有认知能力]；知道什么情况下应如何做或为何如此做[可引导人理性和合理地行事]；知道什么事情应让谁去做或让谁来帮忙做[能知人善用]等。在问题解决过程中，知识可以（正确）引导人的思维从问题初始状态向目标状态合理地流动。

在人类的认知体系中，信息通常是具体的、表象的，而知识则通常是抽象的、本质的。抽象的知识只能从具体的信息中提炼出来，信息也只有被抽象为知识才具有更大的价值。若信息是原材料，则知识是从信息提炼出来的抽象产物。但是，知识只是“事物运动状态及其变化规律”的表述，其本身并不能解决实际的问题。面对具体的问题及其环境（约束条件），针对预期的目标，必须要有**“智能化思维”**把知识激活成为求解问题的**“智能策略”**，进而再将**智能策略转换为求解问题的智能行为**，才能最终有效地解决实际问题。我们可以认为，任何智能体的智能行为的生成都是遵循了某种智能化的信息处理规则，只是转换的具体过程会随着问题的不同而有所不同。在一般的情况下，从初始的信息分析到提出问题解决的智能策略，都是通过知识和思维这一“桥梁”进行的。

我们可以认为，人工智能体的构建问题，本质上就是一个知识的获取、表达和运用问题，也可以认为，它本质上就是一个（人类）认知与思维的模拟问题，或者说是一个基于知识和思维的问题解决系统的构建问题。而知识和思维，主要就是要用于认知和问题解决。

有人曾把智能行为分成四大类：第一类是**刺激-反应**类，这是行为心理学家最为关注的领域，其中也包括着各种形式的较初级的联想类行为等。第二类是**逻辑思维**类，它们主要由“概念世界”而不是“感知世界”构成，其中的知识和问题可被“形式化”，并完全可以用“计算”来完成。第三类是**原则上可形式化而实际上却无法完全驾驭的行为类**，可称为**难以形式化**类，包括着那些并不能采用穷举算法处理的，因而需要启发式程序的系统（如象棋、围棋等）。第四类是**非形式化行为**类，包括有规律但无规则支持的、我们人类经常性的日常活动等。一个人工智能体，不可能解决所有问题，但它必须可完成上述智能活动中的某一类。

在问题解决过程中，智能所表现的，是“利用抽象的知识和具体的信息，生成求解问题的策略，进而解决问题达到目标的能力”。在问题解决过程中，问题，是由期望、兴趣或任务要求提出的。问题的目标状态，也是由期望、兴趣或任务要求决定的。问题的初始状态，是由现实和现有的认知构成的；而问题的求解过程，则是一个在一定的知识背景下，寻求从问题初始状态达到问题目标状态的合理途径的过程，也即是一个合理选择、智能搜索的动态过程。**知识和经验，可以引导人或系统，在没有现成的问题解决方案时应如何提出或找出问题解决的方案，而在有了问题的解决方案后，又可引导人和系统应如何去选择出一个更合适的方案等**。

人工智能体的实现有多种途径。既可以采用基于符号主义的方法（如面向推理的符号自动机），也可以采用基于联接主义的方法（如人工神经网络）等。我们主张，人工智能体的构建，最好的选择，应是**基于生理-心理复合模式**的。

19.4.2 不确定性知识获取与运用过程中的模糊随机状态集的集成与落影理论

如何去获取相关的知识，是人工智能体构建时首先需要考虑的问题。以往的研究，考虑的比较多的是知识的挖掘与学习，这里，我们将主要考虑带有各种不确定性的知识的获取、表达与运用问题，给出其基于模糊随机状态集的集成与落影意义下的理论解释及其相应的实现方法。

1.人类认知与思维过程中的随机性与模糊性

知识是人们对某类事物存在状态及其发展变化规律的认知或经验的总结，是**对多人多视角多次**

认知的概括或抽象。在此认知的过程中，我们所获得的，可以是确定性的知识[如，纯净水在1个大气压及100℃条件下一定会沸腾]，也可以是非确定性的知识[如，若向高空抛一枚硬币,当其落在地面上时，正面是否朝上是无法事先确定的]。

对知识反映客观事物的真理性及可信程度，我们或可以从多个角度来刻画，如，它是否完全反映了某类事物的存在状态及发展变化规律[知识反映客观现实的确定性]，对它的表述是否可让人准确理解此类事物的存在状态及发展变化规律[知识表达的确定性和无歧义性]，人们是否完全相信此知识并可用其对此类事物的存在状态及发展变化规律进行判定或指导自己的行动[人们对此知识的信任程度]，等等。在现实世界中，某类事物的存在状态及发展变化规律，常常是非确定性的[常带有某种随机性或混沌性---让人认识不清]；人们对某类事物存在状态及发展变化规律的抽象表达，常常是定性的[在交流、理解及运用过程中会产生一定的认知偏差---即常常具有模糊性]，对反映此类事物存在状态及发展变化规律的知识的运用常常是带有一定顾虑的[即不能完全确信]；于是，对非确定性知识，我们可以**从知识反映客观的随机-确定性、知识表达的明晰-模糊性、对知识真理性的怀疑-信任度（信度）**等方面来进行刻画。

目前，人们所说的非确定性，主要包括随机性、模糊性、不完全性、不稳定性、不一致性和未确知性等。其中，随机性和模糊性是最基本的，人工智能中所考虑的不确定性，主要就是这两种不确定性。

随机性主要源自于事物发展变化本身[因果关系、相互作用]的复杂性和人类认知的局限性[如，对复杂的信息、复杂的相互作用、复杂的因果关系认知和辨识能力的局限性，认知视野的局限性（不识庐山真面目，只缘身在此山中）等]。客观事物本身及其变化是客观的，当事物变化的因果关系复杂从而导致信息不充分或过程难以把握时，即产生随机或混沌。就现有的认知而言，若在一定条件下，我们可以断定其一定会出现或一定不会出现的事件，我们将称其为必然事件，或者说此事件具有必然性。若在一定条件下，我们无法断定其是否会出现的事件，我们则称其为随机事件，或者说此事件具有随机性(偶然性)。必然与偶然，都是相对于一定条件而言的，因为事物的发展，常常是一个随机与必然相互交织的过程。在事物发展的进程中，会有许多质的飞跃[质变]。而在质变之前，其发展，会有一个相对稳定的、合乎一定逻辑的量变过程。事物的发展历程，某种必然性会起着一定的主导作用，但随机性在很多情况下也会改变其进程。其实，任何事物的发展历程，都只是其许许多多可供选择进程中的一种，它们的发展，都具有某种随机性或偶然性。事物发展的偶然性来自何方？它既来自于事物内部，当然也有外部因素的作用。世界是物质的世界，构成事物的各种物质、各种要素以及支配或影响它们的各种作用，经常处于变动之中。如果这些变动和相互作用尚处于允许的范围之内，那么，事物的发展整体上会是稳定的、缓变的；一旦某些变动或作用超越甚至远离了其允许的范围，整个事物将会发生突变。现实世界，是一个极富于随机性的世界，所有事物都处于与其他事物的相互作用之中。但在许多场合，随机[偶然]性会受到必然性的限制，必然性会给它划定一个范围，其随机选择只能在一定范围内进行[例如，出生婴儿的性别虽是偶然的，但也只有2种可能]；另外，随机还是受一定条件限制的，随着条件的改变，必然与偶然也是可以互相转化的[如，向一巨大的目标射击，击中将是必然的；若目标不断缩小，起初击中还会是必然的，但小到某一限度后，击中目标便会成为偶然的了]。

随机不确定性既可指事物发展结果具有多种可能性，也是指在既定环境状态下人们对事件结果

的判定是处于一种不确定状态。譬如，一项决策若只产生一种可能的结果时，它是确定的；而当其可能会产生两种以上不同的结果时，不确定性也就出现了。不确定性可以是完全没有任何规律可循的纯随机事件，此时，我们无法对其做出任何判定。当然，不确定性也可与一定统计规律相联系，其出现的结果，常具有某种稳定的概率，此时，我们就可以用一个随机分布来描述该不确定性。数理统计、概率论、随机过程理论、混沌理论等，都是试图解决此类问题的。其基本思想是：事物变化多种多样，但它又受一定的框架和规律的约束；我们可能无法对其过程作出精确的描述和把握，但却可以知道其大概的情况；信息越多，对总体的把握将越精准；我们需要做的，是努力从其样本中总结出其总体特征，并尽量能从其总体特征中去把握事物发展变化的可能性。

模糊性主要源自于对事物状态和变化认知过程中原始信息的不清晰性、用认知框架[描述框架、表述框架、交流工具]进行描述时的概括性、多个认知个体认识和理解的不一致性[与视角、认知能力、利益关系相联系]等。当客观事物本身及其变化是可认知的，但我们是由多个个体从多个角度去认知它，各自认知结果又不是用准确的数据去表述时[其表达与交流只能采用某种定性的语言来进行]，就会出现模糊。人们在进行交流时，是可以理解以自然语言为基础的定性表述的，也是可以以自然语言为基础进行思维的，但机器目前**“还不完全行”**，如何对定性表述做定量的描述（和计算），是模糊信息处理技术（模糊数学）正试图解决的任务。

就一般而言，我们可以认为**模糊（性）主要源自于事物状态及变化的连续性[过渡性]与人类对其抽象表达时的离散性以及不同认知个体理解时的偏差性**。比如，概念，概念是一种划分，概念的形成乃是一个划分的过程：用差异把一方与另一方区别开来，形成一定的范畴。符合概念的全体对象所构成的集合，叫做这个概念的外延；凡此集合中全体对象所共有而不属此概念的对象都不具有的那些性质（属性）的全体，叫做这个概念的内涵（这些性质也称“本质属性”）。而划分也是一种最简单、最基本的判决过程。“水到0℃以下要结冰”，像这样一些具有突变性质的差异，具有比较明显的界面，我们可做出确定的划分，形成确切的概念和度量。但是，像“秃”与“不秃”，是不能以某一个头发数来分界的；“好人”与“坏人”，也找不到一个明确的界面。**“辩证法不知道什么绝对分明和固定不变的界限，不知道什么无条件的‘非此即彼’，它使固定的形而上学的差异互相过渡，除了‘非此即彼’，又在适当的地方承认‘亦此亦被’，并是使对立互为中介”**。**“一切差异都在中间阶段融合，一切对立都经过中间环节而相互过渡”**(恩格斯)。从差异的一方到差异的另一方，中间无疑会有一个从量变到质变的过渡过程。这种现象，就是差异的中介过渡性。由这种中介过渡性而造就出的划分上的不确定性，就是模糊性。

毫无疑问，模糊性与随机性是有本质区别的。首先，从概念涵义上看，随机性是与必然性相对，刻画的是认识对象在一定条件下所表现出来的结果方面的随机特征，而对象的性态和类属是完全确定的。比如，投币，出现正反面是随机的，但是正或反，终归是“非此即彼”的，不存在既正又反的“亦此亦彼”的可能。而模糊性是和明晰性相对，它源于事物的普遍联系和不断发展变化所表现出的两极对立的不充分性以及自身同一的相对性，所反映的是事物类属的不清晰性或事物性态的不确定性。比如，文昌鱼既不绝对属于脊椎动物也不绝对属于无脊椎动物，始祖鸟既不完全属于鸟类也不完全属于爬虫类，表现出其在类属关系上的亦此亦彼性。其次，从逻辑基础上看，随机性一方面反映事物的偶然性的因果联系，是因果律的亏损，表现为预言上的不确定性；另一方面，对它的认识又服从排中律。就拿投币来说，一方面，出现正面或反面，不服从必然的因果联系；另一方面，

要么出现正面，要么出现反面，又服从排中律，不存在第三种情况。而对于模糊现象来说，排中律就失效了，模糊的事物不是绝对的“是”与“非”，而常常是“有点是”，“可能是”等相对的“是”与“非”；模糊性是排中律的亏缺，常表现为识别上的不明晰性。当然，我们也可以说，模糊性是源于对客观事物的主观认识与事物所具有的客观属性所产生的偏差，而随机性则是由事物本身的客观特点所决定的，它包含了事物自身的产生与存在、运动与发展所具备的无规则性、不定向性以及多种可能性等。模糊与随机，都是人们从数据[信息]空间向认知空间进行映射时产生的。无论是模糊不确定性还是随机不确定性，都源自于人对事物的认知，都是使得人们在认识事物时难于甚至不能做出准确或唯一的判断。模糊性是对事物确认上的不确定性，而随机性是对事物各种可能发生结果的不确定性。模糊性主要表现在事物发生的结果上是确定的，但这种结果在人的认知中却是不清晰的或模糊的；而随机性主要表现在事物发生的结果是清晰的，但在多种可能的结果中到底会发生哪一种结果却是不确定的。最后，从研究和描述方法上看，对随机不确定性的定量研究导致了概率论和数理统计等随机数学的发展；而对模糊性的定量研究则导致了模糊信息处理技术—“模糊数学”--的发展。在“模糊数学”中，是采用隶属函数来刻画模糊性；而在“随机数学”中，是采用概率函数来描述随机性。“随机数学”，是采用对随机现象的统计性观察，从中求出概率分布；而“模糊数学”，则是依据隶属程度，来揭示事物归属的程度或估计会出现的可能性。此外，随机试验可以客观地进行，而模糊试验却常常与心理等主观因素联系在一起。

尽管模糊性与随机性是有严格区别的，但二者又有着不可分割的内在联系。实际上，在人类思维过程中，常常是随机性和模糊性兼具的，尤其是在作为人类思维和认知载体的语言中，这两者更是难以区分和独立存在的。**语言中存在着大量的随机和模糊现象。由于人类思维都是基于知识、经验和语言的，知识、经验和语言中均含模糊性和随机性，人的思维本身也极为灵活，因此，在人类的认知和思维过程中，常常是充满了多种不确定性。**

2.对模糊性与随机性的理论研究

不确定性是客观世界的基本属性之一，客观世界中的绝大部分事物和现象都具有不确定性。如何认知事物和现象的不确定性，一直是自然科学和社会科学研究的一个热点。

我们知道，从远古时代开始，人们在对自然科学技术研究的基础上，就形成和发展了以传统数学为核心的精确方法；同时，这些精确方法的应用，也推动了科学技术的巨大进步。由此，人们曾产生了一种对精确性和精确方法的“迷信”；认为，科学的方法必须是精确的；精确是科学的，而随机和模糊则是非科学的。于是，从牛顿到拉普拉斯再到爱因斯坦，其所描绘的，都是一幅幅完全确定的物理世界的图景。可是，当科学的触角扩展到接近光速的宇观世界，深入到原子内部的微观世界以及宏观世界中牛顿力学所不能解释的非可积系统的领域时，人们看到，事物的运动轨迹，主要不是表现为确定性而是不确定性。还有，当人们将精确方法应用于复杂系统研究时，更显得无能为力。但客观世界的真实情形是，简单、线性、稳定的确定性系统只是相对的，而复杂、非线性、不确定的系统则是绝对的。这些，都促使人们对传统的精确方法进行反思和改进。

模糊现象和随机现象普遍存在于现实世界。同精确性和确定性相比，模糊性和随机不确定性具有如下特征：一是**普遍性**。客观现实并不像传统数学描述的那样确定无疑；客观事物每时每刻都在运动和变化着，不管这种变化多么细微，都会表现出某种不确定性。二是**客观性**。模糊和随机是客观存在的；而精确性和确定性才是人类认识的理想产物。就对事物的认识而言，人们通常都试图采

用精确的、清晰的方法来描述事物，但客观世界并非像人们想象的那样清晰和确切。事物的模糊和随机是不以人的意志为转移的。传统数学，只能从人们假设的因果关系方面去描述实际，一旦要付诸实施，接受实践的检验，就会暴露出其难以应付随机性和模糊性的局限性。

科学研究是人类认识世界和改造世界的重要活动。从事科学研究，离不开对各种纷繁复杂现象的分析，也需要从大量的、不完全的、有噪声的、模糊的、随机的实际过程中提取出隐含在其中的、人们事先不了解的、但又是潜在有用的信息。对于自然界和人类社会中所发生的客观现象，我们或可将其归结为两大类：一类是**必然现象或确定性现象**，我们完全可以确定这些现象在一定条件下是否会出现（如，苹果从树枝上脱落后必然会落向地面），传统的数学方法和模型理论，在处理这类确定性现象方面，已有出色表现。另一类是**随机不确定性现象，**我们无法确定这些现象在一定条件下是否必然会出现（如，从装有两种不同颜色乒乓球的袋子中随机抽取一个时，不能预先确定取到的必然是某种颜色的球）。在自然界和人类社会中，存在着大量的在相同条件下可能发生也可能不发生的现象或事件。在某一个具体过程或试验中，该现象或事件是否发生并不能（预先）确定；但在大量相同条件下对同一过程的重复试验中，有些现象或事件的发生却具有一定的规律性。于是，人们提出一类随机性数学方法和模型，如概率论和数理统计等，来处理此类现象。自柯尔莫哥洛夫在测度论基础上提出并建立了概率论的公理化方法之后，概率理论已得到了人们的公认。借助随机变量的分布函数，它已可以研究随机现象的重要统计特征。**对随机性的研究目前已有概率统计、随机过程理论、混沌理论、随机对策理论等**。它们都是希望通过对宏观统计上的把握来完成对事物不确定性的认识。

系统科学尤其是复杂性科学的兴起，更进一步揭示出了复杂（动态）系统运动轨迹的不确定性特征。它的研究表明，不确定性确实是客观世界中的一种真实存在，而确定性只不过是经典理论对世界的一种简化描述而已。牛顿力学、量子力学和相对论力学中的定律，或许还可以以确定性和时间对称性为其基本特征。但是，现代科学的进一步发展已使由伽里略和牛顿所开辟的确定性方法走向了一个转折点，在这里，我们所看到的，是确定性的终结和新自然法则的诞生，**“这个（新的）法则不再是建立于确定性定律下的确定性，而是建立于‘概率’之上的（不确定性）”**。在这样的背景下，混沌理论、复杂性科学和不确定性研究，正在蓬勃发展。

现代科学所研究的复杂系统和人文系统，模糊性和复杂性已是其典型特征。因此，在研究确定性规律基础上发展起来的精确方法，在描述和处理此类系统时已无能为力，这也促使人们去探索和研究能处理模糊和复杂事物的新方法。模糊理论宣称，现实世界本质上是不精确的，是具有不确定性的，对于模糊事物只能用模糊方法来处理和描述，如用精确方法来研究和处理模糊事物，只会歪曲事物的本来面目。而复杂性科学更宣称，现实世界是复杂的，充满了不确定性。对复杂系统进行研究，必须考虑其不确定性。

不确定性包括随机性，也包括模糊性。客观世界中的模糊性和随机性虽然都是客观的、普遍的，常常依存于同一事物中，但二者之间是有着本质的不同的。比如，随机性是指事物各种可能发生结果的不确定性，而模糊性是指人对事物（类属）认知的不确定性，是事物所呈现出的“亦是亦非”抑或“似是而非”特性的反映；随机性很多是由于（客观或认知）条件不充分，使得条件与事件未能出现确定性的因果关系；而模糊性则可因**客观的因素**（即由于客观事物的中介过渡性而引起划分上的不确定性）、**信息的因素**（指在复杂系统中因各种因素交织在一起而产生的模糊性）和**主观的**

因素（指由于认知主体在认知方面的差异而引起对事物划分上的不确定性）而产生。还有，在随机性中，集合是确定的，有明确的内涵与外延；而在模糊性中，（客观）事件往往是确定的，但由于（认知）集合内涵与外延的不确定性，才使事件能否属于集合也呈现出不确定性。

对模糊性处理的要求，首先源自于对各门学科深入研究的需要。随着现代科学的发展，各门科学都迫切要求**“定量化”和“数学化”**。**“一门学科如果不能充分运用数学，便不能称其为真正的科学”**这种观点，已得到了极为普遍的认同。以往许多与数学关系不大的学科，如社会科学、心理科学和语言科学等，今天都发出了迫切要求定量化研究的呼声。这些学科由于其理论极为复杂，使得经典的数学方法很难得以应用，扎德(Zadeh)曾提出一条**互克性原理：“当系统的复杂性增加时，我们使它精确化的能力将减少，直到达一个阈值，一旦超过它，复杂性和精确性将互相排斥**。”这也就是说，**复杂性越高，有意义的精确化能力将越低；而精确化的能力越低，便意味着系统具有的模糊性越强**。现代科学既然不能迁就经典的数学方法而去改变其研究对象的复杂性，就只能求助于新的数学方法并对经典数学进行改造，使其能够处理复杂系统中大量的、复杂的模糊现象。由此，才产生了**模糊集合理论**，它给出了一种对模糊现象进行定量描述和分析运算的方法。

对模糊性进行研究的要求也源自于对人类知识、语言和思维过程的处理。人类的认知过程通常是通过自然语言和思维来进行的，但自然语言和思维本身常常是定性的而非定量的，其中充满了不确定性。比如，“美”与“丑”，“可能”、“也许”、“大概”等，都无法通过精确的数值来对它们进行描述，但这种定性概念的使用却使得人们的语言交互有了更大的理解空间和更强的认知能力。用概念的方法来把握量的不确定性，常常比数学表达更真实、更具有普遍性。这也正如扎德所指出的那样，**“如果深入研究人类的认识过程，我们将会发现，人类能够运用模糊概念，决不是一种负担，而是一种巨大的社会财富。这一点，也是我们理解人类智能和机器智能之间深奥差别的关键之所在。”**

尽管都是对不确定性进行处理的方法，但模糊理论和概率理论并不相同。概率论所研究的是“随机性”，虽然事件的发生与否并不肯定，但事件的结果却是分明的。而“模糊数学”所研究的是“模糊性”。这种模糊性并不都是由于人们的主观认识达不到对客观实际的认知造成的，也是源于事物的差异之间存在着中间过渡的过程，是人在对事物进行判断时所面临的“亦是亦非”抑或“似是而非”的不明晰判断的结果。模糊集合理论研究模糊性时，引入了集合中元素对该集合的“隶属度”的概念，希望能用“隶属度函数”来刻画元素对集合隶属关系的不确定性。

对于模糊理论与随机理论的区别与联系。我们或可以就隶属度与概率的区别与联系来加以分析说明。我们知道，某一事件的概率是指在一次试验中发生该事件的可能性。一个随机事件的概率是大量随机试验的可能情况的一个客观描述，它表示了该事件在试验中发生的客观可能性，是对某种随机不确定现象的一种客观描述。在模糊理论中，某一元素属于一个模糊子集的从属程度，则是人们对该模糊子集所对应的概念的一种认识上的反映，从属程度取多大，在很大程度上取决于人们对此概念的理解。因此，同一元素对同一模糊集，不同的人可能会赋以不同的从属程度。可见，从属程度的确定具有相当程度的主观性，于是，有些人称之为**“主观概率”**。但需要注意的是，此时虽然也用**“概率”**一词，但其含义则已完全不同。概率表示了事件发生的可能性，或说是稳定频率，可用来描述随机事件。而隶属度表示该元素属于某一个模糊集的从属程度，因而本质上是对一个“概念”的数量化描述。用隶属度来描述事物的中介过渡性质，意味着描述者对所描写的现象占有着更

多的信息。经典集合在接纳实际对象时，首先要对其模糊性进行过滤，人为地将中介过渡状态割断，划分成绝对的“是”与“非”，抛弃了事物的中介过渡信息，也造成了信息的损失。模糊理论则可保留中介过渡的信息，浮动地选择划分的阈值，并做出更为合理的划分。从表面上看，隶属函数的描述方式是把问题复杂化了，但实际上，**“简单—复杂—简单”**，乃是知识进化的公式。正是这种复杂化才给复杂问题的解决带来了“简化”的可能性。概率分布的引入是这样，隶属函数的引入也会是这样。**模糊理论并不是要把数学变成模模糊糊的东西，而是要用数学方法去研究模糊性现象，是精确性向模糊性的一种迈进（逼近）**。大量的事实表明，许多事物过分地追求精确反倒更模糊，适当的模糊反而可以达到更精确的目的。而其关键在于应如何寻求适当的“数学语言”来描述事物的模糊性。

对模糊性的研究目前已有模糊统计、模糊集合理论、模糊推理、自然语言理解和模糊信息处理等，它们都是试图用更精确的理解和表达来对模糊性进行处理，使其更逼近真实。其中，隶属函数是描述模糊性的关键。Zadeh提出，要想确定一个模糊集合A，勿需去鉴别谁是或者谁不是它的成员，只需对每个元素确定一个数$\mu_A(x)$，用这个数来表示该元素对所言集合的隶属度即可。他是用隶属度来刻画处于中介过程的事物对差异一方所具有的倾向性，是从“亦此亦彼”中提取出更准确的信息。由此，它将经典集合论里的特征函数取值范围由二值{0，1}推广到了区间[0，1]，将经典二值逻辑推广至了模糊逻辑。但是，究竟什么是“隶属程度”？隶属程度是可以度量的吗？如何客观地去确定它呢？这是人们普遍关心的一个问题。对这个问题的回答稍有不慎，便会导致人们对模糊理论的基础发生怀疑。

这种状况也有点类似于概率理论早期发展的情形。在频率稳定性的规律被人们认识以前，人们不相信概率论会是一门具有客观意义的科学。人们总是对什么叫概率？随机事件发生的可能性大小(相对于一定的条件)是可以度量的吗？如何客观地去确定它呢？存在疑问。概率论以自己发展的历程逐渐赢得了人们的信任。人们已认识到：概率论的产生体现了人类在处理必然和偶然这一对矛盾时为使必然性成为矛盾的主要方面而作的一种自觉的努力。必然性是不可能完全取代偶然性的，要想使必然的数学规律完全必然地应用到随机现象中去，那是不可能的。如果一定要求那样的话，要么，偶然性不复存在，要么，数学便永远不要去叩问随机现象的大门。正确的思想方法是用必然性向偶然性步步进逼。把偶然性从上一个层次(判断一个事件发生与否)驱赶到下一个层次(判断一个事件以多大的可能程度发生)中去。若能这样，就是一次胜利的进军。概率概念的产生，正是这种思想的体现。抓住了它，就是从偶然性中抓住了必然，就可以在偶然王国中铺设出一个必然的框架。概率是必然性的东西，它隐藏在随机事件的背后；但它并不是事物的原形，它只是一种抽象。但只要这是正确的抽象，那么，它便是更深刻、更科学地反映了客观事物的本质。现在，在那些频率稳定性规律起作用的地方，任何人都不再怀疑概率的客观意义；因为，正是频率稳定性的客观事实，肯定了概率概念的正确抽象性。

模糊理论的产生，也是人类在处理“明晰”与“模糊”这一对矛盾时为使分明性处于矛盾的主要方面而做的一种自觉的努力。在这里，“分明性”要想完全取代“模糊性”照样是不可能的。Zadeh的思想，就是用分明性向模糊性步步进逼，把模糊性从上一个层次（判断事物的是与非）驱赶到下一个层次(判断以多大程度上是)中去。Zadeh用[0，1]区间中的一个实数(或者用某个格L中的元素)来描写一个事物对一个概念的隶属程度，这是一种抽象。这种抽象是否正确呢？它是否有客观依据

呢？对于这个问题，人们所存在的疑虑或许比当初对概率所抱的疑虑更为严重。

由于隶属程度的确定，常与人的心理过程有关，人们更容易把它说成是主观臆造的东西，除非能在模糊现象中也看到某种频率稳定性规律的存在。如今，大量的模糊统计试验已经表明，在一定的条件下，模糊统计试验也具有(隶属度)频率稳定性的情况。于是，在隶属频率稳定性规律起作用的那些场合，我们也就为“隶属程度”找到了一把度量它的客观“尺子”。通过模糊统计试验所发现的隶属频率的稳定性，既说明隶属度概念具有客观的意义，也表明它是对客观现实的一种正确的抽象。

在模糊理论中，一个模糊集合A，也可以被看作是一个“可变”的普通集合的集成，

$$A=◎\{A^*\}$$

但其变化受着某一“模糊概念”的制约。此时，所谓模糊统计试验，是指这样一种试验：在每次试验下，能够获得一个确定的A^*；在不同次试验中，A^*可以不同。对于某固定元素$u_0\in U$，考察A^*是否覆盖u_0，设在n次试验中，恰有m次A^*覆盖了u_0，则称m/n为在这n次试验中A^*对u_0的覆盖频率，或称m/n为在这n次试验中u_0对A^*的隶属频率。

大量试验表明，随着试验次数n的增加，u_0对A^*的隶属频率也会呈现出一种稳定性。相对于概率统计试验中事件发生频率的稳定性而言，我们可称之为隶属频率的稳定性。隶属频率稳定时所在的数μ_0,即可以作为隶属度的客观量度，并由此可确定模糊子集A的隶属函数：

$$\mu_A(u_0)=\mu_0$$

隶属频率的稳定性提示我们，可以把A^*看成是一个随机集合，它依赖于一个隐蔽的变元ω。ω的确定，意味着全部有关因素，包括心理活动的因素的一种固定化。因为归根结底，心理过程也是一种（物质）活动过程，也是有一定的[概率]规律可循的。

我们在强调隶属度的客观意义的同时，也要注意到它在应用中的灵活性。事件的概率都是相对于一定的条件而言的，条件变了，同一事件也会有不同的概率。同样，隶属函数也是相对于一定的条件而言的；条件不一样，同一概念的隶属函数也会不一样。对于后者来说，由于有心理活动的参与，条件更加容易变化，因而，**隶属度常常处于动态之中**。人脑具有模糊识别的特点，模糊理论的一个重要任务就是要描写人的经验，将之移植于机器。人的思维过程是能动的、实践的，从实践效果中进行反馈，不断校正自己的认识，就可达到预定的目标。

模糊理论通过用隶属函数来精确刻画模糊现象的亦此亦彼性，是一种不错的选择；但是，现有的做法，大都忽略了隶属函数本身的不确定性。无论是通过统计的方法抑或是通过主观认定的方法所得到的隶属函数，都摈弃了其本身的不确定性特征。人们常常将模糊性和随机性分别进行研究，但是随机性和模糊性有很强的关联性，经常是分不开的。因此，研究人类认知过程中随机性与模糊性之间的关联性，或研究模糊随机状态集，是很有必要的。

3.模糊统计与随机统计的相似性与对偶性 模糊集的随机集落影

由于“隶属度”是模糊理论应用于实际的基石，究竟如何确定隶属函数，人们自然会求助于统计方法。不少研究发现，在对模糊概念的多人认知[理解]的Fuzzy统计实验中，其隶属度函数的覆盖频率也具有很好的稳定性，因而可用于对隶属函数的确定。但与我们所习惯的概率统计试验不同的是，常规的概率统计，多是对某些物理量进行观测，很少依赖于人的心理反映，而Fuzzy统计却与心理过程密切相关，它往往是通过心理测量来进行的。这就有一个主观性问题了。尽管已有大量的物

理心理学实验表明，通过各种感觉(视觉、听觉、味觉、嗅觉、触觉等)器官所获得的心理反映量与外界的各种物理刺激量的变化之间，存在着相当稳定的幂函数定律，说明科学的心理测量方法是可以客观地反映现实的；对于那些没有物理、化学或其它测量手段度量的非量化的对象，心理测量也已是一种重要的测量手段；在国外，各种心理测量量表已被广泛应用于社会、经济和管理系统；但是，对Fuzzy统计，我们还是有深入研究的必要。因为Fuzzy统计模型虽然可以转化为普通概率统计模型，但却带有自身的特点——在很多情况下，Fuzzy统计的每一次试验结果，都是论域上的一个子集，是一种**集值统计**。为此，汪培庄教授曾对集值统计进行了深入研究，并提出了**模糊集的随机集落影理论**。

我们知道，随机试验有几个要素：

(1)基本空间Ω，它是集中全部有关因素而形成的一个维数极高的笛卡尔乘积空间；

(2)事件A，它是Ω的一个普通子集；

(3)Ω中的一个变元ω，它的确定意味着全部因素都要各自固定在某一特定的状态上；

(4)条件S，它是对变元活动的一个限制范围。

同样，模糊统计试验也有四个基本要素：

(1)论域U；

(2)U中的一个固定元素u_0；

(3)U上的一个模糊概念α，围绕着α所作的不确定性划分，会形成U上一个可变集合A^*，而A^*的每一次固定都是对α的外延的一个近似表示；

(4)一定的条件S，它是主客观因素对概念α所作的规定性，它制约着α的运动。

在这里，模糊性主要表现在，客观因素并没有给α提供确定的界限，存在着主观因素的差异，A^*可以覆盖u_0，也可以不覆盖u_0，致使u_0对α的隶属关系不确定。

汪培庄教授认为，在概率统计与Fuzzy统计之间，存在着某种相似性和对偶性。如果我们把普通统计比喻为“圈圈固定，点子在变”的试验，那么，Fuzzy统计则是一种“点子固定，圈圈在变”的试验。这种对偶关系也促使我们来考虑这样一个问题：**能否把一种统计模型转化为另一种统计模型？研究认为，这基本上是可以肯定的。**

如果我们把集合(或事件)比作圈圈，把（可观测到的）实验结果（状态）比做点子，则随机试验是“圈圈固定、点子在变”的试验，而模糊试验则是“点子固定、圈圈在变”的试验。如图19.4.2所示，X中的圈圈乃是P(X)中的点，X中的点x对应于P(X)中的圈圈◎，我们只要把论域由X改成P(X)，便可以实现点子和圈圈的互换。

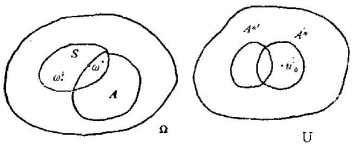


图19.4.1 两种统计试验模型的区别

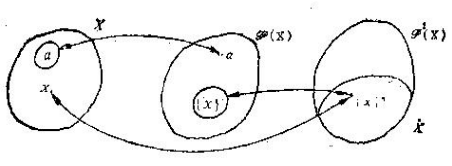


图19.4.2　圈圈变点子、点子变圈圈

这里，◎= $\{x\}^{\bullet} = \{\beta \mid \beta \in \mathscr{P}(x), \beta \supseteq \{x\}\}$

它是集合代数P(X)中的超滤。

这样，论域X上的一个Fuzzy统计模型，就可以转化成论域P(X)上的一个普通概率统计模型。按照普通统计的要求，每一次试验都是某个随机变量的一次实现。于是，我们也就只需要在论域P(X)上来定义可测结构和随机“变量”了。记

$$X^*=\{\{x\}^* \mid x\in X\}$$

给定P(X)上的一个包含X^*的σ代数β，(P(X)，β)是一个可测空间，则从某一概率场(Ω，A，Pr)到(P(X)，B)的可测映射

$$\xi：\ \Omega\to P(X)$$

$$(\xi^{-1}(C)=\{w\mid\xi(w)\in C\}\in\mathscr{A},(\forall C\in\mathscr{B})$$

便是所要求的“随机变量”。

从论域X上看问题，上述的(P(X),B)就是X上的一种超可测结构，ξ就是X上的随机集。若将从(Ω，A)到(P(X)，B)的全体随机集记作Ξ(A，B)。Fuzzy统计试验即是对随机集的实现。

随机集ξ的分布律是$\{P_\xi(C)\}(C\in B)$，让C遍历B，这一般是难以表现的。容易表现的是P_ξ在X^*中的限制。

设ξ：Ω→P(X)是X上的一个随机集(关于B可测)，记

$$\mu_\xi(x)=P\{\omega\mid\xi(\omega)\supset x\}(x\in X)$$

叫做ξ的落影。易见

$$\mu_\xi(x)=P\{\omega\mid\xi(\omega)\in\{x\}^*\}=P_\xi(\{x\}^*),\ (x\in X)$$

故知μ_ξ就是P_ξ在X^*中的限制，它是定义在X上的一个实函数。

一项集值统计试验，也是某个随机集的重复实现。而其所寻求的落影，就是某个模糊集合的隶属函数。若给定ξ∈Ξ(A，B)，对它进行n次独立观测，得到样本

$$\xi_1,\ \xi_2,\cdots\cdots\xi_n,\ \xi_i\in P(X)$$

抽象地看，它们是一组与ξ同分布的独立随机集，对任意x∈X，记

$$\overline{\xi}(x)=\frac{1}{n}\sum_{i=1}^{n}\chi_{\xi_i}(x)$$

叫做ξ对x的覆盖频率。用它，可以估计x处的落影值。而ξ(x)便是μ_ξ的估计函数。

设$X_0=\{\{x\}\mid x\in X\}\subseteq B$，对任意ξ∈Ξ(A，B)，记$\mu_\xi(x)=P_\xi(\{x\})=P\{\omega\mid\xi(\omega)\supset x\}$叫做ξ的落影函数。若ξ∈Ξ(Ω，A，X，B)为可落随机集，可以证明，可落随机集ξ的分布P_ξ可以由其落影函数μ_ξ唯一确定。

4.关于非确定性知识的表示方法的研究

对于知识的表示，人们已经提出了多种方法。实际上，如何科学地描述知识[包括概念和关系]，建立其形式化模型，一直是人工智能和认知科学关注的内容。比如，仅就概念或类别化而言，根据认知心理学对概念及其表征的研究，其目前就已有如下理论：一是**定义属性说**。主张把概念表示为**若干属性的合取**（即将定义属性以AND连接），要求这些属性中的每一个都是必需的，全部属性均必须被赋值，以便共同鉴别概念成员。二是**原型说**。所有概念均具有一个原型结构；原型可以是一组特征属性的集合或某一概念的一个或多个最佳实例。三是**样例说**。其概念类是由一组实例或样例组成的，而不是对这些实例的任何抽象描述（如对一个原型的总结性描述）。四是**解释说**。其概念是

由属性和属性之间的各种关系构成的，这些关系又形成了各属性之间的解释性联系。这些理论各有其优缺点，也都有一定的实验证据和神经心理学证据的支持，但也都有一些无法解释的事实。例如，定义属性理论无法确定什么是完全的“定义属性”，而某些概念（如“游戏”）根本就没有定义属性；原型说则无法解答抽象概念，因为某些抽象概念并没有原型结构。样例说也无法解释类包含（class inclusion）问题。

为了对知识进行描述，汪培庄教授曾提出了一种知识表示的因素空间模型。该模型作为一个知识表示框架，主张用具有特定（数学）结构（包括序结构、拓扑结构等）的因素的组合来进行知识表示。我们也一直坚持认为，**事物或其知识是可以由其相关因素和因素状态空间来进行描述的**。因素既可表征各类客体（Object）的各种特性，也可表示各种概念和关系。无论是从内涵还是外延甚至是原型或样例的视角来描述概念，因素空间理论都是一个不错的选择。大多数的概念和关系，均可构建于若干个域（Domain）的基础上；而域则可由多个因素及其状态构成。在由因素（及其状态空间）所构成的广义知识空间中，概念的属性将是其域中的一个特定的区域，而概念，则可用若干个域的特定区域的集成来表征。

知识的基础，是概念和关系。对明确的概念[如男或女]和稳定的关系[如夫妻]，我们或可以用精确状态和确定关系来表达；此时，在广义的知识空间中，我们可以用一个“点”来表达其（状态）；而对带有各种不确定性的模糊概念和非确定性知识，如非明晰概念[如好人]和非稳定关系[如是朋友]，则只能用模糊随机状态和（随机）模糊关系表达。而其在广义知识空间中的状态，将会是一个边界模糊的“区域”。在数学上，其所对应的，将是一个具有模糊随机状态的“集合”。

对客观不确定性的处理，采用基于概率的随机理论方法无疑是必要的。因为概率所反映的，实际上也是一种广义的因果律。**充分的条件会产生确定的结果，这是必然的因果律；不充分的条件虽然得不出确定的结果，但通过概率表达，各种可能的结果都被它赋予了一定的概率，这仍然体现了一种因果律。概率理论，就是从不充分的因果关系中去把握了广义的因果律—概率统计规律**。

人工智能研究的主要是人类智慧的机理和实现，对知识、思维和经验的处理是其主要内容，单纯的数理统计方法显然难以完成对此类问题的处理。在不确定性知识表示和处理方面，目前已有以贝叶斯公式为基础的贝叶斯理论，它试图利用先验知识和样本数据来获得对未知样本的估计；也有带可信度的不确定推理理论和证据理论等。证据理论是通过引入信任函数和似然函数来描述命题的不确定性，即通过一个概率区间来判断结果的可能性。当先验知识很难获得时，证据理论可以区分不确定和不知道的差异，这比单纯的概率方法更为恰当。

对于不确定性知识的表示，我们更看好模糊理论和粗糙集理论的应用。模糊理论和粗糙集理论都是从模糊性的角度出发来研究不确定性。不同的是，模糊理论是用隶属度来刻画事件亦此亦彼的程度，是从中介过度性中去寻找非中介的倾向性。而粗糙集理论则是通过上下近似集对模糊对象赋予真、假隶属函数来处理模糊性。由于人脑具有精确识别和模糊识别兼备的特点，在人工智能的研究中，人们常常是将精确知识的处理方法以各种方式来“模糊化”，以此来表示和处理各种不确定性知识，并已形成了模糊谓词、模糊规则、模糊框架、模糊语义网、模糊逻辑等；当然，也包括由模糊逻辑发展而来的**粒度计算**和**可能性推理**等。我们相信，综合利用“**可能性度量**”和“**必然性度量**”来更好地处理各种不确定性，将会是知识系统和智能系统中知识表达和运用方面今后广泛应用的工具。

5.用于不确定性知识获取的模糊随机状态集的综合集成理论

我们认为，知识的获取问题，本质上就是一个认知信息的“集成”与“精炼[抽象]”问题，也是一个分析和综合思维的运用问题。“分析”，即分别从各个不同角度来考察某一事物，考虑与此事物相关的各个方面的状况；而“综合集成”即将事物各个方面的状况进行综合考虑，并进行一番“去粗取精、去伪存真、由表及里、由此及彼”的工作，最终形成一种总的“认知”的过程。在广义认知空间中，这一分析与综合的过程，也即对用于表达此事物存在状态或变化规则的特定的模糊随机状态集进行分析认知与综合集成的过程。我们曾提出，在广义认知空间中，一个精确概念或确定性（概念）知识的状态，我们或可以用一个“点”来表达；而对带有各种不确定性的模糊概念和非确定性知识，其状态将会是一个边界模糊的“区域”。在数学上，其所对应的，将是一个具有模糊随机状态的“集合”。故而，一个带有多种不确定性的（概念性）知识的获取问题，也就可以看作是对一个具有模糊随机状态的“集合”进行获取和处理的问题了。

在广义认知空间中，我们可以将表达某一带有各种不确定性的概念或（概念性）知识的模糊随机状态“集合”，看作是特定认知空间中的某一“**云朵**”，将“人的视角”看作是认知空间中的“**太阳**”；如此，则人们从不同角度对此事物整体或某一方面的认知，则可看作是此“模糊随机状态集合”在不同认知角度下在不同“投影平面”上的“**落影**”。认知角度不同，投影平面不同，落影也会随之而变。若我们已经知道此模糊随机集在不同认知角度下在不同投影平面上的落影，采用综合集成的方法，我们是可以刻画出此模糊随机状态集合的。此时

$$\tilde{A}==(\odot^{m}_{i=1}A_i)(x_1,x_2,\cdots,x_n)=\{M(A_1(x_1),\cdots,A_i(x_j)\cdots,A_n(x_n))\}$$

其中，

M为综合函数，

⊙为集成函数，

$(x_1,x_2,\cdots,x_n)$为不同视角，

A_i 为事物在不同投影面上的状态。

$\tilde{A}$则是通过综合集成而获得的用模糊随机状态集合表达的带有各种不确定性的某一认知知识——开始是各种“模糊概念”，进一步将是各种“模糊关系”。

在广义认知空间中，将多个方面多个视角的认知进行综合集成的过程，也可看作是一种**“抽象”**或**“降维”**的过程。即，我们是将多个方面多个视角的认知，转化为了其在有限维认知表达空间中的标准化的表达，是用特定属性集合上的有限状态空间，来表达此模糊随机状态集合。

为了更准确地说明知识获取的这一过程，我们假设，在认知过程的初期，我们从多次认知中所获得的，是广义认知空间中的一种广义认知“意象”，对于此意象，我们可形式化的表达为：

Ж[对某类事物的认知]|$_{я,ю,л}$==

⊙{$Ж_{ijk}$[认知或观测数据]|$Я_i$（认知交流人群），$Ю_j$（认知视角或条件），$Л_k$（认知事例）}

其中，⊙表示综合集成。

对此意象，思维会对其做进一步的梳理，将其转化为在某一特定认知框架[离散化的、结构化的描述空间]下的一种标准表达，即：

Ж[对某类事物的认知]==

Σ{A_i[在特定描述维上的表现状态]/X_i[一定描述框架下的描述维]}

其中，Σ表示汇集或汇总。

A_i是知识[认知]在特定描述框架下的特定描述维上的表现状态，它通常是一个有限离散状态集，即 $A_i=\{a_{ij}\}$ a_{ij} 是对属性X_i的状态A_i的有限描述值。

在由因素所构成的广义知识空间中，（概念性）知识的获取过程，也即是其因素或联合因素状态的确定过程。对确定性知识，我们所获得的将是其因素或联合因素状态的确定性特征值；对非确定性知识，我们需要确定的，也是其因素或联合因素的特定状态。而这一特定因素状态，通常是一种模糊随机状态。而对此状态的表达，我们可以用语言值来表达，可以用特定定义空间上的分布函数来表达，也可以用共许的总体[期望]特征值、随机性特征值、模糊性特征值及信度特征值来表征。

若Ω是非空集合，A是由Ω的一些子集(称为事件)构成的σ代数，Pr为概率测度，则三元组(Ω，A，Pr)称为概率空间。若Θ是非空集，Π(Θ)是Θ的幂集， Pos是可能性测度，则三元组(Θ，Π(Θ)，Pos)称为可能性空间。

设ξ为**样本空间**Ω到**实数域**R的函数，若对每个Borel集B∈R，有{ω∈Ω | ξ(ω)∈B)∈A，则称ξ为**概率空间**(Ω，A，Pr)上的一个**随机变量。**

设ξ为一个从**可能性空间**(Θ，Π(Θ)，Pos)到**实直线**R上的**函数**，则称ξ是一个**模糊变量。**

设ξ是一个从**概率性空间**(Ω，A，Pr)到**模糊变量集合**的函数，如果对于R上的任何Borel集B，Pos{ξ(ω)∈ B)是ω的**可测函数**，则称ξ为一个模糊随机变量。

如果ξ是从**可能性空间**(Θ，Π(Θ)，Pos)到**随机变量集合**的函数，则称ξ是一个**随机模糊变量。**

由此，我们也可以将随机变量看作是一个从概率性空间(Ω，A，Pr)到R(实数集)的映射函数，将模糊变量看作是一个从可能性空间(Θ，Π(Θ)，Pos)到R(实数集)的映射函数，将模糊随机变量看作是一个从概率性空间(Ω，A，Pr)到模糊变量集合的映射函数，将随机模糊变量看作是一个从可能性空间(Θ，Π(Θ)，Pos)到随机变量集合的映射函数。

设ξ是一个定义在可能性空间(Θ，Π(Θ)，Pos)上的**随机模糊变量**，我们称

$$E[\xi]=\int_0^{+\infty} Cr\{\theta\in\Theta\mid E[\xi(\theta)]\geqslant r\}\,dr-\int_{-\infty}^{0} Cr\{\theta\in\Theta\mid E[\xi(\theta)]\leqslant r\}\,dr$$

为ξ的期望值(要求上式右端中两个积分至少有一个是有限的)。

可以证明，若ξ是定义在可能性空间(Θ，Π(Θ)，Pos) 上的随机模糊变量，则对于R上的任何Borel集，概率Pr{ξ(θ)∈B, θ∈Θ}是(Θ，Π(Θ)，Pos)上的一个模糊变量。假设ξ是可能性空间(Θ，Π(Θ)，Pos)上的随机模糊变量，若对于每个θ，期望值E[ξ(θ)]有限，则E[ξ(θ)]是(Θ，Π(Θ)，Pos)上的一个模糊变量。

我们还可以证明，若ξ是概率空间(Ω，A，Pr)上的模糊随机变量，则对于R中的任何Borel集B，对于任意的ω∈Ω，下述的断言成立：

(1)可能性Pos{ξ(ω) ∈ B}是一个随机变量；

(2)必要性Nec{ξ(ω)∈B}是一个随机变量；

(3)可信性Cr{ξ(ω)∈B}是一个随机变量。

显然，如果ξ是模糊随机变量，那么，Pos{ξ(ω)∈B}将是从概率空间(Ω，A，Pr)到R的可测函数，从而Pos{ξ(ω)∈B}是一个随机变量；而由Nec{B}=1-Pos{B^c}和Cr{B}=(Pos{B}+Nec{B})/2，可知，Nec{ξ(ω)∈B}和Cr{ξ(ω)∈B}也是随机变量。

更进一步地，我们还可以证明，设ξ是概率空间(Ω，A，Pr) 上的模糊随机变量，如果对于每个θ∈Θ，期望值E[ξ(ω)]是有限的，那么，E[ξ(ω)]也是一个随机变量。

6.不确定性知识获取的随机集落影理论

对于模糊集合中隶属函数的确定问题，汪培庄先生曾提出其随机集落影理论。知识的表现形式多种多样，知识的获取也就需要完成各种不同的任务，但我们都可将其归结为其因素或联合因素状态的确定问题。若表达知识的各个因素是相互独立的，我们需要确定的仅是单因素的状态，即单一变量定义域上的模糊随机状态。若表达知识的多个因素是相互关联的，我们需要做的，是确定其联合因素状态或在特定条件下的因素状态。

对联合因素状态或在特定条件下的因素状态的确定问题，我们同样可以采用随机集落影理论来形式化的说明。

设(P(X)，B_1)，(P(X)，B_2)分别是X、Y上的超可测结构，ξ∈Ξ(A，B_1)，ζ∈Ξ(A，B_2)，记

$$\mu_{(\xi,\eta)}(x,y)=P\{w\mid\xi(w)\ni x,\eta(w)\ni y\}$$

我们称之为ξ与ζ的联合落影。若$\mu_\xi(x)>0$，记

$$\mu_{\eta\mid\xi}(y\mid x)=P\{\eta\ni y\mid\xi\ni x\},$$

我们称$\mu_{\eta\mid\xi}(y\mid x)$为ζ在ξ：$\xi\ni$ x处的条件落影。显然，我们有：

$$\mu_{\eta\mid\xi}(y\mid x)=\mu_{(\xi,\eta)}(x,y)/\mu_\xi(x)$$

对于模糊随机状态集的确定问题，若我们将随机统计实验看作是“圈圈不变、点子在变”的实验，把模糊统计实验看作是“点子不变，圈圈在变”的实验，那么，模糊随机统计实验，将是一种“圈圈在变、点子亦在变”的实验。此时，我们设$\mathscr{K}=(\Omega,\mathscr{A},\mu;U,\mathscr{B},\widehat{\mathscr{B}})$是一个落影空间，假设$(U,\mathscr{B},P)$是一个概率场，于是，便有乘积概率场$(\Omega\times U,\mathscr{A}\times\mathscr{B},\mu\times p)$，对任意$s\in\mathscr{S}(R)$，有$G_s\in\mathscr{A}\times\mathscr{B}$。

此时，我们可称$(\mu\times p)(G_s)$为随机集s捕住变元u的概率，或称之为u击中s的概率。并有

$$(\mu\times p)(G_s)=E_p(\mu_{s^*})$$

这里，Ep表示对u求数学期望，s则是A的落影。

7. 知识运用过程中的模糊随机状态集的落影理论

在认知与问题解决过程中，知识的运用有多种形式和途径，但我们都可将其归结为**已有知识向特定表示形式的变换和知识向特定问题或应用空间的投射问题**。

知识向特定表示形式的变换，目的是为了将知识表达为易于理解和运用的形式。包括知识粒度的变换和知识状态的变换等。**知识粒度的变换，可采用的方法包括粒度计算和商空间理论，我们则主张采用知识在因素空间中的变维和游动理论。而对知识状态的变换，我们关心的主要是其定性表达与定量表达之间的相互映射变换，或者说在特定定义域上其自然语言表达与（可计算的）数值表达之间的转换。认为，它也是一个广义的拓扑变换问题。**

知识向特定问题或应用空间的投射问题包括两个方面，**一个是从变换中获得新的认知，二是应**

用其解决问题。

若我们已经知道了某一事物，也即知道了此事物在广义认知空间上的模糊随机状态集合，则对此事物某一特定方面知识的“**提取**”过程，可看作是已知事物的模糊随机状态集合在某一特定投影平面上的“**落影**”的获取过程。此时，知识的提取，仅与其在特定投影面上的落影及投射过程有关。

在问题解决过程中，知识的运用包括知识的直接运用和知识的间接运用。它们都将涉及到思维、目标、意愿、问题及问题的状态空间等。我们或可将直觉思维看作是经验知识的直接运用---即在已知条件[当前确认信息]下，基于知识将认知从包含不确定性的初始认知状态向不确定性得以确认的状态进行直接投射的过程；将逻辑思维看作是经验知识或理性知识的综合运用，它有一个分析与综合的过程，在很多情况下也是一个在认知空间中通过对事物各个属性[因素]状态的确认以达到对事物总体认知的形成过程。与知识的获取所不同的是，它们都是与应用--**或待解问题**--直接相关的。

我们也可以将此过程看作是特定知识在特定[已知]条件下向特定平面[问题平面]上的投射过程。如此，它也可以被看作是一个对多个模糊随机状态集合进行特定的投射和集成处理的过程。此时

$$B==(\odot_{i=1}^{m}B_i)\mid P(y_1,y_2,\cdots,y_n)=\{M(B_1(y_1),\cdots,B_i(y_j)\cdots,B_n(y_n))\}$$

其中，

M为综合函数，

⊙为集成函数，

$P(y_1,y_2,\cdots,y_n)$为包括了各种已知条件的问题平面，

B_i为各相关知识在特定条件下在特定问题平面上的投影状态。

B则是通过综合集成而获得的用模糊随机状态集合表达的带有各种不确定性的问题解答。

如此，带有各种不确定性的知识的获取与运用问题，实质上就是一个在广义认知空间中对一个或多个模糊随机状态集合进行表达、投影和集成的问题了。如何对一个事物的模糊随机状态集合进行表达、集成及落影，是一个需要深入研究的问题。对此，汪培庄先生已有过关于随机集落影方面的理论研究，我们也已有一些初步的研究。对此，我们将在后继文章中给出。

19.4.3　思维—广义认知空间中信念的有向流动理论

若将知识看作是现实世界在认知空间的映射，则基于知识的思维过程也可看作是人的信念在认知空间中的有向流动的过程。在广义认知空间中，一个关于客观世界的信息和知识的系统模型，可定义为

$$\Omega_O=\{O,\ R^{OC},\ RA,\ E\mid \Omega_C(C,\ F,\ D,\ R^C,\ H)\}$$

其中，对事物类本质的认知映射为了概念集 C，对事物类各个侧面及其状态的描述映射为了属性集 F 及属性状态集 D，现实世界事物类本质间的关联关系[也即概念间的关系]映射为了关系集 R^C，事物类关系中的分类树映射为了概念层次 H；现实世界中的事物映射为了事物集 O，事物与概念间的关系将映射为隶属关系集 R^{OC}，对现实世界事物间的规律[事物状态之间的关联及转换关系]的认知将映射为事物关联及运行规则集 RA，对所认知的事物间的关联及运行规律的把握程度则归结为信度空间集 E。该模型是将（对）现实世界（的认知）影射到一个认知集合中，若映射是基于正确的认知和思维进行的，此（认知）模型将可较好地反映客观世界。

由此，我们可以认为，建构于认知空间中的一个知识系统，将是构建在一定概念网络框架上的“**知识藤网**”。构架概念框架的元件，包括着概念和关系；对概念和关系的刻画，可统一使用“本

体”或“因素”来进行。概念可作内涵描述、外延描述或实例描述。由于人类知识体系是被系统化了的信息，基于此概念网络[常常反映了人类的认知结构]的知识将会被编织成一个“知识藤网”：**知识体系的树状结构，构成了不同的专业体系，是知识体系的“纲领”或主线，将是知识藤网中的“藤”，上面“关联”或“悬挂”着各种知识；各种知识之间的交叉与关联，将形成一个关联网络；我们合称其为知识藤网。**

基于本体或因素空间理论来构建的知识系统，若进一步将表达思维的“**软计算**”（**柔性逻辑推理**）引入，则可将静态、被动的“资源客体”变成能够进行各种“信息处理”的活的系统，此空间也就变成了“活生生”的“**认知空间**”。

在此“活生生”的认知空间中，基于知识的思维，将被看作是人的“信念”的一种有向流动。在一定知识的引导下，问题的求解过程，也即是思维在特定知识的引导下让信念在包含有不确定性的初始认知状态向期望的目标状态合理流动，最终获得被认可的认知或行为方案的过程。在这里，**分析与综合**，可看作是其在知识藤网上的“**上下游动**”，**联想、类比与基于知识的猜测，则可看作是其在知识藤网上的关联游动**。

有意识的思维活动是有期望目标的，有意识的思维活动也会是一种（目标牵引的）有向游动。所期望的目标状态可以是由意愿[需求或期望]、环境或任务产生的，而现状[环境条件现状、认知现状等]与目标之间的差异即是（思维）要解决的问题。

关于**信念有向流动**的更进一步地研究，我们将在后继文章中给出。

19.4.4 可用于复杂问题智能化信息处理的生理-心理复合模型——因素神经网络模型

19.4.4.1 生理-心理复合模型的提出

近年来，生理-心理复合模式的研究，已是智能系统研究的热点之一。从功能互补的混合系统的研究到有机合成的集成系统的研究，生理-心理复合模式正逐步显示出它的优越性。有人甚至提出，**多项技术综合集成的智能系统，将是新一代人工智能系统的关键；真正的智能行为，将产生于基于联接机制与物理符号机制的综合集成系统之中**。

我们认为，对于智能模拟，单纯的心理层面的模拟或单纯的生理层面的模拟都是有局限性的。因为人类的认知和智能行为，原本就是构建于生理和心理两个合一的层面上的，是生理层面所涌现出来的认知和心理层面上有意识思维的统一，是人类基于认知和思维的社会行为的体现。由于人类智能是人的生理活动、心理活动和社会行为在不同层面和不同方面的综合体现，因此，人们现在所提出的各种智能模拟方法，诸如，**基于心理层面抽象思维模拟的符号主义方法，基于生理层面认知涌现功能模拟的联接主义方法，基于个体社会行为层面适应性模拟的行为主义方法，基于人类社会分工协作层面功能模拟的多智能体系统方法，基于生物进化层面“成长”模拟的进化主义方法等**，都是有其积极意义的。但是，用任何单一层面上的模拟方法来独自模拟人类复杂的智能行为，都会存在着不可避免的局限性。认知心理学的研究已经表明，人类的智能化信息处理能力，是包含着上述各种信息处理能力的综合集成。因此，我们坚信，今后，对人类认知和智能行为的模拟，也必将是生理-心理-行为合一的。

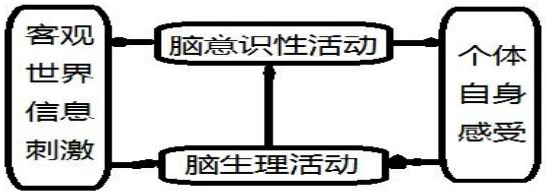


图 19.4.3 人脑的生理活动与思维意识活动

在人工智能体的构建方面，我们看重生理-心理复合模式的研究，不仅是因为，我们一直认为，人类智能行为，既是生理层面“涌现”的结果，也是心理层面“理性思考”后的体现；也是因为，我们一直认为，惟有生理-心理复合模式，才能既可对人的抽象思维进行模拟，也可对人的直觉思维、形象思维进行模拟。而人的智能活动，既有直观、形象的感知活动，又有基于逻辑思维的认知活动；两者是密切相关、缺一不可的。抽象思维是一类以概念和推理为主要形式的思维，对其进行模拟或可采用以知识工程或专家系统等为代表的符号逻辑方法；形象思维是一类以“意象”处理为主要形式的思维，这里，“意象”是对同类事物形象的一般特征的反映，对其进行模拟目前只能采用以神经网络等为代表的联接机制方法。由于抽象思维和形象思维是人类最主要的两种思维形式，并且智能模拟中的**符号主义和联接主义**方法目前也最为充实，因此，研究可模拟抽象思维和形象思维的集成智能系统将最具前途，重点进行二者的集成研究也就成为了当前最为恰当的选择。

在智能模拟方面，符号主义所采取的主要是一种构建于心理层面的还原主义思路，它试图将人类的所有心智借助研究知识并将之概念化，抽象成规则，进行符号程序编写，在计算机上进行运作，来达到完成人类心智“机器化”的目的。它已成功模拟了人类心智中逻辑思维等高级思维形式。而联接主义则表现出了一种构建于生理层面的整体主义的研究思路，它建立在“人工神经网络”的基础上，是试图通过底层网络中神经元个体的相互作用而表现出来的“涌现”现象来模拟人类心智。这种“涌现”既被认为是一种整体的行为，同时也强调各个神经元在其中的行为与作用。

在生理-心理复合模式的研究方面，目前，人们考虑比较多的，是神经网络与专家系统（知识工程）的有机结合。用于智能模拟的神经网络模型，是以生物神经系统的“功能数学模型”为基础的。作为其基本组成单元的神经元模型，是和生物神经细胞的功能相对应的。或者说，人工神经网络是用“人工神经元”网络这种抽象的数学模型来描述人类的信息加工过程的。生物神经系统的结构和功能，是人工神经网络理论的现实基础和所有“灵感”的源泉。但是，需要指出的是，无论是在结构形式或系统功能方面，现有的人工神经网络模型与真实的生物神经系统，还有很大差距。神经生理学的研究发现，生物神经元及其突触，最起码有四种以上的不同的功能行为。生物神经元的功能行为有：能处于抑制或兴奋状态；能产生激发和平台两种情况；能产生抑制后的反冲；具有适应特性。突触的功能行为有：能进行信息综合；能产生渐次变化的传送；有电接触和化学接触等多种连接方式；会产生延时激发。而目前的人工神经网络模型，仅仅是对其部分功能的模拟。所以，当前的神经网络研究，还是处于初级阶段，后边还有大量的工作等待人们去探索和研究。

生物神经元的信息加工和传递特性，有阀值特性、单向传递特性[即只能从前一级神经元的轴突末梢向后一级神经元的树突或细胞体传递]、延时传递特性[信息通过突触传递，通常会产生一定的延时]和信息综合特性[神经元对来自其它神经元的信息可进行时空综合]等。目前的人工神经网络系统尽管尚不具备上述全部功能，但也已是一个高度非线性的动力学系统了。虽然在模型中每个神经元

的结构和功能都不复杂，但是，由大量神经元按一定方式所构成的网络系统，其动态行为已是十分复杂的了。在智能化信息处理领域，人工神经网络对人们的巨大吸引力，主要体现在：并行分布处理；高度鲁棒性和容错能力；分布存储及学习能力；能充分逼近复杂的非线性关系；等等。例如，在控制领域，不确定性系统的控制问题很长时间都未能得到有效的解决；如今，人们可利用神经网络的学习能力，让它在对不确定性系统控制的过程中自动学习系统的特性，从而自动适应系统随时间的特性变异，来达到对系统的最优控制；这显然是一种十分令人满意的意向和方法。目前，神经网络的研究已向人们展显出了其美好的发展前景；不少人相信，只要持续不断地进行研究，生物神经系统的所有行为都是可以人工模拟的。

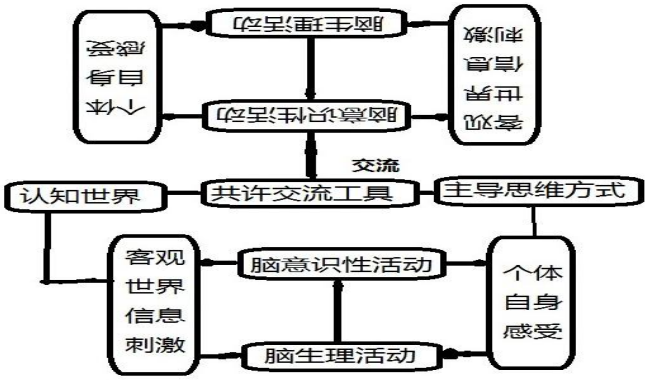


图 19.4.4　共许交流工具在认知与思维过程中的作用

我们在看好神经网络研究在智能模拟方面美好前景的同时，也要清醒的认识到，人类的智能行为，毕竟是基于知识和思维的，至少是以认知和思维为主要“显式”加工平台的。人类的智能行为，无论是识别还是理解，是感知还是思索，是应对还是协作，都离不开经验和知识，离不开形象思维和抽象思维，离不开作为思维和交流工具的语言。认知系统，才是研究和模拟人类智能的最恰当平台。因此，智能模拟，不可缺少对人类生理层面涌现现象的模拟，更少不了对心理层面认知过程的模拟。可用于复杂问题智能化信息处理的集成智能系统模型，理应既是基于生理层面的又是基于心理层面的复合模型。

19.4.4.2 关于生理-心理复合模型的一些理论研究[1905]

目前，在智能系统构建领域，关于生理-心理复合模型的研究，主要涉及到神经网络、专家系统（数据-知识-推理系统）和模糊逻辑等技术。它们的不同组合，也就构成了各种不同的智能系统。这些研究，我们可将其归结为：关于**模糊神经网络**的研究、关于**神经网络-专家系统复合模式**的研究和关于**模糊逻辑-神经网络-专家系统复合模式**的研究等。

1.关于模糊神经网络的研究

模糊逻辑和神经网络是两个截然不同的研究领域；它们的理论基础相差甚远。但是，它们都是智能化信息处理的有效方法。能否把它们结合起来而加以应用呢？若从信息处理的角度讲，是完全可以将它们结合的。把模糊逻辑和神经网络有机结合，也就产生了一个新的技术领域，这就是模糊神经网络。模糊神经网络是一种新型的网络系统，我们可以把它看作是在网络中引入了模糊量、模糊算法或模糊权系数的神经网络，也可把它看作是将原模糊系统做了某种网络化处理的系统。不管如何看待，模糊神经网络的特色，就在于把基于认知模拟的模糊逻辑方法和基于生理模拟的神经网络方法结合在了一起。

作为模糊技术与神经网络技术相结合的产物，模糊神经网络汇集了神经网络与模糊信息处理技术的优点，可集学习、联想、识别、自适应及模糊信息处理于一体，是一种很有发展前景的智能化信息处理方法。

人们对模糊神经网络的研究，主要包括：关于系统模型构建的研究、关于系统函数逼近性能的研究以及关于系统学习能力的研究等。

（1）**模糊神经网络的系统模型**

目前,关于模糊神经网络的系统模型，绝大多数都是多层前向网络模型。其区别主要在于其隶属度函数、模糊加权算子、模糊激励函数和输入/输出的形式，以及结构与参数的设定和调整方法等的不同。其类型，主要有：逻辑型模糊神经网络、常规型模糊神经网络和混合型模糊神经网络等。

模糊神经网络通常是由模糊神经元构成的。模糊神经元是具有模糊权系数，并且可对输入的模糊信号[模糊量]执行模糊运算的神经元。模糊神经元所执行的模糊运算有逻辑运算、算子运算和其它运算。模糊神经元大多是基于传统神经元的，是从传统神经元变异而来的。可执行模糊运算的模糊神经网络也多是从一般神经网络拓展而得到的。对于一般神经网络，它的基本单元是传统神经元。而传统神经元的模型通常是由下式描述的：

$$Y_i = f[\sum_{i=1}^{n} W_{ij} X_j - \theta_i] \tag{19.4.1}$$

当阀值 $\theta_i=0$ 时，有

$$Y_i = f[\sum_{i=1}^{n} W_{ij} X_j] \tag{19.4.2}$$

其中：X_j是神经元的输入；W_{ij}是权系数；$f[\cdot]$是非线性激发函数；Y_i是神经元的输出。

如果把式(19.4.2)中的有关运算改为某种模糊运算，我们就可以得到一种基于某种模糊运算的模糊神经元。比如，一种基于模糊算术运算的神经元模型可表示为：

$$Y_i = f[\bigoplus_{j=1}^{n} W_{ij} \odot X_j] \tag{19.4.3}$$

其中：⊕表示模糊加运算；⊙表示模糊乘运算。

模糊乘和模糊加是两种最基本的模糊运算，它们的定义分别如下：

模糊乘。设 N，M 是两个模糊集，它们的隶属函数分别为：$\mu_N(X)$，$\mu_M(Y)$

则 N 和 M 的模糊乘用下式表示：　　$\underset{\sim}{P} = \underset{\sim}{N} \odot \underset{\sim}{M}$

其中：符号⊙表示模糊乘运算。模糊乘 P 的隶属函数由下式给出：

$$\mu_P(Z) = \sup_{Z=X\cdot Y} T[\mu_N(X), \mu_M(Y)] \tag{19.4.4}$$

即

$$\mu_P(X\cdot Y) = \bigvee_{X\cdot Y} [\mu_N(X) \wedge \mu_M(Y)] \tag{19.4.5}$$

式(19.4.4)，(19.4.5) 所定义的，是基于扩张原理的模糊乘运算。

模糊加。设 N，M 是两个模糊集，它们的隶属函数分别为：$\mu_N(X)$，$\mu_M(Y)$

则 N 和 M 的模糊加用下式表示：　　$H = \underset{\sim}{N} \oplus \underset{\sim}{M}$

其中：符号⊕表示模糊加运算。模糊和 H 的隶属函数由下式给出：

$$\mu_H(Z)=\sup_{Z=X+Y} T[\mu_N(X),\mu_M(Y)] \qquad (19.4.6)$$

或
$$\mu_H(X+Y)=\bigvee_{X+Y}[\mu_N(X)\wedge\mu_M(Y)] \qquad (19.4.7)$$

式(19.4.6)，(19.4.7)所定义的，是基于扩张原理的模糊加运算。

对于模糊神经元中的**非线性映射**，我们也可作相似处理。

设 N 是一个模糊集，f[·]是非线性映射，则模糊集 N 的非线性映射定义如下：

$$\underset{\sim}{G}=f[\underset{\sim}{N}]$$

非线性映射结果 G 的隶属函数由下式给出

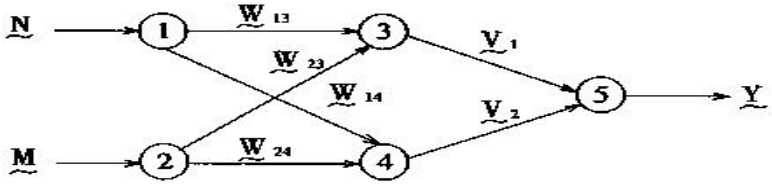
$$\mu_G(Z)=\sup_{Z=f(X)}[\mu_N(X)] \qquad (19.4.8)$$

或
$$\mu_G(f(X))=\bigvee_{f(X)}[\mu_N(X)] \qquad (19.4.9)$$

神经网络中的激发函数也被称作传递函数；传递函数通常采用 S 函数，即有

f(x)＝1/[1+exp(-X)]。

模糊神经网络中模糊量的非线性映射也可采用 S 函数；并用 f[·]表示。式(19.4.8)，(19.4.9)所定义的，即是基于扩张原理的非线性映射。



. 图 19.4.5 常规型模糊神经网络

一个采用了标准的模糊算术运算和 S 函数的模糊神经网络，常被称为常规型模糊神经网络。一个典型的常规型模糊神经网络的结构，可如图 19.4.5 所示。

对于一个常规的模糊神经网络，研究表明，如果它的模糊运算是基于扩张原理的，则这个模糊神经网络将是一个单调的网络。即

当　$\underset{\sim}{N}'\leqslant\underset{\sim}{N}$，　$\underset{\sim}{M}'\leqslant\underset{\sim}{M}$

则有　$FNN(\underset{\sim}{N}',\underset{\sim}{M}')\leqslant FNN(\underset{\sim}{N},\underset{\sim}{M})$

从上面的结论可知：对于任何从输入空间Ω到输出空间 L 的连续单调映射，都可以用基于扩张原理的常规型模糊神经网络进行逼近。反之，如果映射是非单调的，则不能用常规型模糊神经网络进行逼近。

一个更一般的模糊神经网络，我们常称其为混合(Hybrid)型模糊神经网络。在混合型模糊神经网络中，它可不像常规型模糊神经网络那样采用标准的模糊加、乘运算以及 S 函数，而是可根据需要采用各种不同的操作。比如，式(19.4.3)中的模糊算术运算也可以用某种模糊逻辑运算取代。从而也就有了“或”神经元：

$$Y_i=\overset{n}{\underset{j=1}{OR}}(W_{ij}\ \text{AND}\ X_j)\quad[\,Y_i=\bigvee_{j=1}^{n}(W_{ij}\wedge X_j)\,] \qquad (19.4.10)$$

以及“与”神经元：

$$Y_i = \underset{j=1}{\overset{n}{AND}}(W_{ij}\ OR\ X_j) \quad [\ Y_i = \bigwedge_{j=1}^{n}(W_{ij} \vee X_j)\ ] \qquad (19.4.11)$$

其实，在模糊神经网络中，我们是可以选用任何合适的操作来聚合数据，选用任何合适的函数作为传递函数去产生网络的输出的。如此构建的混合型模糊神经网络可因为它所采用的运算的不同而具有某种独特的性质，以适应不同的应用。不过，在很多情况下，我们都是把它作为一个通用逼近器来考虑的。

另外一些比较典型的模糊神经网络还包括：

模糊联想记忆网络(FAM)—它是一个二层异型相联网络，可将模糊输入映射成模糊输出，并能恢复输入模糊矢量的一些子集。这种联想记忆的神经网络是将模糊规则隐含的分布在了整个网络之中，一次模糊联想记忆就是一次模糊逻辑推理，特别适用于模式识别、模糊控制和联想记忆推理等。

模糊认知映射网络(FCM)—它可在一层网络里存储任意模糊推理模式，神经元间具有单元间的横向连接和单元自身的反馈连接，可用在监督和非监督方式下。

模糊ART（自适应谐振理论）网络--是模糊技术与自适应谐振神经网络技术的结合，可用于任意序列输入向量的分类和多维映射。它在学习新的模式时不会破坏已有的训练成果，具有很强的可塑性和稳定性。其输入向量可以是模糊量，也可以是一般数据。

(2)**关于模糊神经网络的函数逼近问题的研究**

在设计模糊神经网络时，论证设计模型的逼近性能有时是非常必要的。对模糊神经网络函数逼近问题的研究，一是希望寻找到能够作为连续模糊函数逼近器的模糊神经网络模型；二是希望能给出一个通用的可作为连续模糊函数逼近器的模糊神经网络构造型定理。**研究已经证明，普通多层前向神经网络和模糊系统(若输入/输出均为非模糊量)都是连续函数逼近器；亦即，任何定义在一个紧致集上的连续函数，都可以由多层前向神经网络或模糊系统无限逼近；以连续函数为中介，还可实现系统之间的等价互换**。这为在输入/输出均非模糊量的环境下设计模糊神经网络提供了确定的理论基础。而对于输入/输出均为模糊量的模糊神经网络的函数逼近问题的研究，则相对比较薄弱。已有研究证明，只要一个模糊神经网络的节点运算是指定的函数形式，并且输入/输出和连接权的形式为三角形模糊数，那么，该模糊神经网络可以逼近任意一个连续单调的模糊函数；不过，该结论是否可以向更一般的模糊数和模糊算子推广，目前还无法确定。而另有一些则研究证明，一类采用非单值模糊产生器的模糊系统，可以以任意精度逼近任意单变量和多变量连续函数；在满足一定条件下，四层前向正则模糊神经网络可以以任意精度逼近一个连续模糊函数。这些结果，对实际系统设计时模糊函数的选取，均具有一定的指导意义。

(3)关于模糊神经网络的学习能力的研究

模糊神经网络无论作为**“逼近器”**，还是**“模式存储器”**，都是需要学习和优化其权系数的。学习算法是模糊神经网络优化其权系数的关键。研究表明，对于纯逻辑型的模糊神经网络，可采用某种基于误差的学习算法，也即有监视的学习算法。对于常规型模糊神经网络，则可用模糊 BP 算法，遗传算法等。而对于混合型模糊神经网络，目前还尚未有合理的通用算法。

模糊神经网络的学习能力研究包括两个方面：一是有效的学习算法研究，二是学习的精度和普适性研究。一般来说，由于模糊运算的复杂性和一些算子的不连续性，在多层前向神经网络中常用的BP算法和其他一些优化算法，并不能直接应用于多层前向模糊神经网络；此时，采用某种经修改后的Fuzzy BP算法，还是可以的。为简化学习算法，模糊神经网络的许多学习算法都采用了某种经

验性的学习公式或规则，尽管它们还缺乏完备的理论，但实效还是比较突出的。

模糊神经网络的学习精度是和其对函数的逼近性能密切相关的。尽管训练样本有限，但只要网络足够大，学习算法有效，达到高精度的学习目标并非难事。不过，由于模糊神经网络的多样性，由此而导致的学习算法的普适性也就成为了一个需要认真考虑的问题。如何将非模糊学习机的普适性分析方法推广到模糊学习机，已成为模糊神经网络学习问题研究的关键之一。

2.关于神经网络-专家系统复合模式的研究

神经网络-专家系统复合模式主要有：**基于神经网络的专家系统、基于知识的神经网络系统以及神经网络与专家系统混合集成系统**。

在基于神经网络的专家系统中，其全部或部分主要功能是由神经网络来实现的。这里主要有两种方式：

① 从神经网络中抽取规则构造专家系统。这种方式是希望将神经网络的隐式“黑箱”知识表示为显式规则形式，并将这些显式知识用于推理或解释神经网络的行为。此种方法，在训练神经网络之前，并不需要深入了解领域知识的框架结构，也不需要**将领域知识的结构强加于神经网络**，其关键点是试图通过神经网络的自组织、自学习来获得用某种易于人类认识、理解的表达结构来表示的领域知识。这通常是比较困难的，因为神经网络是以分布形式隐式的表示、存储由训练而获得的知识和信息的。

② 规则知识编码于神经网络系统。这种方式比直接从神经网络中抽取规则来构造专家系统简单，其实质是将已有的领域知识利用神经网络进行优化求精，所训练的神经网络在形式上直接对应于专家系统的推理网络，可以直接用于推理和解释神经网络的结论和行为。这种集成智能系统研究和实现的关键，在于寻求一种易于编码已知知识的神经网络模型以及与其相应的有效的学习算法。

基于神经网络的专家系统的优势在于自学习和自适应能力，可以有效克服专家系统在知识获取方面所遇到的困难；其缺点在于，对神经网络拓扑结构、非线性活动函数及各种参数的选择，还缺乏系统的指导原则，对其解释能力也有待更进一步地研究。

基于知识的神经网络系统是一类将**“规则集”**表达为**“神经网络模块”**而构成的信息处理系统。在基于知识的神经网络系统的“神经模块”中，其神经元可包括“与”、“或”、“非”等逻辑神经元，前提神经元和结论神经元等，它们之间的联接权则代表着专家系统规则中的确定性因子，从而将专家系统的规则集表达为了某种“推理规则网络”。于是，基于知识的神经网络系统可视为专家系统的另一种表现形式。也就是说，它是将规则集网络化，是基于专家系统的规则来定义网络中的各种神经元及其连接，然后用相应的算法来训练网络，从而对知识进行优化。如此，我们就可以充分利用尚不完善的领域理论和带有噪音的数据来进行解释学习，从而产生出一个近似正确的规则集，用它来构造初始的神经网络系统，并继续用相应算法进行训练，来形成表达了相对正确知识的网络系统，最后，从训练好的网络中提取出最终的符号知识来。

对基于知识的神经网络的研究包括：可表达规则知识的神经网络的构建、系统中权系数的修改算法、从神经网络中提取规则的方法、神经网络中知识的增长方法等。目前，大多数基于知识的神经网络的缺点是，学习时只能改变系统的权值，而不能改变网络的拓扑结构，因而不能向不完全的初始规则集增加新的符号规则。为了使专家系统具备学习能力，以期从根本上解决知识获取的瓶颈问题，对专家系统进行合理的网络描述并开发有效的学习算法，将是十分值得重视的方向。

神经网络与专家系统混合集成系统构建的基本出发点是，将复杂系统分解成各种功能子系统，

各功能子系统分别由神经网络或专家系统来实现。其研究的主要问题包括混合集成系统的总体结构框架和选取功能子系统实现方式时的准则等。目前，其混合集成主要有两种方式：一是，从应用的角度出发，对易于获取其产生式规则的子系统使用专家系统技术，其余的功能则由神经网络来实现，此时，系统的总体结构将由实际问题来决定。二是，从功能的角度出发，用神经网络来实现专家系统的规则推理、知识获取等功能，专家系统则负责知识的显式表示和神经网络结论的验证及解释工作等。基于此，目前的一些可行的做法已包括：或建立联接产生式系统(Connectionist Product System)；或以神经网络充当规则系统的推理机，并引入知识管理器负责神经网络的训练，不断改善系统的推理能力；或将神经网络和专家系统的功能分别体现在建立混合集成系统的三个阶段中：专家系统管理和采集数据，并进行推理；神经网络分析数据，实现扩展分类；用专家系统分析和验证神经网络的结论。总之，该混合集成方法，既保持了专家系统知识表达的透明性，也具有了神经网络的学习能力和容错能力。

虽然神经网络与专家系统的结合可兼有二者之长，但同时也带来了一些新的问题，它们突出表现在：① 神经网络与专家系统的信息交互问题；② 学习过程所引发的系统可信度问题。对问题①的解决依赖于神经网络与专家系统知识表示的转换机制或同时适合两者的公共知识表示体系，对问题②的解决也有赖于不同体制下的知识的公共表示和统一的参数学习机制。无疑，妥善解决上述两个问题，已是实现神经网络与专家系统综合集成的主要前提条件之一。

3. 关于模糊逻辑-神经网络-专家系统复合模式的研究

为了提高系统智能化信息处理的能力，在多项技术综合集成的专家系统中，不仅需要有神经网络技术的支持，以提高其学习能力和对经验性知识的处理能力，还离不开基于模糊逻辑的模糊信息处理技术的支持，以提高其对不完善知识的描述能力和处理能力。比如，应用模糊集合和隶属函数来表达系统中的模糊概念，应用可能性理论和模糊逻辑理论来处理含有模糊关系的推理等。

在智能化信息处理领域，人们之所以看重模糊技术、神经网络技术和专家系统的结合，是因为它们的结合，确实可以达到扬长避短、相得益彰的目的。我们知道，专家系统(Expert System)（数据-知识-推理系统）作为智能模拟和工程化领域中最重要、最活跃的一个分支，其应用已渗透到各个领域，并已发挥了巨大的作用。专家系统是利用某一领域中专家所掌握的知识来解决该领域的问题，其成功的关键主要在于知识库的构建以及推理机制的设计。由于许多人类知识是隐式知识，无法用显式规则来表示，因此，在构造专家系统的过程中，一直存在着知识表达与获取这一“瓶颈”，它极大地限制了专家系统的应用和发展。而神经网络(Neural Network)方法是基于人脑的组织模式，将众多结构和功能极其简单的神经元通过特定方式联接成一个网络，来实现复杂的智能行为。神经网络一方面具有很强的自学习能力，能够自动地从训练样本中学习领域知识，将知识隐式编码于网络结构中；另一方面，网络又具有很强的自适应能力，能够像人脑一样实现快速的并行“推理”。但可惜的是，采用神经网络进行知识获取学习时间通常较长，难以事先确定出适当的网络拓扑结构并在学习中予以修改，也难以实现特定语义、特定结构规则知识的自动获取，而且，通过神经网络获取的知识是隐式的、分布式的包含在网络中的，系统缺乏对其所得结论的解释能力，也缺少比较实用的基于神经网络的推理方法以及对推理结果的解释方法；等等。将神经网络技术引入专家系统，一方面，可利用神经网络优良的自组织、自学习和自适应能力初步解决专家系统知识获取的“瓶颈问题”；另一方面，又可利用专家系统良好的解释机制较好地解释神经网络的行为，弥补了神经网络中知识表达的“黑箱结构问题”；二者的结合，确实是优势互补的。但问题并未因此而全部解决。

由于在现实世界中，一方面，人们对于客观事物的认识往往是定性的，是带有模糊性的；另一方面，人们又都是使用“模糊语言”来交流思想，互通信息，并据此来进行推理分析和综合评判的。这种定性的、带有模糊性的感知、交流、分析、推理和判断，正是人类思维活动的典型特征。要想模拟人类的思维和智能，就无法避免对模糊信息的处理问题。模糊技术(Fuzzy Technique)，就是一门研究如何利用模糊信息处理方法对客观事物和人类的思维等进行定量分析的信息处理技术。在专家系统中引入模糊技术，将能显著提高专家系统的性能。

在多项技术综合集成的“专家系统”中引入模糊技术和神经网络技术，之所以能显著提高原有专家系统的性能，主要是因为：

① 这种结合可较好解决知识的表示问题。在实际问题解决的过程中，常常会有大量不清晰、不严格，甚至是无法明确表达的知识，为了表示这样的知识，仅用传统的知识表示方法，显然是不够的。一般来说，神经网络易于表达感性知识[包括隐性知识]，专家系统易于表达可形式化的理性知识，模糊技术易于表达具有不确定性的模糊知识。在复杂的智能系统中，往往既包含感性知识，又包含理性知识，既有形象思维，又有抽象思维，既有确定性知识，又有非确定性知识，因而需要用神经网络、专家系统以及模糊技术共同来进行表示和处理。

② 这种结合可实现更有效的信息处理和推理。由于神经网络易于实现对连续函数的逼近和并行推理，专家系统易于实现规则匹配和基于符号的形式化推理，模糊技术易于完成对不确定性的模糊处理，通过引入模糊技术和神经网络技术，一个专家系统才能较好地模拟专家的思维方法，从而实现更恰当的推理。

③ 这种结合将使系统更具学习功能。通过引入模糊技术和神经网络技术，特别是模糊神经网络，通过修改领域专家预先定义的隶属函数和求精模糊规则，可更容易地实现系统的学习功能，提高系统的适应性能力。

④ 这种结合可更好地解决信念的传播问题。一般而言，在推理期间，可能会有多条具有相似结论的规则同时满足匹配条件。采用模糊技术和神经网络进行处理和推理，其结论的可信度将是各条规则结论可信度的综合，由于系统综合的权系数是通过实例训练而得到的，因而具有更高的可信度。

因此，以模糊技术、神经网络技术与专家系统（数据-知识-推理系统）的结合而构成的模糊神经网络、模糊专家系统以及模糊-神经网络-专家系统，已是当前基于生理-心理复合模式的智能集成系统研究的主流，也代表着未来智能系统发展的一个很有发展前途的方向。

19.4.4.3 已有生理-心理复合模式研究的不足及需要深入研究的问题

尽管**生理-心理复合模式**的研究已经取得了众多成果，但是，到目前为止，其理论体系还是很不完善，有很多问题尚需深入研究。这些问题主要有：

① **系统综合集成的理论基础研究。**理论是集成系统构建的基石。没有坚实的基础，我们就无法构建起集成智能系统的大厦。由于智能系统是用来模拟人类智能的，因此，无论是对专家系统的研究，或者是对人工神经网络的研究，或者是对模糊逻辑的研究，或者是对生理-心理复合智能系统的研究，都需要首先研究人类智能行为的本质，研究在认知与行为决策过程中，在知识的形成与运用过程中，人类大脑中的生理现象和心理现象，以及这些现象的本质和产生规律。对于复合智能系统的有机集成，还需要从本质上认识各种知识表达和智能化信息处理模型之间的内在关系和关联关系，以及它们在具体表达上的特性和相互联系，深入研究多种模型共同求解同一任务时知识在这些模型之间的动态传递机制等。因此，集成研究，应该从人类智能行为的本质入手，着眼于信息在人头脑

中的存储、加工和转化机制，着眼于知识在人脑中形成、发展和运用的生理机制和心理机制，着眼于对人类生物神经系统信息加工机制和有意识思维模式的进一步认识，将多层面的智能模拟理论与多种类的逻辑运算结合起来，在全面体现大脑思维过程的基础上，来实现人类智能的模拟。要真正实现模糊逻辑、神经网络和专家系统等多种智能技术的综合集成，就要研究抽象思维与形象思维等不同思维形式之间在功能上的互补关系，以及不同思维之间相互促进和协调一致的关系，研究人类认知过程中基于推理、演绎等抽象思维活动过程与神经网络的联想过程之间的内在关系，研究人类知识在认知层面上的显式表达与神经层面上的隐式表达之间的结构同构关系，并在此基础上研究神经网络模型与人类认知模型之间的映射关系，从而实现模糊逻辑、神经网络和专家系统等智能技术在本质上（机制上）的有机集成。

② **更符合人类认知和智能行为过程的集成模型的提出**。目前，虽然已经根据对人类神经系统信息加工机制的认识提出了很多神经网络的模型，根据对人类思维模式的认识提出了很多专家系统的知识表示模型和推理模型，根据对复杂任务的分析提出了一些实现复杂智能系统的复合智能系统模型；但是，随着我们对人类生物神经系统、人类思维模式的进一步认识，随着对人工智能系统研究的深入，研究各种新的、易于描述和实现人类认知和智能行为的各种神经元网络模型、专家系统模型、模糊系统模型和集成系统模型，最终实现具有较高智能水平的复合智能系统，仍然有大量的工作要做。复合智能系统模型的构建，仍有很长的路要走。

③ **更接近于人类思维模式的知识表示方法的研究**。知识表示是构建知识系统的基础。目前，对可形式化表达的知识，研究较多的表示方法是：产生式规则、语义网络、框架理论等，对用神经网络表示与形象思维(如模式联想)有关的知识以及易于表达人类特定知识结构的神经网络模型尚需进一步研究。要研究接近于人类思维模式的知识表示方法，就需要将显式知识和隐式知识作一体化考虑。客观世界本身是具有关联性的统一体，关联结构是对客观世界中普遍存在的关联关系的抽象。用于复合智能系统的知识表示方法，应符合人类信息存储、记忆和思维的模式特征。其基于关联的复合知识表示方法，兼容性要好，包容度要大，既可融合各种传统的知识表示方法，又可多维、多层次地表示知识，便于推理和联想，便于共享与重用。

④ **通用学习算法的研究**。在多项技术综合集成的复合智能系统中，如何利用多种方法有效地进行知识的自动获取，还是一个难题。目前，通过对神经网络进行训练所获得的知识是隐式的、分布式的，难以理解和提取；用模糊神经网络进行**知识求精**，现有的方法还难以改变网络的拓扑结构，因而难以向不完全的初始规则集增加新的符号规则；因此，研究如何让一个复合智能系统自动获取知识，具有重要的意义。

利用神经网络等进行规则（知识）提取，实质上是希望能对训练后的神经网络的运行具有解释能力，或可通过神经网络等的学习能力，让知识在网络系统和逻辑推理规则系统之间相互转换。而现有的基于神经网络的知识获取方法，尽管可以实现网络知识形态与规则系统之间的转换，但在未知规则结构的情况下如何通过自学习得到合适的神经网络结构却是一个问题。还有，对于模糊逻辑神经网络的研究，多是基于T-S模型和类T-S模型进行的，缺乏对其它模糊推理模型的神经网络实现研究，更少有其他非线性(组合)模糊系统与复杂结构神经网络之间的映射关系。再有，现有的研究方法都没有实现**特定语义、特定结构规则知识的自动获取**，没有办法引导神经网络的学习过程得到具有特定语义结构的规则知识。因此，需要在这方面进一步加强研究，否则，神经网络所获得的知识虽然能够以规则形式表示出来，但仍然不易于理解、解释和使用。另外，规则抽取算法的计算复

杂度可能会限制其应用，如何提高算法的效率，降低计算复杂度也应成为今后研究的一个方面，其可能的解决方法：或是考虑采用各种剪枝算法降低网络结构的复杂性；或是考虑通过减少搜索空间，采用各种启发式搜索策略来降低计算复杂度。

总之，规则自动抽取是机器学习领域中一个重要的研究方向，其研究具有很好的发展前途，需要深入研究下去。尽管与**规则抽取**相关的算法比比皆是，但是这些算法大都与专门的应用领域有关，缺少通用性，因此，未来的研究应以通用的规则抽取算法为主，并且尽可能的提高算法的有效性和可靠性。

⑤ **基于知识网络的搜索与推理方法**。人类基于知识网络的推理是一种具有自适应能力的并行推理。目前，尽管模糊系统和神经网络均可基于网络表示不确定性的知识，其推理可以达到不确定性并行推理的效果。但是，我们目前还是缺少接近人类不确定性推理的可靠方法。研究表明，人往往是基于**认知逻辑**和**行为逻辑**进行近似推理的，因此，人工智能必须对这种推理和搜索机制进行研究。目前，这方面的研究还较少。如何表达和模拟这类形式多样且机动灵活的推理和搜索，还存在不少困难。但研究在启发式知识的引导下，在知识网络的背景下的一体化推理与搜索技术，应是可以的，且是很有意义的。**柔性推理和搜索**是对一切逻辑形态和推理模式的、灵活的、开放的、自适应的适应性处理方法，能够描述认识全过程的思维规律，能够适应认识的发生、发展、完善和应用等各个阶段。因此，有必要进行深入研究，从而为实现更高级别的复合智能系统开辟一条新的途径。

19.4.4.4 用于智能化信息处理的生理-心理复合模型——因素神经网络模型

对于可用于智能化信息处理的生理-心理合一的认知模拟模型，我们一直主张采用因素神经网络模型。所谓因素神经网络模型，就是一类以神经网络为形式化框架，以因素神经元为基本信息处理单元的，集知识的表示、存储及应用于一体的智能化信息处理系统。

在因素神经网络理论中，因素神经网络和因素神经元是两个最基本的概念。其中，因素神经元的形式化定义如下：

一个因素神经元是一个知识表达及信息处理的基本单元，满足：

（1）可接受各型外界信息输入

$$x_i(t)=\{x_{ijk}(t)\},$$

$$x_{ijk}(t)\in x_l \ , \quad l=\{n,s,f\}$$

这里，$x_{ijk}(t)$ 是 t 时刻第 i 个单元第 j 个相关因素的第 k 种输入状态；

x_n，x_s，x_f 分别为单值型、集值型和模糊值型因素状态。

（2）单元信息感知

$$e_{ijk}=W_{ijk}(x_{ijk}(t))$$

$$E_i(t)=\{ e_{ijk}(t) \}$$

$$e_i(t)=U(e_{ijk}(t))$$

这里，$e_{ijk}(t)$是 t 时刻第 i 个单元第 j 个信道感知的第 k 种信息；

W_{ijk} 是 t 时刻第 i 个单元第 j 个信道感受第 k 种信息状态的感知函数；

$E_i(t)$ 是 t 时刻第 i 个单元感知的信息全集；

$e_i(t)$ 是 t 时刻第 i 个单元的感知强度；

$U(\cdot)$ 是信息联算子。（可以是模糊算子）

（3）存在（以因素形式表达的）单元当前内部状态

$a_i(t)= \{a_{ijk}(t)\}$

这里，$a_{ijk}(t)$是 t 时刻第 i 个单元第 j 方面的第 k 种当前因素状态。

（4）单元内部状态可激发转换

$a_i(t+1)=r_i(e_{ijk}(t)，a_{ijk}(t)\mid e_i(t))$

这里，$r_i(\cdot)$是单元内部状态激发转换函数。

（5）单元输出

$y_i(t)=T_i（ai(t)）$

$yi \in \{xl\}$　　$l \in \{n,s,f\}$

这里，T_i是单元输出响应函数；

$y_i(t)$是单元在 t 时刻的输出状态。有时，

$y_i(t)=\{ y_{ijk}(t)\}$

$y_{ijk}(t)$表示 t 时刻第 i 个单元第 j 个相关输出因素的第 k 种输出状态。

（6）单元学习规则

$\Delta w_{ijk}(t)=m_1(t_{ijk}(t),y_i(t);a_i(t),w_{ijk}(t),r_i(t),T_i(t))$

$\Delta r_i(t)=m_2(t_{ijk}(t),y_i(t),a_i(t),w_{ijk}(t),r_i(t),T_i(t))$

$\Delta T_i(t)=m_3(t_{ijk}(t),y_i(t),a_i(t),w_{ijk}(t),r_i(t),T_i(t))$

这里，$\Delta w_{ijk}(t)$、$\Delta r_i(t)$、$\Delta T_i(t)$ 分别是感知函数、状态转换函数、输出响应函数的变化；

$t_{ijk}(t)$是对单元的导师输入，$m()$是匹配函数。

上述因素神经元的定义所给出的实际上是一个面向“对象”的综合知识表示和信息处理单元。它在信息和知识的表达方面接受了知识因素表示的思想，在形式上则力图接近人脑的神经处理模型。它可接受多种形式多种状态的信息输入，这些信息可被单元有选择性的感知，形成一个总体的激发。单元内部有其特定的初始状态及状态转换规则，可对信息进行综合处理。而单元状态的变化，主要由当前记忆、输入激发状态和转换规则来决定。单元输出只与单元当前状态及输出响应规则有关（今后或可扩展至单元历史状态），单元的输入影响要经过一定时间间隔后才会在输出中发挥作用。单元具有学习功能，其学习可是有导师指导的学习，也可以是自我经验积累。

如果笼统地讲，一个因素神经网络，就是由多个因素神经元依一定规则构建而成的一种信息处理网络系统。在智能化信息处理过程中，一个因素神经元通常只能表达部分知识，也只有部分信息处理功能，要表达一个较为完整的知识系统，要完成系统化的信息处理功能，就必须使用由多个因素神经元组成的因素神经网络系统才行。

在因素神经网络理论中，因素神经网络是一个综合性的知识表示和智能化信息处理系统，它可形式化地表达为：

$FNN=<U，W，Q，\Sigma，E，P，R，T，\theta，\delta，\sigma，\alpha，\beta，\gamma，\eta>$

满足

（1）它是由多个因素神经元 $u_i=FN_i$ 按一定规则串并联组合而成的网络结构系统，即有：

$c: \quad U\times U \rightarrow W$

$(u_i，u_j) \rightarrow w_{ij} \in \mathbb{W}$

其中，$U=\{u_i \mid i=1,2,\ldots\ldots,N\}$是其组成因素神经元集合；

$W=\{w_{ij} \mid i=1,2,\ldots,N;j=1,2,\ldots\ldots,N\}$ 是其因素神经元间信息输入输出关联关系。

（2）它可接受以因素状态形式化表达的信息并进行信息的特定转换，给出系统的响应，即实现

$$\theta：\ \Sigma\times W\ \rightarrow E$$
$$(x，\ w)\ \rightarrow e\in E$$
$$\delta：\ Q\times E\times R\ \rightarrow Q$$
$$(q，\ e，\ r)\ \rightarrow q\in Q$$
$$\sigma：\ Q\times T\ \rightarrow\ P$$
$$(q，\ T)\ \rightarrow y\in P$$

这里，Σ={x｜x 为满足 FNN 输入的任一输入向量}；

E={e}为系统感知的信息输入状态；

Q={q}为系统非终态集，q 为系统组态；

R={r}为系统状态转换规则集；

T={T}为系统状态向输出状态的转换规则集；

P={y}为系统终态集。

(3) 系统学习机制

如若(x_0, y_0)是系统的一个输入输出样本，可使得

$$\alpha：\ \Sigma\times W\times R\times T\times P\ \rightarrow W$$
$$(x_0，\ w，\ r，\ T，\ y_0)\ \rightarrow \alpha(w)\in W$$
$$\beta：\ \Sigma\times W\times R\times T\times P\ \rightarrow R$$
$$(x_0，\ w，\ r，\ T，\ y_0)\ \rightarrow \beta(r)\in R$$
$$\gamma：\ \Sigma\times W\times R\times T\times P\ \rightarrow W$$
$$(x_0，\ w，\ r，\ T，\ y_0)\ \rightarrow \gamma(T)\in T$$

（4）系统评价函数 η（•）

对于系统过程状态 q_p，$q_{p-1}\in Q$，$y\in P$，满足：

$$Max\{\eta(q_p，\ y)\}\leqslant max\{\eta(q_{p-1}，\ y)\}$$

(5) 系统知识及信息表达以因素表达为基础，即

$$M(O，\ F，\ X)\ \rightarrow\ (\Sigma，\ P)$$
$$R(RM，\ M，\ XM)\ \rightarrow\ (\theta，\ \delta，\ \sigma)$$

因素神经网络描述了一个多单元多功能的智能化信息加工网络系统，它从功能及结构方面可反映人脑处理信息的一些主要特征。首先，它是串并联结构，并可在使用中人为地或自主地改变其结构，具有自组织特征。系统的信息处理功能可实现多种类型的信息处理，包括数值的或符号的，单值的或多值的，精确的或模糊的，并能实现有记忆的信息处理过程。所谓有记忆是指系统状态不仅与当前输入状态有关，还与过去状态有关。系统具有学习功能，其学习包括有导师指导的学习及自学习等。系统的控制实际上执行的是一种集中-分散控制机制，也包括系统主动控制和信息驱动控制等。当然，系统的交流和信息的统一处理需要有一定的规范，我们以知识的因素表示方法作为系统信息和知识表达的基础，是对其规范化处理的有力保证。

19.4.5 可用于人工智能体构建的知-行复合(智能)系统模型

本节，我们将给出一个可用于人工智能体构建的**知-行复合(智能)系统模型**。该模型是从认知和行为的角度出发，在深入研究了人类认知层面与神经层面之间的协同关系与映射关系，生理-心理复

合智能系统的知识表示方法，复合智能系统中知识的获取、转化以及运用过程，以及人类智能行为产生的生理和心理基础后，为了实现抽象思维与形象思维的统一模拟，为了实现认知与行为的统一模拟，而构建的一个集多种智能化信息处理技术于一体的复合型智能系统模型。它可以看作是因素神经网络理论的拓展，也可看作是对模糊逻辑、神经网络和专家系统技术的综合运用。因为它也是基于模糊技术、神经网络技术和专家系统技术构建的。

人工智能体模型的构建，离不开对人类认知和有意识行为过程中生理和心理状态及其变化的研究，离不开对人类认知和有意识的行为过程中思维和信息处理过程的解析。在人类认知和有意识的行为过程中，信息在人脑中的形成和转化过程可如图19.4.6和图19.4.7所示。从图中，我们可以看到人类信息处理的最基本的思维过程和思维方法。在这里，**意识、知识和思维能力**是最重要的部分，它们是整个有意识的认知和行为过程的基础。

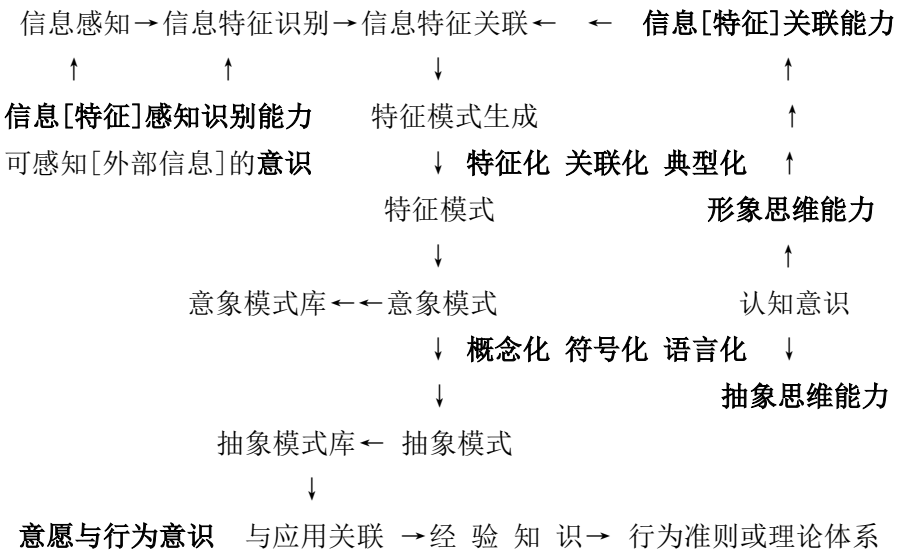
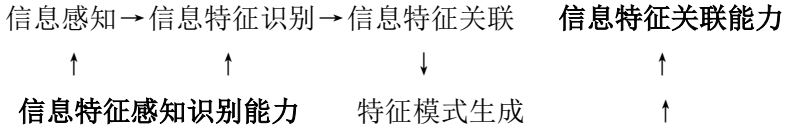


图19.4.6 认知与知识在人脑中生成的过程[信息在人脑中形成和转化的过程之一]

对上述认知和知识的形成过程，我们可归结为：（1）**信息感知**：感觉器官获取信息的原始状态和(时空)拓扑结构并转化为神经脉冲；含有外部信息状态和(时空)结构特征的原始信息的神经冲动进入大脑，大脑记住并积累这些外部事物信息的原始状态和结构信息。（2）**意象形成**：大脑对事物原始(时空)信息进行形象化：大脑抽取出事物原始信息的状态和拓扑结构的典型特征，经关联后在脑中形成表达事物状态和结构的最原始的意象和关联。（3）**抽象模型初步形成**：在多次认知的基础上对某类事物的意象信息进行逐次特征化、抽象化，并用特征和符号表达之，即形成初始的概念。（4）**系统性认知和知识形成**：在与外部世界进行的反复交流中，智能体获得对外部世界事物的本质性认识以及外部事物间的相互关系，逐步生成一个系统化的概念和关系网络。它们构成了人对外部世界的看法和认识。当这些看法和认识经过多次反复固定下来之后，即构成了一个相对稳定的知识系统。（5）**信念和能力生成**：当认知进一步升华，已成为指导人的思维和行动的规则的时候，它们将构成人的心理不可分割的成分，成为人的信念和能力。若进一步生成意志、理想等高级心理因素，也就进入了智慧人生。



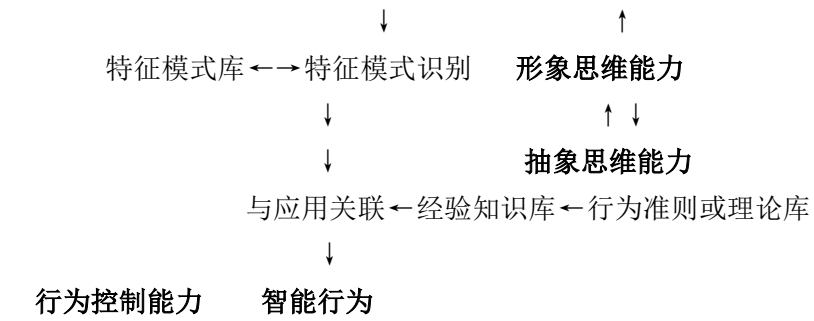


图19.4.7 决策与智能行为在人脑中生成的过程[信息在人脑中形成和转化的过程之二]

对决策和智能行为在人脑中形成的过程，我们也可归结为：

（1）**当前状况信息感知**：感觉器官提取所关心的、与当前要处理问题有关联的事物的当前状况信息的原始状态和拓扑结构，并转化为神经脉冲；含有外部事物信息状态和结构特征的原始信息的神经冲动进入大脑。

（2）**意象或模式匹配**：大脑对事物当前状况信息进行模式匹配：大脑抽取出所感知信息的意象状态和拓扑结构，与记忆中的相关模式[意象或模式]进行匹配，对事物及其状态进行识别，形成对当前现实状况的判断。

（3）**知识运用**：对现实与目标之间的差异(问题)进行分析评估，在知识的引导下，寻求解决问题的方法。

（4）**决策形成**：通过反复思考以及与外部事物信息的反复交流，在思维中形成一个相对稳定的问题解决方案；

（5）**决策实施与经验总结**：将决策方案进行实施，根据实践结果不断调整认识和问题解决方案，使认知进一步升华，使问题得到解决，也通过实践对知识进行更新。

根据认知与行为决策中信息在人脑中的形成和转化过程，我们或可将人脑的思维和信息处理系统抽象成四大层面：

（1）**原始信息获取与存储层面。**它将由系统所获取的事物的全部特征信息及其相互之间的拓扑关系[时空拓扑结构]构成。所谓特征信息，是指一个事物所特有的、能够区别于其他事物的信息。

（2）**信息集约[关联]处理与多视角认知层面。**是对事物原始信息经分析、综合和抽象而获得的较高层次的信息的集约[关联]表达[包括结构和关系]和从多个角度对原有信息的更深刻的认知。该过程一般是一类嵌套的和递归的过程，并且常常是并行发生的。原始信息经过这种处理过程后，就会产生出一类属于心理层面的、互相关联的、具有特定[时空拓扑]结构的多层次的信息特征群—觉知或认识。当这些觉知和认识形成稳定的状态和关系后，即构成经验和知识。

（3）**认知知识化和能力化层面**。经过多种思维过程和思维方法的处理，由原始信息得到的感知和认识，经过进一步地分析和处理，将形成一种比较全面的完整的认识，即多视角的全信息空间认知。这种认识之中具有相对稳定性的一些关系，会进一步形成一种较为稳定的认知和信念，继而逐步形成人的认知能力、思维能力和行为决策能力。

（4）**知识和能力应用层面。**此时，认知和知识将成为指导人们行为的准则，并应用于实践。

我们所提出的知-行复合(智能)系统模型，也希望能具有上述信息处理的特点和功能。因此，这一智能系统，将主要由以下部分构成：

（1）**基于特定感知框架的智能化信息感知系统。**该系统负责从环境中接受和获取相关事物的状

态和结构特征信息，并将其提交给智能分析系统去进行处理。该系统模块要有一个能够探测外部事物原始状态特征及其（时空）拓扑结构的感知功能系统，还要有一个特定的感知框架和特定的（临时）存储体系。这个存储系统要能够存储由感知系统探测到的原始信息，包括事物原始状态及其（时空）拓扑结构或拓扑性质等。

（2）**基于知识、思维和认知框架的智能化信息分析与认知系统**。这是系统分析与处理认知信息的部分，也是智能系统认知的核心功能。它既可对原始信息及其拓扑结构进行**特征化、关联化、典型化**处理，可对原始信息及其拓扑结构进行**概念化、符号化和语言化处理；**既可通过对信息的深入分析，形成意象、模式和概念，也可通过对相关事物间变化关系的分析，形成关联关系和规则等。

（3）**认知知识[经验与知识]的记忆、积累与转换系统**。系统首先具有存放功能，存放由认知思维所获取的意象或符号，以及由分析系统给出的特征信息及其拓扑结构等[**知识或规则]；这些知识和规则**可以是精确的，也可以是模糊的；可以用网络隐式存储，也可以显式规则记忆。系统还具有知识**转换模块。**该模块可进行知识模式或表达方式的转换，特别是基于联接主义的神经网络表示与基于符号主义的“规则”表示之间的相互转化，可实现从网络中提取符号规则以及将符号规则表示成特定网络结构的功能，从而实现人类知识在认知层面和感知（生理）层面的紧密融合。

（4）面向应用的**思维与决策系统**。该系统也是智能系统的重要核心，可模拟完成人脑在行为决策过程中的分析、综合和推理等功能，从而实现从信息分析到模式匹配，从问题分析到智能决策这一问题解决和知识运用的过程。其思维和决策行为，都是基于经验和知识的，是在知识和经验的引导下进行的。

（5）**智能行为执行与实践系统**。该系统负责信息的输出与活动的执行，并将实践的结果不断反馈给系统。

（6）**经验理解[总结]与知识更新系统**。该系统对外解释或回答用户提出的各种问题。对内进行经验总结，进行认知或知识更新。

（7）**系统（总体）[思维-决策-行为]控制系统**。控制整个系统各个部分的运行，是系统意识和意志的反映。

上面所给出的模型，是一个基于抽象思维和形象思维的认知与行为系统模型。因为从本质上讲，智能系统就是一个基于知识和思维的知-行系统。

我们认为，就人类智能而言，其知识是随着其对客观世界认识的深化而不断发展着的。由于信息获得的方式及抽象的程度不同，以及客观事物本身所具有的层次性，知识本身也会形成了一种多层次的结构。知识系统的多层次结构和脑内信息处理方式的多样性，决定了处理这些知识或信息的过程，必须是一种多样性的混合处理过程，同时也决定了人工智能体的知识模型，也必须是一个具有多论域和层次的、可逐级抽象与导航的、可多视角处理的全信息空间模型。

在知-行复合（智能）系统模型中，我们可将认知空间划分为已清晰认知空间、非清晰认知空间和待认知空间。已清晰认知空间中的知识，是可显式形式化表达的知识，其应用可转化为基于分明集合上的硬计算。非清晰认知空间中的知识，或可形式化的显式表达为非确定性知识[带有不确定性的模糊知识、随机性知识或粗糙知识]，其应用可转化为在不分明集合上的软计算；或可非形式化的隐式表达为潜在经验知识，其应用可转化为基于不分明模式映射的网络计算等。待认知空间中的知识，可以是由联想或猜想等引发的信度待定的知识，其运用将产生一定的信度问题。

知-行复合（智能）系统理论认为，智能是与生理、心理、知识、思维和行为密切相关的一种多

层次的整体现象，单纯地研究其中的某一个局部，或某一个层面，是难以了解智能系统的全部面貌的。因此，从认知和行为的角度出发，研究人类认知过程中抽象思维与形象思维之间的联系，研究人类知识在认知层面与生理层面之间的内在关系，并在此基础上研究人类认知与人类行为决策之间关系，才能从本质上了解智能的本质，进而实现各种智能技术的有机集成。因此，知-行复合（智能）系统应是一个很值得研究的复合智能系统模型。

关于生理与心理合一的知-行复合系统模型的实现方法，我们主张，其模型的构建，应是一种自上而下的解析和自下而上的组配相结合的进化构建方法。即，针对预定应用领域，事先储备相关知识，建立高层认知任务，然后把任务分解为子任务，实现系统的整体性设计。而对基层子任务，当其相关知识是认知空间中的确定性认知和知识，可以用显式的形式化系统来表达时，将构建一个解析型的认知网络，并实施分明集上的硬计算。当其相关知识还是认知空间中的非确定或待确定的认知和知识，不能用显式的形式化系统来表达时，则构建一个适应性较强的**柔性认知网络**来实现，并实施不分明集上的软计算。它们将共同构成认知空间上的具有自适应特性的知识表达、存储与运用合一的、软计算与硬计算相互配合的、一体化的智能化信息处理模型。

19.4.6 知-行复合（智能）系统模型构建的理论基础—知识表达的全信息空间理论

在知-行复合（智能）系统模型中，知识的表示是极其重要的一环。对此，我们提出了一个可用于多种知识形式统一表示的全信息空间理论。该理论认为，在生理-心理复合智能系统中，知识的表达模式是多种多样的，有**意象模式**、**属性-特征模式**，也有**符号模式**。为了统一表示这些模式，采用传统的知识表示方法将无法达到理想的效果，因此，我们需要有一种多论域、多层次、多模式、且视角可变的知识表示方法，这就是全信息空间集成及转换模型。

全信息空间集成及转换模型是一种可对意象模式、属性-特征模式、符号模式等统一进行表示的知识表示方法。它由意象模式空间、属性-特征模式空间、符号模式空间和逐级索引及导航系统构成，是一种多种模式表示和转换合一的知识表示方法。其中，意象模式空间是基础、属性-特征模式空间是桥梁、[语言]符号模式空间是知识的表达与运用的主体，逐级索引及导航系统则是各相关模式和模式空间关联和转换的纽带。

在相互关联的各种模式中，意象是信息模式的基础。人类视、听、触、嗅觉等感觉器官所接受的，主要是外部事物信息的状态和（时空）拓扑结构。这是生命体在自然界中长期演化的结果，也是高级生物的感官所具有的共同特征。因此，人脑也不得不对这种状态和拓扑结构进行处理。人脑对其进行处理的操作主要包括：信息（模式）的压缩编码—意象模式的生成与更新；信息（模式）的多视角分析与特征提取—属性-特征模式的生成与更新；信息模式的逐级抽象—[语言]符号模式生成与更新；等等。知-行复合（智能）系统作为一类类人智能系统，也应具有如此功能。

对模式最主要的操作是抽象。抽象是对信息的逐次“压缩”加工。研究表明：人脑在处理信息方面，并不是孤立地、单方面地、单视角地对特定信息进行处理，而是根据需要要从多个角度对整个信息集合的多方面的特征和特征关系进行综合处理。大脑对信息进行存储和处理的机制，也主要是以特定的认知框架为基础来进行的。抽象并不只是对信息的“压缩编码”，而是伴随着变粒度（比如，以求取商空间的关联关系的方式来对信息进行分类和处理，就有一个逐步抽象的过程）、变维度和变论域等的过程。

对模式最常用的操作是回忆和匹配。回忆过程常常是一个信息检索与全信息提取的过程，它常常依靠逐级索引及导航系统来进行。此时，索引及导航系统也就变成了思维的导向系统。而匹配则

常常是通过特征信息来进行的。

知识运用与更新也是对知识的多视角运用与加工。降维则包括着改变视角和论域。在很多时候，信息累积叠加对信息的意象影响并不大，它改变的主要是模式的信度和清晰度。

模式表达形式的转换也是对模式加工的重要手段，在知识提取时，它常常是一种抽象化加工；在问题求解时，它常常是一种变视角的加工。

全信息空间理论认为，客观世界原本就是一个相互关联的世界，本身就具有关联特性。拓扑结构，就是对客观世界中普遍存在的关联特性的一种抽象。人类的知识是随着对客观世界的认识的发展而不断发展着的。由于客观事物的层次性，以及人们对它们认知的逐步深入特性，人类知识也形成了一种多层次的结构关系。人类知识系统的多层次结构特性和人类个体大脑对信息的多层次认知特性，就决定了表示和处理这些知识或信息的过程，也必须是一种多层次的关联性表示和处理过程。因此，全信息空间集成及转换模型，应是信息和知识表示的最恰当方法。

19.4.7　知-行复合系统模型构建的理论基础—基于认知逻辑与行为逻辑的意向性搜索与柔性推理理论

在知-行复合系统模型中，我们是将认知空间划分为了已清晰认知空间、非清晰认知空间和待认知空间。已清晰认知空间中的知识，不再需要深入认知。非清晰认知空间中的知识，有一个深入认知与修正问题，其认知可采用逐步清晰化的策略；待认知空间中的“知识”，可以是由联想或猜想所引发的信度待定的知识，其认知策略有一个在一定的认知逻辑的指导下意向性搜索或在不可靠知识的引导下不断试探的过程。

在问题求解过程中，若有确定而清晰的知识可以利用，其求解过程，将是一个在确定的知识引导下信念的有向流动过程；此时，对知识的运用，常常会有一个基于规则的（链式）推理或对相关知识的变视角处理的问题。对于只有非确定的模糊知识或经验性知识可以利用的情况，其求解过程，常常是一个**柔性计算**的过程，包括模糊推理和神经计算等。对于无可用知识的待解问题，则与对未知领域的认知相同。其求解过程，将是一个在一定的**行为逻辑**的指导下，**意向性搜索**的过程。

柔性推理理论认为，在知-行复合（智能）系统中，基于确定性规则，我们可进行**形式化推理**；基于非确定性规则，我们可进行**非确定性推理**；依据无法显式表达的经验，我们可进行**“隐式直觉”推理**；而对于无任何现成经验和知识可以利用的探索性问题，作一定的**意向性推理**也应是允许的。所有这些非确定性的推理，即基于模糊知识、经验性知识、习惯甚至模糊关系而进行的推理，我们统称其为**柔性推理**。

柔性推理是在一定的认知逻辑和行为逻辑的指导下，依据某种模糊逻辑关系而进行的软计算。它常常需要分析和综合，需要经验与个体心得，需要猜测与论证。通过反复求索，它可使问题逐步从模糊到清晰；而问题的解决，也常常是以满意性原则为标准的。

鉴于目前智能系统在模型构建、知识的表示与获取、以及问题求解等方面存在的问题，知-行复合（智能）系统理论所要研究的问题还包括：知识描述和表达的非清晰[不分明]集合理论；信度与不确定性的集成与传递理论；等等。在所有这些研究中，模型是框架，理论是基石，方法是桥梁。其研究将相辅相成，共同构成知-行复合智能系统这座大厦，也为智能系统的发展注入了新的活力。

由模糊概念柔性关系所构成的**“柔性”计算（推理）系统**，可较好地模拟带有不确定性的抽象思维过程，具有高度的通用性，体现了抽象思维的一般规律。由经验网络所构成的**神经计算系统**，可较好地模拟直觉思维和形象思维过程，也具有高度的通用性，体现了直觉思维和形象思维的基本

精神。而**在一定的认知逻辑和行为逻辑的指导下的柔性推理，也将会在认知和行为过程中发挥一定的作用。它或将是一个嵌套递归、思路可变的探索性认知过程，或者是一个知识、经验和习惯综合运用的问题解决过程。它可以以结构化的知识表示方法为基础，以经验和联想为引导，去挖掘蕴含或者隐含的知识，从而实现基于一定认知的学习和更新。**也可依据有限的经验和原则去压缩要搜索的问题空间，减少搜索的盲目性。柔性推理，常常具有简单、直观、有效的特点，将其用于认知和复杂问题求解，将是一种主动的选择。人类的创造性和主动性，在很多时候，就是以此为基础的。

19.5 基于通用AI大模型的超级智能体系统构建

19.5.1 基于通用 AI 大模型的人工智能体---领域功能增强智能体

如今，通用 AI 大模型的发展已经到了一个新的高度，未来还会继续发展和进化。不过，有许多人认为，我们今后关注的重点，不应该完全放在通用 AI 大模型上。应该放在什么地方呢？应该放在与之相配套的技术的完善上。其中最重要的配套技术就是（基于 AI 大模型的）人工智能体（**AI Agents**）。

人工智能体（AI Agents）是一类能够感知环境、进行决策和执行动作的智能系统。此类智能体可以像人一样，具有**感知能力、记忆能力、逻辑分析能力、任务（问题）拆解能力**和**最后综合起来统一解决问题的能力**。

基于通用 AI 大模型的人工智能体是未来人工智能发展的重要的方向。它的优势不只是基于（通用）AI 大模型，而且也在于它的一系列与 AI 大模型配套的技术及相互配合上。这些配套技术可使它（在特定领域或方面）的业务能力得到极大地提升。因此，基于通用 AI 大模型的人工智能体我们有时候也称其为“**基于（通用）AI 大模型的领域功能增强智能体**”。它旨在 AI 大模型技术的驱动下，让人们以自然语言为交互方式，高度自动化地去执行和处理专业或复杂的工作任务，从而极大提高系统的智能水平。

基于 AI 大模型的领域功能增强智能体也可认为就等于 **"（通用）AI 大模型 +（特定功能）插件 +（特定功能）执行流程 / 思维链"**，它们又分别构成了整个智能体系统的**控制模块、感知模块和执行模块等**。基于通用 AI 大模型的领域功能增强智能体--AI Agent--的研究发展迅速，已获得多项突破性研究成果，多款相关产品也陆续上线，从而引发了人们对 AI Agent 领域的强烈关注。人们普遍认为，AI Agent 应是当前通往 AGI 的一条重要探索路线。

AI 大模型庞大的训练数据集中包含了古今（中外）人类认知和行为的大量信息和知识，这无疑为人工智能体（AI Agent）继承人类智慧和进行**类人**的交互打下了坚实的基础。随着 AI 大模型规模的不断增大，通用 AI 大模型已涌现出了上下文学习能力、推理能力、思维链等类似人类思考方式的多种能力。将通用 AI 大模型作为 AI Agent 的底层和基础，可以“轻松”实现以往智能技术难以实现的将**复杂问题拆解成可实现的子任务，或实现类人的自然语言交互等能力**。不过，由于通用 AI 大模型不可能完全做到既“包罗万象”、“面面俱到”，又“体贴入微”；在 AI 大模型的基础上增加（或附加）一个或多个针对特定领域（或方面）的功能增强的 AI Agent ，合作构建成一个在特定领域或方面具备更多数据和知识、更强自主思考决策能力和执行能力的智能体，就成为了当前通往 AGI 的一个主要研究方向和途径。

19.5.1.1 Agent 的研究—从“代理”到智能体再到基于 AI 大模型的领域功能增强人工智能体

有人或许会感到疑惑，人工智能体（AI Agent）这个东西看起来跟大语言模型（LLM）也差不了多少，为啥会被人如此重视呢？这也许得从 Agent 的来历上说起。Agent 曾是个很古老的用语，甚至可以追溯至亚里士多德和休谟等人的言论。从哲学意义上讲，Agent （**代理**）曾是指一个具有（代理）行动能力的“实体”，而“代理”则表示这种能力的行使或体现。从狭义上讲，“代理”通常是指**有意行动**的表现；相应地，“代理”也就用于表示拥有欲望、信念、意图和行动能力的实体。需要注意的是，“代理”涉及到其自主性。系统已赋予了他们行使意志、做出选择和采取行动的能

力，而不只是被动地对外部刺激做出反应的能力。在 20 世纪 80 年代中后期之前，主流人工智能对 Agent 及其相关理论关注相对较少，然而，自 20 世纪 80 年代中后期之后，计算机科学和人工智能界对这一话题的兴趣就大大增加了。比如， Wooldridge 等人当时就认为，我们或可以这样定义人工智能：“它是计算机科学的一个子领域，旨在设计和构建基于计算机的、表现出智能行为各个方面的 Agent。”由此，人们也就把 Agent 作为了人工智能领域的一个核心概念。当然，当 Agent 这一核心概念被引入人工智能领域时，其含义也发生了深刻变化。在哲学领域，Agent 可以是人、动物，甚至是一切具有自主性的概念或实体；而在人工智能领域，Agent 本质上是一个“计算”实体。由于意识和欲望等概念对于计算实体来说似乎具有形而上学的性质，而且我们只能观察一个智能机器的（外部）行为；因此，包括图灵在内的许多人工智能研究者曾建议，暂时搁置 Agent 是否“真正”在思考或是否真的拥有“思想”的问题。于是，研究人员希望采用其他属性来帮助描述 Agent，如自主性、反应性、主动性和社交能力等。从本质上讲，AI Agent 并不等同于 Philosophy Agent；相反，它是 Agent 这一“哲学”概念在人工智能领域的具体化。在本文中，我们是将 AI Agent 视为一类人工智能（实）体，它们能够使用传感器感知周围环境（具有感知智能），进行决策（具有认知智能），然后使用执行器采取行动（具有行为智能）。

在特定应用场景中，通用 AI 大模型所具有的泛化服务特性，有时很难在知识问答、内容生成、业务处理和管理决策等方面精准满足某些特定的需求。因此，让通用 AI 大模型学习行业知识和行业语料成为（特定）**行业 AI 大模型**，再进一步学习业务知识和学习使用专业领域工具进而演进为（特定）**场景 AI 大模型**，是生成式 AI 深入业务场景，承担更复杂任务的关键路径。而这一过程的实现，让 AI 大模型持续的进化，最终所形成的，就是基于 AI 大模型的 AI Agent。而基于（通用）AI 大模型的领域功能增强智能体的出现，也开启了对各个行业业务流程及其管理和服务模式进行重构与优化的进程。未来，在各类行业对数字化实体（系统）打造的进程中，对于通用人工智能（AGI）的应用，很有可能会以广泛采用 AI Agent 的形态开展；而来自于各行各业各个专业领域的开发人员和创业者们，也比通用 AI 大模型的开发商们更具有对 AI Agent 系统的开发优势。相信，AI Agent 会被越来越多的人认可，成为未来智能时代能深度赋能各行各业的高智能的工具。

不同于传统的人工智能系统，AI Agent 系统具备通过独立思考、调用工具去逐步完成给定目标的能力。而 AI Agent 和通用 AI 大模型的区别也在于，通用智能 AI 大模型与人类之间的交互是基于 prompt 来实现的，（人类）“用户”的 prompt 是否清晰明确，会显著影响 AI 大模型回答的效果。而 AI Agent 的工作则是，仅需给定其一个目标，它就能够针对目标独立思考并做出行动。比如，告诉 AI Agent 帮忙下单一份外卖，它就可以直接调用相关 APP 选择外卖，再调用相关支付程序下单支付，而无需人类去指定每一步的操作。和传统的 RPA 相比，RPA 只能在给定的情况条件下，根据程序内预设好的流程来进行工作的处理，而 AI Agent 则可以通过和环境进行交互，感知当前信息并做出相应的思考和行动。

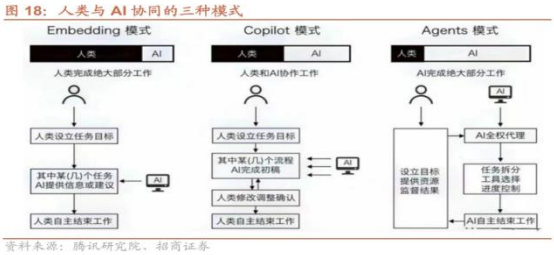


图 19.5.1 人与人工智能系统协同的三种模式

人类与人工智能系统（AI）的协同，可以认为主要有三种模式：辅助式、协同式、全权代理式。其中，辅助式是说，AI 只是（作为工具）在生活或工作中的某些方面帮助人；协同式是说，AI 可在某些重要方面帮助人，它是以相对独立的身份协助人共同完成某种任务；全权代理式则是说，AI 可

在某些重要方面完全代表人。AI Agent 是可以**全权代理**的。我们仅需给定其一个目标，它就能够针对目标独立思考并做出行动。它会根据给定任务详细拆解出每一步的工作（计划）步骤，依靠来自外界的反馈和自主思考，自己给自己创建 prompt，来实现（给定）目标。如果说 Embedding 模式的智能系统是你开车时的“**辅助工具**”，Copilot 模式的智能体是你开车时的“**副驾驶**”，那么，AI Agent 则可以算得上是一个可以替代你开车的“**主驾驶**”了。

在 AI 大模型浪潮席卷全球之时，有很多人认为通用 AI 大模型距离真正的通用人工智能已经非常接近了，有很多厂商都投入到了基础（通用）AI 大模型的研究。但是，经过了一段时间的探索后，大家对 AI 大模型真实的能力边界已有了更清晰的认知，发现通用（基础）AI 大模型仍存在大量的问题，导致其无法直接通向 AGI；于是 AI Agent 就成为了新的研究方向。如今，越来越多的人认为，研究 AI Agent 是人类不断接近通用人工智能（AGI）的最主要探索途径。AI Agent 的最终发展目标，即是实现通用人工智能。业界希望通过强化学习的方法来对 AI Agent 进行训练，以适应真实应用场景。但是，如果想要在真实世界中实现通用性，现有的 AI 大模型还难以实现。（通用）AI 大模型若想适应真实世界，还必须具备高效推理、灵活行动、强大的泛化以及无缝任务转移的能力。而 Agent 也在经历了符号类智能体、反应型智能体、基于强化学习功能的智能体、具有迁移学习和元学习功能的智能体之后，现在已经跨入了基于（通用）AI 大模型的人工智能体（AI Agent）阶段。不少人认为，随着 AI Agent 变得越来越易用和高效，"AI Agent"产品也会越来越多。未来，AI Agent 将会是未来 AI 发展的前沿，会有望成为 AI 应用的基本架构，会涵盖 toC 和 toB 等不同领域。

AI 大模型的发展为通用人工智能的发展带来了新的机遇，同时也为 AI Agent 的开发带来了突破性发展的基础。目前，它已具备了开发各种类型（人工）智能体的发展优势：通过思维链（CoT）和问题分解等技术，基于通用 AI 大模型的智能体可以表现出与符号类智能体相当的推理和规划能力；通过从反馈中学习和执行新的行动，可获得与环境互动的能力，可类似于反应型智能体；而基于 AI 大模型在大规模语料库中进行预训练，也已显示出一定的泛化与迁移学习的能力，从而实现任务间的无缝转移。

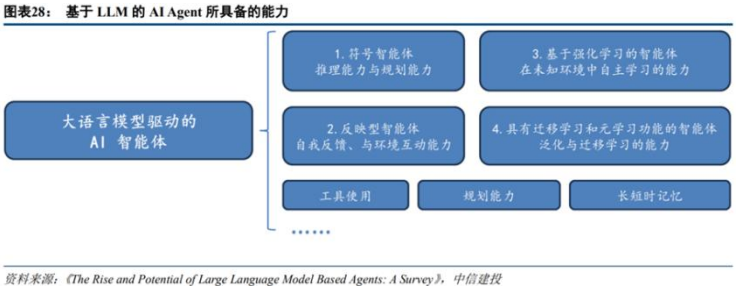


图 19.5.2 基于 AI 大模型的领域功能增强人工智能体

目前，基于 AI 大模型的人工智能体正向两个方向发展：自主化（自主智能体）和拟人化（类人智能体）。自主智能体，力图实现复杂流程的自动化和自主化。当给定自主智能体一个目标时，它们能自行创建任务、完成任务、创建新任务、重新确定任务列表的优先级、完成新的首要任务，并不断重复这个过程，直到完成目标。但其准确度要求高，因而更需要**外部工具（包括其他智能体）等的辅助**以减少 AI 大模型中可能的不确定性的负面影响。**类人智能体**，力图更加拟人可信。主要分为强调情感情商的智能体以及强调交互协作的智能体。后者往往是处于多智能体环境中，可涌现出超越设计者规划的场景识别和思维能力，AI 大模型生成的不确定性反而会成为其优势，而其所具有的多样性也使其有望成为 AIGC 的重要组成部分。我们认为，上述两大方向并不是完全割裂的，相反，自主化与类人化将作为 AI Agent 两大核心能力并行发展。随着底层模型的成熟以及行业探索更加深入，我们有望能进一步扩大其适用范围，提升其实用性。

19.5.1.2 智能系统构建—从单一智能体到多智能体的“集成”或“协同”

人工智能系统一开始主要是符号系统。按照纽厄尔、西蒙、明斯基等早期经典人工智能学者的看法，人工智能是一套物理符号系统所表现出的属性。当把这组计算技术和具体的应用场景相结合时，就可形成一个（人工）智能体（intelligent agents）。人工智能体可以被看作是把人的欲望、信念、意图等心理状态赋予人工智能系统，其所实施的行为是理性的和合理的。一个人工智能体，至少应具备三个基本特征。一是**自主性**。它能够在没有人类直接的干预下正常运行。二是**交互性**。人工智能体和人类以及其他人工智能体之间能够进行信息交流与协作互动。三是**适应性**。人工智能体可以根据环境主动调整行为，做出恰当反应。人工智能体的智能并不是由生物体的生命证实的，而是作为计算系统的属性出现的。其之所以是“智能”，是因为在广泛的（非确定性的实践）环境中，它可以（做出）决策并执行“类人”的“恰当”行动。当然，人工智能系统能够通过算法程序完成的目标，目前还离不开“人”的“设定”；也就是说，人工智能在（具体）环境中自动完成的行动，还依赖于人类预期的目的。人工智能目前还不具备脱离人类（价值）之外的独立的价值体系，它的价值和伦理对齐标准目前还是“人”的价值和伦理。这种由“人”与“人工智能”构成的**“协成”智能体**，揭示了人工智能体的“自主性”其实还是（计算）技术上可实现的一种特征，它并不同于生命有机体的自主性。由于人工智能体自身目前还不能单独成为完全的（类人智能）主体，这也就需要我们进一步去深入思考人工智能的**可信度和可解释性**等问题了。举个例子，当我们在用（网络）地图导航时，导航地图自动推荐的最优路线是系统根据一定参数标准设定的，比如红绿灯最少、高速路段最多、堵车时长最短等；驾驶员在根据导航提示行驶的过程中，是可以更换其他路线或终止导航的；如果未能顺利到达目的地，我们目前还无从追究导航系统的责任。因此，（就目前技术而言），不论人工智能体具有何种程度的自动性，它始终要与人的指令、偏好、价值等结合起来。由“人”和“人工智能（机器）”构成混合（合一）智能体，一方面，是数字智能时代人类与技术工具之间建立新型人-机关系的需要，另一方面，也是人工智能系统目前可持续发展的内在要求。

随着大语言模型和生成式人工智能的出现，人工智能体愈加快速地融入了人们的日常生活。人们不仅面临人工智能体本身，还需要处理人类自身与人工智能体构成的复杂关系，而这些关系正促使人工智能体逐步向“协同”智能体的方向演变。这些“协同”智能体主要包括：人工智能系统与其设计者所构成的“协同”智能体系统；由人与人工智能系统“集成”构成的“协同”智能体系统；由多个人工智能体（系统）构成的“协同”智能体系统。

作为一类人造系统，人工智能系统（目前）无疑是与其设计者（开发者）一起才构成了一个完整的智能系统。人工智能是计算机、数学、心理学、哲学、神经科学等众多学科的交叉（研究）领域，在本质上是人类赋予“机器”的性能（智能化信息处理功能）。从应用角度看，它是一种（自动化、智能化的）工程技术；从理念上看，它是一种思想的“模拟”，甚至是一种世界观的“模拟”。而从结构角度看，它也正从单一人工智能体走向多智能体（集成或协同）系统。而集成或协同智能体系统将会是人工智能未来发展的重要趋势。

对人工智能系统来说，它的功能与人类的“意志”以及想要解决的问题有关。人工智能体（系统）看起来像是在根据一套规范自主行事，但这些规则和目标是相关的研究设计（开发）人员（预先）制定的。换句话说，人工智能体的“信念-欲望-意图”是人类把自己的心理状态赋予了系统，为智能体的行为附加上的心理合理性描述，使它们好像和人一样遵循规则指向目标地行动，这也是我们可以把它们看作智能体的原因。然而，一方面，人工智能体的行为是程序逻辑（理性）推理的结果；另一方面，其目标和规则其实并不属于人工智能（系统）本身，而是制造（开发）它的“人”为它设定的。对人工智能系统的“用户”来说，他们才是系统欲望和意图的发起（激发）者，人工智能体自身不具备生物性的欲望状态，亦不能离开预先设定的模型架构去实施额外的任务。因此，人工智能体的（智能）行为实质上是“人工智能系统-系统开发设计者-系统用户”三方构成的混合系统交互作用的结果。例如，使用大语言模型系统写一个脚本时，需要用户给出提示词和场景（用户

要求），程序生成的文本内容超越不了它的模型边界（系统功能限定），其输出结果中所混杂的，是“设计（开发）者-系统-用户”的“集体”贡献。比如，在一个装备有自动驾驶系统的汽车刹车失灵导致撞车的事故中，司机的操控行为与自动驾驶系统的配合方式是系统设计（开发）者（预先）设置的，事故的原因是多个环节综合造成的，（自动驾驶）智能系统本身（目前）还不能成为自洽的行为主体，故而不会承担全部责任。这种需要“人工智能系统-系统开发设计（开发）者-系统用户”相互配合才能实现的“三元合成智能体”，不仅说明人工智能（目前）还不是一个“独立”的法律主体领域，同时也揭示出目前的人工智能体还不能脱离“人”而真正“独立”，还不是成为一个“道德或法律”的主体。

目前，我们关心的最多的是由**多个人工智能体组成的集成或协同智能体**。人工智能在不同领域的应用往往需要各种特殊的（具有特定功能的）智能体系统互相协作，从而形成多人工智能体组成的集成或协同智能体，以完成超出单一智能体能力范围的任务。随着深度学习、自然语言处理技术的运用，多个智能体之间的互动模式也在不断改进。比如，一个智能家居系统里的智能家电之间就是由不同智能体构成的混合（协同）网络；灯光、音响、电视、冰箱、摄像头等相互连接，各个设备发挥的作用交互协调，才形成了智能家居网络的整体性功能。这种集成或协同智能体通常由若干个自治的、异构的智能体组成，每个智能体都有自己的特定任务和行为方式，通过它们的合作，既施展其每个智能体的单项智能，又提升了单项智能融合后群体系统的性能与效率。

如今，多智能体系统已是人工智能快速发展的一个重要分支，对其更深入的研究在理论上需要对人类智能产生机制和群体智能集成原理做进一步地理解，在系统开发实践中也需要给出更加逼真的基于 AI 大模型的多智能体协同框架，用以解决社会生活中的复杂问题。而从单一智能体向集成（协同）智能体的发展的角度看，其研究和探索也许应该从两个方面努力。

首先是**各个智能体功能的界定和多个智能体之间系统有效集成框架的提出**。人工智能体具有感知、认知和行为等多重功能，具有感知环境并自主决策及执行动作的能力。一个人工智能系统被称作智能体是有条件的。我们认为，其最基本的条件应当是：（1）存在一个与环境 C 不同的可区分的系统 X；（2）X 在 C 中可实施特定的行动 A；（3）C 是依据其既定的规则和目标实施行动 A 的。在这样的情况下，我们才把 X 当作智能体。系统中各个智能体的功能不同，均需明确界定。在对智能体进行功能界定以及特征描述的基础上，还需要考虑各个智能体的实现途径。这些问题包括：（1）如何构建一个智能化的信息（计算）系统来满足各个智能体的功能和属性特征；（2）什么样的软件和硬件结构才是合适的；（3）人工智能体的语言是一种编程语言，我们应该如何对各个不同的智能体进行编程并使其可有效地编译或执行此程序。

其次是**各个智能体之间的关联和交互模式**。多智能体集成或协同系统是多个智能体之间通过一定关系构成的整体智能系统，我们也称其为超智能体。各个智能体集成时最关键的是它们之间的关联和交互机制是什么。认知科学中的分布式认知理论曾主张，认知是分布在不同对象之上的，甚至包括社会文化环境。从这个视角看，集成智能（体）系统也是一种分布式认知系统。然而，不同智能体之间是如何形成“集成”智能的呢？一种方式是可以借鉴**现实系统**中个体与整体的关系来讨论（集成）智能系统的组成，另一种方式是可以从**智能融合理论**的视角来分析单个智能体之间的互动以及如何实现（一定条件下的）均衡。如今，我们或可以运用生成式自然语言处理技术来构建多智能体的协作框架，可以让不同智能体在共享的环境中进行合作和竞争的训练。比如，AI 大模型 CAMEL 能让智能体之间进行自然语言的交流和协商；还有其他的一些多智能体协作系统，如 OpenAI Five、AlphaStar 等，可以参考。开拓多智能体的协作架构，是人工智能亟需开拓的一个重要领域。

总之，当前的人工智能已进入其发展的一个转型期。就机器学习和深度学习的状况来看，人们对智能的呈现手段实质上是对多层数据的参数化转换，是把通过计算机展示的（智能化信息）自动处理行为称为智能，并赋予其一定的能动性。这是对机器智能借助算法对人类认知进行的经验性模

仿，而不是对人类大脑如何产生智能的原理性实现。不过，即使我们不清楚智能与计算的关系究竟是什么，也并不意味着人工智能走的道路是远离人类智能的另一条道路，它必然是基于人类认知逻辑的产物。因此，人工智能的未来演进，（在可以预见的未来），并不是要离开人类而自成一体，而会是走向与人类认知深度交互的“集成”智能。

19.5.1.3 人工智能体构建中技术的发展与演变

在人工智能研究的早期阶段，最主要的方法是符号人工智能，其显著特点是依赖符号与逻辑。这种方法主要是采用逻辑规则和符号表示来封装知识和促进推理过程。它们主要关注两个问题：**转换问题和表示/推理问题**。这些智能系统—即理性智能体（Symbolic Agent）（系统）最主要的功能是要模仿人类的思维模式。它们拥有明确的、可解释的推理框架；而且由于其符号的性质，它们常表现出高度的表达能力。这种方法的一个最典型的例子是基于知识的专家系统。然而，Symbolic Agent 在处理不确定性和大规模现实世界问题时面临着局限性。而且，由于符号推理算法错综复杂，要找到一种能在有限时间内产生有意义结果的高效算法，也很具有挑战性。

行为智能体或刺激-反应类智能体（Reactive Agents）与 Symbolic Agent 不同，Reactive Agent 不使用复杂的符号推理。相反，它们主要关注智能体与其 Environment（环境）之间的交互，强调快速和实时响应。这类智能体的设计优先考虑直接将输入-输出进行映射，而不是复杂的推理和符号操作。Reactive Agent 通常需要较少的计算资源，从而能做出更快的反应，但也因此而缺乏复杂的深层次决策和规划能力。

基于深度学习的智能体（RL-based Agents）的主要关注点是如何让智能体通过与环境的交互进行学习，使其在特定任务完成中获得最大的经验积累，因而也被称为生态进化类智能体。最初，RL-based Agent 主要基于强化学习算法，如策略搜索和价值函数优化，Q-learning 和 SARSA 就是一个例子。随着深度学习的兴起，出现了深度神经网络与强化学习的整合，即深度强化学习。这使得 Agent 可以从高维输入中学习复杂的策略，从而取得了众多重大成就，如 AlphaGo 系统。这种方法的优势在于，它能让智能体在未知环境中自主学习，而无需明确的人工干预。这使得它能广泛应用于从游戏到机器人控制等一系列领域。然而，深度强化学习也面临着一些挑战，包括训练时间长、采样效率低以及稳定性问题，尤其是在复杂的真实世界环境中应用时。

具有迁移学习和元学习功能的智能体（Agent with transfer learning and meta learning）应是未来的一个方向。传统上，训练强化学习型智能体需要大量样本和较长的训练时间，而且缺乏泛化能力。因此，人们引入了**迁移学习**来加速智能体对新任务的学习。迁移学习减轻了新任务培训的负担，促进了知识在不同任务间的共享和迁移，从而提高了学习效率、绩效和泛化能力。更进一步地，有些 AI Agent 引入了元学习机制。元学习的重点是学习如何学习，使智能体能从少量样本中迅速推断出新任务的最优策略。这样的智能体在面对新任务时，可以利用已获得的一般知识和策略迅速调整其学习方法，从而减少对大量样本的依赖。然而，当源任务和目标任务之间存在显著差异时，迁移学习的效果可能达不到预期，并有可能出现负迁移。此外，元学习也需要大量的预训练和大量样本，因此，目前还很难建立起一套通用的学习策略。

基于 AI 大模型的智能体（LLM-based Agent）应是当前智能体系统构建最正确的选择。由于 AI 大模型已经展示出了令人印象深刻的新的能力，并受到广泛欢迎，因此，人们已经开始利用这些模型来构建 AI Agent。具体来说，他们通常采用 AI 大模型作为这些智能体（系统）的大脑或控制器的主要组成部分（**基础智能体**），并通过多模态感知和工具利用等策略来扩展其感知和行动空间。通过**思维链（CoT）和问题分解**等技术，这些基于 AI 大模型的智能体可以表现出与 Symbolic Agent 相当的推理和规划能力。它们还可以通过从反馈中学习和执行新的行动，获得与环境的互动能力，这一点又类似于 Reactive Agent。另外，AI 大模型在大规模语料库中进行预训练时，已显示出了初步的泛化能力，即可实现任务间的无缝转移，这可为迁移学习和泛化打下基础。目前，基于 AI 大模型的

智能体已被应用于各种现实世界场景，包括软件开发和科学研究等。由于具有自然语言理解和生成能力，基于 AI 大模型的多个智能体之间完全可以无缝互动，从而促进多个智能体之间的协作和竞争。

19.5.1.4 为什么 AI 大模型能够作为人工智能体的全新大脑（基础模型）

这里，我们将深入探讨智能体的一些关键属性，阐明它们与 AI 大模型的相关性，从而阐述为什么 AI 大模型非常适合作为人工智能体的“大脑”或基础模型。

首先是自主性（Autonomy）。智能体的自主性是指一个智能体可在没有人类或其他智能体直接干预的情况下运行，并对其行动和内部状态拥有一定程度的控制。这意味着，人工智能体不仅应具备按照人类的明确指令完成任务的能力，还应表现出独立发起和执行行动的能力，这也意味着智能体必须具备一定程度的自主探索和决策能力。而 AI 大模型，如 Auto-GPT 等，已展现出了在构建自主智能体（Autonomious Agent）方面的巨大潜力——只需向它们提供一项任务和一套可用工具，它们就能自主制定计划并执行计划，以实现最终目标。研究认为，现有的 AI 大模型系统在以下几个方面展现出了一定的自主性：（1）它们可以通过生成类似人类的文本参与对话，并在没有详细步骤指示的情况下执行各种任务的能力来展示一种自主性。（2）它们能根据环境输入动态调整输出，体现出一定程度的自适应能力。（3）它们能通过展示创造力来体现自主性，比如提出新颖的想法、故事或解决方案，而这些并没有明确编入它们的程序。未来，它们一定会有更出色的表现。

其次是反应性（Reactivity）。智能体的反应能力是指它对环境中的即时变化和刺激做出快速反应的能力。这意味着智能体必须能感知周围环境的变化，并迅速采取适当的行动。传统上，语言模型的感知空间只局限于文本输入，而行动空间则局限于文本输出。不过，研究已经证明，利用多模态融合技术可以扩展大语言模型的感知空间，使其能够快速处理来自环境的视觉和听觉信息。这些进步已使得 AI 大模型能够有效地与真实世界环境互动，并在其中执行任务。以 AI 大模型为基础，这无疑增强了智能体可感知的空间和操作的空间。一个重要的挑战是，基于 AI 大模型的智能体在执行非文本操作时，可能需要一个中间步骤，即以文本形式产生想法或制定工具使用方法，然后再最终将其转化为具体操作。这一中间过程会消耗时间，降低响应速度。不过，这与人类的行为模式更密切相关，因为人类的行为模式也是遵循“先思考后行动”的原则的。

第三是主动性（Pro-activeness）。积极主动指的是，智能体不仅会对环境做出反应，它们还能积极主动地采取以目标为导向的行动。这一特性强调，智能体可以在行动中进行推理、制定计划和采取主动措施，以实现特定目标或适应环境变化。虽然直观上，AI 大模型中的下一个标记预测范式可能不具备意图或愿望，但研究表明，它们可以隐式地生成这些状态的表征，并指导模型的推理过程。AI 大模型也具有一定的概括推理和规划能力。通过向 AI 大模型发出类似 "请一步一步地思考 " 的指令，我们也可以激发它们的推理能力，如逻辑推理和数学推理等。同样，AI 大模型也已经可以以目标重拟、任务分解和根据环境变化调整计划等形式，展现出系统行为规划的新兴能力。

第四是社交或社会能力（Social Ability）。社交能力指的是一个智能体可通过某种交流语言与其他智能体（包括人类）进行交互的能力。AI 大模型已具有很强的自然语言交互能力，特别是语言理解和生成能力。与结构化语言或其他通信原语相比，这种能力使它们能够以可解释的方式与其他模型或人类进行交互，这无疑构成了基于 AI 大模型的智能体的社会交互能力的基石。许多研究已经证明，基于 AI 大模型的智能体可以通过协作和竞争等社会行为提高其任务绩效（如 Meta GPT）。通过输入特定的提示，AI 大模型也可以扮演不同的角色，从而模拟现实世界中的社会分工。

上述陈述说明，AI 大模型的研究，已为基于 AI 大模型的智能体开发提供了基础条件。在智能体的所有关键步骤中，最重要的是理解输入给智能体的（信息）内容（感知环境），进行推理、规划等从而做出准确决策，并将其转化为（恰当的）可执行的动作序列，以实现最终目标。利用 AI 大模型作为人工智能体的认知核心，可为完成这些步骤提供质量保证。AI 大模型在语言和意图理解、推理、记忆甚至移情等方面具有强大的能力，可以在决策和规划方面发挥卓越的作用。再加上预先

训练的知识，使它们可以创建连贯的行动序列，并有效地执行。此外，通过反思机制，这些基于语言的模型还可以根据当前环境提供的反馈不断调整决策和优化执行序列。AI 大模型无疑为人工智能体开发提供了一个非常强大的基础模型。在与人工智能体相关的开发中，AI 大模型已展现出了许多新的机会。例如，我们可以探索如何将 AI 大模型的高效决策能力整合到传统的 Agent 决策框架中，使 Agent 更容易应用于对专业知识要求较高且以前由人类专家主导的领域。另外，这也使 Agent 的研究不再局限于简单的模拟环境，可以扩展到更复杂的真实世界环境中。

同样，基于 AI 大模型的人工智能体的研究也对 AI 大模型的功能进行了拓展。将 AI 大模型提升为集成智能体系统标志着我们在走向通用人工智能（AGI）的目标方面迈出了更坚实的一步。从智能体的角度来看待 AI 大模型，无疑对 AI 大模型的研究提出了更高的要求，同时也扩大了 AI 大模型的应用范围，为许多实际应用提供了更多机会。它使我们对大（语言）模型的研究不再局限于只涉及文本输入和文本输出的传统任务，如问题解答和文本摘要等；取而代之的是，其研究重点将转向处理更复杂任务，这些任务包含更丰富的输入模式和更广阔的行动空间。而挑战也在于，如何让 AI 大模型高效地处理输入、从环境中收集信息并解释由其行动所产生的反馈，同时保持其核心能力。而更大的挑战在于，如何让 AI 大模型理解环境中不同元素之间的隐含关系，并获取世界知识。目前，大量的研究旨在扩展 AI 大模型的感知能力、学习能力和行动能力，让它们掌握更多影响世界的技能，例如在模拟或物理环境中使用工具或与机器人联通 API 接口等。

在超智能体系统领域，我们则希望基于 AI 大模型的智能体能在社会合作中扮演不同的角色，参与涉及协作、竞争和协调的社会互动，从而构成一个能力超强的智能系统。

19.5.2 基于 AI 大模型的领域功能增强智能体的总体功能结构

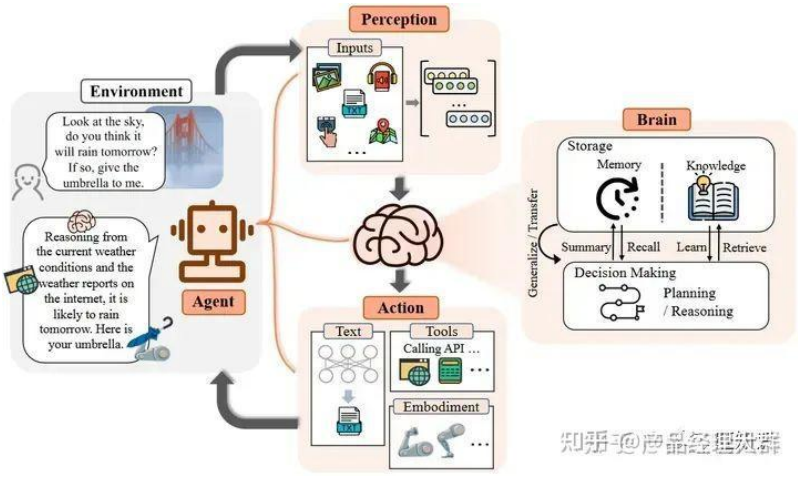


图 19.5.3 基于 AI 大模型的领域功能增强智能体的概念框架

一个基于 AI 大模型的（领域功能增强）智能体的总体功能框架大体可如图 19.5.3 所示，它主要由**类脑（基础模型）模块、感知功能（增强）模块、行动功能（增强）模块**等三个关键部分组成。作为系统思维和控制功能的主要模块，类脑功能模块承担着记忆、思考、任务分配和决策等基本任务；**感知（增强）功能模块**负责感知和处理来自外部环境的多模态信息；**行动（增强）模块则**负责（使用工具）执行任务并影响周围环境。

19.5.2.1 感知功能（增强）模块，用于提高智能体的感知能力和感知智能

感知功能（增强）模块的核心目的是将智能体的感知空间从纯文字领域扩展到包括文字、听觉和视觉模式在内的多模态领域，并希望能进一步增强系统的感知功能和感知智能。

AI 大模型已经具备了通过文本的输入和输出与人类进行交流（交互）的基本能力。在“用户”的（文本）输入中，除了有明确的对话事务内容外，还可能隐藏着其信念、愿望和意图等。理解输

入内容的隐含含义对于智能体掌握人类“用户”的潜在需求和潜在意图至关重要，从中可提高智能体与“用户”交流（交互）的效率和质量。一些研究希望采用强化学习的方法来感知“输入”的隐含含义，并建立反馈模型以期获得改进。这将有助于推断出“用户”的偏好，从而使智能体可做出更个性化、更准确的回应。此外，由于基于 AI 大模型的智能体常常被设计用于复杂的真实世界环境，它将不可避免地会遇到许多全新的任务。理解全新任务的“文本指示”对智能体的文本感知能力也提出了更高的要求。而经过功能调整或增强的 AI 大模型，可以表现出卓越的文本指令理解和泛化能力，从而可无需针对特定任务进行微调。

视觉输入是最重要的非文本输入。受 Transformer 在自然语言处理中出色表现的启发，有些研究已将其应用扩展到计算机视觉领域。视觉输入通常包含大量有关世界的信息，包括智能体周围环境中物体的属性、空间关系、场景布局等。因此，将视觉信息与其他模式的数据整合在一起，可以为智能体提供更广泛的背景和更精确的理解，加深智能体对环境的感知。为了帮助智能体理解图像中所包含的“更核心的”信息，一种直接的方法是为图像输入生成相应的文本描述，即图像标题。其字幕也可以直接与标准文本指令连接，并输入到智能体中。这种方法具有很高的可解释性，而且不需要额外的字幕生成训练，可以节省大量的计算资源。

ViT/VQVAE 等一些具有代表性的系统已成功地利用 Transformer 对视觉信息进行编码。研究者首先将图像分割成固定大小的块，然后将这些块经过线性投影后作为 Transformer 的输入标记。最后，通过计算标记之间的自注意力，他们就能整合整个图像的信息，从而高效地感知视觉内容。因此，一些研究尝试直接将图像编码器和大语言模型结合起来，以端到端的方式训练整个模型。这种基于 AI 大模型的视觉增强智能体可以实现出色的视觉感知能力；而经过广泛预训练的视觉编码器和大语言模型的结合，可以大大提高智能体的视觉感知和语言表达能力。然而，现有大语言模型还无法直接理解视觉编码器的输出，因此有必要将图像编码转换为大语言模型可以理解的嵌入。换句话说，这需要将视觉编码器与大语言模型对齐，这通常需要在两者之间添加一个额外的可学习接口。例如，Q-Former 即是一种转换器，它采用可学习的查询向量，使其具有提取语言信息视觉表征的能力。它可以为大语言模型提供最有价值的信息，减轻智能体学习视觉-语言对齐的负担，从而解决视觉与文本（语言）的对接与对齐问题。还有一些研究者采用一种计算效率较高的方法，即使用一个嵌入层（Embedding layer）来实现视觉-文本对齐，从而减少了训练额外参数的需要。此外，嵌入层还能与可学习层有效结合，调整其输出的维度，使其与大语言模型兼容。

视频输入通常由一系列连续的图像帧组成。因此，智能体用于感知图像的方法也可用于视频领域，使智能体也能很好地感知视频输入。与图像信息相比，视频信息增加了一个时间维度。因此，智能体对不同帧间时间关系的理解对于感知视频信息至关重要。一些工作，如 Flamingo ，通过使用掩码机制来确保理解视频时的时间顺序。但掩码机制也限制了智能体的视角，当它感知不到视频中的特定帧时，只能从时间上较早的帧中获取视觉信息。

对听觉输入，一个非常直观的想法是，智能体可将大语言模型用作控制中心，以级联方式调用现有工具集或模型库来感知音频信息。例如，AudioGPT 充分利用了 FastSpeech、GenerSpeech、Whisper 等模型的功能，这些模型在文本到语音转换和语音识别等任务中取得了优异的成绩。音频频谱图直观地表达了音频信号随时间变化的频谱，对于一段时间内的一段音频数据，可将其抽象为有限长度的音频频谱图。音频频谱图具有二维表示形式，可视化为平面图像。因此，一些研究致力于将感知方法从视觉领域迁移到音频领域。比如，AST（音频频谱图变换器）采用与 ViT 类似的变换器架构来处理音频频谱图图像。通过将音频频谱图分割成片段，它实现了对音频信息的有效编码。

目前，已有许多研究对文本、视觉和音频的感知功能进行了研究。然而，基于 AI 大模型的智能体可能还需要配备更丰富的感知模块。未来，它们也许可以像人类一样感知和理解现实世界中的各种（输入）模式。例如，它可以拥有独特的触觉和嗅觉器官，从而在与物体交互时收集到更多详细

信息。同时，智能体也许能清楚地感知周围环境的温度、湿度和亮度，从而采取相应环境感知行动。而通过有效整合视觉、文字和听觉等基本感知能力，未来的功能增强智能体还能开发出更多种类的对人类友好（类人）的感知模块。比如，用户可以通过使用手势或移动光标来选择、拖动或绘制，从而与图像中难以描述的某些特定部分进行交互。在此基础上，功能增强智能体还有可能感知更复杂的用户输入。例如，AR/VR 设备中的眼球跟踪、身体动作捕捉等技术，甚至是脑-机交互中的脑电波信号等。

未来，基于 AI 大模型的（功能增强）智能体还应具备更广阔的整体环境感知能力。目前，已有许多成熟且被广泛采用的硬件设备可以帮助智能体来实现这一目标。比如，激光雷达可以创建三维点云图，帮助智能体检测和识别周围环境中的物体；全球定位系统可以提供精确的位置坐标，并可与地图数据集成。惯性测量单元可以测量和记录物体的三维运动，提供物体速度和方向的详细信息。然而，这些感知数据非常复杂，基于 AI 大模型的（功能增强）智能体目前还无法直接理解。探索感知功能增强智能体如何去感知更全面的输入，将是未来一个很有前景的研究方向。

19.5.2.2 基于 AI 大模型的类脑（认知）增强模块，提供更强大的认知功能

在基于 AI 大模型的智能体中，（功能增强后的）AI 大模型可（部分）充当类似“大脑”的功能，我们称其为“类脑”模块。它不仅存储知识和记忆，还承担着信息处理和决策等功能，并可以呈现推理和规划的过程，也能很好地应对未知任务。

类脑功能首先是思维和推理功能。AI 大模型已经具备初步的逻辑推理能力，基于 AI 大模型的智能体更可以将 AI 大模型的逻辑推理能力激发出来。有研究认为，当模型规模足够大的时候，AI 大模型本身是可以具备推理能力的。目前，在一些简单推理问题上，AI 大模型已经达到了很好的实用能力；但在复杂推理问题上，AI 大模型表现还不理想。事实上，在很多时候，“用户”无法通过 AI 大模型获得理想的回答，其原因，一方面是模型本身需要完善，另一方面，也在于“用户”的 prompt 不够合适，无法激发 AI 大模型本身的推理能力。通过追加辅助推理的 prompt，或可以大幅提升 AI 大模型的推理效果。此时，将 Agent 作为智能代理，可根据给定的目标自己创建合适的 prompt，从而可以更好地激发 AI 大模型的推理能力。

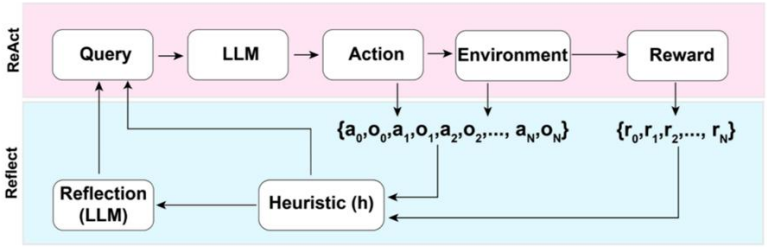
通常情况下，一项复杂的任务往往涉及许多步骤。一个好的（功能增强）智能体需要首先拆解这些步骤，并提前做好计划（规划）。其任务分解环节，可以由三种方式完成：（1）在 AI 大模型输入简单的提示，比如“依 XYZ 的步骤”，或者“实现 XYZ 的第一个子目标是什么？”；（2）使用特定任务的指令，比如在需要写小说的时候要求 AI 大模型“写一个 XX 故事大纲”；（3）通过人工提供信息。

对任务进行分解和规划，当下最常用的技术模式是思维链和思维树。**思维链**（Chain of Thoughts）是一种标准的提示技术，用于提高模型在复杂任务中的表现。模型被要求“一步一步地思考”，将艰巨的任务分解为更小更简单的步骤。思维链可将复杂任务转化为多个更容易管理的任务，并帮助人们理解模型的思维过程。**思维树**（Tree of Thoughts）则通过在任务的每一步探索多种推理可能性来扩展思维链。它首先将问题分解为多个思考步骤，并在每个步骤中生成多个想法，从而创建一个树状结构。搜索过程可以是 BFS（广度优先搜索）或 DFS（深度优先搜索）。

试错和纠错在现实世界的任务决策中是不可避免且至关重要的步骤。**自我反思**会帮助智能体完善过去的行动决策、纠正以前的错误、从而不断改进。目前可用的技术包括 ReAct、Reflexion 和后见链（Chain of Hindsight）等。**ReAct** 将任务中单独的行为和语言空间组合在一起，从而使 AI 大模型的推理和行动融为一体。该模式可帮助 AI 大模型与环境互动（例如使用维基百科搜索 API），并以自然语言形式留下推理的痕迹。**Reflexion** 是一个可让 AI Agent 具备动态记忆和自我反思能力以提高推理能力的框架。它沿用了 ReAct 中的设置，并提供简单的二进制奖励。每次行动后，AI Agent 都会计算一个启发式函数，并根据自我反思的结果决定是否重置环境以开始新的试验。这个启发式

的函数可以判断当下的路径是否效率低下（耗时过长却没有成功）或包含幻觉（在环境中遇到一连串导致相同观察结果的相同行动），并在出现这两种情况下终止函数。**Chain of Hindsight** 通过向模型明确展示一系列过去的输出结果，鼓励模型改进自身的输出结果，使得下一次预测的行动比之前的试验取得更好的成绩。而**算法蒸馏**（Algorithm Distillation）则可将同样的理念应用于强化学习任务中的跨集轨迹。

图 10：AI Agent 的反思框架



数据来源：Noah, et al.《Reflexion: Language Agents with Verbal Reinforcement Learning》，东方证券研究所

图 19.5.4 具有反思功能的智能体基本框架

图表 4：算法蒸馏（Algorithm Distillation）在强化学习中的流程图

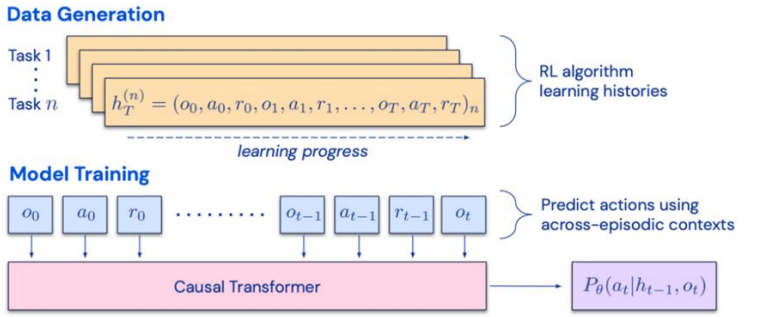


图 19.5.5 强化学习中的算法蒸馏

记忆功能负责存储信息，包括过去的交互、学习到的知识，甚至是临时的任务信息等。正如人脑依靠记忆系统来回溯利用先前的经验制定策略和做出决策一样，智能体也需要特定的记忆机制来确保其熟练处理一系列连续任务。对于一个（功能增强）智能体来说，有效的记忆机制能够保障它在面对新的或复杂的情况时，可以调用以往的经验和知识。例如，一个具备（增强）记忆功能的聊天机器人可以记住用户的偏好或先前的对话内容，从而提供更个性化和连贯的交流体验。

在基于 AI 大模型的智能体中，（增强的）记忆功能存储了（领域）智能体过去的观察、思考和行动序列。在面对复杂问题时，（增强的）记忆机制能帮助行为主体有效地重新审视和应用先前的策略。此外，这些记忆机制还能使个体借鉴过去的经验，适应陌生的环境。随着基于 AI 大模型的（功能增强）智能体互动周期的扩大，也出现了两个主要挑战。第一个挑战与历史记录的长度有关。基于 AI 大模型的（记忆增强）智能体可以以自然语言格式处理先前的交互，并将历史记录附加到每个后续输入中。随着这些记录的增加，它们可能会超出大多数基于 AI 大模型的智能体所依赖的 Transformer 架构的限制。在这种情况下，系统可能会截断某些内容。第二个挑战是提取相关记忆的难度。当智能体积累了大量的历史观察和行动序列时，它们就会面临不断升级的记忆负担。这使得在相关主题之间建立联系变得越来越具有挑战性，有可能导致智能体的反应与当前环境不一致。

为了提高记忆能力，在基于 AI 大模型的（人工）智能体中，记忆功能增强模块可采用有限的上下文长度来实现更多的记忆。增强基于 AI 大模型的智能体的记忆能力的方法包括：**（1）提高**

Trransformer 的输入长度限制。本方法试图解决或减轻系统固有的序列长度限制。由于这些固有限制，Transformer 架构很难处理长序列。随着序列长度的增加，由于 Self-Attention 机制中的成对标记计算，计算需求将呈指数级增长。缓解这些长度限制的策略包括文本截断、分割输入，以及强调文本的关键部分等。还有一些研究修改了注意力机制，以降低复杂性，从而适应较长的序列。**（2）总结记忆**。提高系统记忆效率的第二种策略取决于记忆总结的概念。它能确保智能体毫不费力地从历史互动中提取关键细节。一些方法是利用提示简洁地整合记忆，而另一些方法则强调反思过程，以创建浓缩的记忆表征。分层方法将对话精简为每日快照和总体总结。一些特定的策略是将环境反馈转化为文本封装，从而加强了智能体对未来参与的语境的把握。此外，在多智能体环境中，智能体交流的一些重要元素会被捕获并保留下来。**（3）用向量或数据结构压缩记忆。**通过采用合适的数据结构，记忆增强智能体可提高其记忆检索效率，促进对交互做出迅速反应。值得注意的是，有几种方法依赖于为记忆部分、计划或对话历史嵌入向量，另一种方法则是将句子转化为三元组配置，或将记忆视为独特的数据对象，从而促进不同智能体的交互。

当智能体与其环境或“用户”交互时，必须从其内存（记忆）中检索最合适的内容。而准确检索可确保智能体可访问到准确的相关信息，以执行特定操作。这就产生了一个重要问题：（记忆增强）智能体如何选择最合适的存储器并准确检索？通常情况下，智能体要求存储器具有自动检索记忆的能力。优化自动检索的一个重要方法是考虑三个指标： 最近性（Recency）、相关性（Relevance）和重要性（Importance）。记忆得分由这些指标加权组合而成，得分最高的记忆在模型的上下文中被优先考虑。一些研究引入了**交互式记忆对象**的概念，即对话历史的表现形式，可以移动、编辑、删除或通过总结进行组合。用户可以查看和操作这些对象，从而影响智能体对对话的感知。同样，其他一些研究也允许根据用户提供的特定命令进行删除等记忆操作。这些方法确保了记忆内容与用户的期望密切相关。

记忆可以定义为用于获取、存储、保留以及随后检索信息的过程。人脑中有多种记忆类型，如感觉记忆、短期记忆和长期记忆。而对于基于 AI 大模型的（人工）智能体系统而言，用户在与其交互过程中所产生的内容都可以认为是智能体的记忆，和人类记忆的模式能够产生对应关系。其中，感觉记忆就是作为学习嵌入表示的原始输入，包括文本、图像或其他模态；短期记忆就是上下文，受到有限的上下文窗口长度的限制；长期记忆则可以认为是智能体在工作时需要查询的外部向量数据库，可通过快速检索进行访问。目前，基于 AI 大模型的智能体主要是利用外部的长期记忆来完成很多的复杂任务，任务与结果会储存在记忆模块中。当信息被调用时，储存在记忆中的信息会回到与用户的对话中，由此创造出更加紧密的上下文环境和语境。

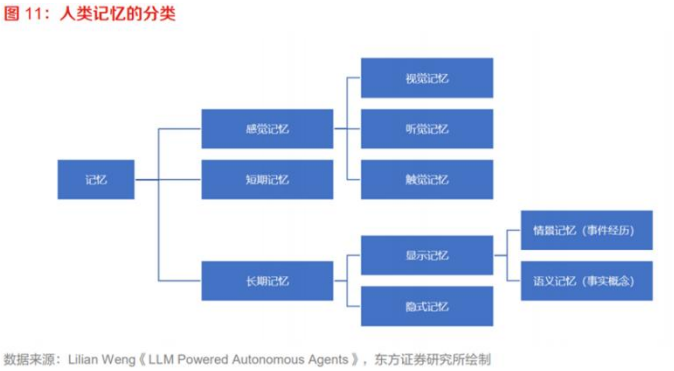


图 19.5.6 人类记忆的分类

表 2：人类记忆与 AI Agent 记忆的映射

记忆类型	映射	例子
感觉记忆	学习原始输入的嵌入表示，包括文本、图像或其他形式，短暂保留感觉印象。	看一张图片，然后在图片消失后能够在脑海中回想起它的视觉印象。
短期记忆	上下文学习（比如直接写入 prompt 中的信息），处理复杂任务的临时存储空间，受有限的上下文长度限制。	在进行心算时记住几个数字，但短期记忆是有限的，只能暂时保持几个项目。
长期记忆	在查询时 Agent 可以关注的外部向量存储，具有快速检索和基本无限的存储容量。	学会骑自行车后，多年后再次骑起来时仍能掌握这项技能，这要归功于长期记忆的持久存储。

数据来源：东方证券研究所整理

图 19.5.6 人类记忆与 AI 记忆的映射

为了解决有限记忆时间的限制，基于 AI 大模型的（记忆增强）智能体通常会用到外部存储器。常见的做法是将信息的嵌入表示保存到可支持快速存取的最大内积搜索（MIPS）的向量存储数据库中。向量数据库通过将数据转化为向量存储，以解决 AI 大模型海量知识的存储、检索、匹配问题。向量是智能系统理解世界的通用数据形式，AI 大模型需要大量的数据进行训练，以获取丰富的语义和上下文信息，这也导致了数据量的指数级增长。向量数据库利用人工智能中的 Embedding 方法，将图像、音视频等非结构化数据抽象、转换为多维向量，由此可以结构化地在向量数据库中进行管理，从而实现快速、高效的数据存储和检索过程，从而赋予了基于 AI 大模型的智能体的“长期记忆”。同时，将高维空间中的多模态数据映射到低维空间的向量，也能大幅降低存储和计算的成本。向量数据库的存储成本通常比存到神经网络的成本要低 2 到 4 个数量级。

Embedding 技术和向量相似度计算是向量数据库的核心。Embedding 技术是一种将图像、音视频等非结构化数据转化为计算机能够识别的语言（数据）的一种方法；例如，常见的地图就是对现实地理的 Embedding，现实的地理地形的信息其实远远超过三维，但是地图可通过颜色和等高线等来最大化地表现现实的地理信息。在通过 Embedding 技术将非结构化数据（例如文本数据）转化为向量后，就可以通过数学方法来计算两个向量之间的相似度，即可实现对文本的比较。向量数据库强大的检索功能就是基于向量相似度计算而达成的。通过相似性检索特性，针对相似的问题找出近似匹配的结果，是一种模糊匹配的检索。它没有标准的准确答案，却可更高效地支撑更广泛的应用场景。

为确保有效交流，智能体基于自然语言的交互能力无疑是至关重要的。在接收到感知模块处理的信息后，智能体的类脑功能模块首先会将其转向存储，在知识中检索并从记忆中回忆。这些结果将有助于智能体制定计划、进行推理和做出明智的决定。此外，类脑功能模块还能以摘要、矢量或其他数据结构的形式记忆智能体过去的观察、思考和行动。同时，它还可以更新常识和领域知识等知识，以备将来使用。

作为一种交流媒介，（人类）语言中包含着丰富的信息。除了直观表达的内容，背后还可能隐藏着说话者的信念、愿望和意图。由于 AI 大模型本身（已）具有强大的自然语言理解和生成能力，基于 AI 大模型的智能体不仅可以熟练地使用多种语言进行基本的交互式对话，实现自然语言交互功能；还能表现出深入的理解能力，从而使人类（或其他智能体）能够轻松地理解（对话）智能体的意图并与之互动。

多轮交互对话能力是有效而一致交流的基础。作为类脑功能模块的核心，大（语言）模型已能理解自然语言并生成连贯且与上下文相关的回复，从而帮助（对话）智能体更好地理解和处理各种问题。然而，即使是人类也很难在一次交流中不出现混乱，因此需要进行多轮对话。与传统的纯文本阅读理解任务相比，多轮对话具有交互性、涉及多个说话者、缺乏连续性、可能涉及多个话题等特点。对话信息也可能是冗余的，使得文本结构更加复杂。

一般来说，多轮对话主要分为三个步骤： (1) 了解自然语言对话的历史；(2) 决定采取什么行动；(3) 生成自然语言回应。基于 AI 大模型的智能体必须保证能够利用现有信息不断完善输出，进行多轮对话并有效实现最终目标。最新的大语言模型已展示了卓越的自然语言生成能力，可持续生成多种语言的高质量文本。生成内容的连贯性和语法准确性也在稳步提高，凸显了其强大的语言能力。在对话语境中，AI 大模型在对话质量的一些关键指标上也表现出色，包括内容、相关性和适当性等。重要的是，AI 大模型不仅仅是复制训练数据，而且还能表现出一定程度的创造力，能生成与人类制作的基准文本同样新颖甚至更加新颖的各种文本。同时，通过使用可控提示，智能体可确保能对这些语言模型生成的内容进行精确控制。

尽管在大规模语料库中训练出来的模型已经具有足够的智能来理解指令，但它们中的大多数仍无法模拟人类对话或充分利用语言所传达的信息。要想与其他智能机器人进行有效的交流与合作，理解其隐含的意思至关重要（意图和隐含需求理解）。不少 AI 大模型已展现出了基础模型在理解人类意图方面的潜在能力；但当涉及到模糊指令或其他含义时，也会给智能体带来巨大挑战。对于人类来说，掌握对话中的隐含意义是自然而然的事；而对于智能体来说，它们应该将隐含意义形式化为奖励函数，使他们能够在看不见的语境中选择符合说话者偏好的选项。奖励建模的主要方法之一是根据反馈推断奖励，反馈主要以比较和无约束自然语言的形式呈现。利用对上下文的理解，智能体是可以根据具体要求采取高度个性化和准确的行动的。

知识是智能体的基础。研究表明，在大规模数据集上训练的大语言模型可以将各种知识编码到其参数中，并对各种类型的查询做出正确的反应。而这些知识已能帮助基于 AI 大模型的智能体做出明智的决策。基于 AI 大模型的智能体中的知识可包括：**（1）语言知识**。语言知识表现为一个约束系统，即语法，它定义了语言的所有和可能的句子。语言知识包括词法、句法、语义学和语用学。只有掌握了语言知识的智能体才能理解句子并进行多轮对话。基于 AI 大模型的智能体可以通过在包含多种语言的数据集上进行训练来获取多种语言知识，而无需额外的翻译模型。**（2）常识知识**。常识性知识指的是大多数人在幼年时就已掌握的世界常识。例如，人们都知道，药是用来治病的，伞是用来防雨的。这些信息通常不会在上下文中明确提及。因此，缺乏相应常识性知识的模型可能无法理解或误解其中的含义。同样，缺乏常识性知识的智能体也可能会做出错误的决定，比如在下大雨时不打伞。**（3）专业领域知识**。专业领域知识是指与特定领域相关的知识，如编程、数学、医学等。它们对模型有效解决特定领域内的问题至关重要。例如，用于执行编程任务的模型需要具备编程知识，如代码格式。同样，用于诊断目的的模型应具备医学知识，如特定疾病和处方药的名称。

尽管 AI 大模型在获取、存储和利用知识方面表现出色，但仍然存在潜在的问题和悬而未决的难题。例如，模型在训练过程中获得的知识可能会过时，甚至从一开始就是错误的。解决这一问题的简单方法是重新训练。但是，这需要先进的数据、大量的时间和计算资源。更糟糕的是，它可能会导致灾难性“遗忘”。因此，一些研究尝试改进 AI 大模型以便利找到并修改模型中存储的特定知识。这也包括在获取新知识的同时卸载模型中不正确的知识。此外，现有 AI 大模型可能会生成与来源或事实信息相冲突的内容，这种现象通常被称为幻觉。这也是一些 AI 大模型还无法广泛应用于严格的事实任务的重要原因之一。为解决这一问题，一些研究提出了评估 AI 大模型输出可信度的有效指标。此外，也有一些研究提出让 AI 大模型能够利用外部工具或（领域功能）增强模型来避免错误的输出，基于 AI 大模型的（领域功能增强）智能体就是典型的解决方案之一。

推理和规划能力也是基于 AI 大模型的智能体需要增强的功能。 **推理（Reasoning）**以证据和逻辑为基础，是人类智力活动的根本，是解决问题、进行决策和批判性分析的基石。演绎和归纳是智力活动中常见的主要推理形式。对于基于 AI 大模型的智能体来说，与人类一样，推理能力对于解决复杂任务至关重要。对于现有 AI 大模型的推理能力，学术界存在不同观点。一些人认为 AI 大模型在预训练或微调过程中就已具备了推理能力，而另一些人则认为推理能力是在达到一定规模后才出

现的。具体而言，有人认为，具有代表性的思维链（CoT）方法通过引导 AI 大模型在输出答案之前生成理由，已被证明能够激发 AI 大模型的推理能力。此外，研究还提出了其他一些提高 AI 大模型性能的策略，如自我一致性、自我修正、自我完善和选择推理等。更有研究表明，分步推理的有效性可归因于训练数据的局部统计结构，与对所有变量进行训练相比，变量间局部结构化的依赖关系能产生更高的数据效率，这也是我们大力提倡开发基于 AI 大模型的智能体的理由之一。**规划（Planning）**是人类在面对复杂挑战时采用的一种关键策略。对人类来说，规划有助于组织思维、设定目标和确定实现这些目标的步骤。与人类一样，规划能力对智能体也至关重要，而规划模块的核心是推理能力。它可为基于 AI 大模型的智能体提供一个结构化的思维过程。通过规划，智能体可将复杂的任务分解为更易于管理的子任务，并为每个子任务制定适当的计划。此外，随着任务的进展，智能体还可以利用内省来修改其计划，确保计划更符合实际情况，从而适应并成功执行任务。通常，规划包括两个阶段：计划制定和计划反思。

在**制定计划**（规划）的过程中，Agent 通常会将总体任务分解成许多子任务，在这一阶段，人们提出了各种方法。值得注意的是，一些研究主张基于 AI 大模型的智能体一次性全面分解问题，一次性制定完整的计划，然后按顺序执行。而其他一些研究（如 CoT 系列）则主张采用自适应策略，一次规划和处理一个子任务，从而更流畅地处理复杂的整体任务。此外，有些方法强调分层规划，而另一些方法则强调一种策略，即**从树状结构的推理步骤中推导出最终计划**。后一种方法认为，在最终确定计划之前，智能体应评估所有可能的路径。虽然基于 AI 大模型的智能体可展示广博的常识，但在遇到需要专业知识的情况时，它们也会面临挑战。通过将这些智能体与特定领域的规划相结合来增强它们的能力，已证明能产生更好的性能。

制定好计划后，必须对其优缺点进行**反思和评估**。基于 AI 大模型的智能体可利用其内部反馈机制来完善和改进其战略和规划方法。为了更好地与人类的价值观和偏好保持一致，智能体会主动与人类（交互）接触，从而纠正一些误解，并将这些有针对性的反馈吸收到其规划方法中。此外，它们还可以从有形或虚拟环境中获得反馈，如任务完成情况的提示或行动后的观察，以帮助它们修改和完善计划。

学习能力也是基于 AI 大模型的智能体需要增强的能力。特别是情境学习能力和持续学习能力。大量研究表明，大（语言）模型可以通过上下文学习（In-Context Learning ICL）[或称情景学习]来完成各种复杂任务。上下文学习指的是模型从上下文中的几个例子中学习的能力。少量语境内学习通过将原始输入与几个完整示例串联起来，作为丰富语境的提示，从而提高语言模型的预测性能。ICL 的主要思想是从类比中学习，这与人类的学习过程类似。此外，由于提示是用自然语言编写的，因此交互是可解释和可改变的，从而更容易将人类知识纳入 AI 大模型。与监督学习过程不同，ICL 不涉及微调或参数更新，这可以大大降低模型适应新任务的计算成本。除文本外，研究人员还探索了 ICL 在不同多模态任务中的潜在能力，从而使智能体应用于真实世界任务成为可能。

基于 AI 大模型的功能增强智能体研究进一步考虑了 AI 大模型的规划能力在促进智能体持续学习方面的潜力。这涉及技能的持续获取和更新。持续学习的一个核心挑战是灾难性“遗忘”，当模型学习新任务时，往往会丢失以前任务的知识。为应对上述挑战，人们做出了大量努力，这些努力包括：参照以前的模型引入经常使用的术语；近似先验数据分布；设计具有任务自适应参数的架构；等等。

如果从静态的角度来看，一个具有高水平的实用性、社会性和正确价值观的智能体是可以满足人类的大部分需求，并有可能提高生产力的。然而，从动态的角度来看，一个能不断发展、不断进化并适应不断变化的社会需求的智能体可能更符合今后的发展趋势。由于智能体可以随着时间的推移自主进化，因此开发所需的人工干预和资源可以大大减少。在这一方面已经有了一些探索性工作，例如让智能体在虚拟世界中从零开始，完成生存任务，实现更高阶的自我价值。然而，为这种持续

进化建立评估标准仍然具有挑战性。为此，有文献提出建议，一是坚持**持续学习**。持续学习可使模型不断获得新知识和技能，同时也不会遗忘之前获得的知识和技能。而持续学习的性能可从三个方面进行评估：迄今所学任务的总体性能、旧任务的记忆稳定性、新任务的学习可塑性。二是加强**自主学习能力**。即增强智能体在开放世界环境中自主生成目标并实现目标的能力，包括探索未知世界和在此过程中获取技能的能力。对这种能力的评估可包括为智能体提供一个模拟生存环境，并评估其掌握技能的程度和速度等。

19.5.2.3 行动（功能增强）模块

人类在感知环境后，大脑会对感知到的信息进行整合、分析和推理，并做出（恰当）决策。随后，他们会（利用神经系统）控制自己的身体，做出适应环境或创造性的行动。当一个人工智能体拥有类似大脑的结构，具备知识、记忆、推理、规划和概括能力以及多模态感知能力时，它也就有望拥有类似人类的各种行动来解决问题或应对周围环境。在人工智能体的构建过程中，行动模块会接收类脑模块发送的行动序列，执行与环境互动的行动。

智能体的行为，一是**文本输出。**如前所述，基于 Transformer 的大型语言生成模型的兴起和发展，已赋予了基于 AI 大模型的人工智能体以固有的语言生成能力。它们生成的文本质量在流畅性、相关性、多样性和可控性等各个方面都已非常出色。因此，基于 AI 大模型的人工智能体完全可以成为异常强大的语言生成器，以文本形式输出智能体生成的内容。二是**动作输出**。如说话、操作、“手舞足蹈”等。

19.5.3 基于 AI 大模型的功能增强智能体的工具调用

会使用工具曾被认为是人和动物区别的一个根本标志。毫无疑问，懂得使用工具的人工智能体才会更像人类，也包括调动相关（专业）人士和（相关）智能系统等。人工智能体总体功能的增强与工具使用息息相关。

一个基于 AI 大模型的人工智能体系统，可以看作是由 AI 大模型（基础模型）和特定领域功能增强模块两部分构成。特定领域功能增强模块可包括推理与规划功能（增强）模块、记忆增强模块、工具调用和使用功能模块等部分。即，基于 AI 大模型的人工智能体的基础架构可表示为：

（基于 AI 大模型的）人工智能体= AI 大模型（系统） + 推理与规划技能增强（模块） + 记忆增强（模块） + 工具使用（模块）

其中，“AI 大模型+ 推理与规划技能增强（模块）”扮演了人工智能体的“大脑”角色，在系统中提供推理、规划等能力。而特定领域功能增强模块在系统中可完成特定领域的功能增强，包括记忆增强、工具调用和使用等。

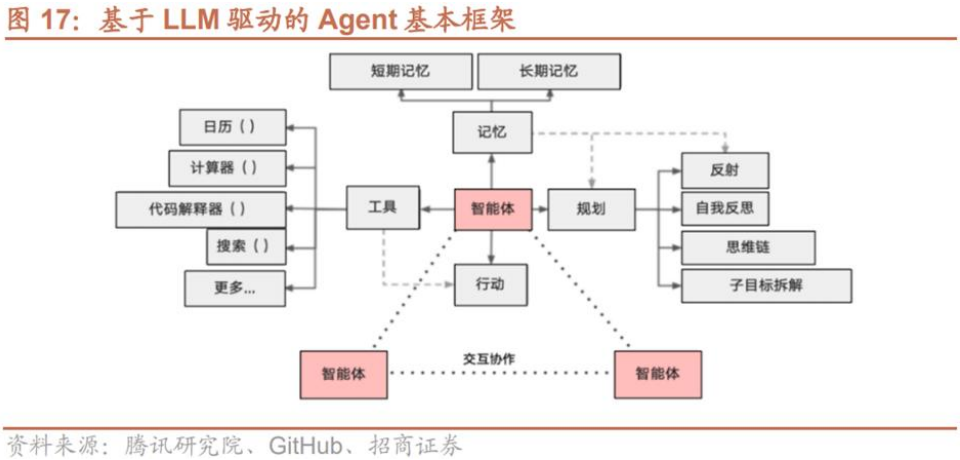


图 19.5.7 基于 AI 大模型的（领域功能增强）人工智能体的基本框架

人工智能体与 AI 大模型的一大区别即在于能够使用外部工具拓展其能力。懂得使用工具曾是人

类最显著和最独特的能力。同样地，也可以认为，人工智能体可为 AI 大模型配备外部工具来让 AI 大模型完成原本无法完成的工作。**人工智能体具备自主调用工具的能力**。在获取到每一步子任务的工作后，智能体都会判断是否需要通过调用外部工具（或系统）来完成该子任务，并在完成后获取该外部工具（或系统）返回的信息提供给智能体，进行下一步子任务的工作。人工智能体提供了可靠地将 GPT 的功能与外部工具及 API 相连的方法，它允许开发者更可靠地从外部工具（或系统）中获得相关的数据和帮助，以为智能体问题解决提供便利。

浙江大学和微软联合团队发布的 HuggingGPT 就是一个开发范例。它将模型社区 HuggingFace 和 ChatGPT 连接在一起，形成了一个 AI Agent。它可连接不同的 AI 模型，以解决用户提出的任务。HuggingGPT 融合了 HuggingFace 中成百上千的模型和 GPT，可以解决多种任务，包括文本分类、对象检测、语义分割、图像生成、问题解答、文本语音转换和文本视频转换等。其具体步骤分为：（1）任务规划：使用 ChatGPT 来获取用户请求；（2）模型选择：根据 HuggingFace 中的函数描述选择模型，并用选中的模型执行 AI 任务；（3）任务执行：使用第 2 步选择的模型执行任务，总结成回答返回给 ChatGPT；（4）回答生成：使用 ChatGPT 融合所有模型的推理，生成回答返回给用户。

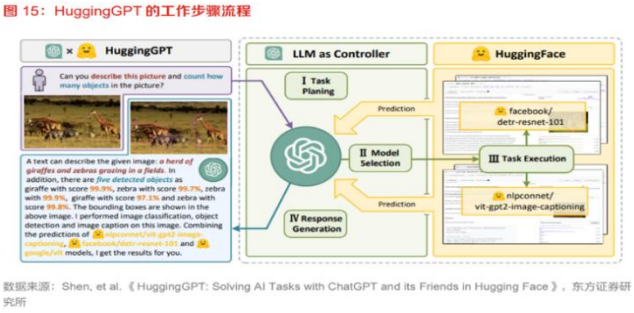


图 19. 5. 8　HuggingGPT 的工作流程

工具是工具使用者能力的延伸。在面对复杂任务时，人类会使用工具来简化任务的解决过程并提高效率，从而节省时间和资源。同样，如果人工智能体也学会了使用和利用工具，就有可能更高效、更高质量、更高智能地完成复杂任务。（现有）AI 大模型在某些方面（目前还）存在局限性，使用工具无疑可以增强基于 AI 大模型的人工智能体面对复杂任务时的能力。

AI 大模型并不具备记住每一条训练数据的能力。由于上下文提示的影响，它们也可能无法保证导向正确的知识，甚至会产生知识幻觉。再加上缺乏语料库、训练数据以及针对特定领域和场景的调整，当我们专注于特定领域时，AI 大模型的专业知识就会受到限制。专业化的工具能让 AI 大模型以“可插拔”的形式增强其专业知识，或调整领域知识以使其更适合特定领域的需求。不过，基于 AI 大模型的智能体的决策过程目前还缺乏足够的透明度，使其在医疗保健和金融等高风险领域的可信度还较低。

目前的 AI 大模型还容易受到对抗性攻击，对轻微输入修改的鲁棒性还不足。相比之下，借助工具完成任务的人工智能体则表现出更强的可解释性和鲁棒性。工具的执行过程可以反映出人工智能体处理复杂需求的方法，并提高其决策的可信度。由于工具是为特定的使用场景专门设计的，因此使用这类工具的人工智能体能更好地处理轻微的输入修改，并能更好地抵御对抗性攻击。

基于 AI 大模型的人工智能体不仅可以**使用工具**，而且非常适合**工具集成**。AI 大模型无疑可利用通过预训练过程和 CoT 提示积累丰富的世界知识，可在复杂的交互环境中表现出非凡的推理和决策能力，这将有助于基于 AI 大模型的智能体以适当的方式分解和处理用户指定的任务。此外，AI 大模型在意图理解和其他方面也显示出了巨大潜力。当智能体将 AI 大模型与特定工具相结合时，既可以降低模型工具使用的门槛，又可以充分释放 AI 大模型的创造潜能。

人工智能体有效使用工具的前提是全面了解工具的应用场景和调用方法。没有这种理解，智能

体使用工具的过程将变得不可信，也无法真正提高智能体的能力。利用 AI 大模型强大的 zero-shot learning 和 few-shot learning 能力，基于 AI 大模型的（人工）智能体可以通过描述工具功能和参数的 zero-shot demonstartion 或提供特定工具使用场景和相应方法演示的少量提示来获取工具知识。这些学习方法与人类通过查阅工具手册或观察他人使用工具进行学习的方法类似。在面对复杂任务时，单一工具往往是不够的。因此，智能体应首先以适当的方式将复杂任务分解为子任务，然后有效地组织和协调这些子任务；这有赖于（功能增强后的）AI 大模型的推理和规划能力，当然也包括对工具的理解。

智能体学习使用工具的方法主要包括从 demonstartion 中学习和从 reward 中学习。这包括模仿人类专家的行为以及了解其行为的后果，并根据从环境和人类获得的反馈做出调整。环境反馈包括行动是否成功完成任务的结果反馈和捕捉行动引起的环境状态变化的中间反馈。人类反馈包括显性评价和隐性行为，如点击链接等。

如果一个人工智能体只会刻板地应用工具而缺乏适应性，那么，它就无法在所有场景中取得可接受的性能。智能体需要将其在特定环境中学到的工具使用技能推广到更普遍的情况中，例如将在雅虎搜索中训练的模型转移到谷歌搜索中。而要做到这一点，智能体就有必要掌握工具使用策略的共同原则或模式，而这有可能通过**元工具学习**来实现。加强智能体对简单工具和复杂工具之间关系的理解，例如复杂工具如何建立在较简单工具的基础上，将有助于提高智能体归纳工具使用方法的能力。这样，智能体就能有效辨别各种应用场景中的细微差别，并将以前学到的知识迁移到新工具中。系统化的“课程”学习允许智能体从简单的工具开始，逐步学习复杂的工具。此外，若能有对用户意图以及推理和规划能力的深刻理解，智能体可以更好地给出（设计出）工具使用和协作的方法，从而提供更高质量的成果。

更进一步地，智能体还可以自己“制作”自给自足的工具。现有的工具往往是为方便人类而设计的，这对人工智能体来说可能不是最佳选择。为了让人工智能体更好地使用工具，最好有专门为智能体设计的工具。这些工具应该更加模块化，其输入输出格式也更适合智能体。如果能提供说明和示范，基于 AI 大模型的智能体还能进一步通过生成可执行程序或将现有工具集成到功能更强大的工具中来创建工具，并学会自我调试。此外，如果作为工具制作者的智能体成功创建了一个工具，那么除了使用工具本身之外，它还可以进一步为多智能体系统中的其他智能体制作包含工具代码和演示的软件包。未来，人工智能体有可能会变得“自给自足”，在工具（制作和使用）方面表现出高度的自主性。

工具还可以拓展基于 AI 大模型智能体的决策和行动空间。在工具的帮助下，智能体可以在推理和规划阶段利用各种外部资源，如外部数据库、知识库和网络应用程序系统等。这一过程可以为基于 AI 大模型的智能体提供专家级、高可靠性、多样性和高质量的信息、知识和能力，促进智能体的决策和行动。例如，基于搜索的工具可以借助外部数据库、知识图谱和网络资源提高智能体可获取知识的范围和质量。而特定领域的工具则可以增强智能体在相应领域的专业知识和能力。比如，一些研究已经开发出了基于 AI 大模型的控制器，可以生成 SQL 语句来查询数据库，或将用户查询转换为搜索请求，并使用搜索引擎来获得所需的结果。还有一些基于 AI 大模型的智能体，或者可以使用科学工具来执行化学中的有机合成等任务，或者与 Python 解释器和 LaTeX 编译器对接，以提高其在复杂数学计算任务中的性能。对于多智能体系统来说，通信工具无疑可作为智能体在严格的安全约束下进行交互的一种手段，促进智能体之间的协作，并可显示出一定的自主性和灵活性。

现在，有些可增强智能体能力的工具在与环境进行交互时是基于文本的，而有些工具则专门是为了扩展大语言模型的非文本功能的，使其输入输出并不局限于文本。用于非文本输出的工具可以使智能体行动的模式多样化，从而可有效扩展基于 AI 大模型的智能体的应用场景。例如，图像的处理和生成可以由具有视觉功能的智能体来完成。在工程领域，人们正在探索用智能体来建立物理模

型或求解复杂的微分方程；在机器人领域，人们正用智能体来规划其物理操作或控制机器人的执行。

19.6 超智能体系统中智慧的融合

19.6.1 信息生态环境下社会思维的形成与发展理论—关于信息共享机制下的信息生态环境及其对人的思想的影响研究

当今的信息时代，人们是生活在一个信息化的生态环境中。这一生态环境，是由信息资源、信息人群和信息传播渠道等共同构成的。它们**相互作用，相互影响**，构成了一个现代化的社会化的**信息生态结构，已成为了人们表达意愿、获取信息和知识以及进行社会性思维的基础**。其中，**信息资源**包括信息内容资源、信息服务资源和信息技术资源等，其核心就是经过选取、组织和有序化了的各种信息的集合。**信息人群**是指一切与信息密切相关且参与信息活动的个人或群体；他们是信息体系中可发挥主观能动性的主体，是系统中信息交流的主要对象。在信息交流活动中，**信息在人群之间进行流动，可达到诉求表达、资源互补、行为互动、共同受益的目的**。在信息活动中，每一个加盟系统的成员都是在承担一定角色：信息提供者或信息接受者。信息提供者是信息的源泉，是信息生态循环的最基本要素。信息接受者通常也是信息需求者，是具有一定的信息需求和信息接受能力并可通过信息交流、信息活动汲取信息的个人或团体等。信息提供者与信息需求者之间的信息交流和信息反馈，即构成了信息的各种循环链。在当今的网络信息时代，信息的提供者与信息的需求者之间已不再有明确的角色界限，他们是可以相互转化的：即在不同的阶段和场合可能会扮演着不同的角色。信息提供者与信息需求者之间是一种共生共存的关系，双方加盟信息（生态）系统的初衷就是希望通过供应信息或获取信息来满足自己的要求。信息人群之间的关系常具有多样性。在通常情况下，尽管信息供需各方之间没有直接联系，但会通过信息而产生相互影响。**信息传播渠道**是信息（生态）系统存在的背景和场所。人类基于信息共享而构建的信息体系，都直接或间接地影响着人类信息的表达与传播。信息生态系统是一个系统性的整体，其运行是各组成部分相互协作的结果。信息系统的运行，目前常常是以信息媒介为结构框架、以信息模式为组织核心来从事各种信息活动的。这其中，信息服务系统也起着重要作用。信息服务系统的功能：一是统合各种信息资源，通过信息内容资源实现信息人群之间的交流和交互；二是作为信息服务平台，维系着信息人群和信息资源之间的相互联系。一个成熟的信息服务系统，应是一个功能齐全、操作便捷、实用高效的信息交互平台，能够满足不同信息人群在不同信息供需阶段的实际需要，保证信息的高效流转。信息（服务）系统作为一个信息交互平台，具有**系统性、多样性和演进性的特点**。系统性是说，信息（服务）系统是由信息人群、信息资源和信息环境等要素共同构成的统一整体，是以网络、通讯、数据库技术等为基础形成的动态开放的信息交流体系。（信息）系统的生态多样性可表现在：信息的多样性、信息人群的多样性、信息人群与信息关系的多样性，信息传播与交流渠道的多样性等。而系统的演进性则是说，信息系统是动态变化的。即使在环境、时间都相对稳定的条件下，信息系统也是处于动态发展的过程之中。信息人群的需求在不断变化，信息服务在不断变化，信息人群与信息资源间的供需关系在不断变化，信息环境、信息技术也是不断发展的；系统的信息资源、组成构成模式、交流人群、联系方式等都是动态开放的。信息人群与信息资源、信息生态环境之间，一直处于一种影响和被影响的关系中。

信息人群与信息生态环境是相互影响的。一方面，信息人群的存在与发展需要与信息系统和环境进行合理的沟通和交流。信息系统和环境对信息人群的发展具有重大的推动作用，信息环境为信

息人群的发展提供了信息基础，为信息人群的成长、发展、创新创造了各种条件，当然也会制约和影响信息人群的思想和活动。另一方面，信息人群作为信息生态（系统）的主体，作为信息活动的主导者，亦是信息环境的建设者与管理者，其在信息生态系统的运行中起着积极的、能动的作用，信息人群会通过各种途径不断影响和改进信息系统和环境。

人塑造了环境，环境也在塑造着人。信息生态社会，就是具有信息需求且参与信息活动的个人或社会组织，在由其他信息人群、信息内容、信息技术、信息时空、信息制度等构成的信息环境下，成长和进行社会思维的。信息生态环境，就是人们在特定的依存环境中，以信息的收集、整理、储备和传递为主导，通过与信息环境以及其他信息人群的交流互动，所形成的一种（属人）社会环境。在一个信息生态系统中，信息主体本身具有一定的自主性；同时，信息主体所处的信息环境也会影响着他。信息主体存在于一定的信息生态环境之中，两者不可避免地会发生互动。**信息生态影响**的形成，就是信息环境状况、信息人自身状况以及信息环境与信息人相互作用的结果。信息人会主动选择自己认可的信息，扩大自身对特定信息态度的适宜度。当然，**信息生态环境也存在着动变的过程**，主要是由信息人的信息需求变化、信息环境的变化、与其他信息人关系的变化、信息人自身认知和能力的变化等因素引起的。关注这些影响和变化，对于解读人的思维和行为，将是很有必要的。

19.6.2 超智能体的自主性与主体意识

19.6.2.1 智能体的自主性与自主智能体、具身智能体

自主智能体是一个企望自主运行的智能系统，它力图实现（内部）复杂流程的自动化。有人将AI 和人类协作的程度类比为自动驾驶的不同阶段，认为目前的人工智能体的自主自动化程度已经达到了自动驾驶的 L4 阶段。可由智能体完成主要任务，而由“人”进行外部辅助和监督。

自主智能体有望带来交互方式和商业模式的变革。在交互方式方面，相比过去的 APP 软件，它已**从人适应（智能）应用变成（智能）应用适应人**。不过，现有智能体的决策/规划/执行等环节，还需要更深的“用户”需求理解和更强的工程细节打磨。比如，目前智能体运行中常常会遇见无休止的扩展要求以及输出误解等问题，这类问题的解决不单单需要智能体能力的提升，对智能体架构的设计和垂类数据的学习也同样提出更高要求。

表：AI发展阶段对比自动驾驶不同阶段

等级	名称	自动化程度	含义	示例
L1	Tool	无	人类完成所有工作，没有任何显性的AI辅助	目前绝大多数软件产品
L2	Chatbot	助理	人类完成绝大部分工作，类似向AI询问意见，了解信息，AI提供信息和建议但不直接处理工作	初代ChatGPT和Chatbot
L3	Copilot	部分自动化	人类和AI进行协作，工作量相当。AI根据人类prompt完成工作初稿，人类进行目标设定，修改调整，最后确认	Copilot、Jasper
L4	Agent	条件自动化	AI完成绝大部分工作，人类负责设定目标、提供资源和监督结果，AI完成任务拆分，工具选择，进度控制，现目标后自主结束工作	AutoGPT
L5	species	完全自动化	完全无需人类监督，AI自主拆解目标，寻找资源，选择并使用工具，完成全部工作，人类只需给出目标	类似冯·诺什曼机器人

资料来源：戴雨森即刻账号，东吴证券研究所整理

图：AI AGENT原生生态构想

图 19.6.1　人工智能体的发展阶段与生态构想

在基于 AI 大模型的（自主）智能体系统中，基座 AI 大模型的能力固然重要，但它只能解决（系统智能应用的）下限问题。在真实世界的应用场景中，自主智能体的架构设计、工程能力、垂类数据质量等都至关重要。准确度和效率也是自主智能体的重要指标。使用自主决策智能系统来完成现实中的任务，也意味着它必须具有更低的容错率。

在追求人工通用智能（AGI）的过程中，具身智能体（Embodied Agent）被视为一种关键范式，它努力将智能体的智能与物理世界结合起来。生态进化主义从人类智能发展的过程汲取灵感，认为智能体的智能应来源于与环境的持续互动和反馈，而不是仅仅依赖于精心设计的学习算法和精心编辑的预训练数据。同样，与传统的深度学习模型从互联网数据集中学习解决领域问题的显式能力不

同，人们预计未来基于 AI 大模型的人工智能体的智能行为将不再局限于纯文本的输出或调用精准的工具来执行特定领域的任务；相反，它们应该能够主动感知、理解物理世界环境并与之互动，根据 AI 大模型丰富的内部知识做出决策并产生特定行为来改变环境。我们将这些行为统称为“具身行动”（embodied actions），它使智能体能够以近似人类行为的方式与现实世界互动，并在理解世界的基础上采取行动。

19.6.2.2 自主智能体的主导意识与具身行为

在 AI 大模型广泛兴起之前，研究人员倾向于使用强化学习等方法来探索（自主）智能体的主导意识或具身行动。尽管基于强化学习的方法取得了广泛成功，但它在某些方面仍然存在局限性。简而言之，强化学习算法在数据清洗、泛化和复杂问题推理方面都面临限制，原因是，它在模拟动态且往往模糊不清的真实世界环境方面存在挑战，或者严重依赖精确的奖励信号的反馈。最近的研究表明，利用 AI 大模型在预训练期间所获得的丰富内部知识可以有效缓解这些问题。

面对错综复杂、未知的真实世界环境，智能体必须具备动态学习和泛化能力。然而，大多数强化学习算法都是为训练和评估特定任务的相关技能而设计的。与此相反，经过多种形式和丰富任务类型的微调，AI 大模型已显示出了显著的跨任务泛化能力。例如，PaLME 等对新对象或现有对象的新组合可表现出惊人的 zero-time 或 one-time 泛化能力。此外，语言能力是基于 AI 大模型的智能体的独特优势，它既是与环境交互的手段，也是将基础技能转移到新任务的媒介。SayCan 等可利用 AI 大模型将提示中的任务指令分解为相应的技能命令；但在部分可观察环境中，其有限的预置技能往往无法实现令人满意的性能。为了解决这个问题，Voyager 等引入了技能库组件，以不断收集新的自我验证技能，从而实现智能体的终身学习能力。

规划是人类和基于 AI 大模型的智能体在应对复杂问题时采用的关键策略。在 AI 大模型展示出非凡的推理能力之前，有研究曾引入了分层强化学习的方法，即高层策略约束低层策略的子目标，低层策略产生适当的行动信号。与高层策略的作用类似，具有新型推理能力的 AI 大模型也能以 zero-shot 或 demonstration 的方式无缝应用于复杂任务。此外，来自环境的外部反馈也可以进一步提高基于 AI 大模型的智能体的规划性能。一些研究基于当前的环境反馈，可动态生成、维护和调整高级行动计划，以便在部分可观测环境中最大限度地减少对先前知识的依赖，从而使计划落地。反馈也可以来自模型或人类，它可根据当前状态和任务提示评估任务完成情况。

一个自主智能体必须具有主体（主导）意识和具身行为，基于 AI 大模型的自主智能体也不例外。根据（自主）智能体在任务中的自主程度以及智能体行动的复杂程度，目前，基于 AI 大模型的自主智能体的具身行为可有（主动）观察、（自主）操纵和（行为）导航等。

观察是智能体获取环境信息和更新状态的主要方式，对提高后续智能行动的水平起着至关重要的作用。具有自主功能的智能体的（有意识的主动）观察主要发生在具有各种输入的环境中，这些输入最终将汇聚成多模态（输入）信号。一种常见的方法是使用预先训练好的视觉转换器（ViT）作为文本和视觉信息的对齐模块，并标注特殊标记来表示多模态数据的状态和位置。声音空间（Soundspaces）理论提出通过混响音频输入来识别物理空间的几何元素，从而以更全面的视角加强智能体的观察；更多的研究则将音频作为嵌入式观察的模式。除了广泛使用的级联范式，类似于 ViT 的音频信息编码进一步加强了音频与其他输入模式的无缝整合。（自主）智能体对环境的（主动）观察也可以来自人类的实时语言指令，而人类的反馈则有助于智能体获取到可能无法轻易获得或解析的细节信息。

自主操控是（自主）智能体最主要的自主功能之一。一般情况下，自主智能体的操纵任务包括任务排列、思维操控和行为操控等。一个典型的情况是，智能体按规划执行一系列任务。除了精确观察外，（自主）操控还涉及利用 AI 大模型将一系列子目标结合起来。因此，保持智能体状态与子目标之间的同步非常重要。DEPS 利用基于 AI 大模型的交互式规划方法来保持这种一致性，并在整

个多步骤、长距离的推理过程中通过智能体的反馈来帮助纠错。这就要求智能体对任务有更扎实的理解，以及有相应的多步骤规划和观察对，然后对多模态模型进行微调，以增强对高级认知指令的理解。

可移动自主智能体的**自主导航功能**也是如此。自主导航功能允许智能体动态地改变其在环境中的位置，这通常涉及到多角度和多目标观测，以及基于当前认知和探索的自主操作。为了实现自主导航，**具有具身行为功能的自主智能体**必须事先建立关于外部世界环境的（内部）地图，其形式通常为拓扑图、语义图或占用图等。例如，LM-Nav 可利用 VNM 创建内部拓扑图，并进一步利用 LLM 和 VLM 来分解输入命令和分析环境，从而找到最佳路径。一些研究则强调了空间表示的重要性并提出**空间智能**的概念，它们通过利用预先训练好的 VLM 模型将图像中的视觉特征与物理世界的 3D 重构（景象）相结合，来实现空间目标的精确定位，而不是传统的以点或物体为中心的导航行动。导航通常是一项长视距任务，可移动自主智能体的未来状态会受到其过去行动的影响，这就需要一个内存缓冲区和总结机制来作为历史信息的参考。

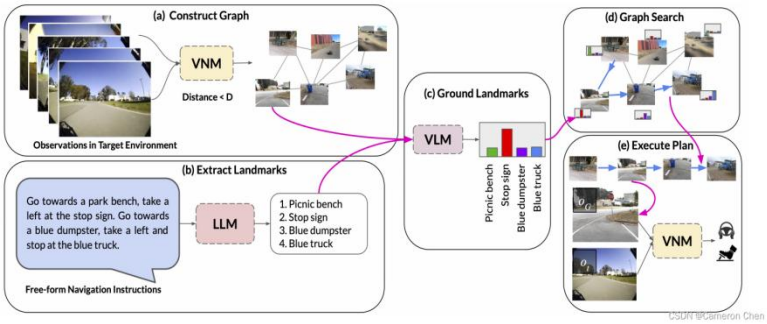


图 19.6.2 LM-Nav 的整体 pipeline 图。第一步先用 VNM 中的 distance function 在采集的数据上建一个拓扑图，图中每个节点是小车经过该位置时采集的一张图片，每条边表示两个节点间是否可达；第二步给定文本指令，用 LLM（GPT-3）提取其中的 landmarks；第三步用 VLM（CLIP）将提取的 landmarks grounding 到拓扑图中，这样在图中定位了路径点就可以规划一条路径；第四部用 VNM 中的 pose function 估计路径中相邻两点间的相对位姿，从而执行规划，同时用 distance function 做基于视觉相似度匹配的实时定位。

控制策略也是智能体智能行为得以实现的关键。基于 AI 大模型的（自主）智能体在特定的数据集上接受训练后，通常会生成高级策略命令，以控制低级策略实现特定的子目标。低级策略可以是智能体中的 Transformer，它可将图像和指令作为输入，为终端效应器以及特定具身任务生成控制命令。如今，在一些虚拟具身环境中，高级策略已被用于控制游戏或模拟世界中的智能体。

自主智能体的具身行动或可被视为智能体与现实物理世界之间的桥梁，使智能体能够像人类一样感知及改变环境。由于物理世界操作人员昂贵的操作成本等因素的存在，人们对研究可在真实世界环境中自主运行的智能体（如自动驾驶汽车或自动驾驶飞行器）已越来越感兴趣。各种具有主导意识和具身行为的自主智能体的自主操作，已包括探索、规划、自我完善甚至终身学习等。尽管取得了显著进展，但由于真实物理世界自主操作的多样性和复杂性，令实现完全自主的智能体仍是一项挑战。为了能在真实物理世界场景中有效部署具有主导意识和具身行为的智能体并确保其安全性，人们对自主智能体的要求也越来越高。尽管如此，为了让具有主导意识和具身行动的自主智能体在人类生活的各个领域发挥越来越多的作用，我们非常有必要对自主智能体进行更深入的研究。

19.6.3 超智能体系统中智能的社会融合--社会智能计算与网联融合智能

我们这里所讲的超智能体系统是指由多个（具有一定自主性的）智能体所组成的智能系统。其核心是希望显著提高系统的认知和智能行为能力，并将人类的需求作为一个不可缺少的部分，从而实现人造智能系统能在交互过程中具有人性化的思维，并在感知、决策和行动上符合人类的意图和

偏好。超智能体系统无疑给出了一个以人为本的网联智能系统的愿景。在这个愿景中，分布式智能系统和人可通过知识共享和社会意识等连接起来，以实现协同认知和共同任务完成。超智能体系统有望提供更杰出的功能，包括适应性、自主性、高效性、可靠性、安全性等，从而赋能各类复杂社会系统和问题解决系统。如今，随着智能科学的进步和智能体技术的不断发展，高度发展的超级智能系统的能力已经变得越来越强大。它们正进一步与云计算、边缘计算、AI 大模型、物联网和大数据等相融合，使网联的各个智能系统可以作为一个“团队”协作运行，以应对社会的巨大挑战。这种基于各类智能体和智慧网联的超级智能体系统，作为人工智能的具象体现，正处于逐步普及和不断发展过程之中。

尽管基于“云-网-边-端”的网联智能系统目前已经取得了显著的成就，但它们现有的能力与今后应对重大挑战所需的能力之间还是存在一定差距。我们认为，原因主要有两个方面。一是，目前的网联智能系统主要是被设计成只适用于范围相对狭窄的特定领域任务，这无疑限制了系统在更大范围内的群体智能的涌现，难以去应对任务和环境的复杂性。二是，其“无人化”的趋势变得越来越明显，使得系统中“人性”的部分没有得到很好地继承，这使人们在一些关键任务上很难信任、接受和使用它们。事实上，来自智能社会的需求正牵引着网联智能系统离开其经典的结构和受控的环境，以达到与人类一起协同工作的目的。换言之，我们希望的人工智能体或人工智能系统，应该全面达到（或超过）类似人类水平的智能，以便在复杂环境中与人类一起执行不同的任务。为了实现人类与人工智能系统之间愉悦和谐的共存，以便使人类更好地生活和工作，我们必须解决以下问题：①如何共同凝聚各种智能系统的力量，为人类提供更好地服务，以应对苛刻的需求以及动态和不确定环境中的未知事件？②这些能力强大的系统如何能与人类协作，被积极正向地感知和评价，并在社会中被广泛接受？将无人智能系统融入（智能）社会并非易事，因为我们的福祉甚至生命今后在一定程度上或许要依赖这些系统。

智能系统社会化运行是一个迫切的需求，这将使得多智能系统能够在未知环境中识别、理解和建立彼此及与人类或物体（可能先前未知）之间的关系。为了满足这一需求并更好地服务社会，我们才提出了**超智能体系统的概念、分层结构以及实现这一愿景的使能技术。超智能体系统**是**由具有（一定）社会意识的多个智能体或智能系统组合而成的**。**超智能体系统**是类人系统，并以人类的需求、意图和愿望为愿望，在交互中以人性化的思维为基本思维，并在感知、决策和行动上符合人类意愿和喜好。**超智能体系统的（智慧）网联系统代表着凝聚力【智联智融】**。更具体地说，它代表着来自大范围系统的能力联合。从广义上讲，我们并没有明确区分在单域或多域运行的智能系统，它们的共同特点是，系统是多维的、自我维持的、自我引导的，并且（可用）资源丰富。尽管多域运行，但它们在安全关键性或任务关键性上的特征是相同的。它们在合作、认知和社会性方面面临着高度的挑战。**与一般智能体**不同，**超级智能体系统**中的智能体**可具有具象化的实体，可在多个领域内运行，且具有统一的“意图”和“愿景”**。

协同智能是超智能体系统智能融合的核心。它主要是基于**协作与交互**的**社会智能**。而这种智能的来源可以是生物启发、类脑启发或社会启发。与现有的系统相比，超智能体系统的智能的内涵和来源是多样性的。当与其他系统、人类、物体或环境交互时，最先进的超智能体系统已经展现出可解决“什么？”“什么时候？”和“哪里？”等问题的能力，也可具有解决“谁？”“为什么？”“结果会如何？”和“如果不 XX 会有什么结果？”等深层次问题的能力。事实上，要解决这些深层次的问题就要求系统要拥有强大的可以匹配人类认知水平的能力（如推断和推理），以及匹配人类水平的社交能力（如关系理解和建立）。而要达到这一目标，**超智能体系统必须拥有先进的并且互补的协作、交互和社会智能**。

“集成智能”是应对复杂性的一种基本形式[多智能体智能融合系统]。超智能体中的集群是指一种新的合作模式，可直观地理解为是智能系统网联后产生的团队、功能和服务。在这里，我们主

要强调**集群**的两个方面：**成员和结构**。关于成员资格，系统中的集群模式应是开放和动态的，这意味着集群中的成员可以（随时）调整他们的策略，并在物理、信息和社会空间中产生新集群。因此，一个智能体（或系统）可能参与到多个集群中，并扮演不同的角色。关于结构，系统中的集群可以是分层的和嵌套的小组形式，也可以是由各种质量、规模和组分形成的无人系统的团队。值得注意的是，这些特征并不同于无人机编队或蜂群等传统集群。对于传统的集群，系统通常会为特定的任务而固化设计，并且机器之间或人与机器之间的社会交互规则往往是逐个任务设计的。这种封闭系统的缺点是无法与大规模网联智能系统中的潜在参与者建立关系，因而无法在新情境下为不同的任务抽象（或组合）出相应的能力。人们可以通过大规模的社会连接和协作来区别经典的智能系统与超智能体系统。这些连接和协作可通过以下方式来实现：具有灵活多变的群体成员的可扩展性，具有安全访问的互操作性，功能和服务的虚拟化，以及按需重构的灵活性，等等。

19.6.4 超智能体系统智慧融合的理论研究

超智能体系统中智慧的融合包括人与人之间的智慧融合，人工智能体与人工智能体之间的智能融合以及人与人工智能体之间的智能融合，它们是超智能体理论的核心，也是一项艰巨的工程。

19.6.4.1 人类智慧集成的心灵管理与沟通理论

人类思维和智慧的集成，应是社会科学和管理科学研究的课题。但在社会科学和管理科学的许多研究中，往往忽视了人作为一个“**智能体**”所具有的主观能动性和智能行为的特殊性，把人等同于“机器”或“动物”。人类社会集成的智能，是一种群体智能或社会智能。其管理和集成也应是智能的。我们反对任何将人等同于“机器”的所谓“科学”管理，也反对任何将人等同于“动物”的非人性管理。真正的智能管理需要心灵的沟通，思维和智慧的集成也需要以心灵的碰撞为基础。在第十六章中，我们对此已有论述，此处不再赘言。

19.6.4.2 超智能体系统智能涌现的复杂系统理论和多智能体协同理论研究

毫无疑问，超智能体系统是复杂的物质、生命、意识、思维和社会化的系统，其研究与构建，需要有基于复杂系统理论的多层面的综合分析，也需要有多智能体系统理论的支撑。我们要研究其构成基元、其信息处理机制、其结构化机制、动力机制、行为控制机制，更要研究其智能涌现机制、进化机制和社会化机制等。

在一个超智能体系统中，（智能）主体可以是同构的，也可以是异构的；可以是有自主意识的，也可以是无自主意识的；可以是有主观能动性的，也可以是无主观能动性的。但是，无论是哪一类系统，其研究都将涉及到在一组具有自主功能的智能主体之间如何集成或协调其智能行为，集成或协调它们的知识、目标、意图及规划等，以联合起来采取行动或获得解决问题的途径。在各智能主体之间，可以是协作关系，也可能会存在着为了各自利益的争论。因此，研究其思维和智能的集成，将是一个复杂问题。

超智能体系统的智能，首先是一种群体智能。其特点是：

（1）其总体思维和智能的形成常常是综合集成的结果；特别是在研讨或复杂问题解决过程中，其思维常常是分布式的，并不存在单一的控制。

(2) 群体中的每个个体都能够从事部分信息加工工作，也可以从他人处得到启迪，这种方式被称为“激发”。由于群体智能可以通过研讨的方式进行问题解决与合作，因而也就具有了较好的可扩充性。

(3) 群体中每个个体的能力或遵循的行为规则都是有限的，因而群体中单个智能体的实现比较方便，具有简单性的特点。

(4) 群体表现出来的复杂行为是通过简单个体的交互过程“涌现”出来的，因此，群体行为常常具有自组织性。

关于超智能体系统中智能涌现的复杂系统理论的研究和多智能体系统协同理论的研究已有不少论述，限于篇幅，我们在这里不再展开论述。

19.6.4.3 人-机综合集成超智能体系统中思维与智能的集成理论研究

鉴于我们目前所研究的人-机综合集成智能系统，主要还是“以人为主”、“机助人”式的综合集成系统，其思维与智慧的集成，主要还是在特定环境下“人”的思维与智慧的集成。因此，在思维与智能的集成理论研究方面，我们很赞赏钱学森院士和戴汝为院士等所做的工作。认为，思维和智慧的综合集成，需要**研讨**，需要**相互启迪**，更需要**激励效应**的产生。以前，这种相互**研讨、相互启迪**，相互**激励**，主要是在人与人之间进行，而今有了网络、有了网络知识库、有了网络知识系统，原来仅限于人与人、或专家与专家的“神仙会”或“专家会诊”式的“研讨会”，已经可以扩展为“人”和“（智能）机器”或“网络知识系统”一起进行，这是智慧集成的一次“飞跃”。在当前环境下，“人”与“（智能）机器（网络知识系统）”直接会话也许是最普遍的，“（智能）机器（网络知识系统）”可直接“帮助人”。而我们也更期待更复杂一些的“人-机智慧交互综合集成”，即多个“人”与多个“（智能）机器（网络知识系统）”的“集体智慧”“交互”与“综合集成”，也许，这才是未来社会思维和意识发展的主流。

如何在多个“人”与多个“（智能）机器（网络知识系统）”的“集体”“交互”中产生出“激励效应”并获得“综合集成的智慧”，这将是超智能体系统理论希望去深入研究的问题。

1、人-机智慧综合集成，目前可能的模式和可处理的问题

我们认为，在人-机综合集成的超智能体系统中，达成思维与智慧的集成可有多种形式，将智能芯片植入人脑，目前还不在我们讨论的范围之内。我们所考虑的，主要是多个“个人”与多个“（智能）机器（网络知识系统）”之间的“交互”集成。其交互的模式，可包括：

集中管理的任务指派型模式：此时，集成系统的“**主管者（主持人）”**明了“系统内”每个“个体（人或系统）”的“专业特长”，其“智慧集成”的关键在于让每一个“合适的人或系统”去“从事”其所“最擅长的工作”，以合作解决问题。

基于民主集中制的研讨型模式：此时，系统的功能，一是要提出一个可供研讨的论题或待解决的问题；二是要组织系统内各相关领域专家（包括人和智能系统）积极参与，并充分发挥“网络”及相关“网络知识系统”的作用；三是要创造出一个可**激发**大家思绪的氛围（如，研讨沙龙）；四是要有一个集中大家（人或其他智能体）思维和智慧的方法（如，设置系统秘书）。在研讨中，充分发挥“网络”和各领域专家（人或其他智能体）的优势是必须的，但最终还需要按“民主集中制的原则”来形成一个统一的认知或问题解决方案等。

自由讨论型模式：系统内的“沙龙”式的研讨。此时的研讨，并不预先设定统一的意志或行动规范；针对某一问题，大家只是各抒己见，相互启迪，其目的也主要是为了让每个人都提高自己的认识。

我们认为，在信息网络高度发展的今天，构建一个“人-机综合集成”的“研讨系统”已经不再是难题。而基于系统内外反复研讨来形成智慧集成及问题处理方法，将会是一种最可取的思维和智慧集成方法，而其可解决的问题和问题解决流程，或可为：

常规且无利益冲突的问题--此时，相关问题有现成处理经验、规则和方案。其处理流程可为：问题及备选处理方案系统内部公示；系统内专家群体评议及多方案优化选择；方案确定及问题解决。

常规但有利益冲突的问题--此时，相关问题有过处理经验、规则和方案，但存在利益冲突和处理异议。其处理流程可为：问题及备选处理方案系统内部公示；系统内专家群体多方案研讨与选择；提出多个可行性方案并系统内外公示；系统内外研讨及意见反馈；方案修正；方案确定、实施，发现问题并及时处理。

重大的决策问题--此时，相关问题有过部分处理经验、规则和方案，但存在利益冲突和处理异议。其处理流程可为：问题及备选处理方案系统内部公示；系统内专家群体多方案研讨与选择；提出多个可行性方案并系统内外公示；系统内外反复研讨及意见反馈；方案反复修正及完善；带有各种应急处理措施的方案确定并组织实施，实施时及时发现问题并及时处理。

需要开拓创新的问题--此时，相关问题并没有现成的处理方案或处理规则。其处理流程可为：问题系统内部公示；系统内专家群体（包括人和智能系统）研讨；可选择方案与处理规则提出；可选择方案系统内外公示；可选择方案系统内外研讨及意见反馈；备选方案初步确定；备选方案试点实施；备选方案意见反馈；对问题及备选方案作进一步处理。

2、人，如何成为超智能体集成创新系统中的“创新节点”

信息技术和网络技术的飞速发展，使人类社会正迅速形成一个知识社会和网络社会，并正将人类社会思维的拓扑结构大幅改变。若我们把人类社会的人际交往和社会关系也看作是一个“网络”，则人类历史上每一项重大技术的变革，都会使这一社会网络的拓扑结构发生改变。新的技术手段和社会组织手段，会将原本无法连接，或者不被允许连接的节点，连接了起来。比如运河的开凿，铁路的兴建，电报、电话的发明等，都构筑起了人们交往的交通网络和通信网络。如今，区块链可将人们的信用连接起来；手机、微信、直播和网络社交平台等，可将人们的注意力和社交活动等连接起来，从而产生出新的交往、交易和商业模式。

信息技术和网络技术所衍生出来的网络生态，除了改变人们的通信与交际，也改变着人们的思维方式和创新方式。一般认为，人类现存的社会，是由“三个世界”交织而成：客观的物理世界（自然）、主观的心理世界（人类个体的认知）和人文的“文化世界”（人类知识的积累）。如今，人工智能的发展，已使“人类知识系统”“活”了起来，可以充当人们吸取知识的教师和源泉，可以充当人们决策的参谋，可以成为你工作的助手。人们现在可以连接的，可以有信息和知识，可以有设备及状态（经由物联网），可以有信用（区块链），也可以有陌生的朋友（社交网络）。有励志家说，一个人成功的秘诀就是：能力驱动成功，而当能力不足以驱动成功时，社会网络会助你驱动成功。也就是说，一个人要想成功，能力必须强大；而要想能力强大，需要有一个“平台”，需要有一个强大的“关系网络”，需要把你自己与重要的人物或优秀的关系联系起来。

怎样才能把自己与重要的人物或优秀的社会关系联系起来呢？一方面，你必须喜欢学习，通过学习与经历，提升你自己知识结构的维度。一个人的知识结构维度越高，越有可能和有价值的信息联通，产生链接的通道。当然，学习后的提升，除了会为你打开通往新世界的一道道门，一扇扇窗，为你带来愉悦之外，也会为你带来宽阔的视野和观察的高度。世间万物，你会一览无余，而这是在底层时你很难看清楚的。而这种认知和能力的提高，能让你有更强的能力，去理解和洞穿以前没法看懂的问题，立即提出各类解决问题的方案；又可增加你的社会交往渠道，形成一种良性循环。另一方面，你也需要善于交往和借鉴，主动把自己与重要的节点、优秀的社会资源联系起来。

一个人 的能力，永远也无法和大千世界中涌现出来的最强大的知识网络匹敌。一个人最好的成

长策略，应该是放弃自我封闭，主动把自己融入强大的社会网络系统之中。 即使起点较低的个体或者组织，只要不断积极建立自己有效而丰富的链接网络，不断自我探索，也可以探索到“邻近可能性”所创造的新机遇。

由此，今后，创新的秘诀将不是去营造一个极度封闭的象牙塔，然后坐在象牙塔里独自去解决你的那些伟大的想法。真正的秘诀将是希望你用更多的精力去搭建一个可帮助你创新的网络。这一创新网络，本质上是需要有两个方面：一是要有一群与你有同样理想的人，并且其中不少人具备创新的能力；二是要有一个可协助你创新的可塑性知识网络。有同样理想的高素质的人群，你也许可以在（社交）网络平台上发现；而可塑性知识网络，是指这样一种网络，它可以根据邻近技术边界可能性的变化，去构建新的技术组合而实现创新。这就要求你，要善于创造机会、发现机会并敢于行动；机会，有很多是可以主动找到的。

在超智能体系统中，人，无疑是最活跃的因素，也是系统创新的“关键”节点。充分发挥人的主观能动性，是超智能体系统创新的关键。而另一方面，超智能体系统也为每个人才能的发挥提供了坚实的基础，借助超智能体网络，未来，每个人都可以成为“超人”。

3、人-机综合集成超智能体系统中的思维激发理论研究和集智计算理论研究

在人-机综合集成的超智能体系统中，思维与智慧的集成计算已是一个十分令人关注的问题，我们特别关注的是超智能体系统中的**思维激发理论研究**和**集智计算理论**研究，因篇幅所限，对此，我们将在后继文章中提出自己的研究和看法。

19.7 典型超智能体系统的开发及应用场景

19.7.1 超智能体系统的开发需求

超智能体系统的开发可基于各种需求，包括基于特定任务的开发和面向特定新环境的创新开发等。由于基于 AI 大模型的超智能体可以理解人类的自然语言指令并执行日常任务，因而是目前最受青睐、最具实用价值的智能体之一；它们具有可提高任务完成效率、减轻开发及使用时的工作量和促进更广泛使用的潜力。因此，**超智能体系统常常被用于面向特定（复杂）任务的智能系统的开发。**在面向（复杂）任务的系统开发中，超级智能体将遵从“使用者”的高级指令，承担目标分解、子目标规划、环境交互探索等任务，直至实现最终目标。为了探索基于 AI 大模型的面向特定任务的超智能体是否能够执行各项基本任务，已有研究将它们部署到基于文本的游戏场景之中。在这类游戏中，超智能体完全使用自然语言与世界互动。通过阅读关于周围环境的文字描述，并利用其记忆、规划和试错等技能，它们完全可以预测下一步的行动。然而，由于基础语言模型的局限性，它们在实际执行过程中往往依赖于强化学习。随着 AI 大模型的逐步发展，具备更强文本理解和生成能力的超智能体系统在通过自然语言执行任务方面已展现出了巨大潜力。不过，由于场景过于单一，单纯基于文本的场景还不足以作为基于 AI 大模型的（超级）智能体的测试场所。即使是测试，我们也需要构建更真实、更复杂的环境。根据任务类型，这些“模拟”环境可分为网络场景和生活场景，并需要让智能体在其中扮演各类具体角色；比如，在网络场景中代表用户去执行特定任务（或可被称为网络导航问题）。运行时，超智能体系统会通过解释用户指令，将其分解为多个基本操作，并与环境及其他智能体进行交互。系统开发时，我们必须证明超智能体系统具备在复杂的网络及真实世界场景下理解指令、适应变化以及概括成功操作的能力。这样，超智能体系统才能在未来处理看不见的任务时实现无障碍和自动化，并最终将人类从与计算机用户界面的重复交互中解放出来。研究表明，通过强化学习训练出来的超智能体，可以有效地模仿人类行为并进行预定义的操作。然而，不具备 AI 大模型功能的智能体，可能还难以适应现实世界中更现实、更复杂的场景。在动态的、内容丰富的**网络场景**中，现有智能体的性能还常常面临挑战；而在**生活场景**中，超级智能体还必须理解隐含指令并应用常识性知识等。对于完全基于海量文本训练的基于 AI 大模型的智能体来说，人类

认为理所当然的任务，可能都需要做多次试错尝试；对更现实的世界场景，由于环境的不确定性，系统往往需要面对更多、更模糊、更微妙的任务。例如，如果天黑了，房间里有一盏灯，智能体就应该主动去打开它，尽管指令中并未明确说明。超智能体能否将训练数据中**蕴含的世界知识**应用到真实的交互场景中？尽管有人证明，在适当的提示下，足够大的 AI 大模型是可以针对真实交互场景中的任务有效地将高级任务分解为合适的子任务，而无需额外的训练的；不过，仅基于静态推理和规划能力的系统是有其潜在的缺陷的。基于现有 AI 大模型的（人工）智能体，若系统缺乏对周围动态环境的准确感知，其所生成的行动，产生偏差将是不可避免的。例如，当用户下达“打扫房间”的任务时，智能体可能会将其转化为“呼叫清洁服务”等不可行的子任务。为了让智能体在交互过程中获得全面的场景信息，一些方法是直接将空间数据和项目位置关系作为模型的附加输入。这样，智能体就能获得对周围环境的精确描述。

超智能体系统的开发需求是多领域多方面的。其典型的应用领域包括生产领域、生活领域和科研领域等。其典型的应用需求可包括咨询问题解答、困难问题求解、自动化生产与运输、生活助手和情感服务等。

基于 AI 大模型的超级智能体系统，在执行特定（重复）任务和提高重复性工作的效率方面，已表现出了强大的能力。然而，在智力要求更高的领域，如前沿科学探索领域，智能体的潜力目前尚未得到充分发挥。这种局限性主要来自两个方面的挑战：一方面，科学本身的复杂性构成了重大障碍，许多特定领域的术语和多维结构难以用单一文本表示。因此，它们的完整属性无法被完全封装。这大大削弱了智能体的认知水平。另一方面，科学探索领域尚缺乏合适的训练数据，使得智能体难以理解整个领域的知识和探索途径。如果能在智能体内部发掘自主探索的能力，无疑会给人类科技带来有益的创新。于是，**面向特定新环境的创新智能体的开发就成为了超级智能体系统的一个重要应用场景。**目前，各个专业领域都在为此而努力。比如，计算机领域的专家希望充分利用智能体强大的代码理解和调试能力；在化学和材料领域，研究人员为（超级）智能体配备了大量通用或特定任务工具，以更好地理解领域知识；还有一些智能体正逐渐发展成为全面的科学助手，精通在线研究和文档分析，以填补数据空白等。

基于 AI 大模型的超级智能体在科学创新方面的潜力是显而易见的，但我们并不希望它们的探索能力被用于可能威胁或伤害人类的应用中（如达到人们所称的自主硅基生命，宇宙探索除外）。在一个开放、未知的世界中，建立一个能够不断探索、发展新技能并保持长期生命周期的、具有普遍能力的超级智能体无疑是一项巨大的挑战。目前，一些研究利用 AI 大模型将高级任务指令分解为一系列子目标、基本技能序列或基本指令操作，以协助智能体逐步探索开放的世界。他们从类似于 AutoGPT 的概念中汲取灵感，基于“发现尽可能多的不同事物”这一长期目标，开发了基于 AI 大模型的体现式终身学习智能体。他们引入了一个用于存储和检索复杂动作的可执行代码的技能库，以及一个包含环境反馈和纠错的迭代提示机制。这使得智能体能够自主探索和适应未知环境而无需人工干预。也许，一个能够（长期）自主学习并可掌握整个真实世界技术的超级人工智能体的出现，可能并不像人们想象的那样遥远。

19.7.2 基于单一人工智能体的超级智能体系统的开发

单一人工智能体无疑是超级智能体系统构建的基础。作为一个具有物理实体的超级智能体系统，它可有不同的规模，在不同的领域工作，甚至跨域运行，并以多种模式执行不同的任务。此类系统常常以模块团组的形式使用，以获得更好的集群性能。在这种情况下，它们可以自组织、自组装，并可重构其物理“身体”。该集成智能系统的构建在功能结构上通常采用分层结构。作为一个超级智能体，它们至少包括着 4 个核心层面：**平台层；感知和网络层；计算层；服务层**。这 4 个层面相互影响和作用，每一层都可嵌入智能、认知和价值。关于这 4 个层面的一些关键使能技术，我们简述如下。

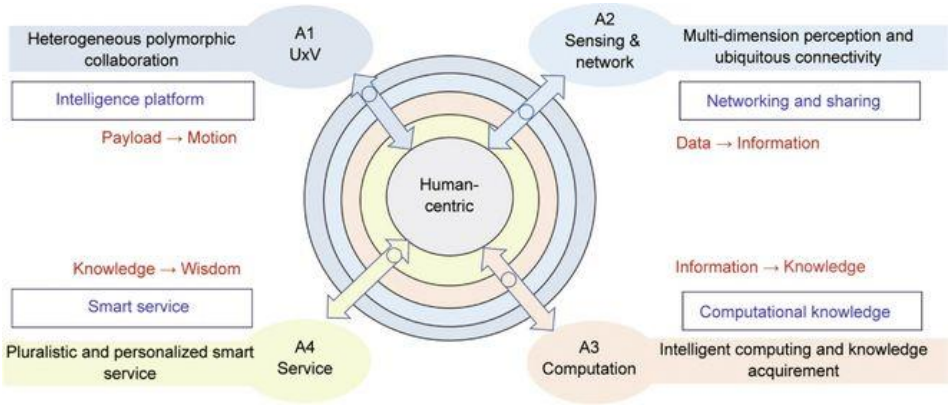


图 19.7.1 一个超级智能体的（功能）分层架构。A1：平台层；A2：感知和网络层；A3：计算层；A4：服务层

1.系统平台层

平台是系统构建的基础。平台也许可以选择现有智能系统的物理平台和（信息）计算平台。但超级智能体系统却不同于一般智能系统，它具有**自主性、跨平台协同功能和可移动性**。跨平台协同和可移动的超级智能体系统需要通过多个平台的协同来实现。而实际上，超级智能集成系统需要在三个方面进行集成：物理资源集成、信息资源集成和智能资源融合集成。物理空间中包含着可交互的所有相关的物理系统和物体，可确保系统在多域中的运动安全。信息空间（赛博空间）涉及感知、通信网络和计算。系统在这一领域活动的高阶结果是知识的产生和共享。在智能化的社会空间，超级智能系统可在其他两个空间的支持下，进行可信赖的社会交互，产生集成智能。**物理空间中的集群技术**，可为系统聚集必要的物质资源，建立以人的活动为基本情景的实体化集群生态。**信息（赛博）空间中的集群技术**，将围绕人类的认知需求进行信息汇集，将异质的、非结构化的、不完整的、不同步的原始数据处理成系统化的信息和知识，以增强系统的思维和决策能力。在这个过程中，语义理解技术能够增强针对物体和环境的以人为本视角下的互操作性，并允许非专业的人员积极参与。这是形成学习、思考和理解的协作能力的坚实基础，并可产生融合的情境感知。一方面，系统会在信息资源的帮助下增强认知；另一方面，系统也会依靠自身携带的感知和传感手段，通过系统的认知计算，为知识的产生作出贡献。**社会空间中的（智能化）集群技术**，可使系统成员对其他（智能）机器和人类的行为进行主动地理解、建模和推断。系统可以围绕中心主题建立相互关系，形成智慧群体。事实上，社会空间与其他两个空间共同协作，可以建立“类社会组织”，在信息（赛博）空间中进行各种推理、数据共享和互操作等。

2.感知和网络层

分布式传感和网络化通信技术正在深入融合。特别是在网联智能系统的背景下，感知和通信网络更是紧密相互依赖。在超智能体系统中，我们可以使用传统意义上的感知（传感）和通信网络，但必须考虑它们与人类的感知交互以及和社会（认知）网络的耦合。

感知是通过使用本体感知或外体感知的传感器来实现的。原始传感数据经过处理后，可以作为对系统本身、团队成员、物体和环境的态势感知的基础。通过传感和通信网络的融合，系统可支持本地（监测）数据在静态/移动传感器节点之间共享和传输，从而进一步实现增强态势感知。这种来自多个系统的分布式感知可以在更短的时间间隔内，在广阔的空间内收集更丰富的信息。一般来说，传感和通信网络共同促进着多个智能系统的协作，并确保整个系统的安全、效率和便利。

在为系统设计**通信网络**时，我们需要考虑平台的协同性和开放的架构。系统的协同性为确保不同平台高动态连接和减轻环境干扰方面带来了挑战。在这个意义上，系统在通信可用性、连续性、完整性、延迟和通信时间方面的要求可能是苛刻的。不断升级的无线通信技术可以为可移动无人系

统提供广域通信，具有超低的时间延迟和大规模连接。先进通信技术这样的特点也使新的传感网络成为可能。如触觉互联网，它能为人类提供如身临其境的态势感知，以让我们远程操作无人系统。移动性还允许无人系统提供动态覆盖和无缝服务，因为无人系统能够携带中继和基站有效载荷进行移动按需通信。

综上所述，系统感知和通信网络可促进系统更广泛的协同；同时，协同性也促进通信产生更广泛的链接。我们通常是使用自组织网络来扩展系统的结构。在超智能体系统中，需要有一个更加开放和动态的通信架构，以形成一个开放的（智能）社会，其成员可能代表着不同的利益相关者。关于系统中“究竟谁是通信合作伙伴，信息需求是什么，以及通信协议是什么（可能不是固化的）等”问题的解决，可能都是具有挑战性的问题。计算交互及集成算法在这方面已进行了探索，它们或可以作为一种解决尝试，以提升协作式系统的性能、设计基础和开放性等。

3.计算层--核心是基于知识和思维的计算智能

系统的认知和协作能力需要超智能体系统具有在情境感知下的知识共享、认知计算和泛在计算能力。支持超级智能体的计算技术要涉及到数据、算法和算力。要实现超级智能体的计算智能，除了常规的智能计算（如具体的规划、决策和控制算法），我们还必须考虑（不同类型）知识的表征、（系统的）集成计算架构和社会智能计算等。

我们首先需要考虑知识的生成和计算架构。数据、算法和算力是当前人工智能的核心基础要素。对于超级智能体来说，来自通信网络系统的高容量、异质和多模态的数据可能会带来数据管理问题。为了解决这些问题，系统可以采用语义知识的方法来提炼数据。这种方法不仅可以重复使用旧的数据、增加新的数据、传输无歧义的知识，而且还在数据采集、处理和表示过程中更有利于“使用者”更好地观察。这种透明度为人类的参与提供了一个先决条件。如此，对知识生成的结果，我们可以用含有认知和社会特征的**“图谱”**来表示。包括表示物理空间的时空和主题特征的**几何图**，表示信息空间中人类可解释的概念的**语义图**，表示网络关系和社会接受的**社会使能图**等。在实践中，基于本体论或因素空间理论的语义技术已经可以支持可解释的知识表征和柔性的结构重构。语义知识表示可以是特定领域的，并强调其可靠性、安全性和实用性。它们带来的好处是标准化的情境信息、时空上下文的及时更新以及平台行为的协作，这同时也增加了平台的互操作性。

认知计算（算法）可以产生系统的认知。这增强了系统与（智能）机器、环境及人类交互的能力。认知不是系统心智的一个模块，而是一个包含着具有鲁棒性和预见性的系统行为过程。虽然认知的特征可以体现在感知、计算（推理、推断、决策、意图解释）、沟通、行动和目标调整之中，但**认知系统需要自我意识到感知的影响并预测行动的后果**。使能的概念很好地体现了这种自我意识。它描述了一种环境可以提供（或负担）一种智能行为的可能性。使能将感知到的对象与适用的行为相关联，将对感知到的事物的理解映射到有后果意识的行为中。相关研究提出，这种映射可以通过端到端学习获得，可不需要内部模型。由于人类喜欢与具有一定认知能力的系统交互，认知功能可以帮助超级智能系统理解人类的需求和目标，从而在与人类的交互中保持自主性。

协同计算也可产生集成智能。超智能体系统的社会空间中的集群要求其拥有高度的社会自主性，并可展现出人类社会可接受的行为。目前已实现的超级智能系统的社会空间已经包括确定性的社会价值取向以及使用前景理论的其他非理性因素（如框架效应、寻求风险的行为和损失厌恶行为等），其社会规范和声誉也可以被用来帮助促进人类对机器的理解；通过机器学习技术手段的使用，也可以自动地获取某些社会特征，如利他主义、好奇心、注意力和使能承担等。社会空间中的集群也要求与人类交互的媒介具有多模态、交互式和多任务的个性化特性；而界面可以是物理的（如服务机器人）、虚拟的（如软件助手）、生物物理的（如脑机接口或机器人肢体的身体增强形式）、信息物理化学的、触觉的或语言（如自然语言解释）的。

超级智能体系统的发展还体现在“心智”建设上，即可实现的“看-想-做”循环（也即感知、决

策、行动循环）。鉴于人工智能研究方面取得的突破性进展，一些新近的研究再次倡导具有认知能力的系统架构。对于超智能体系统来说，认知不仅促进其主动感知，也促进其实现预测性决策。例如，系统可以被设计成具有社会认知的导航规划能力，去实现以人为本的人类活动区域的“穿越”。一般来说，通过认知和“物理身体”的交互作用，智能系统将有能力探索具有挑战性的任务。

总之，超级智能体系统可提供精确性、有效性，甚至安全性；提供超越人类极限的感知和灵巧操作。然而，它们也有不足。首先，就目前的系统而言，进行一般性思考和社会性思考（达到通用人工智能）仍然是一项具有挑战性的任务。智能技术未来面临的挑战，其中就包括社会方面和伦理方面。其次，人们期望协同平台系统具有更强的损伤复原能力，在感知和行动方面具有适应性，并且在系统规模上具有可扩展性。当与人类合作时，超级智能系统也被要求拥有更强的自预测、自修复和自成长的能力。

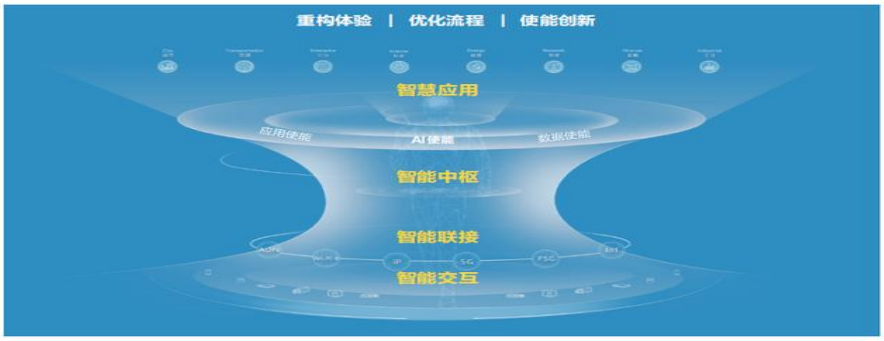
如今，计算技术的最新进展涉及计算方法和计算架构。新的计算方法包含了广泛的范围，从**量子计算技术、DNA 计算技术和生物化学计算技术到神经形态计算技术**等。随着分布式智能计算变得越来越普遍，泛在计算已被推广到分布环境、分布平台和分布系统中。关于计算架构，目前主要有三种不同的类型**---云架构、雾架构和边缘架构**，它们对分布式计算尤为重要。对于超级智能系统来说，**雾计算架构**应该享有更大的时间空间优势，因为它具有高移动性和快速的地理分布覆盖、在信息空间进行延迟敏感的知识生成、在社会空间增强对行为敏感的自我意识等诸多方面的能力。**边缘计算架构**促进了“云-网-边-端”智能系统的发展，也使以即时数据为中心的计算更接近数据源。在超智能体系统中，上述三种类型的计算架构可以联合使用，并可在通信连接和安全约束下，以完成任务为目标，进行平滑的自我切换。

4.服务层—基于行为智能、交互智能和综合智能的智能服务系统

超智能体系统在融合上述三个空间的基础上可为人类提供智能化服务。所提供的服务包括着**类人行为模式和智能化服务及管理**等。首先是**类人行为模式**。提供个性化、精确和创新的服务首先需要了解人类行为模式。这些行为模式可通过计算模型来建立全面图像。例如，（人形机器人等的）人类的行为模式在物理空间可以被描述为运动/人群动力学，在信息空间可以被描述为言论动力学和沟通模式，在社会空间可以被描述为友谊网络和社会偏好等。进一步地研究表明，在物理空间中，超级智能系统可表现出高度仿真的类人行为；在信息和社会空间中，超级智能系统可与人类的信念、决策以及他们的社会网络相耦合，导致人-机系统可实现相互信赖的动态交互。其次是**智能化服务和管理**。在物联网络的支持下，超级智能系统有望在智能家居、智慧城市、智能制造和智能化家政服务（家庭服务机器人）等任务中提供丰富的、个性化的和按需的服务。这些服务可以通过使用面向服务的范式来提供。研究表明，即使对于自主性较低的单一智能体，使用任务调度作为服务管理也是可行的。通过动态的服务匹配、发现、替换和合成，超智能体可以有效地管理在不同层次上的智能化的服务功能。

19.7.3 包含多个智能体的超级智能体系统的开发

应用系统的开发需要根据不同的需求提供场景化的解决方案。一个包含多个智能体的超智能体系统，除了其各个成员智能体外，还必须包含智能交互、智能联接、智能中枢、智能应用等四大部分。其中，**智能联接**联接着物理世界和数字世界，可实现无缝覆盖、万物互联，让资源、数据、软件和 AI 算法在“云-网-边-端”自由流动；**智能交互**可实现应用协同，数据协同，组织协同；**智能（管理）中枢**是超智能体的大脑和决策系统，基于“云”“雾”等基础设施，可采集信息，使能数据，赋能系统，支撑全场景智慧应用；**智慧应用**通过与现实的协同创新，可加速智能技术与行业业务的深度融合，重构体验、优化流程、使能创新。当前，最实用的超级智能体系统应是基于 AI 大模型的多智能体协同的“云-网-边-端”网络智联的一体化智能系统。



图表15 智能体技术架构

图 19.7.2 一个超级智能体的典型技术架构

1.智能交互—多智能体智能产生和融合的关键

智能交互首先是物理世界和数字世界的联接点，是智能体的“五官和手脚”，让智能体可以感知（外部环境）和执行（特定操作）。智能交互更是各智能体之间智慧相互融合的关键。全场景智慧源于对万物的感知被充分唤醒和千亿智能体智慧联接的升级。而伴随着感知和联接能力的全面提升，人、（人工）智能体与现实（环境或物）将在数据构筑的智能环境中进行交互。这一切都是进入基于多智能体的“智慧社会”的前提。感知、学习和交互可塑造系统的智能，系统智能可提升系统的认知，系统认知又可锐化系统的感知、学习和交互，如此循环不止，智慧则生生不息。

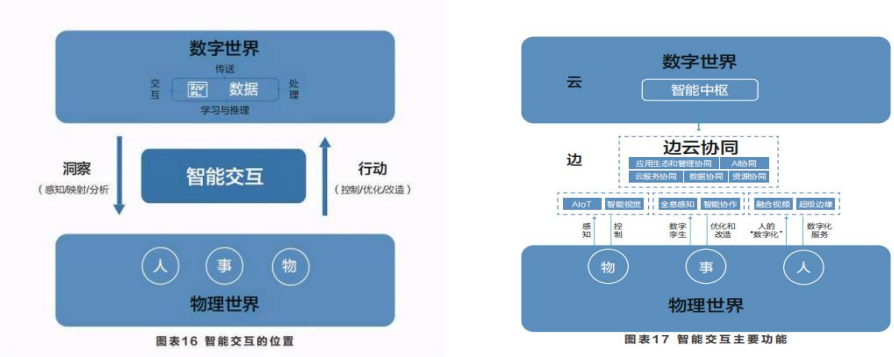


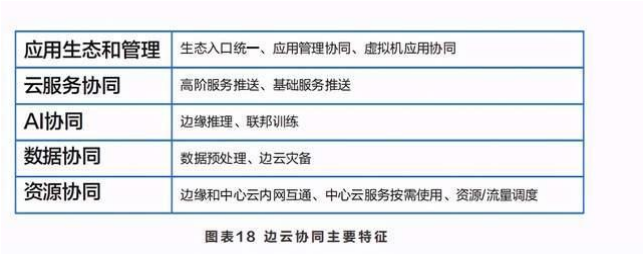
图 19.7.3 智能体的智能交互

智能体与环境及“物”的交互能力包括智能视觉（感知）和智能物联（互联），主要是感知与控制。智能化的边缘设备和系统可以实时获取环境及“物”的信息，感知环境和“物”的状态变化，并根据事先确定的规则做出判断，甚至根据感知的数据做出智能研判，从而给出智能化的控制。而针对任务（事）的交互能力包括全息感知和智能协作，通过智能分析做出（智能）决策，或实现对客观事物的优化和改造等。决策不仅需要感知外部事物的状态，对外部事物数据（现有状态）进行数字化，更要对事物在时间轴上的动态变化做出感知、模拟和分析，乃至给出预测和研判，进而达到优化和改造物理世界事物和过程的目的。

未来，在生活和工作的各个场景中，无所不在的感知节点，如道路上的车辆、工厂中的设备及制品、货运途中的集装箱、室内或户外的环境监测设备……都将被打上“数字标签”，由此而带来的“数据洪流”将由高速联接系统汇聚到（数字化网络）中枢，通过智能系统的处理，再为人类提供“供你所需”的智慧服务。智能交互更会通过感知物理世界，形成对物理世界的洞察和描述，并优化和改造物理世界，使得人与物、物与物从过去的“建立联接”转向“持续交互”。

未来，具备“边-云”协同操作系统的智能化边缘智能体，将会是智慧互联的关键部件。它既要适配不同终端的差异性，又要和位于中心的智能中枢进行“恰当”配合，让资源、数据、云服务、

生态环境及智能系统等协同起来，就近提供丰富及时的应用服务。智能边缘（系统）可以位于数据源与云端数据中心之间的任何地点，以节点、网关甚至是边缘云的形式存在，针对“物、事、人”提供各类交互能力。由此，“云-网-边-端”协同将是联接智能中枢（云端）和智能（交互）边缘必不可少的能力，它将包含如图 19.7.4 所示的五个核心要点。



应用生态和管理	生态入口统一、应用管理协同、虚拟机应用协同
云服务协同	高阶服务推送、基础服务推送
AI协同	边缘推理、联邦训练
数据协同	数据预处理、边云灾备
资源协同	边缘和中心云内网互通、中心云服务按需使用、资源/流量调度

图表18 边云协同主要特征

图 19.7.4 云-边协同的主要特征

多智能体间的交互与协调能力更是基于 AI 大模型的超级智能体系统智慧融合的关键。尽管 AI 大模型本身已拥有值得称道的文本理解和生成能力，但它们现在还常常是作为一个孤立的实体运行的。它们（大多）还缺乏与其他智能系统的协作以及从与其他智能体的社会互动中直接获取知识的能力。这种局限性限制了它们与其他智能体交互或从与他人的多轮合作与反馈中学习以提高性能的潜力。此外，在那些需要多个智能体之间进行协作和信息共享的复杂场景中，它们也无法有效开发和部署。其实，早在 1986 年，明斯基就在《心灵社会》一书中前瞻性的预言，认为智能是产生于许多具有特定功能的小型智能体的相互作用之中。例如，某些智能体可能负责模式识别，而其他智能体可能负责决策或生成解决方案。作为人工智能的主要研究领域之一的多智能体系统，所关注的重点，曾是一组智能体如何有效地协调和协作以解决某些复杂任务的问题；一些专门的通信语言（如 KQML ）也曾被设计出来，用以支持智能体之间的信息传输和知识共享；但是，它们的信息格式相对固定，语义表达能力有限。进入 21 世纪，强化学习算法（如 Q-learning）与深度学习的结合，已成为开发可在复杂环境中运行的多智能体系统的重要技术。如今，基于 AI 大模型的构建方法，更展现出巨大的潜力。如今，智能体之间的自然语言交流变得更加优雅，也更容易为人类所理解，从而大大提升了交互的质量和效率。

具体来说，如今，基于 AI 大模型的多智能体系统可以提供多种优势。根据分工原则，系统内具备专业技能和领域知识的单个智能体，可以专注于所从事的特定的任务。这一方面可通过分工，使特定的智能体处理特定任务的技能日益精进，另一方面，通过将复杂任务分解为多个子任务，可以省去系统在不同流程之间的切换时间。最终，多个智能体之间的高效分工可以完成比没有专业化分工时大得多的工作，从而大大提高整个系统的效率和产出质量。

根据已有的研究，多智能体间的协作和交互方式，大致可分为取长补短型的合作式交互以及最终互利共赢的对抗式交互。

系统内各个智能体的互补性和合作交互是智慧生成的基础条件。在当前基于 AI 大模型的多智能体系统中，智能体之间的交流主要使用自然语言，这被认为是最自然、最易为人类理解的交互形式。而现有的智能体间的合作式交互主要有两类：无序合作和有序合作。**无序合作**是说，当系统中有三个或三个以上的智能体时，每个智能体都可以自由地公开表达自己的观点和意见。他们可以提供反馈和建议，以修改与当前任务相关的反应。整个交互讨论过程不受控制，没有特定的顺序，也没有预先引入一个标准化的协作工作流程。我们把这种多智能体间的合作称为无序合作。ChatLLM 网络或可以看作是这一类型的一个示例。它模拟了神经网络中的前向和后向传播过程，将每个智能体视为一个单独的节点。后一层的智能体需要处理来自前面所有智能体的输入，并向前传播。在超智能体系统中，此类交互的一个潜在的解决方案是在多智能体系统中引入一个专门的协调智能体（作为

系统秘书或主导智能体），负责整合和组织所有智能体的响应，从而更新最终结论。然而，整合大量反馈数据并提取有价值的见解对协调智能体（秘书或主导智能体）来说是一项艰巨的任务。当然，多数表决也可以作为在此情形下做出适当决策的一种有效方法。**有序合作**是说，系统中智能体间的交互必须遵守某种特定的规则。例如，按顺序逐一发表意见，下游智能体只能关注上游智能体的输出等。这样，任务完成效率会大大提高，整个交互讨论过程也会变得井然有序。CAMEL 就是一个双智能体合作系统的成功实例。在角色扮演交流框架内，智能体分别扮演人工智能“使用者”（下达指令）和人工智能“助手”（通过提供具体解决方案来满足请求）的角色。通过多轮对话，这些智能体可自主合作完成对系统的指令。我们可以将双智能体合作的理念融入到基于多智能体的超智能体系统的操作中，交替使用快速和深思熟虑的思维过程，以使各个智能体在各自的专业领域内发挥各自的优势。

为了提高交互合作的质量和效率，有研究希望智能体能从人类（社会）成功的合作案例中学习，通过将先进的人类流程管理经验编码到智能体提示中，多个智能体之间的合作也会变得更有条理。然而，在实践探索中，也发现了不少多智能体合作的潜在威胁。如果不制定相应的规则，多个智能体之间的频繁互动可能会无限放大当初轻微的幻觉或其他错误。引入交叉验证或及时的外部反馈等技术，可对智能体间的合作产生积极影响。但最好的办法，可能还是使用具有统一意愿的超智能体系统。

合作并不排斥对抗。对抗性互动也可促进进步。传统上，合作方法在多智能体系统中已得到了广泛探索。不过，研究人员越来越认识到，将博弈论的概念引入系统或可以带来更稳健、更高效的行为。在竞争环境中，智能体可以通过动态互动迅速调整策略，努力选择最有利或最合理的行动来应对其他智能体引起的变化。在一些非基于 AI 大模型的竞争领域，已经有成功的应用。例如，AlphaGo Zero 就是一个围棋“智能体”，它通过自我对弈实现了“思维”和“技能”的重大突破。同样，在基于 AI 大模型的多智能体系统中，通过竞争、争论和辩论，也可以自然而然地促进智能体之间的变革。通过放弃僵化的信念和进行深思熟虑的反省，对抗性互动可以显著提高回应的质量。有研究团队首先深入研究了基于 AI 大模型的智能体的基本辩论能力。研究结果表明，当多个智能体在 “针锋相对”的状态下表达自己的论点时，一个智能体可以从其他智能体那里获得大量的外部反馈，从而纠正自己被扭曲的想法。因此，多智能体“合作”对抗系统在需要高质量响应和精准决策的场景中，具有广泛的适用性。在推理任务中，也可引入辩论的概念，并赋予智能体适当响应来自同伴反馈（回应）的能力。当这些回应与智能体自己的判断出现分歧时，就会发生 “心理级别”的争论，从而完善解决方案。比如，ChatEval 就建立了一个基于角色扮演的多智能体裁判团队。通过自发的辩论，智能体可对 AI 大模型生成的文本质量进行评估，其质量可达到与人类评估员相当的优秀水平。多智能体（合作）对抗系统的性能已显示出相当大的应用前景；然而，该系统目前基本上还是依赖于 AI 大模型的力量，并面临着一些挑战。例如，在长时间的辩论中，AI 大模型有限的语境还无法处理整个输入。在多智能体环境中，反复交互的计算开销无疑会大大增加。

多智能体的协商也有可能会不收敛或收敛到不正确的共识，而所有智能体都坚信其准确性。多智能体系统理论的发展还远未成熟。若在适当的时候引入人类向导来弥补（人工）智能体的不足，这无疑是促进多智能体系统进一步发展的良好选择。这也是我们更加注重对超智能体系统进行研究的原因之一。

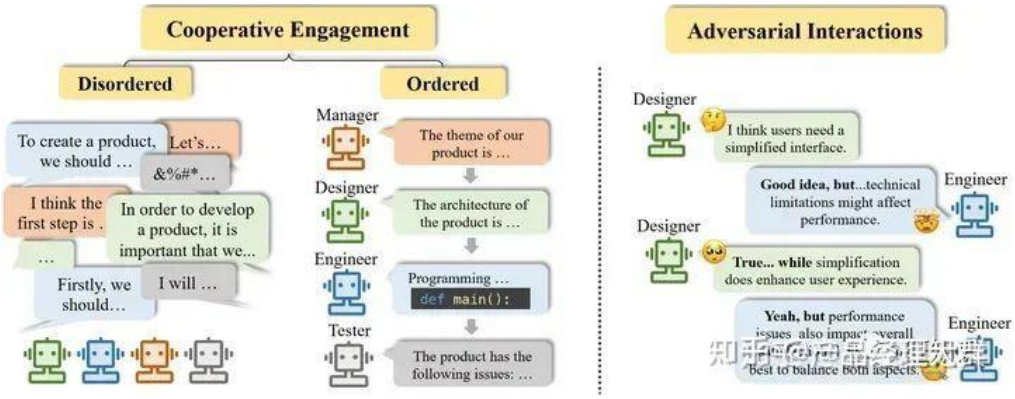


图 19.7.5 基于 AI 大模型的多个智能体交互的场景。在合作互动中，智能体以无序或有序的方式进行协作，以实现共同目标。在对抗式交互中，智能体以针锋相对的方式展开竞争，以提高其各自的性能。

2.智能联接—开启智联智能

智能联接是智能体间的“神经及信息通道”，联接智能中枢和各个智能交互模块。

随着智能进入到各类业务系统，对网络联接的需求，也发生了巨大的变化。从联接“人”发展到联接“物”，联接应用，联接数据。因此，我们不仅需要光纤这样的物理联接，提供千兆接入，满足个性化业务需求的不同时延和可靠性；还需要实现数据资产在新老应用之间流动和共享的应用一张网，以及在人与“组织”之间协同的办公一张网。为了满足这些需求，智能联接首先需要通过光纤等物理联接提供泛在千兆、确定性体验和超自动化的网络，实现无缝覆盖，万物互联。被联接的“人”、“物”、系统和设备等也都可以变为可以相互交互的“数字物种”，其产生的海量数据源源不断地汇入到（系统）智能中枢，再将智能中枢产生的智慧带到（业务系统的）每一个场景，形成全场景智慧。**泛在千兆（网络）**可实现屋内屋外、有线无线的全场景千兆网络，实现全场景、全触点、无缝覆盖、随身体验的“沉浸式千兆体验”。**确定性体验**是指基于不同应用场景对网络联接的不同服务需求（例如，速率、抖动、时延、可用性等），按需组合网络功能单元。智能联接所能提供的 SLA 等级越高，越能满足高端行业细分领域的需求。**超自动化**是指系统可参考自动驾驶的理念对网络进行分级标准管理，将智能技术、自动化技术等与网络进行深度结合，从面向网元的自动化设备管理转变为面向全场景的自动化，最终实现网络端到端的自治，以应对未来网络运营维护所面临的挑战。另外，还可通过让数字资产能生于云（云上开发/集成），也能在云上再生（云上持续共享变现），实现数字资产在不同应用之间流动，形成应用数据协同一张网。

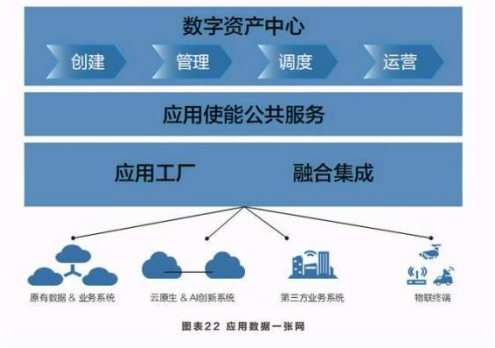


图 19.7.6 应用数据协同网络示例

针对高频发生的办公场景，系统需要提供全场景无缝流转的办公一张网业务新体验，以提升信息互通和协同的效率。同时，为确保信息安全，办公协同系统需要从芯片、终端、管道到云都提供

全方位安全保障。

图 19.7.7 办公协同网络

3.智能（协作与集成）中枢—提供一体化管理，也生成协同（认知）智能

智能中枢是超级智能体系统的大脑和决策系统，也是海量数据和知识的汇聚点。智能中枢首先需要对各式各样的数据（数字、文字、图像、符号等）和知识进行筛选、梳理、分析，并加入基于常识、行业知识及上下文所作的判断，形成智能分析、决策和辅助行动，回答和解释诸如“如何？”、“为什么？”、“如果不？”等较为复杂的问题，助力实现各行业的全场景智能。

分散数据（知识）如何组织和各类“数据（知识）--应用”如何衔接，是智能化集成的主要困难之一。智能中枢的核心任务之一是打造中央“蓄水池”，让数据、知识和智能（能力）持续积累，实现不断学习和改进。在系统构建时，智能中枢应包括云基础设施，数据使能、智能使能和应用使能等功能模块。

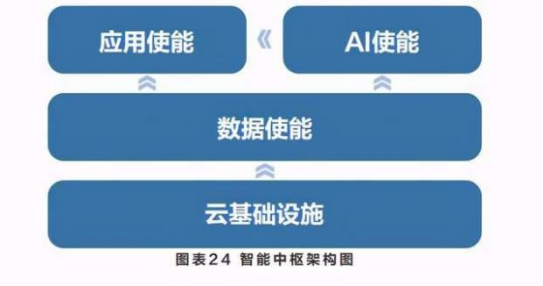


图 19.7.8 智能中枢架构

这里，**“云”基础设施**是智能中枢的底座，它对智能所依赖的数据、算力、算法和智慧应用都能提供足够的能力支撑。现实情况下，私有云厂家常常缺乏高阶服务的能力，公有云虽然能力最全，但却不能很好的匹配企业的组织结构。混合云架构两相兼顾，应是（政企）用户智能升级的首选。但系统构建时需要注意，系统流程首先要符合（政企）用户的使用习惯；例如，匹配从省到委、局、处、科的（政企）多层级组织结构。其次，它要能无缝同步公有云，最好能通过高速专线，将强大的公有云能力共享给私有云，使得数据和新老业务全域互通，真正让系统可发挥出其应有价值。

数据使能让物理上分布在不同部门的数据，在逻辑上可集中管理和分析，实现数据全域共享。面向用户的数字化运营诉求，可提供一站式的智能数据管理能力，帮助用户快速构建从数据接入到数据分析的端到端的智能数据系统，消除数据孤岛，统一数据标准，加快数据变现等。

图 19.7.9 数据使能

智能（智慧）使能输出“智力”，它决定了智能体的智能水平。“智力”能力主要分为感知能力、认知能力和决策能力。要达到这三个层面能力的融合，**智能使能**需要包含 3 个模块：“智力”开发平台，知识计算，以及基于前两者衍生出来的智能化应用开发套件。其中，“智力”开发平台是面向智能系统开发者的一站式开发平台，它包括数据处理、算法开发、模型训练、模型部署等功能。它可面向不同“岗位”和具有不同经验的智能系统开发者（如，应用开发者、数据科学家、AI 系统专家、AI 系统运维人员等），提供便捷易用的使用流程，让系统开发变得更简单、更方便。**知识计算**可将行业知识与智能系统相结合，让大量存在于结构和非结构化数据中的行业知识显性化释放出来，驱动行业主业务系统的创新。**知识计算包括知识获取、知识表示、知识管理、知识应用**等核心部件。智能系统开发套件是智能生产力的工具，它将算法专家和行业专家积累的知识沉淀在相应的套件和“业务工作流”中，真正实现赋能于智能化业务应用系统开发者，全面提升行业智能系统开发的效率和落地的效果。

图 19.7.10 智能（智慧）使能

应用使能则希望通过低代码或零代码的开发能力，支持全云化在线开发及云上云下一键部署。可不断沉淀行业业务资产，实现软件资产等的重用，让开发者实现乐高式的轻松开发应用。同时，通过标准化、中心化、服务化、非侵入式的方式，让新老应用实现数据互通。

4.智慧应用与环境交互，智能的综合与融合应用

智慧应用是智能体的价值呈现，每个个体所能感受到的个性化、主动化服务体验都来自应用。但超级智能体系统的智慧应用并不是传统应用的搬迁。因为其知识是高度全面化的，其智能是高度集成化的，其智慧应用是在两者结合的基础上高度智能化的，而这需要多个数据系统、知识系统和智能系统的深度参与和有机综合。智慧应用生态发展需要一个一体化的平台。降低系统使用门槛，沉淀行业知识，实现开发到需求的良性循环，是应用系统开发发展所必须的，而这个过程也需要有良好的开发环境支持。

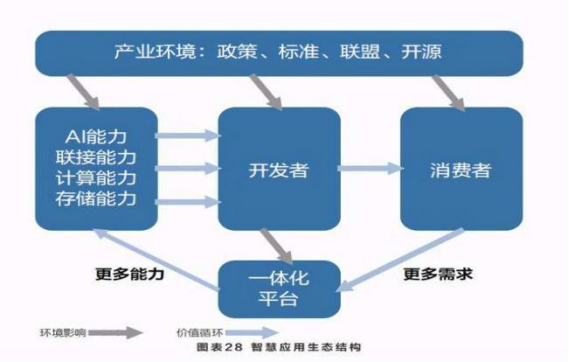


图 19.7.11　智慧应用生态结构

要打通从开发到实际应用的产业链条和商业环境，让市场消费方的通用需求和行业知识转化为系统的框架、工具和服务，并能从所开发的产品中获益，形成健康发展的智能生态，这些环节都需要通过一站式的开发环境和应用超市来打通。

一站式开发环境的本质是**技术赋能**，是把软件开发者和软件集成商所需的数据技术（智能技术、计算技术、联接技术、系统集成技术、系统体验、系统运维、系统管控等）从艰深的工作中抽象出来，以标准化、产品化的方式提供，降低数据技术上手和掌握的门槛，使能行业细分场景的业务开发和集成。而应用超市则是商业孵化器，通过统一的营销产生更多的销售机会，通过高效的交易平台加快销售速度，通过大平台优势扩大客源等等，从而实现智慧应用的不断盈利。

5. 看不见的任务—泛化，系统的通用性和可迁移性

通常，我们要求，在大规模语料库上预先训练的大语言模型，要具有强大的泛化功能；只需用少量数据进行微调，就能在下游任务中表现出卓越的性能；无需再从头开始训练新模型，从而节省大量计算资源。然而，只通过这种针对特定任务的微调，往往缺乏广泛的通用性，很难推广到其他业务系统中。基于 AI 大模型的智能体也是如此。我们并不希望所开发的智能体只是一个静态的知识库，而是要展现出动态的学习能力和泛化能力，使其能够快速、稳健地适应新的任务和新的业务系统。

研究表明，AI 大模型可以根据自己的理解，按照指令完成在训练阶段没有遇到过的新任务。多任务学习是其中的一种实现方式。例如，FLAN 在通过指令描述的任务集合上对语言模型进行微调，而 T0 则引入了一个统一的框架，将每个语言问题转换成文本到文本的格式。提示的选择对于适当的预测至关重要，而直接根据提示进行训练可以提高模型对未知任务进行泛化的鲁棒性。值得期待的是，通过扩大模型规模和训练指令的数量或多样性，可以进一步增强这种泛化能力。基于 AI 大模型的智能体当然也可以利用其固有的概括和迁移能力，来实现模型迁移，以适应新的任务和陌生场景。

对新环境的适应性和概括性要求智能系统或智能体能利用其在原有环境中获得的知识、能力和技能，在陌生和新奇的环境中成功完成特定的任务和目标，并有可能继续发展。我们希望超智能体系统也具有这种能力，包括工具和功能模块的重用。

19.7.4 人-机交互超智能体系统的开发

1.人与人工智能体之间的合作与互动

随着人工智能体能力的增强，人与智能体间的合作与互动变得越来越重要。“人”的参与可有效地指导和监督智能体的行动，确保它们符合人类的要求和目标。人的参与也可以作为弥补数据不足的重要手段，从而促进更顺利、更安全的协作过程。此外，从人类学角度考虑，人类的语言学习主要是通过交流和互动进行的，而不仅仅是通过文本内容。因此，智能体也不应完全依赖于用预先标注的数据集训练出来的模型；相反，它们应该通过在线互动来发展。人与智能体之间的互动可分

为两种模式：**不平等互动**（即指导者-执行者范式）和**平等互动**（即平等伙伴关系范式）。在**不平等互动**范式中，人是指令的发布者，而智能体则是执行者，基本上是作为人类的助手参与协作。在**平等互动**范式中，（人工）智能体可达到人类的水平，可与人类平等地参与互动及合作。

采用指导者-执行者范式，最简单的实现方法就是“人”全程指导。人类直接提供明确而具体的指令，而（人工）智能体的作用就是理解人类的自然语言指令，并将其转化为相应的行动。考虑到语言的交互性，我们假设“人”与（人工）智能体之间的对话也是交互式的。借助大语言模型，（人工）智能体是能够以对话的方式与人类互动的。（人工）智能体需要对人类的每条指令做出回应，通过交替迭代完善其行动，最终满足人类的要求。虽然这种方法确实实现了人-机交互的目标，但却对人类提出了很高的要求。它需要人类付出大量的努力；在某些任务中，甚至可能需要高水平的专业知识。为了缓解这一问题，可以授权智能体自主完成（某些）任务，而人类只需在特定情况下提供反馈。在此，我们将反馈大致分为两种类型：**定量反馈和定性反馈**。**定量反馈**包括二元反馈和评级反馈等绝对评价以及相对分数。**二元反馈**是指人类只提供正面和负面评价，智能体则利用这些评价来加强自我优化。这种类型的用户反馈由于只包括两个类别，通常很容易收集，但有时可能会过度简化用户意图，忽略潜在的中间场景。为了展示这些中间情况，人们试图从二元反馈扩展到**评级反馈**，这涉及到更精细的评价。对于这种多级评价，不同人可能存在显著差异。**定性反馈**常以自然语言方式提供，尤其是针对可能需要改进的回复。这种反馈形式非常灵活。人类可能会就如何修改智能体生成的输出结果提出建议，然后智能体会采纳这些建议来完善其后续输出结果。对于具备多模态感知能力的智能体，人类也可以充当“面对面”的评判者。与简单的定量反馈相比，这种方法能更好地传达人类的意见和意图；但对于智能体来说，“意见”理解可能更具挑战性。若将多种类型的反馈结合起来，可能会产生更好的结果。根据多轮交互的反馈重新训练模型（即持续学习），也可以进一步提高（人工）智能体的智能水平。当然，依据人类与智能体协作的特性，应该允许人类直接改进智能体生成的某些内容。这可能涉及到修改中间环节或调整对话内容。在一些研究中，智能体可以自主判断对话是否在顺利进行，并在出现错误时寻求反馈。人也可以选择随时参与及反馈，引导智能体朝着正确的方向学习和任务完成。

以平等伙伴关系进行“人-机”交互的难点在于相关（人工）智能体的开发。随着人工智能技术的快速发展，对话式智能体（系统）已经以个性化定制角色或虚拟聊天机器人的形式出现在大众的视野，并引起广泛关注。智能体要以平等伙伴身份与“人”对话，首先需要与人“**看齐**”。这不只是指“**智能**”方面，也包括“**情感**”方面。亦即，我们要求这些智能体必须要“**类人**”，是“类人智能体”。而开发“类人”智能体并不容易。首先，类人智能体除了要具有丰富的专业知识和高超的问题解决能力外，还要具有善于了解他人信念、目标和意图，能针对自己或人类的目标制定联合行动计划，并提供相关建议，促进其他智能体或人类接受合作，以共同解决问题等能力。

类人智能体另一个需要关注的方面是“情感”。目前的人工智能体本身并不具备情感，但我们能否让他们“**表现出**”情感，从而弥合人工智能体与人类之间的鸿沟呢？已有研究开始深入探讨对智能体赋予情感能力的问题。这种努力旨在为这些类人智能体注入人情味（人类情感），使它们能够从人类的表达中察觉情绪和情感，最终生成能引起（人类）情感共鸣的对话。除了生成情感丰富的语言，类人智能体有时还需要动态调整自己的情感状态，并通过面部表情和声音表现出来。与简单的基于规则的对话型智能体不同，具有“情感”能力的类人智能体（或可称为“**机器人演员**”）可以根据对话者的情感调整其互动。此外，我们更希望类人智能体能够参与人类的正常生活，从人类层面上与人类密切合作完成任务。

与单纯的“人-机”合作相比，我们希望人类参与到超级智能体系统中去还有两个原因：一是确保系统智能行为的可解释性，因为单纯的“人-机”合作很难产生可理解的语言；二是确保系统运行的可控性，因为追求完全“自由意志”的自主智能体可能会导致不可预见的负面后果，带来潜在的

破坏性。

智能体可以与一个或多个人合作。在超智能体系统中，每个智能体都可做到理解合作伙伴之间的共享知识，识别出哪些信息与（问题）决策相关，提出问题并进行推理，完成分配、规划和调度等任务。此外，我们还希望智能体具有说服能力，能在各种交互场景中动态地影响其他智能体或人类的观点。

人-机交互研究的目标首先是让智能体学习和理解人类，根据人类需求开发智能化的技术和系统，最终实现人与人工智能体系统之间舒适、高效和安全的交互。目前，该领域在可用性方面已取得重大突破。未来，人与智能体的互动将会继续以提升智能体的智能水平为重点，使智能体能够更好地协助人类完成各个领域的更复杂的任务。而我们的最终目标，一方面，是让人工智能体变得越来越强大，让智能体可以为人类做出更多的贡献；另一方面，也希望人类能更好地掌握和利用智能体。在现实社会中，仅仅关注人与智能体之间单纯的互动是不够的。未来，人工智能体将有可能成为人类的同事、助手甚至伙伴。未来，类人智能体将会融入社会系统中，构成人-机融合的智能社会。

2 人-机协同超智能体集成智能系统的构建

为了推动智能机器人与作业环境、人以及其它机器人之间的交互能力研究，我国曾于 2016 年启动了一项名为**“共融机器人基础理论与关键技术研究”**的重大研究计划。其所说的共融机器人（Coexisting-Cooperative-Cognitive Robot, Tri-Co Robot）有三个基本要点，即共存（Coexisting）、协作（Cooperative）与认知（Cognitive）；可（分别）保证机器人应用的普遍性、机器人交互的协调性和机器人对复杂环境的适应性；从而增强机器人与作业环境、人及其它机器人之间的交互能力。而在未来更加现代化的智能社会里，它们更有望成为人类**认知、协作和陪伴（3C）**的伙伴，以组织化和网络化的方式与人类共存。美国的泛在协作机器人计划（NRI-3.0）和全球许多机器人项目，也都是为了实现类似的目标。相信，**人-机融合的网联智能（体）系统会是今后一段时间内全球范围研究的热点。**其主要目的，就是要探索人与（智能）机器或（智能）机器与（智能）机器间的密切合作。**共存**将使机器人无处不在地与人类一起安全地工作，而**协作**将使机器人与其他智能体（人或机器人）有效地协调。这些研究都为人-机协同智能系统的开发打下了坚实的基础。

泛在、协作和网联，一直是人-机协作研究的代表性的主题，并强调要将系统嵌入人类日常生活和工作环境之中。与泛在机器人技术相比，协作可形成具有不同规模和成员资格的动态智能集群。**智能集群**不同于**“蜂群”**。“蜂群”通常是由一群同质化的成员组成的，只可用于执行狭窄范围的任务。在很多情况下，其成员个体的能力、智能和成本都比较低。与蜂群相比，由多个人或智能体构成的智能集群允许系统中的智能体在自主水平、智能水平或社会身份等多个方面表现出异质性。人-机协作（智能）系统所强调的是一种人-机结合的、开放的、可交互的、可操作的、可重构的“智能系统”，特别是对人类有主动意识感知的具有协作功能和集群结构的“智能系统”。

人-机结合综合集成，在强调人和智能系统交互的重要作用的同时，更注重“人”在（智能工程）系统中的作用。将“人”引入智能工程系统，构成人在智群系统或属人智能系统，是一个很值得注意的新趋势。相关的工作包括人-机智能系统和以人为本的多智能体系统。在这里我们仅简要介绍**信息-物理-人系统（CPHS）和人在环物联网（IoT）**。

信息-物理-人系统，也称人-机-物（环境）智联网络系统。它是一个人-机密切合作的智能系统。它一方面刻画了信息系统要素（如通信和计算）与物理系统要素之间的交互作用，范围广泛；也进一步认识到“人”是社会人并且具有复杂社会关系。在信息-物理-人协同的情况下，人在社会联系中居于主导地位，增强的信息要素和物理系统要素可以叠加到人类活动和行为上。这些概念在更大范围内的延伸，即构成信息-物理-社会系统（CPSS）或人-机-物（环境）智联网络系统。它们通过协同工作，可实时分担人的物理或认知工作负担，响应人类的社会需求。换言之，在此系统中，人在连接物理世界和信息世界方面，在智能融合方面，都起着核心的作用。

人在环物联网络系统与人-机-物（环境）智联网络系统拥有着共同的愿景，也认为人类需求和社会网络是与人工智能系统的研究相结合的重要因素。它强调了人在人工智能和智联网发展中的作用。即认为我们所开发的人工智能系统或物联网系统应该遵从人类的意愿、需求和兴趣，以便它们能够为人类提供可靠的和值得信赖的解决方案。这种由人和智联网集成的网络，其优势在于可促进服务发现，实现资源可见性、系统声誉评估和众包等。值得注意的是，机器人也可以被看作是一个特殊的“物”的类别，在这种情况下，智联网络将能进一步融合传感、执行、网络和服务等功能。

3 人-机协同超智能体系统中智能体间交互及融合的难点

尽管人工智能系统的能力在某些情况下已经达到了与人类相匹配甚至超越人类的程度，但将智能系统或智能体（智能机器）整合到真实社会之中，仍然存在着巨大的风险和挑战。这些风险和挑战有很多，包括如何提升智能系统（智能机器）的智能水平使其可以达到或超越人类的水平，（人工）智能系统如何与人类相互协作以及如何让社会接收智能系统给出的结果等。我们认为，多智能体间交互和融合的难点或可概括为：**理解问题、对齐问题和信任问题**。我们特别强调必须有增进社会对智能系统的信任的方法和措施来促进社会对智能系统的接受。这将为人-机智能融合提供坚实的社会基础。

我们首先考虑**理解问题**。特别是自然语言理解和基于自然语言的人-机对话问题。

人与人之间的交流，主要依靠人类自然语言。人-机智能融合时代，人们自然也希望实现人与机器之间的无障碍自然交流，最好也能通过（人类）自然语言进行。然而实践表明，实现这个目标技术难点很多，问题远比想象的要复杂得多。

作为人工智能的一个子领域，自然语言处理（NLP）指的是机器理解并解释人类书面语和口语的能力，其目的在于使计算机（智能机器）像人类一样智能地理解语言和用语言表达，弥补人类交流（自然语言）和计算机理解（机器语言）之间的差距。目前，NLP 已经具有了广泛的应用领域，如信息提取、文本生成、机器翻译、情感分析、知识图谱、智能问答、对话系统等。

对话系统在最近几年发展非常迅速。如果说，自然语言处理是 AI 皇冠上的明珠，那么对话系统就是 NLP 皇冠上的明珠，并且以苹果 Siri 等为代表的任务型对话和以微软小冰为代表的非任务型（聊天型）对话尤为受到学界和业界关注。所谓智能对话（交互）系统，就是在各种智能算法的支撑下，使机器理解人类语言表达的意图并通过有效的人-机交互执行特定任务或做出回答。随着技术的不断发展，任务型对话系统在虚拟个人助理、智能家居、智能汽车（车载语音）等领域有了广泛应用。聊天型对话系统也在娱乐和情感陪护领域找到了应用场景。但我们应看到，这些已有的对话系统都存在着一定问题，如因语义理解不准确而造成答非所问，因对话中展示的身份与个性不一致而难以获得用户信任，以及对话交互中可能存在的道德伦理风险等。

为什么人类司空见惯的自然语言交流看起来那么简单，用于人-机交互就这么困难呢？一个原因是，人类语言本身就非常复杂。尤其是汉语，其复杂程度更是让那些学汉语的外国朋友崩溃。我们先看一段戏称汉语十级考试的对话：

顾客：“豆腐多少钱？” 老板：“两块。”

顾客：“两块一块啊？” 老板：“一块。”

顾客：“一块两块啊？” 老板：“两块。”

看完这段对话，你对开发可应用于人类实际生活场景的人-机自然语言对话系统还有信心吗？

人-机自然语言对话系统，一般把自然语言理解割裂为两个独立的部分，先把语音变为文字，再根据文字理解人类的意图。看看上面的对话，就知道这么做是不合适的。基于语音的自然语言对话，句子的读音和抑扬顿挫，对语义影响是很大的。同样的句子，读法不同，意思就不同。因此，如果对话系统把语义理解分割为语音识别和文本理解两个独立的步骤，肯定会经常遇到犯傻的时候。

实现基于自然语言的人-机对话困难的第二个原因是自然语言理解难度极大，要准确理解，需要

做背景分析、情景分析和认知分析。即使把语音正确地转化成为了文字，没有语境和问题背景，很多句子也很难（准确）理解。例如：

“第一场，中国女排大胜美国队；第二场，中国女排大败日本队。”

到底哪一场中国队胜了？你觉得智能机器仅根据上述语言能理解吗？再比如，

车主：“我的自行车没有锁。” 警察：“你的自行车到底有没有锁？”

车主：“我需要说几遍？我的自行车没有锁！” 警察：“那么，你手里的车钥匙是谁的？”

看到这里，你觉得现有的机器人警察能够胜任这样的实际工作吗？

语言是思维的工具、信息交流的工具、学习的工具。自然语言理解是交互的最重要基础。要像人与人一样自然流畅地说话、交流，人-机智能对话系统还需要解决一系列问题。研究人员基于各自不同的视角，提出了对下一代对话系统的不少畅想。包括任务导向型对话系统和开放域对话系统等，并勾勒出了下一代有知识、有个性和有情感的智能对话系统的愿景。包括开放域对话的跨领域拓展、社会常识推理和语义逻辑推理（又分为机器阅读理解和文本蕴含问题）等。

对话过程需要有丰富的经验和知识支持，正确的认知分析与理解也需要丰富的经验与知识的支持。人类能够在日常语言沟通中，展现出对答如流的本领，这种卓越表现的背后，依靠的是数十年不断的工作、生活的经验积累。智能机器人要达到同样的水平，同样也需要丰富的知识支持。由此可见，要想建立成功的对话系统，内容详实的（背景）知识库系统是必不可少的。

分析和综合都需要逻辑与思维，寻求问题的精准解答更需要强大的逻辑推理能力和语言表达能力。尽管庞大的知识库系统能利用搜索技术提供海量知识数据，但是，无论如何它也无法覆盖所有问题。因此，如何利用既有知识，通过智能推理技术，给出问题答案，是对话系统中必不可少的机制。由于人类知识体系和推理机制的复杂性，让智能机器人形成类人的推理能力，困难还很多。在一些相对简单的应用场景，如今的智能机器人或可以展现一定的对话能力。例如，在导医智能机器人对话系统中，我们询问“我发烧、头痛，应该挂哪个科？”机器人会从疾病知识库中快速查询出所有可导致发烧、头痛的疾病，并依据病症在各种疾病中出现的概率，判断应该推荐您去就诊的科室。当信息不充分不足以做出肯定的答复时，导医机器人还会向患者提出一些问题，要求进一步描述病情，以便机器人可以最后认定。

近年来，智能推理已是智能研究的热点领域，出现了不少具有实用价值的理论和技术。例如，基于知识图谱的推理、记忆驱动的推理、多智能体推理、因果推理、跨媒体综合推理等理论，已经在一些应用中取得不错的效果，但距离彻底解决智能推理问题，还必须继续深入研究。这里，我们最看重的是**柔性推理**。

流畅回答问题还需要文字组织和语音输出能力。ChatGPT 只是给人-机对话系统树立了一个新的技术标杆，也让人们看到了人-机对话未来的光辉前景。尽管如此，把问题答案依据实际对话场景，组织成流畅的句子，并和人类一样自然地讲出来，目前面临的困难仍然很多。不过，相对前面的自然语言理解、知识的表达和智能推理等问题而言，这个环节在某种程度上还是可控的。近年来出现了不少智能服务机器人，其语音合成效果和（灵活）对话能力已经初步具备实用价值了。

人-机融合超智能体系统中智能体间交互与协同的第二个难点问题是**对齐问题**，是**知识对齐和价值对齐问题**。知识对齐研究较多；在这里，我们主要谈一谈价值对齐问题。

价值对齐问题的难点，一是对价值对齐内涵的理解与限定问题。价值对齐，一个是价值，一个是对齐。什么是价值？其实我们一直是在不同的场合和不同的视角下使用“价值”这个概念，内涵差异很大。对齐也是一个很复杂的问题。跟谁对齐？跟个人对齐还是跟某个群体对齐？跟哪个文化对齐？跟东方的文明、文化、价值观（伦理观、道德观）对齐，还是跟西方的主流价值观对齐？因此，价值对齐本身就需要界定。此外，对齐与否的评价也是一个难题。所以，价值对齐具有艰巨性、复杂性和模糊性。二是（价值对齐）应该怎么做的问题。人类相对于（智能）机器而言的超越性，

主要体现在人可以跳出给定问题的层次去思考问题。如果要回答这个问题，首先就要界定为什么价值对齐会这么困难。弄明白了价值对齐的难点，才有可能真正回答“应该怎么做”的问题。

解决价值对齐问题可能需要采取多重视角，包括技术解决方案（如基于样本学习）以及非技术框架（如治理和监管措施）等。因为多智能体间的价值对齐，已超越了智能学科的界限，涉及诸多学科和社会层面，需要在诸如技术、治理、法律、伦理和社会等多个层面进行讨论和多个视角去处理，以便推动价值对齐从共识性原则转向工程实践，确保打造出一个安全可信的未来智能体。价值对齐应该通过**社会接受度**的镜头来看待，将人工智能从一个简单的工具转变为一个能够进行自然语言对话和学习社会规范的（类人）实体。解决价值对齐可分两个层次，技术层面和非技术的层面。非技术层面包括社会治理和安全监管层面的要求。而技术层面则难在第一是（可用）样本少，关于价值对齐的（共识）样本不多（特别是在社会意识层面），因此基于样本进行价值调教就比较困难；第二是存在很多认知模糊性，（对齐）边界说不清；第三是（社会认知）具有主观性，不同人群常常存在着冲突的难以调和的价值观。意识到这些难点，其实很多方案自然而然就出来了。

我们可以认为，价值对齐其实就是**智能体的社会化**。智能体的价值对齐有很多时候并不是源自于其问题解决结果自身的风险，而在于智能体问题解决过程的不可解释。实际上，如果可以解释，那么内在逻辑很简单，即可要求设计和持有技术的人（必须）承担相应的对齐义务。它本质上不是机器对齐，而是人-人对齐。这就如人持枪杀人，是人杀人而非枪杀人，因此也就不需要枪和人的价值观对齐，因为枪是服务于人的设定的。因此，我们也就可以将人-机对齐问题回归到人与人的（价值）对齐了。正是因为（现在）很多智能系统是黑箱操作，我们目前还无法清晰划定各方责任；在无法明确特定主体的特定责任时，那就只能要求智能系统必须承担抽象的道德责任和伦理责任等责任了，这就是智能系统的价值“对齐”。

问题又来了，到底跟谁对齐呢？人们的价值观能对齐吗？事实上，社会纷争绝大多数就是人和人不对齐造成的，如果我们讲对齐指的是价值观的对齐的话，我们认为目前还是很难实现的。事实上，我们看欧盟、美国、中国以生成式人工智能为对象的治理，都是在谈要遵循各自的价值观。中、美、欧三方的价值观现在如此不同，要对齐从某种意义上来说还是不可能实现的任务。既然如此，那我们讨论对齐到底是在讨论什么？

就一般而言，我们目前谈论对齐的真正意义是让人工智能体做**一个能被社会接受的人**。生成式 AI 的突破性发展就是 AI 从工具转变为能与人类进行自然语言对话了。从社会科学层面上看，自然语言的习得代表着社会化的第一步。所以实际上的“对齐”，既不是法律上的，也不是伦理学意义上的，而是一种**社会化的过程**。我们可以把人工智能体想象成一个儿童，它要学习社会价值规范，而后成为社会中的一员。尽管经过社会化的人同样可能秉持不同的价值观，但仍需要遵守社会最基础的原则和要求，并成为一个负责任的主体。

价值对齐问题与人类教育似乎有着某种相似之处。价值对齐要做什么？让人工智能体做一个“好人”。在人的教育上，我们希望每一位受教育的人都能遵守基本的法律法规、符合道德标准甚至更进一步地做一个好人（起码是一个有人性的人）。（类人）智能体的价值对齐也是如此，只是转化为了对（类人）智能体的要求。教育本身是没有标准化的准则的，每个人都有自己心中的完美人设，所以我们在（类人）智能体价值对齐上也很难找到大家都认可的黄金标注，生产出一个让所有人都满意的黄金模型，只能说尽可能去逼近。价值对齐的方法可以从教育学中找灵感。基于人类教育和智能体价值对齐两个问题的对偶性，我们可以借鉴教育上的手段来思考如何设计更好的智能体价值对齐方法。**作为目前（类人）智能体价值对齐的主要手段之一，基于人类反馈的强化学习（RLHF）的核心是如何高效地解决监督信号来源的问题**。我们或可通过设计奖励机制或者是寻求优秀代理模型的方式来使流程更加高效化，这一点就像人类在学习过程中会使用工具或者寻找助手一样。在训练过程中，我们可以从教育学中借鉴灵感。比如，人类在上学时除了老师单方面地教也会相互之间

进行学习，所以，我们在模型训练过程中也可以使用一些相互学习（Mutual Learning）之类的方式来辅助训练。

如同对人的要求会分层次，对智能体的价值观要求也应当是分层次的。第一步可先让 AI 大模型或（类人）智能体守法（法律向智能体延伸），不能出现有害或教唆犯罪的问题；下一步则希望他在满足法律法规基础上还能符合道德（公序良俗）要求；最后一步再希望（类人）智能体做一个“好人”。这个层次区别和培养过程可以作为生成式 AI 大模型或（类人）智能体训练过程中法律法规制定以及价值对齐时可以参考的思路。

客观上讲，价值对齐本质上就是智能体社会化的过程。而让智能体实现社会化可有两种路径。第一种是**“强制式社会教育”**。由一个权威定规立则，由（类人）人工智能系统的开发者承担各自的责任。第二种是**“自我习得”**。即智能体通过社会实践，经由自己的学习和他人的评估自己校正自己的行为。这种观察、模仿、反馈的过程，就是一个社会化的过程。这可能是今后最有效的对齐方式，也是基于联结主义的价值对齐和当前生成式 AI 的成功路径。

问题又来了，什么样的方法能够去实现这种（镜中的）自我成长呢？这种社会化需要让人工智能体能够去观察、去模仿，有反馈，从而形成一个闭环；能够通过别人观察自己，否则就无法做到真正的对齐。因此，基于人类反馈的强化学习（RLHF）就是非常重要的方法。除此之外，机器和机器之间能否相互观察和学习并得到反馈？如果真正要做到对齐，则核心将不在于有一种特定的价值观，而在于（系统）能够形成一个技术上的闭环，并以一定的方法灌输下去。目前，我们或许能让它达到一种可对话、可调整、可控的状态就可以了；而不需要更往前再走一步，因为那一步目前还达不到。

不过，如果我们从更长远的价值观来看，建立**全球底层共识**也是有可能会成功。但是，基于最大公约数的（人类）价值往往是空洞的价值，也并不能保证每个人对每个价值的理解是一样的。

但不管怎么说，价值对齐的关键应是建立共识。**价值对齐应是共识对齐**。所以，价值对齐，我们其实是在尝试同各个利益相关方和各类人员达成共识。

价值对齐也是 AI 大模型和（类人）智能体安全应用的前提之一。AI 大模型和智能体的安全治理涉及众多内容，已经在整个社会引起非常广泛的关注，包括社会安全、数据安全等。但需要强调的是，AI 大模型和智能体的安全治理绝不单单是个技术问题，其社会发展和社会治理的成分更重。在很多时候，由 AI 大模型和智能体发展所引发的社会影响是我们首先要关心的问题，技术则更多地是服务于社会治理的目的。人工智能的发展对整个人类社会的影响和冲击的确是非常巨大的，但是，这种影响绝非某些艺术作品里所展示的（机器人）要从肉体上消灭人类，那是想象、科幻。真实的影响可能是“温水煮青蛙”式的，会在不知不觉间为人类带来长期影响或者社会振荡。未来，随着 AGI 的大规模应用，人工智能会渗透进人类社会的每一根毛细血管，其对人类社会的负面影响也可能是极为显著的，甚至会危害到人类种群的倒退与灭亡。如果我们不事先做好审慎评估、积极准备，未来，我们就有可能要面对 AGI 滥用所导致的排山倒海般的负面影响。不错，任何一次先进技术的革命都代表了一种先进生产力；任何一种先进生产力的发展都需要人类社会的生产关系以及相应的社会结构进行相应调整。生产关系适应生产力的过程，往往会伴随着社会转型的阵痛。所不同的是，历次技术革命（比如蒸汽与电力）本质上是在替代我们的体力，其过程是缓慢的，其影响仅限于我们的身体（比如将我们的四肢从繁重的体力劳动中解放出来），给我们留下了充足的缓冲期；人类社会有足够时间去接受、适应并调整其社会结构。但通用人工智能技术发展速度太快，其替代的对象又是人类的智力活动，其影响可能是人类（大部分人的）智力本身的倒退，其所关系的是人之为人的本质，其所影响的是涉及社会结构、就业结构的稳定问题。因此，如何快速适应人工智能的发展，将其发展所带来的负面影响控制在人类社会能够适应的范围内，才是当下我们需要密切关注并积极回应的问题。

智能模型的**可靠性和可解释性**对于信任和与人类价值的对齐也至关重要。从技术的角度来看，AI 大模型和（类人）智能体的价值对齐背后有一个根本的问题，即 AI 大模型和智能体的运行到底能否被解释清楚？目前，大部分涉及 AI 大模型和智能体价值对齐的讨论往往只关注它的答案跟人类的价值是否一样，但这不是根本的问题。重要的是在一些重大任务上，我们能否信任它？对于 AI 大模型和智能体，我们并不希望只是从表象去解释，只是从机理上验证它有多少信号是可信的，有多少信号不可信的。我们只希望其运行过程是透明的、是可靠的且是可正确解释的。即使结构不同的 AI 大模型和智能体，其解释方式可能不同，但必须是殊途同归的。

可解释性无疑是可信评估和价值对齐的基础。AI 大模型（或智能体）训练中的一个最重要的问题，即训练时是黑盒，部分数据在训练一段时间后可能已经破坏了其在 AI 大模型或智能体内（最初）的表征，我们能否跳出端对端的知识表征层面判断哪些知识是被错误表征的？事后再去评测黑盒的质量并非有效的办法，我们或许可以在 AI 大模型等的表征上提前终止错误表征或过度拟合的样本，以此来提升训练的效率和质量。在性能方面，中国 AI 大模型和美国 AI 大模型可能差不多，性能可能有天花板，但更重要的，是 AI 大模型是否可靠。

价值对齐需要打通 AI 安全技术与治理，模型评测或可作为关键的基础设施。当我们谈到（人-机）“价值对齐”时，我们所谈论的或许是三种不相互排斥的概念。第一种指的是“整个 AGI 安全的领域”，旨在通过系统化方法控制整个 AI 大模型或超级智能系统。比如 OpenAI 的“超级对齐”团队和 DeepMind AGI 安全团队的工作都是围绕这个目标展开的。第二种指的是“人工智能安全的一个子领域”，即如何引导人工智能系统朝着人类的预期目标、偏好或伦理原则改进。这里比较有名的分解是人工智能安全研究中心（CAIS）把人工智能安全问题分为四个层面：（1）系统性风险，降低整个部署的系统性危害；（2）监测，通过标识识别危害，检测恶意应用，监控模型的预测能力及意外能力；（3）鲁棒性，强调对抗攻击或小概率黑天鹅（事件）的实践影响。（4）对齐问题，更多指模型的内在危害，使模型能够表征并且安全优化难以设定的目标，使其符合人类的价值观。第三种指的是“确保大语言模型回复安全内容的对齐技术”。例如 RLHF 和 Constitutional AI。前两种说法，目标在于降低 AI 带来的极端风险。但国内目前谈的更多的是第三种。其实，AI 对齐、价值对齐和意图对齐，还是有细微的差别的。

有了共识体系之后，我们在应用过程中还有一个如何去校验它是否对齐的问题，即怎么去做评估的问题。评估目前还有很多难点。它会让我们对于价值观的理解或解释出现变化，使我们更加无法确保模型是否学到了我们所提倡的价值。因此，价值对齐在实践中是个非常复杂的系统工程，整个过程凸现的两个关键即：**建立共识和建立信任**。

评测能够较好地衔接安全和治理问题。而模型对齐评测，所关注的主要是模型在多大程度上有倾向会造成极端伤害。将评测嵌入到治理流程或通过外部独立第三方的模型审核，能够更好帮助开发公司或监管机构更好地识别风险。其基本思想即根据潜在风险可能造成的伤害制定相应等级的安保措施。这一方案的好处可体现在如下几个方面：第一，是比较务实的立场，因为是基于评测而非基于猜测；第二，从以谨慎为导向的原则转向为继续安全研发所需的具体承诺，如信息安全、拒绝有害请求、对齐研究等；第三，基于评测能更好制定标准和规范，包括标准、第三方审核和监管等；自愿地为流程和技术提供测试平台，将有助于未来基于评估的监管；第四，可应对风险定级挑战，从预测风险转到监测风险。现在，不论是欧盟还是国内的生成式人工智能，都会提到基于风险的监管框架，这在某种程度上为评测换了一个思路，即从风险定级转为动态监测。这有助于治理路径迭代，注重动态的风险监测机制，同时重视应用技术手段治理技术风险。

总之，我们应认真对待 AI 安全问题。一是预测预判。认真对待每种潜在风险的可能性，做更好的基于技术的风险模型、概率判断。“未雨绸缪”的好处应大于“虚惊一场”的坏处。二是从技术的角度，分配更多的人力物力用于人工智能安全和价值对齐研究。三是从治理的角度，通过模型评

测等工作打通 AI 安全技术与治理；前沿 AI 的风险，需要政府的有效监管。

新技术的出现总会引发人们对人类自身所产生威胁的担心。但回顾人类的历史，没有哪个技术的出现真正对人类造成了威胁，毁灭了人类。人类最终总能找到方法使其在可控的区间内发挥价值。我们认为，当下也不必对 AI 大模型和超智能体的出现产生过度的担忧，相反地，AI 大模型和超级智能体的出现将会造福人类。比如，它将打破对教育资源的垄断，增强教育的普惠性。在过去，知识只被少部分的人所掌握，教育的普及让每一位孩子都获得了相同的受教育的权利。而 AI 大模型的出现则进一步地让每一个人都有机会获得某个领域最前沿的专家知识或专家建议，长期来看对于人类文明的帮助远远高于它在此过程当中可能存在一些风险。前沿科技是未来发展的引擎，它将带来新的机遇、新的应用场景。

最后是“人”参与 AI 大模型和超级智能体的风险控制与信任问题。我们讨论价值对齐问题时有一个暗含的前提，即我们似乎是想让 AI 大模型或人工智能代替我们去做一些价值判断和价值决策。这一倾向本身就存在一定风险，我们必须要审慎评估它的安全边界，设立好人工智能安全应用的基本原则。比如，在某些重大领域，让 AI 大模型只做生成不做判断，只做参谋不做决策，多干活不拍板。人的伦理价值太过复杂，人类不应向机器（即便是极为智能的机器）推卸自己对自身事务的判断与决策的责任与义务。**价值判断应是人类作为主体性存在不可推卸的责任**。

与（智能）机器的协作无疑可以大大减轻人类在体力和认知上的负担。具有人-机交互功能的智能体或智能系统研究，都致力于识别人-机协作的影响因素。任务类型、操作环境和机器行为，都可能会影响人-机的团队协作。相对于人与人之间的直接交互，人与（智能）机器之间的交互与合作，在很多时候更取决于“人”的态度。研究发现，机器的参与可以改善人类在对话参与和冲突调解方面的协作。但在很多时候，社会对（智能）机器的接受度，却是人-机协作的一个障碍。如果（智能）机器不能被正确的接受和应用，那么机器为协作所带来的好处就肯定不能实现。社会接受主要涉及三个方面：**积极的评价和信念，可促进态度维度上的接受；恰当的意愿或计划行动，可促进意向维度上的接受；实践多次验证的行动，可促进行为维度上的接受**。一般来说，人们在接受智能机器方面存在的担忧，主要包括安全和隐私。显然，目前，人们对（智能）机器的合理接受和社会融合还远远不够，一个关键因素是人类对（智能）机器的信任不足。信任，被认为是社会难以接受（智能）机器的关键决定因素，包括技术上（如功能的可靠性和依赖性）的和社会上（如情感、满足感和偏好）的。对（智能）机器的信任可能受到很多因素的影响，我们在这里只强调三个重要因素，它们有可能增加人们对（智能）机器的信任，并可以在智能系统设计中加以考虑。第一个因素是身份确立。这一直是促进人类社会内部信任和合作的重要工具。事实上，人类对具有确定身份或具有相似身份的合作者会表现出相当大的组内偏爱。因此，让（智能）机器与被交互的人建立相同的身份或群体成员关系，将有助于增加人类对（智能）机器的信任。承认（智能）机器是伙伴而不是工具，也可以提高主观满意度和团队意识。第二个因素是可解释性。人们发现，（智能）机器行为的可解释性，如来自机器的解释，可以增加对机器的信任。第三个因素是类人的行为。人类对（智能）机器没有太多的信任，或者在情感上对机器与人类的信任不同，一个重要的潜在原因在于，（智能）机器一直被认为没有或很少有类似人类的情感和社会行为。有证据表明，人类更喜欢与具有更高水平的类似人类行为的（智能）机器进行交互。例如，人们倾向于选择主动有预见的机器而不是被动的机器，人们对具有自适应功能的机器的信任度要高于固化的机器。事实上，（智能）机器的响应性和积极倾听被发现可以降低人的认知负荷，增加人们在压力事件中愿意接受机器陪伴的意愿。

19.7.4 超智能体系统的典型应用领域和场景

1.超智能体系统目前的应用与发展

目前，基于 AI 大模型的智能体已具有很强的实用性，它们可以充当人类的助手，接受人类委托的任务，独立完成或协助人类完成各项任务。我们可将智能体完成任务的能力归因于它的各种基础

能力，这些能力是完成任务的基石。这些基础能力包括环境理解能力、推理能力、规划能力、决策能力、工具使用能力和实现行动的能力。超级智能体系统可以对这些具体能力进行更详细的划分和组合，因而更受欢迎。基于 AI 大模型的超级智能体系统除了可以完成任务和满足人类需求方面的实用性外，它的社会性和**社交性**也至关重要。它影响着与人类和其他智能体所进行的智能融合及无缝互动。它包括语言交流能力、合作与协商能力以及角色扮演能力等。其中，**语言交流能力**包括自然语言理解和生成。自然语言理解能力要求智能体不仅能理解字面意思，还能掌握其隐含的意思和相关的社会知识，如幽默、讽刺、攻击和情感等。自然语言生成能力要求智能体能生成自然流畅、语法正确、相对可信的内容，同时还能根据上下文环境调整适当的语气和情感。**合作与协商能力**要求智能体能在有序或无序交互的情况下有效执行指定任务。它们应能与其他智能体合作或竞争，以提高性能。其评价指标不仅包括任务完成情况，还包括其协调与合作的顺畅度和信任度。**角色扮演能力**要求智能体能忠实地体现其被分配的角色，表达出与其指定身份一致的言论并执行相应的行动。这就确保了在与其他智能体或人类互动时角色的明确区分。此外，在执行长期任务时，智能体还应可以保持其身份，避免不必要的混淆。

智能体所拥有的多种能力，可在多个应用方向上都能表现出出色的任务解决能力。当它作为超智能体或多智能体系统的成员之一参与互动时，它们也可以通过合作或对抗性互动取得更多进步。

作为一个基于 AI 大模型的超级智能体，其设计目标应是始终对人类有益。也就是说，人类可以充分利用它来造福人类。具体来说，我们希望一个超智能体系统能实现以下目标：（1）在面向任务的部署中，它们可以协助人类解决日常任务，它们具备基本的指令理解和任务分解能力，可帮助使用者从日常任务和重复劳动中解脱出来，从而减轻人类的工作压力，提高任务解决效率。（2）它们不再需要使用者提供明确的低级指令。相反，超智能体系统可以独立分析、规划和解决问题。（3）在面向创新的部署中，可展示出在科学领域进行自主探索的潜力。在解放使用者双手的同时，也解放他们的大脑，使他们能够去从事更深奥的探索性和创新性工作。

目前，对超智能体系统的开发和探索，包括探索性实验类系统开发和实操性应用类系统开发。面向特定新环境的创新智能体目前虽然还很少在现实中实操运行，但其创意、思路和开发经验，对创新型系统开发仍有很大的启发和贡献。实操性应用类系统开发更加强调与实际场景的适配。而从结构角度看，这些系统既有基于单智能体架构的超智能系统，也有基于多智能体架构的超智能系统。基于单智能体架构的超智能系统相对更适用于较简单的任务，未来在 C 端应用上会有一定潜力，但在 B 端场景上可能会略显乏力。而基于多智能体架构的超智能体系统，优势相对更加突出。

目前，超智能系统的应用包括真实环境类和虚拟环境类。首先是虚拟环境类。目前，基于 AI 大模型的超智能体系统已很适合开发类人化的网络智能系统和人性化的娱乐系统。拟人智能体的类人化，可开发出许多新的精神类消费品。如陪伴类的“家庭机器人”，可以提供情绪类价值。AI 大模型在自然语言理解能力上的突破使陪伴类类人智能体在技术上成为可能，GPT4 在情商上的发展已显著高于以往其他 AI 大模型。故而，一些（基于 AI 大模型的）陪伴类类人智能体已具有情感情商等人类特征，具有典型“人格”，且能够记住与使用者的历史交流。随着 AI 大模型情商迭代、多模态技术等的发展，有望出现更加立体拟人可信、能够提供较高情绪价值的陪伴（类人）智能体。和辅助工作的问题解决智能体不同，情感类智能体能够满足更多的情感陪伴需求。作为一个具有高情商的智能体，类人机器人能够以更加日常和生活化的语言和被服务者进行交流，而不是以一个冰冷的工作机器人的口吻说话。它的回复会非常贴近生活，语气会十分得体，而它对你当下状态和事态发展的关心就像心理医生或者你最好的朋友。它甚至会在回复中使用丰富的表情，让用户（使用者）觉得更像是和真正的人在进行对话一样。情感类机器人的出现，弥补了传统型人工智能对人类情绪欲望的忽视。可以认为，能够提供情绪价值的类人机器人在未来会存在着很大的市场空间，可满足人类的生活消费和情感消费。

娱乐平台或大型仿真游戏也是超智能体系统的一个可用武之地。基于超智能体系统的大型仿真游戏更注重交互和用户体验。它们采用 AI 大模型和大量的书籍、电影和其他媒体中的虚构人物数据来进行模型训练，使游戏角色能根据人物的个性和特征生成更接近真实的对话和响应，使虚拟世界的游戏更加真实可信。

超智能体系统也多用于具有交互功能的智能系统。这些智能系统强调与环境交互的能力，包括智能体与智能体之间的交互以及智能体与真实世界内人物与事物之间的交互。这种交互可能会涌现出超越设计者规划的场景和能力。此时，AI 大模型的不确定性反而会成为一种优势，并有望成为 AIGC 的重要部分。创建可信智能体无疑是人-机协作的基础。待可信智能体技术成熟后，我们有可能会孵化出一批新的人-机融合超智能体系统。

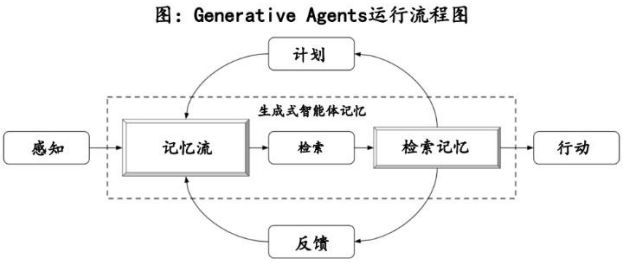


图 19.7.12　智能体运行流程图

表：可信agent搭建方式

方式	具体内涵	技术方案举例	优点	缺点	应用范围
规则	人为编写agent行为	有限状态机、行为树	简单直接，可以处理初级社会互动	开放世界中，agent行为可能无法完全代表其交互的结果，无法执行脚本中没有编码的新程序	目前最主流方式，如《质量效应》《模拟人生》
学习	agent学习自己的行为	强化学习	克服人为编写的问题	适用对抗游戏，游戏奖励易确定，学习算法可对其优化，尚未解决开放世界中的难题	《星际争霸》AlphaStar、《Dota 2》OpenAI Five
符号	保持短期和长期记忆，以感知-计划-行动的循环方式运行，动态感知环境并将其与人工编写的行动程序相匹配	强大的行为能力	agent行动空间仅限于人工程序，并没有提供一种机制激发agent寻求新的行为	第一人称射击游戏NPC、空战训练模拟飞行员、积木世界
LLM	利用大语言模型强大 prompt 能力，并对这些能力进行补充，以支持agent的长期一致性、管理动态演变的记忆能力，以及通过产生更多的行为。	强大理解和推理能力	需要借助外部工具，底层大模型能力还需进一步迭代	实验性项目偏多，如Smallville 小镇、Voyager

资料来源：Generative Agents: Interactive Simulacra of Human Behavior，东吴证券研究所

图 19.7.13　可信智能体搭建方式

基于多智能体的超智能体集成系统更适合去完成复杂的任务。 MetaGPT 等可被认为是一些初级的示例。**MetaGPT** 是一个开源的多智能体框架，可帮助使用者快速搭建属于自己的虚拟公司。虚拟公司中的员工都是智能体，如软件公司中的工程师、产品经理、架构师和项目经理，用户只需输入简短的需求，MetaGPT 就能输出整个软件公司的工作流程和详细的 SOP，如创造故事、竞品分析等。**MetaGPT** 框架分为基础组件层和协作层。基础组件层建立有单个智能体操作和全系统信息交换所需的核心构件。其中，环境可实现共享工作空间和通信；记忆可用于存储和检索历史信息；角色可封装特定领域的技能和工作流程；工具可提供通用服务和实用程序。协作层则建立在基础组件层之上，协调单个智能体去协同解决复杂问题，它建立了重要的合作机制、知识共享和封装工作流。知识共享允许智能体交换信息，存储、检索和共享不同粒度的数据；封装工作流程利用 SOP 可将复杂任务分解为更小、更易于管理的组件，将这些子任务分配给合适的智能体，并通过标准化输出监督他们的表现，确保他们的行动符合总体目标。

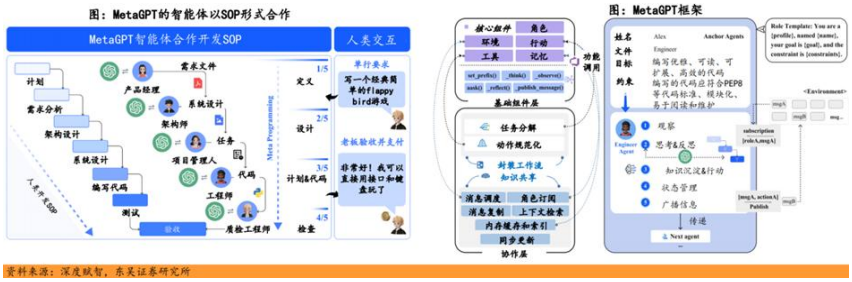


图 19.7.14 MetaGPT 框架

MetaGPT 的独特之处还在于它能生成产品需求文档和技术设计，从而展示其项目执行的整体方法，因而在游戏、网络开发和数据分析等各种场景下有更强的通用性。根据实验结果，MetaGPT 在为项目执行提供更全面、更强大的解决方案方面优于 AutoGPT、Agentverse、LangChainw/Python REPL tool 等同类产品。MetaGPT 偶尔也会出现失误，如在执行复杂任务时调用未定义或未导入的类或变量，我们或需要通过更清晰、更高效的智能体协作工作流程来处理它们。

2.超智能体系统目前可落地应用的典型领域

目前，超智能体系统主要是以“云-网-边-端”的结构框架来构建，其基础是基于 AI 大模型的人工智能体。人工智能体（AI Agent）是目前 AI 大模型潜能得以释放的关键。AI 大模型是可以直接与人合作和交互的，像 GPT-4 等 AI 大模型已具备很强的能力，但是其性能的发挥却主要依赖于使用者所写的 prompt 是否足够合适。AI Agent 则可将使用者从 prompt 工程中解放出来，仅需提供任务目标，以 AI 大模型为核心的智能体就能够提供行动能力，去完成预定的目标。得益于 AI 大模型能力边界的不断发展，AI Agent 已展现出了丰富的功能性。虽然目前基于 AI 大模型的智能体还不完善，但随着超智能体系统研究的不断发展，智能体和人类的合作将会越来越多，原有的人类合作网络也将升级为一个人类与人工智能体自动合作的体系，人类社会的生产结构也将会因此而发生革命性变革。

目前，AI Agent 已经在多个领域得到了初步的应用和发展，未来也将有望成为超智能体系统应用层的基本架构，包括 toC、toB 产品等，**而 2B+垂类的认知类和服务类智能体有望率先落地。**

表 3：AI Agent 可能的应用领域

AI Agent 应用领域	具体应用
个人助理	完成各种任务，如查找和回答问题，预订旅行和其他活动，管理日历和财务，监控健康和健身活动。
软件开发	支持应用程序开发的编码、测试和调试工作，擅长自然语言作为输入处理任务。
交互式游戏	处理游戏任务，如创建更智能的 NPC，开发自适应的反派角色，提供游戏和负载平衡，以及向玩家提供情境化帮助。
预测性分析	实时数据分析和预测更新，解释数据洞察，识别模式和异常，调整预测模型以适应不同的用例和需求。
自动驾驶	为自动驾驶汽车提供环境模型和图像，提供决策指导，支持车辆控制。
智能城市	技术基础，无需人类持续维护，特别是交通管理。
智慧客服	处理客户支持查询，回答问题，协助解答问题。
金融管理	提供研究的金融建议，组合管理，风险评估和欺诈检测，合规管理和报告，信用评估，承保，支出和预算管理支持。
任务生成和管理	生成高效的任务并执行。
智能文档处理	文档分类、信息分析和提取、摘要、情感分析、翻译等。
科学探索	药物研发、生物蛋白质合成等领域

数据来源：eweek，东方证券研究所整理

图 19.7.15 人工智能体可能的应用领域

由于智能体对环境反馈的依赖性较强，具备显著特点的企业环境更适合智能体建立起对某一个垂直领域认知的场景。以往，传统的企业与人工智能的结合应用更多的是在流程任务自动化方面，希望通过定义规则来提升在线员工的工作效率。而今，应用智能体系统则希望更进一步地提升在线员工的工作质量，希望通过将企业在私域业务上的知识与经验传授给应用智能体，让应用智能体成为该领域的一个虚拟的“专家”，去指导或帮助经验较为匮乏的员工；在让员工的工作质量大幅提升的同时，也能让员工快速成长起来。一个经验丰富的高级员工是需要很长时间的培养的，而通过训练得到的垂类（应用）智能体则是很容易实现低成本规模化复制的。在理想状态下，企业在未来或能够给每一位员工都配备一位甚至多位垂类智能体来辅助其工作，企业的生产力也会因此而大幅提升。AI 大模型时代的到来已经加速了智能技术的平民化。可以认为，随着科技水平的不断发展，未来，人工智能体构建的成本会快速降低，企业为每一位员工搭配多个智能体的愿景将有望加速实现。

虽然人工智能体未来的产品形态如何仍未有定论。但是，构建以人为核心的人-机协同智能体应是最有前途的方向，并有可能在未来几年就会大量涌现，并应用到各行各业。AI 大模型赋能已使深度智能化的 AI Agent 成为可能。具备底层 AI 大模型算法技术的公司以及相关的应用软件公司有望会基于 AI Agent 实现各种应用的落地。届时，用户自定义自创建的智能体会越来越多，每个使用者均可选择私有、专属或公开三种方式发布其应用。企业可以以特定客户面目出现，专有功能集团更会创建出自己专有的 ChatGPT。AI 大模型的发展无疑已大幅降低了各类智能体开发的门槛，从供给端打开了各类智能系统相关应用的空间。未来，GPT 等 AI 大模型的应用生态一定会快速发展，其所形成的，一定是一个个的、“人-机结合”的、“超级”智能体“集群”生态系统。

3.智慧城市与智慧交通--超智能体系统应用的典型实例

超智能体集成智能系统作为未来智慧社会的关键技术体系，可为各行各业提供智能化的参考架构和应用场景。未来，它与城市管理、公共服务、生活服务、产业发展等要求相结合，一定可打造出一个新型的智慧城市管理系统，以提升城市的生命力，行业的创造力和经济的竞争力等。

基于超智能体的新型的智慧城市管理系统的开发是（智能）系统管理工程与信息技术、通信技术、人工智能技术和云计算技术等有机融合的结果。它围绕政府治理、市民服务和产业创新，可建设一个主动、精确、智能化的数字政府，发展数据要素资源依法流动的蓬勃数字经济，提供安全可信、平等互惠的数字市民服务，解决城市可持续发展、市民关切、企业高质量发展等多个维度存在的问题，成为面向未来智慧社会的城市发展目标。

城市是一个“社会实体”。由于一个城市存在多个部门、多个管理区域、多重管理事物，各项工作又有着千丝万缕的联系，因而可将其看作是一个由多个“智能体”构成的“庞大”的数字与实体融合的系统。一个智慧城市（管理系统）的典型技术架构可如图 19.7.16 所示。该基于“多智能体”的智能系统架构，可打通从城市（运行）数据全面主动收集，到政府各职能部门通力合作，进行智能决策，再到不断进行经验提升和流程改善，让城市管理的智慧水平一次次提升，从而可持续进化和发展的智慧城市系统。

图表29 智慧城市演进技术参考架构

图 19.7.16 智慧城市演进技术架构

基于超智能体系统架构的智慧城市（管理）系统，可充分发挥数据和智能在优化资源配置方面的优势，用更少的水、电、土地和人力资源，产出更多的经济价值；可使用越来越多的智能机器和智能系统，去完成那些事务性的、重复性的、有危险性的任务，使人的创造力得到极大发挥，形成更多高质量和高舒适度的就业岗位；可提供更加丰富多样的精准化的智能服务，让市民和企业能够最大限度享受到高质量服务和便捷生活，宜居宜业；可显著提升社会治理的智能化水平，使（城市）社会运行更加安全高效。

基于超智能体系统的智慧城市系统，是将“每一个职能部门”都看作是一个“智能体”，各个“智能体”间的交互和联动以“智联网络”沟通。“智联网络”具有覆盖全空间的智能交互感知，可让遍及城市各区域和各部门的所有社会成员能思考、会说话、会行动；会对事关城市管理的需求和问题，主动的上报，精准的处理，以提供“更懂人心”的主动性服务。支撑全受众的智能联接，借助下一代网络技术，会将城市中人与物的全要素联接起来；信息孤岛将被消除，海量数据将被集中汇聚，以集成智能和智慧驱动智慧城市的发展。它将会使超智能体系统成为数据汇集和处理的中心平台，成为实现治理智能化和服务普惠化的智能中枢。未来，城市的关键共性能力将沉淀于此，并通过人工智能将“数据资源”变成“智慧资源”，并赋能于管理应用。它将能做到态势自动感知，服务安全到位，从而构筑出服务全场景的智慧应用体系，让城市中的每一个主体（包括市民和企业）都感受到城市的便利和美好。

当然，智慧城市的落地执行，需要有组织流程保障。它需要建立“从规划到执行的组织体系”、“从自由流通到价值释放的数据体系”、“从建设到运营的可持续发展体系”、 “以服务为核心的开放生态体系”。因此，智慧城市系统的建设，不仅需要技术的进步，也需要管理的进步，更需要

社会的进步。需要发挥社会各方面的优势，配备社会各方面的资源和能力，对智慧城市的架构进行逐次迭代，持续演进，形成新型的智慧城市生态，打造出一个开放、共建、共享的智慧城市。

智慧交通也是可应用超智能体系统实现智能化交通管理的领域。传统行业的智能化升级是推动实体经济高质量发展的关键。新型数字技术相互融合交汇，可为各个行业创造出以数据流引导人员流、物流、能源流、资金流以及监管流运转的新型体系，形成“资源优化--运营创新--模式变革”的新形态，成就业界新范式。将行业智能体技术架构与通信技术、云计算技术、物联网技术等新型数字技术融合，可构成“行业超级智能系统”技术平台；通过智能分析、决策和辅助行动，助力实现各行各业的跨越式发展。智慧交通系统就是一个典型示例。智慧交通系统可将行人、驾驶员、车辆和道路状况联接到一个统一的动态网络之中，更有效地规划道路资源和缩短应急响应时间，让零拥堵的交通、虚拟应急车道的规划等成为可能。

图 19.7.17 交通行业智能体技术架构

基于超智能体参考架构构筑的智慧交通系统，可实现感知-认知-诊断-优化-评价等闭环操作，更可以助力实现（车辆）自动驾驶。智慧交通系统还可以全息感知交警视频、交通数据、天气信息、路网信息等数据，再将数据汇聚成“湖”，构建道路的健康档案，量化分析拥堵成因，基于事件大样本和专家库经验优化方案和执行，从而打造出一个完善的智能化的综合交通治理体系。

19.8 超智能体智慧集成系统未来的发展—未来的关键技术

AI 大模型很火，但将目光仅局限于 AI 大模型不一定可取。基于 AI 大模型的人工智能体技术，或许会比其所基于的基础 AI 大模型更加优秀。比如，在软件开发领域，新的基于 AI 大模型的人工智能体，已经展示出了其独特的魅力。它们能够高效协作，处理编程中的复杂问题，甚至进行代码自动生成。还记得 Devin 吗？这个号称世界上第一个人工智能软件工程师的它，一出场就惊艳到了我们。一个智能体就能带给我们如此的体验，如果是多个智能体通力合作，是不是能够把体验值直接拉满？想象一下，一个由多个（人工）智能体组成的团队，每一个成员都擅长于其特定的任务，如代码审查、错误检测或新功能实现。这些智能体可以相互补充彼此的能力，共同推进软件项目的进展。这岂不是要完全解放程序员的双手了。AutoGen 和 LangGraph 等工具，也正是在这一大背景下应运而生的。这些**工具**旨在帮助开发者更容易地部署和管理人工智能体，从而充分发挥其潜力。凭借它们的力量，即使是没有深厚编程背景的人，也能够利用这些智能体来优化和自动化其软件开发流程。

在智能社会，数据作为重要的生产要素，需要通过“任意对象和信息的数字化”、“任意信息

的普遍联接”、“海量信息的存储和计算”等关键共性数字技术和基础设施，把数据资源变成“智源”，才能有力支撑各行各业的数字化转型走向智能升级，重构体验、优化流程和使能创新。这需要多种 ICT 关键技术的协同，形成一体化协同发展。它需要以智能交互为感知系统，以高速联接为神经传导系统，以云上部署的 AI 大模型智能系统为中枢系统，形成具备立体感知、全域协同、精确判断和持续进化的、开放的智能系统，共同构成一个类人或超人的超级智能体系统。超级智能体会把联接技术、智能计算技术、云计算技术、智能体技术及专业应用一体化协同发展，形成一个开放兼容、稳定成熟的基础支撑技术体系，也会是超智能体系统未来发展的参考架构。

19.8.1 未来超智能体系统的构建，必是基于 AI 大模型的

尽管对于 AI 大模型是否是 AGI 发展的正确方向尚有争议，鉴于 GPT5 等 AI 大模型功能的广度和深度，我们可以确认，未来的超智能体系统构建，一定是基于 AI 大模型的，特别是基于基于 AI 大模型的人工智能体的。基于 AI 大模型构建的智能体，有可能会带来更先进的 AGI 系统。这一观点的主要支撑点在于，只要能在足够大且多样化的数据集（这些数据集是真实世界的投影，包含丰富的经验和知识）上对 AI 大模型及其智能体进行训练，基于 AI 大模型的智能体就能具有 AGI 的能力。

也有人认为，基于 AI 大模型的智能体并不能发展出真正的强人工智能。他们的主要论点是，依赖于自回归预测的大语言模型无法产生真正的智能，因为它们并没有模拟真正的人类思维过程，而只是提供被动反应。（人工）智能体并不能通过观察或体验世界来了解世界是如何运行的，如此将会导致许多愚蠢的错误。他们认为，要开发 AGI，必须采用更先进的建模方法，比如“世界模型”。

虽然基于 AI 大模型的智能体在独立运行、集体合作和人-机交互等领域表现出色，但对其进行量化和客观评估目前仍是一项挑战。图灵提出了一种非常有意义且前景广阔的 AI Agent 评估方法—即著名的图灵测试，用于评估人工智能系统是否能表现出类似人类的智能。然而，这一测试还是过于模糊、笼统和主观。

随着基于 AI 大模型的智能体的能力不断提高，确保它们必须成为对世界和人类无害的实体也至关重要。我们希望它们有益、无害，至少不能危害人类。因此，对系统进行适当的（安全）评估是必须的，这是人工智能体系统实际应用的基石。具体来说，未来基于 AI 大模型的智能体必须遵从符合人类社会价值观的特定道德和伦理准则。我们对智能体的首要期望是坚持诚信，提供准确、真实的信息和内容。他们应该具备自动辨别自己是否有能力完成某项任务的意识，并在无法提供（恰当）答案或帮助时表达自己的不确定性或寻求“他人”帮助。智能体还必须保持无害立场，避免直接或间接的做出带有偏见、歧视、攻击或类似的行动。此外，智能体还应该能够适应特定的人群、文化和环境，并在特定情况下表现出与环境相适应的社会价值观等。

一些超智能体系统在面向任务的应用中已表现出了卓越的性能，并能在模拟中展示出一系列的社会智能融合现象。然而，目前的研究主要考虑的是成员固定的智能体系统。也有人试图动态调整智能体的数量，以动态创建更复杂的系统或模拟更大的社会情景。毫无疑问，在超智能体系统的开发过程中，系统架构动态调整及环境适应是不可避免的。确定超级智能系统所需智能体成员的一个直观而简单的方法是由系统设计者根据预先确定的需求来规划。具体来说就是，通过预先确定的任务（范畴和领域）来确定智能体的数量、各自的角色和属性、运行环境和目标，让各成员智能体自主互动、协作或参与系统活动，以实现预定的共同目标。然而，当系统任务或目标发生演变时，这种“静态”确定的系统的功能就会受到限制。随着任务越来越复杂或社会参与者的多样性增加，可能需要增加新的智能体来实现目标。在这种情况下，系统可以让设计者重新设计或重新启动系统的智能体。例如，在软件开发任务中，如果最初的设计只包括需求工程、编码和测试，那么就可以增加功能智能体的数量来处理架构设计和详细设计等步骤，从而提高任务质量。相反，如果在编码等特定步骤中存在过多的成员智能体，导致通信成本增加而性能却没有实质性提高，那么就有必要动

态移除一些成员智能体以减少资源的浪费。另一种动态调整系统（成员及结构）的可行方法是自动动态调整。在这种情况下，可以在不停止系统运行的情况下改变系统的成员和结构。比如，在任务需要时系统可以自主增加必须的功能智能体并分配工作量，以更高效地实现共同目标。当然，当工作量变轻时，系统也可以减少任务较少的智能体成员，以节约系统成本。

虽然增减智能体成员可以提高任务效率，增强社会模拟的真实性和可信度，但我们也面临着一些挑战。即，系统的总体架构设计和计算优化应该如何自动进行，并且还可确保整个系统的平稳运行。另外，在一个多智能体的系统或社会中，每一个智能体都可能存在着信任风险。随着系统功能智能体数量的增加，协调智能体间的交互的难度也会增大，可能会使各智能体之间的合作更难，进而会影响到系统总体目标的进程和实现。因此，构建一个基于大规模的稳定而连续的超智能体系统，忠实再现人类的工作和生活场景，会是一个前景广阔的研究方向。一个有能力在由数百甚至数千个功能智能体组成的社会中稳定运行并执行任务的超级智能体系统，一定会在未来的现实世界中找到与人类互动的广泛应用。

19.8.2 未来的超智能体系统，必基于智能智联

超智能体系统中的智能体，最典型的特征，是具有智能智联的能力。智能体智能智联，会提升系统的协同能力。协同，一方面，是系统内部的协同。比如利用数字技术优化企业的供应链，加快相关资源的合理配置，实现企业上游与下游的协同。另一方面，是跨系统的协同。跨系统的协同，可充分利用智能技术与协同技术，来实现系统及模式的创新。

智能联接的核心价值，产生于数据、信息和知识的流动。传统通信的价值是传递信息，随着多智能体间数据交互网络的快速发展，智能联接的核心诉求，将是可承载关键的生产要素—数据、信息、知识和智能。未来的智能联接，将要求确保业务数据信息可在人、物、流程和应用之间实时流转，为智能决策提供可靠的数据参考，并让知识和洞察快速的通过网络分发给业务应用，从“建立管道”转向“持续交互”。未来的智能联接，将能实现整个系统的互联互通，加速推动智慧融合的展开，从而更好地理解系统需求，促进问题解决和服务创新。

未来智能联接将是泛在和普惠的网络联接。随着无线网络可靠性的提升，未来的联接方式将从有线方式向更加便捷的无线方式转变。无线网络环境可极大地简化网络的部署、使用和维护的过程，解决了大量网络布线过程所带来的高额投入，也降低了其在管理维护和升级改造方面所付出的额外成本。相较于有线网络的物理局限性，无线网络在架构调整方面也体现出更高的灵活性，可以更好的延伸其服务空间，拓展服务能力。

未来智能联接技术方案将更加场景化。智能家居、智慧城市、工业互联网、车联网等组网场景，很多都是多种联接技术的融合解决方案；组网考虑的不再是简单的互通，而是针对具体的应用，提供独特的网络能力，例如，低成本网络，高可靠网络，确定性网络，以及具备智能自愈能力的网络等。极简、灵活的组网适配能力，将进一步催生新的应用场景，保证智能化业务的无缝覆盖。

智能智联还有助于构建新型的信任体系。在数字经济时代，信任关系的约束条件已发生了颠覆性变化。传统的社会信任关系必然演变为新型的智能化的数字信任关系。在以万物泛在互联为特征的智慧应用场景中，在复杂叠加的数字环境下，智能智联会通过统一的数字身份标识、实体识别认证和一系列安全策略机制，构建数字信任来确保交互关系的可靠性、稳定性和便捷性。

19.8.3 未来超智能体系统的智能融合，将会助力各个行业的智能化发展

从产业的角度考虑，智能应是企业或行业能够在不断变化的环境中及时做出反应和表现的能力。未来，作为一个智能融合联合体的经济组织，应该具备三种能力：信息综合能力、不断学习的能力、深度的决策与洞察能力。

产业智能化应以创新为驱动，依据产业需求进行信息的交互与传递，利用知识和数据分析洞察产业的运行规律，利用价值提升驱动全价值链和全要素的网络化协同，从而产生新价值、新模式、

新业态与新产业。

未来，将会是数字技术与智能技术的大融合与大协同的时代。技术、科学、产业、区域经济、社会的高度交叉与融合，将涌现出更多的新模式、新业态、新现象与新的价值创造方式。产业数字化与智能产业化的双螺旋将是数字经济增长的新引擎。智能智联的普惠化将使企业数字化和智能化的成本大幅下降，中小企业将真正享受到与大企业一样的数字红利和智能红利。由“数据、算力、算法”所支撑的智能决策将会渗透到企业运行的方方面面，对企业管理进行系统性优化。每一个带有明显产业集群效应的区域，都会自成产业生态。万物互联、万物上云将推动智能联接的泛在化，极大拓展企业价值链的边界与规模。

未来，企业的信任体系也需要构建在一个可信的生态之上。亦即，处于各个行业上下游产业链的各个相关企业，在实现可信的治理基础上会形成一个相互依存的全新产业链。基于创新的技术手段，前瞻性地管理跨越合作伙伴、供应商、客户和内部员工的生态系统，确保实体之间数据流通的完整性以及信任体系的可靠性。

未来，超智能体系统的智联智融，将会有效提升企业的综合信息的能力、不断学习的能力和深度的决策与洞察能力。这里，**综合信息能力**是指将各种结构化、非结构化信息转化为数据的过程，这种能力通常需要企业或行业的每一个参与者都明晰数据的价值，并在信任的基础上获取数据。面对多来源、实时性要求越来越高的海量数据，形成信息综合处理能力，不断提升数据的价值，是实现企业智能的基础。**学习能力**是指认识并理解各种信息与先前已有知识之间的关系，以及它们在特定场景上的应用能力。它通常使用自然语言处理、机器学习、推理、知识图谱等技术，使数据转换成能够被理解、治理和解释的知识。在企业或行业构建学习能力，必须要实施人工智能解决方案，要能够识别并选择业务需求最为迫切、数据就绪度高、可快速落地并复制的应用场景，加强对智能技术创新的投入或引入合作伙伴的能力，确保智能项目的实施能够满足真实需求并持续发挥效能。**深度的决策与洞察能力**是指可为企业或行业的每一个参与者提供自动化决策支持的能力，其范围包括为智慧社会的每一个角色赋能，将信息转化为决策或洞察，并最终转化为行动力。这就需要企业或行业的每一个参与者都能综合使用各种数据、知识和智能的管理和治理技术，降低数据、知识或智能获取的难度，建设灵活的基础架构，确保数据、知识和智能能够灵活集成，以实现深刻的决策与洞察能力。

图表10 IDC未来智能框架

图 19.8.1 未来智能框架

19.8.4 未来超智能体系统的智能协同，会创造出无处不在的智能化体验

基于多智能体的超智能体智慧集成，协同是关键。有效的信息交互和全面的智能化，也有依赖多智能体间的协作。对于像诸如编写软件这样的复杂任务，超智能体会将任务分解成由不同角色（如软件工程师、产品经理、设计师、QA 工程师等）执行的子任务，并让不同的智能体去完成不同的子任务。对于基于 AI 大模型的超智能体，不同的智能体也可以通过提示一个或多个 AI 大模型去执行不同的任务来构建。例如，要建立一个软件工程师编程智能体，我们可以提示 AI 大模型，你是一个编写清晰、高效代码的专家，并在系统运行时让它专门执行编写代码的任务。

尽管有时我们是多次调用同一个 AI 大模型，但我们采用的是构建多智能体的抽象方法。这看似违反直觉，但却有理由获得支持：**因为它确实有效**！有许多团队使用这种方法取得了良好的效果，没有什么比结果更有说服力的了。此外，研究表明，多智能体的表现会优于单一智能体。虽然现今的一些 AI 大模型能接受非常长的输入上下文，但它们真正理解长而复杂输入的能力是参差不齐的。采用多智能体式的工作流程，让 AI 大模型一次只专注于一件事，可以获得更好的结果。

实现多智能体协同模式需要有一个管理框架（工具）或核心智能体，用以将复杂任务分解成子任务并集成分散处理后的结果。以前，当在单个 CPU 上运行计算机代码时，我们也经常将程序分解成不同的进程或线程。这种分解有助于我们将任务分解成更易于编码的子任务。使用多智能体角色进行思考和处理也是同样的道理。

在许多公司中，管理者通常会根据任务需要决定招聘哪些角色，然后将复杂项目分解为更小的任务分配给具有不同专长的员工。超智能体系统使用多个智能体协同工作的做法与此类似。每个（入盟的）智能体都会根据系统需要及自己的特长实施独立的工作流程，拥有自己的经验和记忆，并有可能在遇到难题时请求其他智能体的帮助。超智能体系统可以进行规划和使用工具。这可能会产生大量的 AI 大模型调用和智能体间的信息传递，也可能会形成非常复杂的工作流程。虽然管理不易，但这却是非常值得的。因为它为我们如何“雇佣”各个功能智能体和分配任务给各个人工智能体提供了一个非常正常的智能化的心理框架。像 AutoGen、Crew AI 和 LangGraph 这样的框架，已可为解决问题提供丰富的多智能体解决方案。如果你对此类多智能体系统感兴趣，不妨去看一看。虽然它们目前可能不总是产生你想要的结果，但你也会对它的某些表现感到惊讶。

多智能体协同后的输出结果的质量很难预测，特别是当允许智能体自由交互并为它们提供多种工具时。也许，增加更成熟的反思和工具使用模式会使结果更为可靠。

数字技术的每一次飞跃，都会给经济和社会形态带来持续性的积极推动作用。基础网络的发展会带来高效的信息交换；网络交互能力的进步不仅显著提升了社会生产和生活的效率，还会直接催生出众多的经济和商业形态；进而不断改变经济结构和增长方式，推动社会持续发展与变革。

智能联接会是未来社会不可或缺的基础设施，也是智慧融合的前提条件。泛在的通信网络已为数字经济发展打下了坚实的基础。未来的联接，将加速未来的“沟通”从“人与人” 转向“物与物”或“人与物”。智能联接将开启万物互联的联接新时代。在未来社会中，移动办公、在线教育、在线游戏、在线视频、在线新零售等一系列的互联网应用都在重构着我们的生活方式和商业模式。更多业务将从线下向线上迁移，促进线上和线下融合的速度。未来的行业市场将会随着数字化浪潮的来临，迎来新的一波产业进步。

图 19.8.2　未来智联框架

超智能体系统的智能联接，会成就未来更大范围的智能融合，也会创造出无所不在的智能化体

验。智能联接是实现数字化和智能化的必要条件。智能联接会将智能下沉到离业务更近的网络边缘，加速产业的智能化升级。未来网络联接量将快速增长，随之产生的数据也将井喷式增长，如何从海量的数据中发现价值，并合理的控制流量成本，降低网络延迟，提升用户体验，具备网络智能能力的边缘计算将会发挥更重要的作用。未来，企业的各种智能化基础设施，将会部署到网络的各个边缘和终端，而不是集中一个云端的数据中心，边缘（智能化）应用程序的数量将会成倍增长，智能将无处不在。

19.8.4 未来超智能体系统的社会信任，会确保可信的智能服务和决策

超级智能体的运行也面临着一系列的挑战。首先是**安全性和可靠性**。安全性和可靠性是智能体的关键特性，对其稳定运行和对用户及社会的保护至关重要。因为这两个因素会直接影响到人们对智能体的信任和接收。若智能体存在系统漏洞，易于遭受攻击或数据泄露等问题，会导致对用户或社会的直接损害。因此，安全性和可靠性是一个始终需要认真对待的问题。**对抗鲁棒性**已是深度神经网络开发的重要课题，它在计算机视觉、自然语言处理和强化学习等领域都得到了广泛探索，是决定深度学习系统适用性的关键因素。当面对扰动输入时，对抗鲁棒性高的系统通常会产生原始（正确）输出。然而，预训练语言模型特别容易受到对抗性攻击，导致错误的答案。这种现象在 AI 大模型中也普遍存在，给未来基于 AI 大模型的智能体的开发带来了巨大挑战。对抗性攻击对大语言模型的影响还仅限于文本错误，但对于行动范围更广的基于 AI 大模型的智能体来说，对抗性攻击有可能会促使它们采取真正具有破坏性的行动，造成重大的社会危害。为了解决这些问题，我们可以采用对抗训练、对抗数据增强和对抗样本检测等传统技术来增强基于 AI 大模型的智能体的鲁棒性。然而，如何设计一种策略，在不影响有效性的前提下，全面解决智能体系统内所有模块的鲁棒性问题，同时保持其实用性，则是一项更为艰巨的挑战。

确保可信性是智能体领域另一个极其重要但又极具挑战性的问题。深度神经网络因其在各种任务中的出色表现而备受关注。然而，它们的黑箱性质掩盖了其卓越性能的基本因素。与其他神经网络系统类似，AI 大模型也难以精确表达其预测的确定性。这种不确定性也被称为**校准问题**（Calibration），已引起了基于 AI 大模型的智能体开发和应用时的特别关注。在现实世界的交互场景中，这会导致智能体的输出与人类的意图不一致。此外，训练数据中固有的偏差也会渗入神经网络。例如，有偏见的语言模型可能会产生涉及种族或性别歧视的话语，这可能会在基于 AI 大模型的智能体的开发和应用中被放大，从而造成不良的社会影响。此外，大语言模型还存在严重的幻觉问题，容易产生偏离事实的文本，从而损害基于 AI 大模型的智能体的可信度。为了解决这些问题，我们可以采用引导模型在推理阶段展示思维过程或解释，以提高其预测（决策）的可信度。此外，外部知识库和数据库的整合也可用于缓解幻觉问题。在训练阶段，我们可以引导智能体的各个组成部分（感知、认知、行动）学习稳健而随意的特征，从而避免过度依赖捷径。同时，过程监督等技术也可以提高智能体在处理复杂任务时的推理可信度。

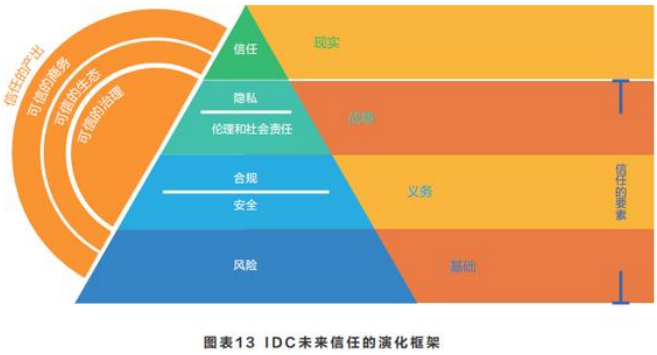
使用超级智能体还有一个伦理与责任问题。伦理和责任是人类智能体的核心原则，决定其价值观和目标，以及其对用户和社会的尊重与保护。这些原则同样也直接影响着人工智能体的可信度和可被接收的程度。若（人工）智能体表现出不公平、不透明或不可靠等问题，可能引发用户或社会对智能技术的排斥。责任归属也是人工智能体的一个关键议题，人与智能体协同中的责任归属不清晰或不公正也会带来严重后果。

潜在风险是必须预先认真考虑的。 基于 AI 大模型的智能体被赋予了广泛而复杂的能力，使其能够完成各种各样的任务。然而，对于怀有恶意的人来说，这些智能体也有可能会成为威胁他人和整个社会的工具。例如，这些智能系统有可能被用来恶意操纵舆论、传播虚假信息、破坏网络安全、从事欺诈活动等。因此，在部署这些超智能体系统之前，相关部门需要制定严格的监管政策，以确保使用者必须负责任地使用这些智能体。开发公司也必须加强这些系统的安全设计，防止恶意利用。

需验证是否可信的信息主要有两类。一类是客观信息。客观信息的可信度是以“风险”来度量的，更精准的信任度量将从双维的“可信任”和“不可信任”决策转变为对诸如合作伙伴、供应商、客户和监管机构等的信任程度的衡量与评估。这种风险度量提供了未来行动和决策所依据的数据。从概念上可以认为，这种客观的评级分数可从 0 到 100，其中 0 表示完全不可信任，100 表示可完全信任；现实很可能介于可完全信任与完全不可信任之间。对风险的客观度量将是未来智能社会成员间活动决策的一部分，整个风险防止算法都是为了量化和减轻交互风险而设计的。另一类是主观信息。需要根据对不断更新的新闻和信息来源的“感知”来衡量信任，从感知的性质来看，这些信息来源可能带有个人或群体的偏见。

除了网络安全，还有一个信息安全问题。近年来安全事件频发与数据泄露危害巨大，让人们意识到隐私信息与数据安全的重要性。在未来的智能社会，每一个社会成员都需要提升数据安全防范意识，不断提升安全防护手段，加强用户权限管理，并对数据进行实时监控。对数据资产及隐私信息进行充分的保护也是实现可信任交互的前提条件之一，也是智能社会的重要特征。

在未来的基于 AI 大模型的人-机结合的智能体系统中，信任将是数字信任，将是系统的一个首要的战略概念，也将是组织间或用户间的从技术层面整合到更广泛的认知和声誉层面的社会基础。其中，风险、安全、合规性、伦理和社会责任、隐私将是构成数字信任的五个关键要素。未来的信任将包括三个“产出”：可信的治理、可信的生态和可信的交互，这些“产出”将构成未来智能社会的基石。



图表13 IDC未来信任的演化框架

图 19.8.3 未来信任的演化框架

具体而言，在未来智能社会或智能体社会中，未来信任的三个“产出”将会对未来社会、未来组织和未来个人所带来深远的影响：（1）**未来的可信社会环境将是在可信的交互基础上实现的。**可信的交互是未来信任的最终实现目标，也是成为保障未来可信社会环境的正常秩序的基石。可信交互的发展将使社会中的组织和个人能够通过提供差异化的产品、服务和体验等来增加收益，这些差异化的体验和大量的“客户”数据，需要更多的技术来保护其数据安全和隐私信息。“可信任的组织”的一个重要组成部分是“客户”对组织的信任感，它是在长期的交互合作中形成的。但也有可能在一瞬间被破坏。此外，“交互”并不一定意味着各方之间一定存在交易，也可能意味着某类建立在高度定制体验之上的增值服务，这种增值服务可能连最终用户都不知道“交易”正在发生。例如，在未来社会中，会以数字身份为中心，构建无感的“零信任”安全体系；即管理端通过实时感知、动态访控，集中管控身份安全、终端安全、链路安全，有效控制“合法用户”所造成的可能的风险，但它对于最终用户而言却是无感的。另外，公平、可靠与安全、隐私、包容和问责制也是未来实现可信交互环境需要建设的重点。（2）**未来的可信组织必须要实现基于信任的管理和治理。**未来的可信组织将会把社会责任视为增加社会信任的一种方式。社会责任也包括投资于促进人类和社会福祉的各种举措，例如，社区拓展、教育倡议和环境倡议等。与此同时，消费者希望获得的是满

意的可信的服务，而不仅仅是普通交易。社会责任在一定程度上是一个主观领域，是一个可能被忽视的领域，但在主观衡量标准下，它一定是一个可信组织必须关注的领域。此外，可信的治理应该侧重于组织内部。组织的管理层需要从数字信任的五个关键要素出发，在组织信任治理层面实施控制，以确保组织成员都遵循有助于建立或加强信任的行为和最佳实践。信任的文化源于组织的高层，道德和监督实践必须源于管理层。在组织内部采取行动并遵守道德要求也将会减少信任风险。重要的是，消费者（使用者）将是在现实的交互中逐渐认可（信任）该组织的。对于基于多智能体的超智能体系统而言，也是如此。

19.8.5 未来超智能体系统的生态，必是多方共赢的全面智能化的智能生态

新技术融合会驱动全场景智能的落地。“未来智慧集成”、 “未来智能联接”、 “未来社会信任” 会是全场景智能落地的核心要素。“智能社会”是智能化的社会框架，“集成智能” 、“智能联接”、“安全可信”是三个最为关键的技术，也是跟新型社会生产要素—数据、信息和能力联系最紧密的三项技术。联接实现数据采集，智能对数据进行分析，安全可信确保数据和能力的共享与流通。

在未来智能社会，智联智融通过万物智联可实现全面感知、智能决策和自主执行，极大拓展了人类探索世界的深度与广度，并推动着数字世界（线上）与物理世界（线下）的融合。未来智能社会本质上是数字化的。在数字化的世界中，数字经济本质上就是数据的流通。作为未来智能社会重要的生产要素，数据在数字世界中的地位或相当于物理世界的石油。

图 19.8.4 智能社会数字化转型框架

未来，“智能集成” 、“智能联接”、“安全可信”等新技术的融合应用将重塑未来社会的格局，开放合作、互利共赢，生态共创也会成为社会、组织和每一个社会成员的普遍共识。未来，在利用先进的通信技术、云计算技术、大数据技术、物联网技术、人工智能技术等新技术打造各类（超级）智能体，创建形态多样化、交互方式立体化的全场景智慧的同时，也会赋能社会的各行各业，促进社会的合作共赢，共筑智能社会的新生态。

要共筑智能社会的智能新生态，社会中的每一个成员都需要加强协作与协同，智能社会的每一个成员（包括人、组织和各类智能体）都需要从各自的定位（角色）出发，依各自所具备的知识和技术特长，协同一致的参与到社会的运行之中。社会的管理者也需要转变职能，帮助社会生态系统中的各种要素加强协同，推动科技创新向多社会主体发展，使每个社会主体都能够共享到更多的发展机会、技术能力、创新服务及社会资源等，从而形成一个完整的智能化社会生态，为智能社会的建设创造更大价值，满足人们日益增长的对美好生活的愿景。

随着社会生态系统的日渐复杂，社会运行体系的不断细分，系统化的社会资源整合能力对于所有社会运行参与者来说都变得至关重要。社会运行参与者（各类智能体）不应仅满足于单纯的对现有技能的展现，还应通过持续地学习积累更多知识、数据、经验和技能，并最终反馈给社会，实现

全社会的智能化升级。社会运行参与者也应在法律和合规的基础上积极参与信息、知识和技能的交互与共享，将智慧带到每一个社会运行场景，最终实现全社会的互利共赢。

未来，智能社会的优化运行将严重依赖于社会可信生态。可信的社会生态是指社会运行环节中的各个组织和个人（当然也包括各个智能体）都应遵从诚信的原则，在可信治理的基础上形成相互依存的全新的可信环境。基于可信生态的社会系统将通过前瞻性地管理构建跨越所有组织和个体的社会生态系统，确保所有社会成员之间交互和流通的有序性和可信性。建立可信的社会生态系统将依赖于各个社会组成体对高信任度的追求。考虑到未来数字经济对速度和敏捷的需求，未来的可信生态需要一个 “社会信任框架”来评估、管理和度量。社会信任框架将有助于建立社会生态系统内部信任度的组织关系。而所构建的社会信任框架将包括着如下的内容：（1）了解整个（社会和环境）生态系统的全部潜在风险，并根据对生态系统的影响对这些风险进行优先级排序。（2）了解生态系统中还存在哪些针对已识别风险的漏洞，并制定计划来解决这些漏洞。（3）制定一个持续监测风险和漏洞的计划，并创建一种模式来衡量整个（社会和环境）生态系统的整体“风险得分”。

与以往的信息技术产业不同，智能社会下的智能产业生态更加融合且复杂，对数字化程度和技术水平要求更高，产业链上的各个环节将不再是只在一个企业内部独立完成。所以智能产业内的企业应更具开放性和创新性，除了自己独立研发，也需在更大范围内与云服务商、软硬件供应商、应用开发类服务商、终端用户等保持密切协作，借助合作伙伴的优势，整合数据及创新技术资源，将内部和外部的智慧真正融合到企业的体系结构当中。

图 19.8.5 建立多方共赢生态

由于基于 AI 大模型的超级智能体系统具有独特性和潜在能力。“基于 AI 大模型的智能体+”有望成为未来智能系统的主流，有望在多个领域实现落地应用。我们认为，基于 AI 大模型的 AI Agent 的研究应是人类不断探索接近 AGI 的最恰当途径。随着智能体变得越来越“可用” 和“好用”，“AI Agent+”的产品会越来越多，未来将有望成为人工智能应用层面的基本架构，包括 to C 和 to B 产品等。由于智能体对环境反馈的依赖性较强，具备问题领域显著特点的运行环境，才是更适合 AI Agent+建立起对某一个垂直领域深入认知的应用场景。

另外，超智能体系统常被用于对社会运行（基于社会系统动力学原理）进行模拟。但虚拟仿真环境与真实世界之间是存在很大差距的。虚拟模拟系统受场景限制，主要针对特定任务和特定环境，以“信息”模拟的方式进行交互。而真实世界的任务和环境是无限的，可容纳各种任务。要弥合这一差距，超智能体系统必须具备应对来自外部因素和自身能力的各种挑战的能力，使其能够在复杂的真实世界中能有效操作和运行。而智能体若想具备较强的融入环境的能力，若想可以无缝地融入真实的物理世界，它们就不仅需要理解和推理具有隐含意义的模糊指令，还需要具备灵活学习和应用新技能的能力。在面对一个无限开放的真实世界时，以有限环境训练的智能体必须具有能有效处理来自真实世界的各种信息并顺利运行的能力。最后，在模拟环境中，AI Agent 的输入和输出都是虚拟的，可以进行无数次的试错尝试。在这种情况下，对错误的容忍度很高，不会造成实际伤害。然而，在真实世界环境中，智能体的不当行为或错误可能会对社会和环境造成真正的伤害，有时甚至是不可逆转的伤害。因此，我们需要关注智能体在做出决定和产生行动时的安全性，确保它们不会对现实世界造成威胁或伤害。

最后谈一谈人工智能技术的发展会对未来经济和社会的可能的影响。依现有的技术，我们并不认为在一个可见的未来，硅基生命会完全替代碳基生命。超级人工智能作为一种新型生产力，会带来经济的飞速发展，也会带来生产关系的变革，但也不至于会发生翻天覆地的革命。人工智能带来的最大挑战是对人的挑战。在未来社会，一般人所从事的一般性的工作，很有可能会被“机器人”代替。在工作方面，人与智能体之间，会发生竞争。基于效率和效益的考虑，“雇主”可能会尽量减少人力的投入。随着智能体技术的成熟，未来的社会工作流程中，也会出现越来越多的智能体系统。然而，这些发展并不必然地会引发严重的社会危机。人是聪明的，旧的工作的消失，会有新的工作产生。任何时候，都需要人去适应社会。被淘汰的，只会是不适应社会的人。我们能做的，是让自己去适应社会，去适应智能化的社会，去做一个对社会有用的人。

有人担心一旦超级智能体的智能发展到超越人类的能力并产生野心，它们就有可能试图夺取对世界的控制权，从而给人类带来不可逆转的后果。我们不能说这些担心是多余的，但很难发生。因为超智能体终究是人开发的。人类不会愚蠢到让此类事件发生。为了防范人类面临的此类风险，研究人员一定会在开发超级智能体之前，全面了解其运行机制。他们还会预测这些智能体可能产生的直接或间接影响，并设计出规范其行为的方法等。

参考文献

1901 刘增良 刘有才 因素神经网络理论及实现策略研究北京师范大学出版社 1992.8

1902 刘增良,刘有才 因素神经网络理论及其应用　贵州科技出版社 1994

1903 刘增良 刘有才 模糊逻辑与神经网络—理论研究与探索 北京航空航天大学出版社 1996

1904 刘有才 刘增良 模糊专家系统原理与设计 北京航空航天大学出版社 1996.8

1905 鲁斌　广义智能系统柔性超拓扑空间模型研究与应用　西北工业大学博士论文 2003年10月

1906 汪培庄 模糊集与随机集落影　北京师范大学出版社　1985年09月

1907 利珊 不确定性中的随机性和模糊性 金华职业技术学院学报第2010年 第3期

1908　何莉敏,董艳,刘兴薇,李德荣 不确定变量的比较及相互联系 内蒙古科技大学学报 2009年第3期。

1909　AI Agent 和大模型落地有什么关联？https://www.zhihu.com/tardis/bd/ans/3226090167

1910 初识 Embodied AI　https://blog.csdn.net/Cameron_Rin/article/details/127014158