Agent思维状态模型*

马光伟 徐晋晖 石纯一

(清华大学计算机科学与技术系 北京 100084)

E-mail: mgw@cst4.cs.tsinghua.edu.cn

摘要 文章综述了20世纪90年代以来多Agent系统中Agent思维状态模型的研究结果.从直观分析、形式化模型、结构模型、合作研究和应用方面来介绍Bratman的哲学观点、Cohen和Levesque的意图模型、Rao和Georgeff的BDI模型等重要成果,最后介绍了正在开展的工作.

关键词 Agent,MAS(multi-agent system),思维状态,信念,愿望,目标,意图,承诺,合作,联合意图.

中图法分类号 TP18

在多Agent系统(multi-agent system,简称MAS)中,Agent思维状态模型用来研究如何描述Agent的思维属性和它们之间的关联,以及与感知、规划、行为、协调、合作等活动的关系.从意识立场出发,一般把信念(belief)、愿望(desire)和意图(intention)当作基本的思维属性(简称BDI).Agent的BDI分析不同于人们所熟悉的知识表示和推理,它是对思维过程的一种深层次描述,为的是适应MAS求解过程中环境的多变,在突发事件发生的情况下,保证理性的、一致的合作行为.

Dennett描述了人们看待系统的立场:(1) 物理立场,着眼于系统的物理特性和规律;(2) 设计立场,着眼于系统的设计目标;(3) 意识立场,把系统看做理性Agent,通过信念、愿望和其他意识属性来预测Agent的行为.Singh认为把Agent看做意识系统的好处是:(1) 对设计者和分析者来说,这样是自然的;(2) 对描述复杂系统的行为提供了简洁的表示,便于理解和解释;(3) 不依赖物理实现就可以得到许多Agent行为的规则和模式;(4) 可被Agent自身用来进行互相推理[1].实践表明,对于许多复杂系统,即使有了详尽的结构和工作机理的描述,也很难从设计立场来预测和解释它的行为,更适合采用意识立场的描述.意识观念是一种抽象工具,为人们描述、解释和预测复杂系统的行为提供了一种方便的方式[2].

另外,有关Agent模型的研究还有两方面的动力:(1) 控制分布式计算的复杂性;(2) 克服人机界面的局限性.伴随分布式计算的广泛使用,需解决Agent之间互操作的复杂性问题.人们试图使用与人类思维属性相对应的概念,在规划层提供一种类似于对低层通信协议的封装.随着任务复杂性的增加,直接操纵界面的优势正在消失,人们试图用Agent技术来实现一种间接管理风格[1].目前,Agent思维状态模型是以模态逻辑为工具来描述的,一般知识逻辑用KTD45公理、信念逻辑用KD45公理、目标和意图用KD公理来分析.Agent的逻辑描述还存在一些问题,如:(1) 逻辑全知问题:$\varphi \supset BEL(\varphi)$,这个要求对资源有限的Agent是不现实的;(2) 无为而治问题:如果Agent认为φ必然总是为真或者必然最终为真,那么它无需把这作为目标或意图;(3) 副作用问题:如果Agent认为$\varphi \supset \gamma$必然总是为真并且它也有意图φ,那么也不能强迫它采纳意图γ.

* 本文研究得到国家自然科学基金资助.作者马光伟,1969年生,博士生,主要研究领域为多Agent系统,社会行为规范机制.徐晋晖,1966年生,博士生,主要研究领域为多Agent系统,联盟机制.石纯一,1935年生,教授,博士生导师,主要研究领域为人工智能应用基础.

本文通讯联系人:马光伟,北京 100084,清华大学计算机科学与技术系

本文1998-07-01收到原稿,1998-10-23收到修改稿

1 直观分析

1.1 BDI 的直观描述

信念描述了 Agent 对当前世界状况以及对为达到某种效果所可能采取的行为路线的估计,属于思维状态的认知方面.愿望描述了 Agent 对未来世界状况以及对所可能采取的行为路线的喜好,属于思维状态的感情方面.Agent 可以拥有互不相容的愿望,而且也不必相信它的愿望是可实现的.目标是 Agent 从愿望中选择的子集,但还没有采取具体行动的承诺.一般 Agent 相信目标是可实现的.由于 Agent 资源有限,不能一次去追求所有的目标,所以,Agent 选择某个目标(或目标集)来作出追求的承诺,形成意图.Agent 的当前意图(或意图结构)是被选的目标集和处理状态,意图属于思维状态的意向方面,其作用是引导并监督 Agent 的动作.承诺表示从目标到意图的转换,它决定了 Agent 对于所追求的意图的坚持程度并控制对意图的重新考虑.在 MAS 中,一般认为社会承诺是多 Agent 群体的胶合剂,Agent 之间的合作、协调等往往归结为建立不同性质的社会承诺.

信念、愿望、目标和意图的关系[3]:

(1) 意图-信念一致性:Agent 应当相信它的意图是可能的,而不相信达不到目标.在正确的条件下,相信会达到目标.

(2) 意图-信念不完全性:Agent 有意图达到某种状况,但不必相信这种状况最终一定会实现,也即 Agent 对其意图持不完全的信念是理智的.

(3) 副作用问题:Agent 有意图做 α,且相信做 α 必须要做 β,那么也不必要求有意图做 β.

(4) 内部一致性:Agent 要避免拥有冲突的意图,但允许拥有冲突的愿望.

(5) 手段-目标分析:意图要求 Agent 在未来某时要思考提出的问题,而愿望则没有必要.

(6) 跟踪成功或失败:意图可被认为是愿望加上行动和实现的承诺.所以,必须对意图的成功或失败进行跟踪,在失败时进行重新规划.

1.2 Bratman 的理性平衡和行为意图

理性平衡是使理性 Agent 的行为符合环境的特性,包含各种客观条件和社会团体因素,对 Agent 的约束表现在:(1) 有限的计算能力和信息资源,Agent 不可能在任何时刻都进行计算,也不可能在给定时限内完成任意数量的计算,环境也会经常发生不可预测的变化,关于世界的知识也不可能总是完全的;(2) 协调问题,Agent 需要协调现有的活动和将来的活动,还要协调不同 Agent 的活动.Bratman 认为,必须使用意图来描述理性平衡,而且意图不能归结为信念与愿望,必须看成是一个独立的思维属性[5].

意图分为将来意图和当前意图.当前意图引导立即就要发生的行为,与行为有密切关系.而将来意图不仅引导行为,同时引导将来的意图,与规划联系紧密.行为意图对于 Agent 行为的制约,一方面表现在自主 Agent 不可以随意改变自己已有的行为意图,尤其是已经付诸实施的意图;另一方面,自主 Agent 不能无视环境的变化,坚持不合实际或不再重要的意图[5].

1.3 联合意图

Tuomela[6]认为联合意图是社会科学的中心概念,包含了参加者的推理和联合行动的产生.其作用是:(1) 与单 Agent 的意图一样,提出问题、限制 Agent 的行为选择;(2) 负责对联合行动的初始、指导和监控;(3) 帮助将个体和群体联系起来;(4) 规范 Agent 的思考和行动.设 Agent 集合 G 的个体 A 有"We-Intends to do X"的个体意图当且仅当在 G 中的所有个体同意联合执行 X 的基础上有:(1) A 打算执行它的部分,(2) A 有对联合行动的前提成立的信念,(3) A 相信 G 中存在对于联合行动前提成立的相互信念,(4) 使(1)成立,部分是由于(2)和(3).A 有"We will do X"的 group-intention 个体意图,当且仅当在 G 中的所有个体同意联合执行 X 的基础上有:(1) A 有"We-Intends to do X",或(2) A 形成一个持续的执行 X 的 group-intention,它是"We-Intends to do X"的意向.

有执行联合行动的 Joint Intention 当且仅当(1) 这些 Agent 有执行联合行动的 group-intention;(2) 对于(1),存在相互信念.

已经形成执行 X 的联合意图(JIP)当且仅当(1) 每个个体接受联合执行 X 的规划,(2) 每个个体将他们的接受信息通知其他个体,(3) 由于(1)和(2)在个体之间产生了相互信念,他们联合承诺去执行 X,且每个个体承诺完成它的子任务.

2 BDI 模型

2.1 思维状态模型

Cohen 和 Levesque 的意图模型[7]基于正规模态逻辑的可能世界模型.它以模态连接词定义信念和目标,依据持续目标定义了两种意图,进而导出了一些有实际意义的性质.形式化为:

(1) $M,\sigma,n,v\models(BEL\ x\ \alpha)$ iff 对所有从 n 通过信念关系 B 可达的可能世界都有 α 为真.

(2) $M,\sigma,n,v\models(GOAL\ x\ \alpha)$ iff 对所有从 n 通过目标关系 G 可达的可能世界都有 α 为真.

(3) $(P\text{-}GOAL\ x\ p)=_{def}(GOAL\ x\ (LATER\ p))\wedge(BEL\ x\ \neg p)\wedge(BEFORE\ (BEL\ x\ p)\vee(BEL\ x\ \Box\neg p))\ \neg\ (GOAL\ x\ (LATER\ p)))$.

(4) $(INTEND_1\ x\ \alpha)=_{def}(P\text{-}GOAL\ x\ [DONE\ x\ (BEL\ x\ (HAPPENS\ \alpha))?;\alpha])$,$\alpha$ 表示一个行为.

(5) $(INTEND_2\ x\ p)=_{def}(P\text{-}GOAL\ x\ \exists\ e(DONE\ x\ [(BEL\ x\ \exists\ e'(HAPPENS\ x\ e';p?))\wedge\neg\ (GOAL\ x\ \neg\ (HAPPENS\ x\ e;p?))]?;e;p?))$,$p$ 是一个谓词.

Cohen 和 Levesque 的意图模型是比较早期的工作,成为后续工作用以比较的基准.但是,我们认为,将意图归结为信念和愿望的做法不适于对 Agent 有限理性的假设.

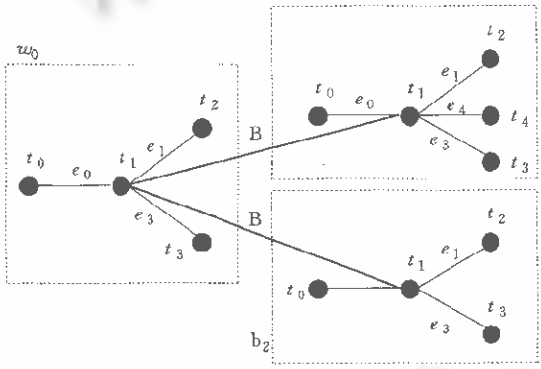

图1 时间树的世界模型

Rao 和 Georgeff 的 BDI 模型[8]的形式化系统是对计算树逻辑 CTL* 的扩充,使用经典的可能世界语义模型.图 1 表示的是 t_1 时刻 Agent 的信念可达关系.每个世界的时间结构是一个时间树,Agent 拥有一个线型的历史和分支型的未来,分支代表 Agent 在相应时刻的选择.在该模型中,公式有状态公式和路径公式两种,以模态连接词定义信念、目标和意图,给出一些模型约束和公理,阐述了思维属性之间的约束关系,特别提出强现实性对于理性的作用,还给出了不同的承诺策略来描述 Agent 的不同性格.在后来的工作中,有很多都是基于 Rao 和 Georgeff 的工作.

Konolige 和 Pollack 的意图模型[9]基于非正规模态逻辑,使用认知结构描述 Agent 的思维状态,阐述语义的工具是集合论意义下的场景概念,引入的意图关系图的概念以直接的方式表示了意图的结构,初步表达了 means-end 的结构.

Bell 的属性改变模型[10]建立了一套持续规则,表征 Agent 会保持它的思维属性直到有改变的理由,显式地引入了时间的概念.形式化语义是对 Hintikka 可能世界语义的自然扩展,用模态连接词定义影响、信念、愿望、意图、义务和理性,且显式地加入了时间特性.

Gaspar 和 Coelho 指出[11]已有意图模型的不足:(1) 采用模态逻辑和理想假设;(2) 没有给出目标和意图是如何产生的线索;(3) 没有提出当环境变化时,目标和意图的修正问题.基于此,他们提出了 Agent 的目标和意图的模型以及修正框架,扩展了以前提出的 Agent 思维状态中的信念的模型.信念和目标模型都是推理模型,带有Agent用于产生信念或目标的推理规则,也给出了怎样从一个不协调的信念或目标集合选出偏好子集.

Linder 和 Hoek[12]的模型很有表达力,在两个层次上讨论了喜好、目标和承诺的性质,对于承诺考虑了静态和动态两个方面.静态指 Agent 已作出的承诺,动态指 Agent 作出承诺的行为.模型对逻辑全知问题有比较完善的处理.

Cavedon 和 Padgham[13]指出意图逻辑的非正规性,在 Kripke 框架中引入"不可能世界",发展了非正规世界

框架,给信念和意图提供了细化的语义,部分消除了正规模态逻辑给意图带来的不合理的逻辑性质.

Georgeff 和 Rao[14]分析了意图修正的语义,通过在原来的 BDI_{CTL}^* 逻辑中扩充了相信、愿望和意图的 only 模态(OBEL,ODES 和 OINTEND),给出了模型约束和公理,刻画了意图修正和信念修正、愿望修正的关系.这种分析为探索思维状态的动态演化指出了一个方向.

文献[15]分析了用模态逻辑框架研究意图的局限,提出了"意图的两维结构理论",基于关系理论从相关性和时序关系两维上对意图进行刻画,探讨了意图的产生和动态变化.提出了意图产生框架和基于 BDI 结构的 Agent框架.

Dongha 的承诺模型[4]基于线性时态逻辑的子集,为承诺提供了语义,提出用若干承诺原则来指导 Agent 对意图的取舍,也为 Agent 如何构造目标规划提供了模型.

2.2　理论与实践分离的问题

Wooldridge 指出,以模态逻辑加可能世界语义阐述的 Agent 理论只能进行抽象的描述,总的来说都无法以常规的方式来付诸实现,原因是:(1) 在所使用的逻辑描述和实际 DAI 系统结构之间缺乏清晰的关系,特别是,可能世界模型对于实现系统过于抽象.(2) 这些逻辑描述对 Agent 的推理能力都作了不现实的设定[16].

Rao 认为,Agent 理论与实践分离的原因是,对描述逻辑的定理证明和模型检查的复杂性问题还没有充分讨论,所以已实现的 BDI 系统都趋向于把 3 个主要的思维属性看做数据结构而不是模态算子.实际的系统往往作出过分简化的假设以致缺乏坚实的理论基础,描述逻辑很少对解决实际问题提供帮助.Rao 还认为,BDI 研究的一个目标就是使用某种合理且很有表现力的语言来阐述模型理论、证明理论和抽象解释器这三者间的一一对应关系[17].

研究 MAS 系统的方法有意识逻辑、时态逻辑和进程代数等,但没有一个是完全适用的,Wooldridge 提出了三阶段法:(1) 针对要研究的 MAS,建立模型;(2) 针对(1)的模型实现可运行模型,形式地描述此系统如何运行;(3) 将(2)中可运行模型的运行历史作为逻辑的语义基础,用来对(1)中的模型进行描述和推导[18].

基于实际系统(dMARS),Rao 发展了 AgentSpeak(L)语言,这个语言及其操作语义是开发实际系统所用语言的简化.AgentSpeak(L)基于一阶语言并带有事件和动作的处理.信念、愿望和意图没有表示为模态公式,而只是对应 AgentSpeak(L)语言中的实体,其意义则是由设计者赋予的.通过描述一套基本信念和规划来描述 Agent.Rao 给出了 AgentSpeak(L)的操作性语义和 BDI 解释器的算法,又进一步地给出了证明理论,它们是一些证明规则,阐述在不同的事件下 BDI 配置的转换,用来证明关于 Agent 的特性[17].

2.3　合作的思维状态

基于单 Agent 思维状态模型,Levesque 和 Cohen[19]给出了 Agent 组/队的联合意图的形式化定义,联合意图定义为特殊的联合持续目标,并导出了几个有实际意义的定理.Castelfranchi[20]认为此定义不是充分必要条件,因为:(1) 缺少相互依赖、合作动机和责任的描述,对有些竞争、交换现象,虽符合定义,但不是合作,因而没有联合意图;(2) 在某些特殊环境下,即使缺少通信和对共同行动方案的共享,也存在合作.

Castelfranchi[21]区分了 MAS 中存在的多种承诺形式,讨论了达成社会承诺的个体的权利和义务.在多 Agent之间存在着客观的社会依赖关系,因而产生了 Agent 之间的各种行为关系.相互依赖产生合作行为,而交互依赖则产生交换行为.

Grosz 和 Kraus[22]基于 Shareplans 模型和 Pollack 的单 Agent 规划,发展和完善了合作规划理论:(1) 使用个体意图去建立合作者对他们的联合活动的承诺;(2) 建立 Agent 对它的合作者有完成子行动的能力的承诺;(3) 在合作的背景下解释帮助行为;(4) 包括承包并将承包与合作进行区分;(5) 通信的需要是导出的,是由于对组活动的承诺;(6) 密切配合的子规划被保证.

Dunin-Keplicz 和 Verbrugge[23]基于 Rao 和 Georgeff[18]的 BDI 模型,提出了集体信念、意图和承诺的定义以及 3 种承诺机制.

Haddadi[3]基于 Rao 和 Georgeff 的 BDI 逻辑,扩充了联合承诺,在描述合作推理和协商方面作了新的探索.对于合作来说,联合目标是被委派或采纳的目标,承诺要求双方同意合作并共同相信它们都希望合作.手段-目的推理描述为信念、目标、意图和决定的状态转移,对应一组规则,说明在什么条件下承诺被形成、修正和维持.

还讨论了通信机制以及如何与这种推理融合的问题.

2.4 BDI结构

PRS是基于BDI模型的Agent结构[24](如图2所示),Rao和Georgeff[25]用BDI逻辑对PRS进行了逻辑描述,并给出了抽象BDI解释器.PRS只考虑了单Agent的情形.

GRATE*是基于BDI模型的结构[26](如图3所示),使用联合意图和联合责任的概念建立合作行为并监视联合行为的执行.GRATE*考虑了多Agent的情形.

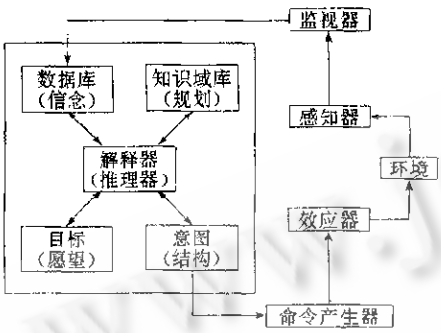

图2 PRS结构

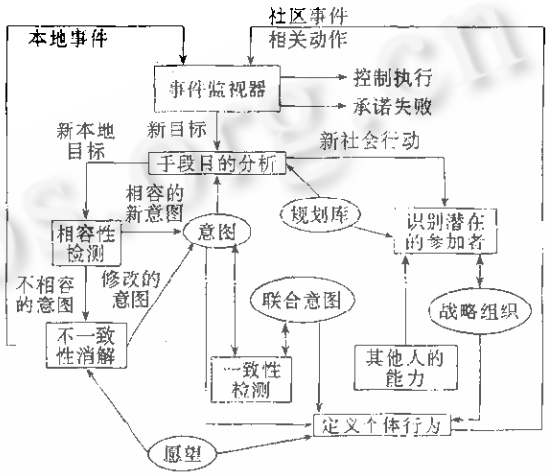

图3 GRATE*的功能结构

2.5 应 用

INTERRAP[27]从工程角度提出了设计Agent的混合式分层结构.Agent包括知识库、世界界面和3个等级的控制层,定义了思维状态的种类和修正这些思维状态的基本功能及感知和行动之间的功能关系.

Jennings[28]分析了复杂环境下Agent合作问题求解中由于缺少联合意图的支持而导致混乱的现象,基于联合意图理论提出了联合责任模型,指导GRATE*的设计.联合意图分为承诺的前提条件和承诺的跟踪协议,该协议定义了承诺的改变条件及相应的活动.

Tambe[29]基于联合意图理论,建立了队模型,由队状态和队算子组成,队状态是关于队构成情况的描述,队算子表示对联合行动的联合承诺,既表示了该队目前采取的联合活动,又指定了每一个成员所扮演的角色,描述了支持队工作的Agent结构STEAM.

Shoham[30]提出了AOP(agent-oriented programming)框架,采用思维状态模型从系统实现的角度来讨论Agent所应有的结构和行为特性.Agent的状态由思维属性组成,基本元素是信念B和承诺OBL,决策是对自己的承诺.对思维模型各元素关系的假定是内部一致、好意和内省的.在AOP中规定了信念的存在或不存在状态、承诺都自动持续.Shoham讨论了信念修正问题,并提出了通用的Agent解释器.

3 结束语

文献[7~9]产生的影响较大,后来的模型有的显式地考虑了时间,有的引入了其他思维属性.近来提出的解决理论和实践脱节问题的模型和研究方法也有新意,但尚需对其理论基础做深入探讨.

研究理性有两类方法.一类认为合乎逻辑的是理性的,为此提出了各种逻辑体系,定义了公理系统和推理规则,来证明一些特定的命题是否成立,认为合理的行为是基于当前信念合乎逻辑地推导出来的.目前对于思维状态模型的研究大都属于这类方法.另一类方法是采用决策理论,其信念模型是描述如果采用一个行动将会发生什么,为每个后果都赋予概率.愿望模型是用实数表示那些可能状态的效用.一个合理的行动是使期望效用最优化的行动,这需要依据信念和愿望使用概率论计算得到.从概念看,逻辑方法实现了理性的推理,决策理论方法

通过最优化主观效用而实现了理性的决策.从技术看,使用符号推理的逻辑理性无法使效用最优化,而使用数值分析方法的决策理论也忽视了推理环节.而对于处于动态环境中资源有限的 Agent 来说,既需要对世界进行推理,也需要作出合理的决策,使它从其行为结果中获益,显然需要融合这两种方法的研究.

思维状态研究的问题是:(1) 内部理论:研究思维状态的演化规律,强调动态性、操作性、可实现性以及演化规则;(2) 行为理论(环境理论):目的是解决问题、提高任务求解效率和可靠性等,行为结果是任务的解决,而行为产生是受思维状态控制的,必须研究思维状态如何产生一致、合理的行为,最终使任务得到解决.

多 Agent 规划与传统规划有着本质的差异.在 MAS 中,Agent 的个性会有差异,系统环境不断变迁,使得规划生成过程和执行过程交织在一起,可能在任何时刻的规划都只是部分规划,这时,思维状态对于引导和监督多 Agent 规划会起到关键性的作用.

意识立场和行为立场、逻辑方法和效用理论、内部观点和外部观点、理论和实践、可解释性和可计算性以及静态和动态,在以 Agent 理论为前沿的 DAI 研究领域中充斥了这类矛盾,研究方法论都应当把调和这些矛盾作为努力的方向.目前我们正在开展的工作有:

(1) 针对已有 BDI 模型中的概念不明确、逻辑全知、缺乏动态关系表达的缺陷,引入假设信念、解释愿望和意图在思维状态认知方面的含义,建立理性 Agent 的动态 BDI 模型.

(2) 提出 Agent 描述语言,并描述 Agent 群体的合作结构和合作思维状态.

(3) 在 BDI 模型中引入相关的社会性思维属性,结合社会规范和协调方法研究,建立面向社会性 Agent 的 BDI 模型.

(4) 提出基于效用分析的承诺机制.

(5) 研究市场机制下 Agent 思维状态的定义与模型.

致谢　本文的研究工作得到国家自然科学基金资助,此项目编号为 69773026 和 69733020.本文是 MAS 讨论班文献阅读的综述.参加讨论班的还有:康小强、李毅、曹子宁、王一川、陈威、路军、杨树林、孙芳、顾宇红、李建民等同志,应明生教授也参加了讨论并给予指导,在此一并表示感谢.

参考文献

1 Bradshaw M. An introduction to software agents. In: Bradshaw M ed. Software Agents. Menlo Park, California: AAAI Press, 1997. 3～46

2 Wooldridge M, Jennings N R. Intelligent agents: theory and practice. The Knowledge Engineering Review, 1995, 10(2):115～152

3 Haddadi A S. Communication and Cooperation in Agent Systems. Berlin: Springer-Verlag KG, 1996. 1～134

4 Dongha P. Toward a formal model of commitment for resource bounded agents. In: Wooldridge M J, Jennings N R eds. Intelligent Agents, Proceedings of the ECAI'94 Workshop on Agent Theories, Architectures, and Languages. Berlin: Springer-Verlag KG, 1995. 86～101

5 Shi Chun-yi *et al*. The advances of distributed artificial intelligence. Pattern Recognition and Artificial Intelligence, 1995,8(supplement):72～92

6 Tuomela R. Philosophy and distributed artificial intelligence: the case of joint intention. In: O'Hare G, Jennings N R eds. Foundations of Distributed Artificial Intelligence. New York: John Wiley & Sons, Inc., 1996. 487～504

7 Cohen P R, Levesque H J. Intention is choice with commitment. Artificial Intelligence, 1990,42(2～3):213～261

8 Rao A S, Georgeff M P. Modeling rational agents within a BDI-architecture. In: Allen J, Fikes R, Sandewall E eds. Principles of knowledge representation and reasoning: Proceedings of the 2nd International Conference (KR91). San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1991. 473～484

9 Konolige K, Pollack M E. A representationalist theory of intention. In: Bajcsy R ed. Proceedings of the 13th International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1993. 390～395

10 Bell J. Changing attitudes. In: Wooldridge M J, Jennings N R eds. Intelligent Agents, Proceedings of the ECAI'94 Workshop on Agent Theories, Architectures, and Languages. Berlin: Springer-Verlag KG, 1995. 40～55

11 Gaspar G, Coelho H. Where do intentions come from?: a framework for goals and intentions adoption, derivation and evolution. In: Ferreira C P, Mamede N J eds. Progress in Artificial Intelligence, Proceedings of the 7th Portuguese Conference on Artificial Intelligence, EPIA'95. Berlin: Springer-Verlag KG, 1995. 115～128

12 Linder B van, Hoek W van der, Meyer J. Formalising motivational attitudes of agents: on preferences, goals and commitments. In: Wooldridge M, Müller J P, Tambe M eds. Intelligent Agents II: Agent Theories,

Architectures, and Languages. Proceedings, 1995. Berlin: Springer-Verlag KG, 1996. 17～32
13 Cavedon L, Padgham L, Rao A *et al*. Revisiting rationality for agents with intentions. In: Xin Yao ed. Proceedings of the 8th Australian Joint Conference on Artificial Intelligence. Singapore: World Scientific Publishing Co. Pte. Ltd., 1995. 131～138
14 Georgeff M P, Rao A S. The semantics of intention maintenance for rational agents. In: Mellish C S ed. Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1995. 704～710
15 朱朝晖.意图的两维结构理论及非单调推理的研究[博士学位论文].南京航空航天大学,1998
(Zhu Zhao-hui. Two-dimensional structure intention theory and nonmonotonic reasoning [Ph.D Thesis]. Nanjing University of Aeronautics and Astronautics, 1998)
16 Wooldridge M. This is MYWORLD: the logic of an agent-oriented DAI testbed. In: Wooldridge M J, Jennings N R eds. Intelligent Agents. Proceedings of the ECAI'94 Workshop on Agent Theories, Architectures, and Languages. Berlin: Springer-Verlag KG, 1995. 160～178
17 Rao A S. AgentSpeak(L): BDI agents speak out in a logical computable language. In: Velde Walter Van de, Perram John W eds. Agents Breaking Away, Proceedings of the 7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World. MAAMAW'96. Berlin: Springer-Verlag KG, 1996. 42～55
18 Wooldridge M. Temporal belief logics for modeling distributed artificial intelligence systems. In: O'Hare G, Jennings N R eds. Foundations of Distributed Artificial Intelligence. New York: John Wiley & Sons, Inc., 1996. 267～286
19 Levesque H J, Cohen P R, Nunes J H T. On acting together. In: Proceedings of the 8th National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press/MIT Press, 1990. 94～99
20 Castelfranchi C, Conte R. Distributed artificial intelligence and social science: critical issues. In: O'Hare G, Jennings N R eds. Foundations of Distributed Artificial Intelligence. New York: John Wiley & Sons, Inc., 1996. 527～544
21 Castelfranchi C. Commitments: from individual intentions to groups and organizations. In: Victor Lesser ed. Proceedings of the 1st International Conference on Multi-Agent Systems. Menlo Park, California: AAAI Press/The MIT Press, 1995. 41～48
22 Grosz B F, Kraus S. Collaborative plans for complex group action. Artificial Intelligence, 1996,86(2):269～357
23 Dunin-Keplicz B, Verbrugge R. Collective commitments. In: Durfee F ed. Proceedings of the 2nd International Conference on Multi-Agent Systems. Menlo Park, California: AAAI Press, 1996. 56～63
24 Georgeff M P, Ingrand F F. Decision-making in an embedded reasoning system. In: Sridharan N S ed. Proceedings of the 11th International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1989. 972～978
25 Rao A S, Georgeff M P. An abstract architecture for rational agents. In: Nebel B, Rich C, Swartout W eds. Principles of Knowledge Representation and Reasoning: Proceedings of the 3rd International Conference (KR'92). San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1992. 439～449
26 Jennings N R. Specification and implementation of a belief-desire-joint-intention architecture for collaborative problem solving. International Journal of Intelligent and Cooperative Information Systems, 1993,2(3):289～318
27 Müller J P. The design of intelligent agents: a layered approach. Berlin: Springer-Verlag KG, 1996. 1～227
28 Jennings N R. Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. Artificial Intelligence, 1995,75(2):195～240
29 Tambe M. Agent architectures for flexible, practical teamwork. In: Proceedings of the 14th National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1997. 22～28
30 Shoham Y. An overview of agent-oriented programming. In: Bradshaw M ed. Software Agents. Menlo Park, California: AAAI Press, 1997. 271～289

About the Mental State Model for Agent

MA Guang-wei XU Jin-hui SHI Chun-yi

(*Department of Compuer Science and Technology Tsinghua University Beijing* 100084)

Abstract This paper summarizes the research of mental state model for agent, which is a major subject in the research on multi-agent system. Most references are published after 1990. From different perspectives, covering informal analysis, formal models, architectures, cooperation researches, and applications, many important results are introduced in this paper, such as Bratman's analysis, Cohen and Levesque's intention model, Rao and Georgeff's BDI model, etc. Finally, some prosperous works are introduced as well. We hope that this presentation can help readers deepen their understanding of agent.

Key words Agent, multi-agent system, mental state, belief, desire, goal, intention, commitment, cooperation, joint-intention.