心理学探新 2025,Vol.45,No.3,195 -202
PSYCHOLOGICAL EXPLORATION

面向自然人智交互的共享心智模型*

张 亮[1,2],王晓宇[1,2],刘 舫[3],马翠霞[4,5],王宏安[4,5]
(1.中国科学院心理研究所认知科学与心理健康全国重点实验室,北京 100101;2.中国科学院大学心理学系,北京 100049;
3.中国传媒大学媒体融合与传播国家重点实验室,北京 100024;4.中国科学院软件研究所,北京 100190;
5.中国科学院大学计算机科学与技术学院,北京 100049)

摘 要:自然而智能的人机交互是未来人机协同的发展趋势,这就要求智能体具备理解人类意图和实现类人协作的能力。然而,现有认知模型在应对动态性、复杂性和社会性交互需求方面存在局限,难以有效支持自然人-智能体交互系统的设计与评估。因此,基于经典认知模型框架,并结合心理理论等社会认知机制,共享心智模型在感知、认知、行为三大模块的基础上,创新性地引入共享空间模块,构建了包含共同经验、心理理论等模块的多维理论框架。通过共享外部空间的情境同步和共享内部空间的认知对齐,该模型能够动态描述人智交互过程中的协同机制,为设计具有社会智能的交互系统提供理论支撑,并为交互自然性的量化评估与优化提供新思路。

关键词:人智交互;自然交互;认知模型;心理理论
中图分类号:B842.5 **文献标志码**:A **文章编号**:1003-5184(2025)03-0195-08

1 引言

人智交互(Human - Agent Interaction,HAI)是人与智能体之间信息的双向交流与协作,是一个崭新的、对交互式智能系统及相关现象进行设计、评估与实现的研究领域。随着人工智能技术的飞速发展,智能体的形态不断扩展,从早期的规则驱动系统、命令行控制的机器人,到基于深度学习的自主决策系统,再到具备多模态交互能力的具身智能体(Embodied Agent),人智交互方式正朝着高度协同化、社会化的方向演进。交互技术已成为智能系统创新的核心驱动力,而自然的人机协作则是人智交互的重要目标。

在人机交互领域,“自然”指交互是“自发的”,“直觉性的”,而且是自然而然、无需过多思考而产生的(Grandhi et al.,2011)。早在上个世纪,Steve Mann 就提出了“自然用户界面”(Natural User Interface)的概念,最初聚焦于无隐喻的物理交互,尤其是在可穿戴设备中的实现(Mann,1998),而如今的人智交互已超越传统界面范畴,演化为多模态、多通道的协作模式。智能体不仅能通过语音、手势、触觉与人类交互,还可借助脑机接口、情感计算等技术感知用户的认知状态与情绪(Besginow et al.,2022)。目前对“自然人机交互”尚且缺乏单一的、一致定义,但普遍认为自然交互需要让用户像与真实世界中的物体交互一样与计算机进行交互(Wigdor & Wixon,2011)。

当前,人智交互领域涌现出大量创新技术(如生成式对话、协作机器人),进一步拓展了人智交互的应用边界,也促使对“自然”交互的要求从传统人机交互所强调的“自发性”与“直觉性”,延伸至用户与智能体的交互需同真实世界中的人际协作类似。人类与智能体在协作过程中需要形成的类社会性默契,例如意图对齐、动态环境适应以及共同目标理解(Marathe et al.,2018;Yuan et al.,2022)。与传统人机交互中精确、离散的指令输入不同,自然的人智交互要求智能体能够理解模糊的上下文、预测人类意图,并动态调整行为策略。例如,在协作装配场景中,机器人需根据工人的手势、语言指令甚至眼神注视点,实时推断任务目标并调整动作规划(Schirmer et al.,2024);在交流对话中,智能体根据用户的行为调整自身的行为特点(Biancardi et al.,2021)。这种交互模式的关键在于双向心理模型的建立:智能体需具备类人的社会认知能力(如心理理论、共情机制),而人类也需逐步适应智能体的行为模式与决策逻辑。

人智交互的设计规范与理论框架仍处于探索阶

* **基金项目**:国家自然科学基金重大项目(T2192932),国家自然科学基金(62272447),国家重点研发计划项目(2016YFB1001201)。
通信作者:张亮,E-mail:zhangl@psych.ac.cn。

段。学界从不同的角度阐述“人智交互”需要满足的设计准则,以人为中心是普遍认同的核心特征(Jiang et al. ,2024;许为 等,2024)。具体而言,人智交互需要以人的体验为基础,以智能体的自主性与适应性为支撑,通过降低认知负荷、增强协作流畅度,实现人智共生的目标。这一准则很大程度继承了人机交互长久以来的研究基础。然而,随着智能体在复杂任务中展现出高度的主动性与自适应能力,传统人机交互理论在解释人智交互显示出局限性。尽管传统人机交互研究已提出以人类处理器模型为代表的多种认知模型(Card,1983),但这些模型多基于被动响应式交互设计,难以描述人智交互中的动态协作、意图协商等复杂过程。因此,亟需构建新的理论框架,以指导具备社会智能的智能体设计与评价。

在人类处理器等传统认知模型的基础上,本文进一步强调了人智协作过程中意图对齐、以及对环境和彼此理解的重要性,提出了一种面向自然人智交互的共享心智模型。该模型旨在为智能体赋予类社会协作能力,使其能够对齐人类的认知状态与行为目标,从而促进自然人智交互的发展。

2 传统认知模型

2.1 人类处理器模型及其发展

传统的认知模型可以用于评估交互方式操作时间,使用模型分析交互过程,以及使用认知模型替代用户。1983 年,Card 提出了人类处理器模型(Model Human Processor,MHP)简洁高效地描述和预测人机交互过程(Card,1983)。MHP 模型将人类的心理加工过程概括为感知、认知、动作三个处理器。

在 MHP 模型的基础上,人们更加精细地提出多种模式模拟人的行为,例如 GOMS 模型、ACT - R(Adaptive Control of Thought - Rational)模型。GOMS 模型同样由 Card 等人于 1983 年提出,用于模拟用户与系统交互时使用的知识和认知过程(Card,1983)。GOMS 模型由 Goals(用户期望达到的目标),Operators(用户执行的基本操作),Methods(实现目标或子目标的一系列操作),Selection Rules(实现目标有多个方法时,选择方法的规则)四部分组成。GOMS 模型是人机交互和界面设计领域中最常用的信息加工模型,尤其在可用性测试领域被广泛应用。ACT - R 是由 Anderson 提出的一种模拟人类任务执行中的思维和行为的认知模型(Anderson et al. ,2004)。该模型包括视觉、运动、陈述性记忆和程序性记忆等模块,并可以通过产生式系统应用产生式规则,协调模块之间的信息交流与任务执行。面向不同的任务或设备,研究者不断改进经典认知模型。例如,为解决多任务情景下的行为仿真,有研究者结合排队网络模型(Queueing network)和 MHP 模型提出了 QN - MHP 模型(Liu et al. ,2006)。

随着人工智能和人机交互的发展,认知模型也需要根据技术的发展不断演化。例如,国内心理学家和计算机科学家提出了面向智能时代的人机合作心理模型(刘烨 等,2018),这一模型认为人与计算机的信息处理系统均包括感知、认知和动作三个处理器,可以实现多模态数据的获取、加工和存储。该模型提出人与计算机的交互本质上是人与人的交互,具有和人与人交互相似的属性和规律。

2.2 传统认知模型在自然人智交互的应用及局限

在自然人智交互中,尽管交互技术和应用场景与传统认知模型提出时已有较大变化,但人的认知与行为规律相对稳定,经典的认知模型在人智交互中仍然可以发挥重要作用。

认知模型在人机交互中的重要应用是模拟人类操作行为和预测人机交互操作时间。Card 等人为感知、认知、动作这三个处理器设定了时间范围,使用这一模型可以粗略估计操作任务的执行时间(Card,1983)。QN - MHP 计算模型可以仿真人的多任务操作,例如,该模型可以较好预测司机进行车辆驾驶任务及阅读地图的行为模型和操作时间(Wu & Liu,2007)。认知模型也可以模拟用户的感知、注意、记忆等认知过程,预测用户在不同条件下的认知负荷,进而对交互系统进行优化。

然而,使用传统认知模型对自然人智交互进行指导时存在明显不足。首先,传统认知模型以模拟人类认知过程为目标,多从个体的角度对感知、认知及动作三个模块进行建模,忽视了人智交互的互动特性。人与机之间的融合在不断增加,人需要理解机器如何看待世界、做出决策,并与机器默契配合(刘伟,2023)。因此,人与计算机的交互过程,不仅需要从个体的角度分析感知、加工与动作,更需要把两者看作一个整体(张警吁,张亮,2018)。面向智能时代的人机合作心理模型在这一方面进行了初步拓展,提出人机交互的本质是人与人的交互,对智能时代更为复杂的人机交互过程进行解释,强调了交互时理解和预测交互对象心理状态的重要性(刘烨等,2018)。但是,已有模型通常建立在静态结构

上，难以有效解释人智交互中随着时间发展产生的变化。在持续交互中，人与智能体的了解程度、信任程度，甚至行为策略会不断变化，这类动态变化尚未被纳入现有模型的框架中。此外，人机交互不再局限于以电脑桌面为中心的人机交互范式，环境复杂性与多样性显著增强，不同的环境对感知输入、信息加工、行为输出会产生重要影响。

因此，为更有效地描述与理解自然人智交互，新的认知模型需要将人机交互视为动态耦合的系统，并在兼顾个体内部的信息加工机制的同时，强调人与智能体之间的信息共享与意图理解。

3　自然人智交互下的共享心智模型

基于人类处理器模型、面向智能时代的人机合作心理模型等经典认知模型，并结合自然人智交互的协作目标，本文提出了共享心智模型的整体结构，如图1所示。

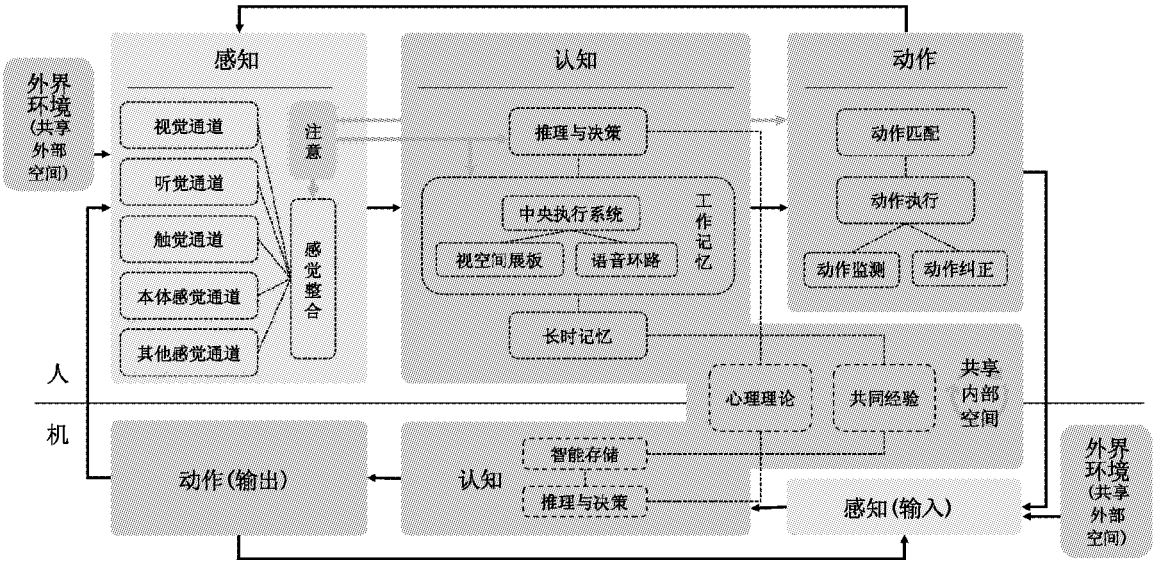

图1　共享心智模型示意图

该模型的基本假设如下：

假设一：自然的人智交互需模拟人类与智能体之间的类社会性协作。智能体不仅是工具，而是具备社会认知能力（如心理理论、意图推理）的协作伙伴。交互双方（人类与智能体）需通过双向理解建立类人际信任。

假设二：交互系统中的个体（人类与智能体）均具备独立的个体加工空间，涵盖多模态感知模块、认知模块以及动作模块。

假设三：交互系统的用户与其他部分存在于共享外部空间中，且有共享内部空间。共享外部空间指个体均在同一情境中进行信息感知、认知处理及动作选择与执行。共享内部空间指因为进化或之前的经验在个体交互空间中形成的重叠部分，存在共同经验模块以及心理理论模块。

假设四：个体加工空间与共享空间具有持续协同演化的能力。共享空间随着人智交互过程可不断扩充共同经验系统、增强心理理论模块。同时，共享空间的优化也使个体加工空间的内部过程得以优化。

假设五：人智交互的自然性主要由共享空间的完备性决定，具体体现为内部共享空间规模，如协作经验库的覆盖广度、心理理论网络的理解深度；外部共享空间一致性，如双方对情境信息的感知对齐度；以及交互双方的处理效能，如智能体的实时推理速度、人类对智能体行为的可解释性等。

3.1　个体加工空间

每一名用户或每一完整的智能体都可以看作一个个体。每个个体的信息接受、处理与动作执行都在各自的个体加工空间中进行。个体加工空间分为多模态感知模块、认知模块和动作模块。

3.1.1　多模态感知模块

多模态感知模块将外部原始信息（如物理信号、其他个体的行为、环境状态）转为认知模块可加工处理的内部表征。该模块通过异构传感器与算法协同，实现多通道信息的同步采集与跨模态信息的融合，最终得出一致的、完整的表征。

交互系统中的个体具有多种感知觉通道来获取信息。例如，人有视觉、听觉、嗅觉、触觉和味觉这五大基本感觉，同时还有其他感知觉，如本体感觉、温

度觉、时间知觉等(Spence,2018)。智能体可以获取语音、文字、图像、物理传感器信息、生物信号等多通道的信息。智能体的感知通道需要不断丰富,以匹配人的多种多样的感知通道。

人的大脑持续接收来自各个通道的信息,并将各个信息加以整合(Lalanne & Lorenceau,2004)。不同感觉通道之间相互有影响。认知心理学实验发现视觉通道可以影响听觉通道的信息处理,例如McGurk 效应(McGurk & MacDonald,1976)。智能体在进行信息采集时,也应该共同加工来自各通道的信息,即跨模态信息的对齐与整合(Martin et al.,2021)。同时,各个通道的信息需要一致,否则将会造成接收信息个体的误解。

多模态感知模块含有注意模块对信息进行选择。注意是心理活动对一定事物的指向与集中,选择注意把与交互系统相关的信息从大量的无关信息中过滤出来(Sanders & Astheimer,2008)。在注意指导下对多模态信息进行的选择包括有自下而上的突显性信息选择(例如,闪烁的光和响度大的警报声会捕获注意)和自上而下的目标信息选择(例如,地铁站的箭头标识引导人们选择正确的路线)。当视觉、听觉、触觉等多类信息同时出现时,注意能在多模态信息中起到调控作用,实现有效的感觉整合。

3.1.2 认知模块

认知模块对来自感知模块的信息进一步加工,转化为可执行的操作,同时受到注意的调控。认知模块包含工作记忆模块、长时记忆模块、推理与决策模块、元认知模块。不同个体在各个模块中的加工特点有差异。

工作记忆模块是一个容量有限的信息加工系统,用来暂时保持和存储信息,是连接感知觉、长时记忆和动作输出的加工平台(Baddeley,2010)。根据 Baddeley 提出的工作记忆模型,工作记忆分为中央执行系统、视觉空间展板、语音环路三个成分。中央执行系统负责在不同任务之间进行转换、注意转移、抑制干扰及保持内容更新等核心控制功能,视觉空间展板主要负责存储视觉和空间信息,语音环路则专门负责存储有关语音的信息。人的工作记忆容量是有限且相对较小的,智能体可以同时处理的信息容量相对较大。感知模块中的注意与工作记忆模块交互密切,注意影响信息在工作记忆中的保持与更新(Oberauer,2019)。

长时记忆模块中存储个体以往学习和经历的所有知识和经验。长时记忆是个体“心理上的过去”,是个体经验积累和认知能力发展的前提。智能体的长时记忆可以显著提升与人类交互时的表现(He et al.,2025)。长时记忆可分为陈述性记忆和非陈述性记忆(Squire & Dede,2015)。陈述性记忆指个体对事实或个人经历的记忆,可以被有意识地提取与表达。非陈述性记忆指个体无法通过有意识的过程而接触的知识,例如程序性技能或启动效应。智能体的长时记忆系统相对人类的长时记忆系统更稳定,人类的长时记忆在提取过程中可能受到干扰或修改(Sinclair & Barense,2019)。

推理指利用已知信息得到结论的过程。现实生活中的个体掌握的信息是有限的,认知能力也是有限,个体通过启发式做决策或判断,如代表性启发式(个体根据事件与过去经验的相似程度来进行判断和预测)和可得性启发式(个体根据容易想象或回忆的事件或事物做决策)(Artinger et al.,2015)。面对同样的信息,不同个体会做出不同的推理。推理与决策在各类经典认知模型中均有体现,例如采用规则系统或产生式系统模拟逻辑推理过程(Baddeley,2010)。这一模块也会受到注意调控,如注意会影响对选项的偏好,进而影响决策结果(Bhatnagar & Orquin,2022)。

元认知指人类对其自身认知活动的认知。元认知包括元认知知识、元认知体验和元认知策略三个成分(Norman et al.,2019)。元认知知识指个体在认知实践活动中积累起来的关于认知活动的一般性知识,元认知体验是个体在认知活动中所产生的认知体验和情绪体验,元认知策略是指个体在调控认知过程中有意采取的行动。元认知使个体面对复杂任务时更灵活,以更好地实现目标。模拟人类元认知能力也是提高智能体能力的重要环节(Conway - Smith & West,2024)。

3.1.3 动作模块

动作模块包括动作匹配、动作执行、动作监测与动作纠正。动作匹配指个体需要通过认知模块的信息加工结果匹配具体动作(Pezzulo & Cisek,2016)。动作具有针对特定目标的层级式表征形式,即最高层为通过认知模块得到的抽象概念,中间层为具体的运动规划,底层则对应着实际的神经肌肉活动等。动作匹配子模块形成的动作序列发送指令至效应器中,动作执行子模块根据指令完成相应动作。个体在交互中可以通过多种方式完成动作执行,但其带

来的交互体验会存在差异。动作监测子模块对动作的执行进行监测。动作监测包括内部控制及外部控制,内部控制通过比较实际的动作和计划的动作进行实时监督,外部控制会借助反馈信息对动作进行调整。动作纠正根据动作监测的结果实时地纠正动作,保证动作目的的实现。动作模块的运动也会收到注意的调控,如注意分散会造成动作精度下降、注意焦点的不同会影响实际的动作表现(Song,2019)。

个体发出的动作是多模态的。在交互过程中,人可以通过传统的鼠标、键盘发出动作,也可以通过笔、语音、眼动、肌电、脑电信号等形式发出动作,例如基于眼动的网页浏览工具(Menges et al.,2017)、通过生理信号告知机器人自身的状态(Rückert et al.,2023)等。智能体也可以通过图像、声音、机械运动等多种形式发出动作。智能体还需要不断拓宽输出通道。目前的交互技术以视觉和听觉为主,辅以少量的触觉,这远远少于人体感知世界的方式。

3.2　共享空间

人智交互中的人与智能体并不是完全独立的个体,而需要双向理解与默契配合来达成特定目标。共享空间描述了独立的个体加工空间之外的空间。共享空间分为共享外部空间和共享内部空间。

3.2.1　共享外部空间

在人智交互系统中,交互不仅存在于人类与智能体之间,还嵌入在一个动态变化的物理与社会环境中,这一环境则为共享外部空间。此空间特征会直接影响到个体加工空间中各个模块的运行效率与方式。例如,在有害环境下,人的内部激素水平会发生变化,认知功能和动作操作等均会发生变化(Metz et al.,2005;Schwabe et al.,2022)。

情境感知是针对共享外部空间的核心任务。智能体需要实时感知环境状态,来实现更佳的交互。人与人进行交互时,外界环境会改变交互的形式。例如,在安静场景,人与人的交流更倾向于文字和手势;在嘈杂的户外场景,人们倾向于言语沟通。人智交互中,个体情境感知结果的一致性影响交互的自然性。例如,人类与机器人完成装配等协作任务,个体的动作规划必须根据当前共享的环境状态进行动态调整。自然人智交互中的适应性行为均建立在系统内个体对情境的共享理解之上。

3.2.2　共享内部空间

共享内部空间指人智交互系统中个体之间感知模块、认知模块与动作模块的重叠部分。共享内部空间的规模决定着人智交互过程中信息传递的效率与意图理解的准确性,进而影响实际的交互体验。共享内部空间主要由共同经验模块和心理理论模块两个关键部分组成。

(1)共同经验模块

共同经验模块是交互系统中交互行为持续性进行的基础,其来自于个体之间的信息共享历史与各类知识的交集,可进一步分为有效信号模块和共享长时记忆模块。

有效信号模块的目标是保障交互过程中各类信息的可达性与可解释性。个体动作执行模块发出的动作信息及个体状态信息需要被其他个体的多模态感知模块接收并成功解析。个体的真实状态或意图往往是通过多种途径进行表现的。例如,个体可以通过对方的面部表情识别相应情绪,但面部表情相同而身体姿势不同时,表达的情绪有差异(Aviezer et al.,2008)。多模态信息的协同处理能力是构建有效信息模块的基础。

共享长时记忆模块具体指交互系统中各个个体均具有的长时记忆内容。这些共享内容可以显著降低沟通成本,提升交互效率。例如,两个具有相似专业背景的个体在技术交流时无需解释基础概念便可系统解决问题。类似地,当智能体与用户有相同的知识图谱或历史对话记忆,共享的长时记忆会提高智能体的上下文一致性,优化人智交互系统的运行效率(Bartoli et al.,2022)。

(2)心理理论模块

心理理论指理解其他个体的愿望、信念、动机等心理状态的能力,是个体解释和预测自己和他人行为的基础(Gallagher & Frith,2003)。理解对方的意图是人智交互的重要目标。目前的大语言模型可以在对话中完成对用户意图的建模,但意图识别层次较浅,仍然理解长指令、模糊指令等(Chang et al.,2025)。共同注意机制是人类的重要特征之一,当别人转移视线时,人类能领会到其他个体在关注某些事情且将注意力转移到同样的目标上。拥有共同注意模块的学习机器人对儿童的学习有促进作用。高级心理理论指在交互过程中习得的关于个体与个体的复杂知识库,这些知识库将帮助个体对其他个体的行为和状态进行理解。

心理理论模块不仅负责建模对方状态,还会在其指导下对个体加工空间各个模块的工作进行指

导,以实现交互视角、语义等方面的对齐。例如,在人机协作搭建任务中,拥有心理理论模块的机器人将按照用户的视角描述物体位置及操作方向,而不是按照机器人自身的视角。此外,心理理论有助于提升人智系统交互的容错性。个体在进行感知、认知以及动作匹配与执行时的差错不可避免。自然人智交互中的差错是系统中各个体共同的经历,心理理论模块可以支持其他个体对差错进行行为解释与修正。例如,当用户一个操作被智能体误解时,用户会立即明白智能体内部的知识架构或觉察到用户自身违背共享空间原则,立即改变动作、执行新的操作。自然的智能体也可以根据上下文判断用户可能为"误操作",从而避免直接执行错误动作。

共享内部空间描述了个体与机器的联系,共享空间在这种联系中不断更新。一方面,共享经验模块随着交互过程不断扩展。人智交互的个体会不断记录对方的语言模型、动作偏好等,个体共同的交互经验将整合进共享长时记忆模块,逐渐形成关于其他个体特征的复杂知识库。另一方面,心理理论模块会在互动过程中持续细化。随着共享经验模型的发展,个体不断更新心理状态的预测模型,心理理论可以更好地进行意图理解与调整自身行为实现协同。个体的心理理论能力是可以动态发展的,人类的心理理论会随着年龄不断发展(Meinhardt - Injac et al. ,2020),智能体的心理理论也可以通过技术发展不断优化(Rabinowitz et al. ,2018)。

4 结论与展望

本文结合人类处理器模型等经典认知模型,基于自然人智交互的特点,提出了共享心智模型。此模型继承与发展了传统的感知、认知、动作三个模块,创造性地提出了共享空间模块,用以描述人类与智能体在交互时存在的重叠,并且随着交互的进行与技术的发展,共享空间可以不断扩展。共享心智模型强调了自然人智交互不仅要求个体具备独立的感知、认知与动作,还需要构建广泛的共同经验与完善的心理理论。

共享心智模型为人智交互提供了一个理论模型,可以使系统研究人员与设计人员从感知、认知、动作以及共享空间的优化等多维度对人智交互的自然性等进行提升。该模型也为评价自然人智交互提供了理论基础,个体加工空间包含的能力与共享空间完备性是评价交互体验的重要维度。同时,该模型可以为包括多用户、多智能体在内的复杂人智交互提供评价思路和优化框架。

共享心智模型试图为自然人智交互提供新的理论指导与发展建议,同时也存在一定局限,未来还需进一步完善与扩展。第一,共享心智模型仅是一个理论模型,从理论模型发展为可计算模型,仍需要大量的数据支撑。此模型还需要结合具体的交互形态和设备,使用真实的用户数据进行验证与修正。第二,该理论模型提供了一个自然人智交互基本框架,但对于不同类型的智能体(如,大语言模型或具身智能机器人),其模块功能与实现机制可能存在差异,模型在实际应用中还需要根据交互目标与智能体特征进行一定调整。第三,共享心智模型赋予智能体社会协作能力,但与面向智能时代的人机合作心理模型不同的是,该模型并未将人智交互与人人交互完全对等起来。至少在目前阶段,人类与智能体交互的心理机制与人人交互存在差异,人类对智能体的交互方式也难以与其对其他人类的交互方式相同(Alarcon et al. ,2023)。人智交互可能沿用部分人人交互的社交模式,但仍旧需要进行修正,发展出人智交互特有的理论模型(Krämer et al. ,2012)。此外,自然交互是人与智能体交互需要实现的重要目标,但正如文初所提,对于"自然性"的定义与评价体系还未形成统一共识。实现理论上的统一与认知模型的落地还需要心理学、人工智能、人机交互等领域的共同努力。

张亮与王晓宇为共同第一作者。

参考文献

刘伟.(2023).人机混合智能:新一代智能系统的发展趋势.*上海师范大学学报*(哲学社会科学版),*52*(1),71 - 80.

刘烨,汪亚珉,卞玉龙,任磊,褚宇明.(2018).面向智能时代的人机合作心理模型.*中国科学:信息科学*,*48*(4),376 - 389.

许为,高在峰,葛列众.(2024).智能时代人因科学研究的新范式取向及重点.*心理学报*,*56*(3),363.

张警吁,张亮.(2018).自然交互的认知机理与心理模型.*中国计算机学会通讯*,*18*(5),30 - 35.

Alarcon,G. M. ,Capiola,A. ,Hamdan,I. A. ,Lee,M. A. ,& Jessup,S. A. (2023). Differential biases in human - human versus human - robot interactions. *Applied Ergonomics*, *106*, 103858.

Anderson,J. R. , Bothell,D. , Byrne,M. D. , Douglass,S. , Lebiere,C. , & Qin,Y. (2004). An Integrated Theory of the Mind. *Psychological Review*,*111*(4),1036 - 1060.

Artinger, F., Petersen, M., Gigerenzer, G., & Weibler, J. (2015). Heuristics as adaptive decision strategies in management. *Journal of Organizational Behavior*, *36*(S1), S33 – S52.

Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., & Bentin, S. (2008). Angry, Disgusted, or Afraid? Studies on the Malleability of Emotion Perception. *Psychological Science*, *19*(7), 724 – 732.

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), R136 – R140.

Bartoli, E., Argenziano, F., Suriani, V., & Nardi, D. (2022, November). Knowledge acquisition and completion for long – term human – robot interactions using knowledge graph embedding. *In International Conference of the Italian Association for Artificial Intelligence* (pp. 241 – 253). Cham: Springer International Publishing.

Besginow, A., Büttner, S., Ukita, N., & Röcker, C. (2022). Deep learning – based action detection for continuous quality control in interactive assistance systems. In *Human – Technology Interaction: Shaping the Future of Industrial User Interfaces* (pp. 127 – 149). Cham: Springer International Publishing.

Bhatnagar, R., & Orquin, J. L. (2022). A meta – analysis on the effect of visual attention on choice. *Journal of Experimental Psychology: General*, *151*(10), 2265.

Biancardi, B., Dermouche, S., & Pelachaud, C. (2021). Adaptation mechanisms in human – agent interaction: Effects on user's impressions and engagement. *Frontiers in Computer Science*, *3*, 696682.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human – Computer Interaction*. Hillsdale: Lawrence Erlbaum Associates.

Chang, Z., Lu, F., Zhu, Z., Li, Q., Ji, C., Chen, Z., Liu, Y., Xu, R., Song, Y., Wang, S., & Li, J. (2025). *Bridging the Gap Between LLMs and Human Intentions: Progresses and Challenges in Instruction Understanding, Intention Reasoning, and Reliable Generation*. No. arXiv: 2502.09101.

Conway – Smith, B., & West, R. L. (2024). Toward Autonomy: Metacognitive Learning for Enhanced AI Performance. *In Proceedings of the AAAI Symposium Series* (Vol. 3, No. 1, pp. 545 – 546). Springer.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind.' *Trends in Cognitive Sciences*, *7*(2), 77 – 83.

Grandhi, S. A., Joue, G., & Mittelberg, I. (2011). Understanding naturalness and intuitiveness in gesture production: Insights for touchless gestural interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 821 – 824.

He, Z., Lin, W., Zheng, H., Zhang, F., Jones, M. W., Aitchison, L., Xu, X., Liu, M., Kristensson, P. O., & Shen, J. (2025). *Human – inspired Perspectives: A Survey on AI Long – term Memory*. No. arXiv: 2411.00489.

Jiang, T., Sun, Z., Fu, S., & Lv, Y. (2024). Human – AI interaction research agenda: A user – centered perspective. *Data and Information Management*, *8*(4), 100078.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17 – 94.

Krämer, N. C., von derPütten, A., & Eimler, S. (2012). Human – Agent and Human – Robot Interaction Theory: Similarities to and Differences from Human – Human Interaction. In M. Zacarias & J. V. de Oliveira (Eds.), *Human – Computer Interaction: The Agency Perspective* (pp. 215 – 240). Springer.

Lalanne, C., & Lorenceau, J. (2004). Crossmodal integration for perception and action. *Journal of Physiology – Paris*, *98*(1), 265 – 279.

Liu, Y., Feyen, R., & Tsimhoni, O. (2006). Queueing Network – Model Human Processor (QN – MHP): A computational architecture for multitask performance in human – machine systems. *ACM Transactions on Computer Human Interaction*, *13*(1), 37 – 70.

Mann, S. (1998). Humanistic computing: "WearComp" as a new framework and application for intelligent signal processing. *Proceedings of the IEEE*, *86*(11), 2123 – 2151.

Marathe, A. R., Schaefer, K. E., Evans, A. W., & Metcalfe, J. S. (2018). Bidirectional Communication for Effective Human – Agent Teaming. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, Augmented and Mixed Reality: Interaction, Navigation, Visualization, Embodiment, and Simulation* (pp. 338 – 350). Springer International Publishing.

Martin, D., Malpica, S., Gutierrez, D., Masia, B., & Serrano, A. (2022). Multimodality in VR: A survey. *ACM Computing Surveys (CSUR)*, *54*(10s), 1 – 36.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746 – 748.

Meinhardt – Injac, B., Daum, M. M., & Meinhardt, G. (2020). Theory of mind development from adolescence to adulthood: Testing the two – component model. *British Journal of Developmental Psychology*, *38*(2), 289 – 303.

Menges, R., Kumar, C., Müller, D., & Sengupta, K. (2017). GazeTheWeb: A Gaze – Controlled Web Browser. *Proceedings of the 14th International Web for All Conference*, 1 – 2.

Metz, G. A., Jadavji, N. M., & Smith, L. K. (2005). Modulation of motor function by stress: A novel concept of the effects of stress and corticosterone on behavior. *European Journal of Neuroscience*, *22*(5), 1190 – 1200.

Norman, E., Pfuhl, G., Sæle, R. G., Svartdal, F., Låg, T., & Dahl, T. I. (2019). Metacognition in Psychology. *Review of General Psychology*, *23*(4), 403 – 424.

Oberauer, K. (2019). Working memory and attention – A conceptual analysis and review. *Journal of Cognition*, *2*(1), 36.

Pezzulo, G., & Cisek, P. (2016). Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition. *Trends in Cognitive Sciences*, *20*(6), 414 – 424.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine Theory of Mind. *Proceedings of the 35th International Conference on Machine Learning*, 4218 – 4227.

Rückert, P., Wallmeier, H., & Tracht, K. (2023). Biofeedback for human – robot interaction in the context of collaborative assembly. *Procedia CIRP*, *118*, 952 – 957.

Sanders, L. D., & Astheimer, L. B. (2008). Temporally selective attention modulates early perceptual processing: Event – related potential evidence. *Perception & Psychophysics*, *70*(4), 732 – 742.

Schirmer, F., Kranz, P., Bhat, B., Rose, C. G., Schmitt, J., & Kaupp, T. (2024). Towards a Path Planning and Communication Framework for Seamless Human – Robot Assembly. *Companion of the 2024 ACM/IEEE International Conference on Human – Robot Interaction*, 960 – 964.

Schwabe, L., Hermans, E. J., Joëls, M., & Roozendaal, B. (2022). Mechanisms of memory under stress. *Neuron*, *110*(9), 1450 – 1467.

Sinclair, A. H., & Barense, M. D. (2019). Prediction Error and Memory Reactivation: How Incomplete Reminders Drive Reconsolidation. *Trends in Neurosciences*, *42*(10), 727 – 739.

Song, J. H. (2019). The role of attention in motor control and learning. *Current opinion in psychology*, *29*, 261 – 265.

Spence, C. (2018). Multisensory Perception. *In Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1 – 56). John Wiley & Sons, Ltd.

Squire, L. R., & Dede, A. J. O. (2015). Conscious and unconscious memory systems. *Cold Spring Harbor Perspectives in Biology*, *7*(3), a021667.

Wigdor, D., & Wixon, D. (2011). *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Elsevier.

Wu, C., & Liu, Y. (2007). Queuing Network Modeling of Driver Workload and Performance. *IEEE Transactions on Intelligent Transportation Systems*, *8*(3), 528 – 537.

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., ... Zhu, S. C. (2022). In situ bidirectional human – robot value alignment. *Science Robotics*, *7*(68), eabm4183.

A Shared Mental Model for the Natural Human – Agent Interaction

Zhang Liang[1,2], Wang Xiaoyu[1,2], Liu Fang[3], Ma Cuixia[4,5], Wang Hongan[4,5]

(1. State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101;
2. Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049;
3. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024;
4. Institute of Software, Chinese Academy of Sciences, Beijing 100190;
5. School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049)

Abstract: Natural and intelligent human – agent interaction represents the future direction of human – agent collaboration, requiring intelligent agents to comprehend human intentions and achieve human – like cooperation. However, existing cognitive models exhibit limitations in addressing dynamic, complex, and socially interactive demands, thereby inadequately supporting the design and evaluation of natural Human – Agent Interaction (HAI). To address this, the Shared Mental Model is proposed based on classical cognitive architectures, integrating social cognitive mechanisms such as the theory of mind. Building upon the traditional perception – cognition – action modules, the model innovatively introduces a Shared Space Module, establishing a multidimensional theoretical framework that incorporates shared experience, and theory of mind. Through situational synchronization in shared external spaces and cognitive alignment in shared internal spaces, this model dynamically describes collaborative mechanisms during interaction processes. It provides a theoretical foundation for designing socially intelligent interaction systems and offers new insights for the quantitative assessment and optimization of interaction naturalness.

Key words: Human – Agent Interaction; Natural Human – Computer Interaction; Cognitive model; Theory of mind