

从人类共情走向智能体共情*

颜志强 苏金龙 苏彦捷**

(北京大学心理与认知科学学院和行为与心理健康北京市重点实验室, 北京, 100871)

摘要 随着机器人在人类生活中重要性的增强, 人们开始关注其传统智能以外的其它能力, 如共情。基于共情的经典理论, 本文以新兴的智能体共情 (artificial empathy) 为主题, 同时结合人工智能领域的最新研究, 系统地阐述了现有与智能体共情有关的理论模型和实证研究, 并结合文献计量学分析, 对该领域研究宏观现状进行可视化呈现。指出在未来研究中, 借鉴心理学、认知神经科学和计算机科学等学科的研究成果, 注重机器人理论模型的建构和研究方法的创新, 对机器人共情系统的建构有着重要作用, 而其间涉及到的伦理问题同样不容忽视。

关键词 共情 智能体共情 机器人 人机交互 情绪理解

科技的进步使得机器人的智能得到了飞速发展 (蔡曙山, 薛小迪, 2016), 但是机器人现有的“能力”显然还不足以让其成为人类理想的生活“伴侣”。其中重要的一点就在于与人类相比, 它们缺少了“人情味”, 难以对人类的情感需求进行恰当的理解并做出合适地反应 (Valuch, Pflüger, Wallner, Laeng, & Ansorge, 2015)。

针对人类情感, 心理学已进行了诸多研究。但如何把这种情感与无生命的机器联系在一起, 却是近 20 年研究者才开始关注的主题, 并催生出了“情感计算” (陶建华, 谭铁牛, 2004)。情感计算通过计算建模的方法对人类语音、图像进行识别, 进而为和谐人机关系的营造提供基础。作为人际交往的粘合剂, 共情深受研究者们的关注 (刘聪慧, 王永梅, 俞国良, 王拥军, 2009; 颜志强, 苏金龙, 苏彦捷, 2017; 颜志强, 苏彦捷, 2017; Shamay-Tsoory, Aharon-Peretz, & Perry, 2009)。然而, 目前共情的研究更多的集中于人际互动场景, 而较少考察人机互动过程中涉及到的共情表达, 而后者无疑对具有更大社会价值的情感型机器人的开发具有重要意义。基于上述背景, 本文拟对与共情这一重要代表性情绪及其

理解能力有关的研究进行系统的梳理, 希望为情感型机器人的开发及相关研究提供参考。

1 传统共情理论模型及其启示

共情 (empathy) 是指个体知觉和理解他人的情绪并做出适当反应的能力 (Decety & Svetlova, 2012)。它既是个体产生亲社会行为的重要动机, 也是道德行为的基石 (Decety & Cowell, 2014a, b)。目前, 具有较大影响力的共情理论模型主要有两个: 俄罗斯套娃模型 (de Waal, 2008; de Waal & Preston, 2017) 和双加工模型 (Decety & Fotopoulou, 2015; Decety & Jackson, 2006; Decety & Meyer, 2008)。

俄罗斯套娃模型认为共情包括三个层次的内容: 第一层是运动模仿和情绪共鸣; 第二层是共情关注和安慰; 第三层则是观点采择和有目标的帮助, 三层依次嵌套 (de Waal, 2008)。其中, 第一层内容最为基础, 也是人类共情的发生基础, 甚至在很多哺乳类动物当中也存在 (de Waal & Preston, 2017)。这意味着现阶段机器人共情的表现可能更多的需要从这一层面开始着手。与 de Waal 不同, Decety 和 Meyer (2008) 从认知神经科学的角度入手, 以脑成

* 本研究得到国家自然科学基金资助项目 (31571134, 31872782) 的资助。

** 通讯作者: 苏彦捷。E-mail: yjsu@pku.edu.cn

DOI:10.16719/j.cnki.1671-6981.20190207

像研究的成果为基础提出了双加工模型。依据该模型,共情加工存在两条信息加工通路:一条是受控制的以前额叶为神经基础的自上而下的通路,强调人类意识和认知能力在共情加工中的作用;另一条是基于镜像神经元系统的自动化的自下而上的通路,强调共情加工的无意识性。在共情加工的过程中,上述两条通路相辅相成,知觉表征使得个体产生自动化的反应,个体的意识和认知能力则会对产生的反应进行调节,从而使个体能够更好的对相应的情绪反应进行管控(黄菁菁,苏彦捷,2010)。

总的来看,从共情加工机制的角度来说,俄罗斯套娃模型(de Waal, 2008; Preston & de Waal, 2002)更多的着眼于共情的发生发展,与机器学习理论中深度学习的概念相契合。深度学习强调机器学习过程中的自组织适应,本质是一个发展过程,而这与人类共情的发展存在很多共通之处。换言之,人类的共情是个体基于与他人进行交互从而不断发展的,这也就是说,如果能够把共情的核心系统赋予机器人,那么借助计算机的深度学习,机器人是有可能借助人机交互的过程进行自我学习,从而发展出和人类相似的共情系统的(Asada, 2015a)。然而,鉴于该领域研究仍然处于胚胎阶段,共情的核心系统究竟是什么,以及即便确定该核心系统之后如何把其赋予机器人在很大程度上仍然是未知的。相比之下,双加工模型(Decety & Meyer, 2008)则从另一个角度为机器人共情的实现提供了指引。具体到机器人方面,其中涉及到的自下而上的自动化加工通路可以借由机器人的内在情绪系统数据库进行快速响应(陆泉,张良韬,2017);自上而下的认知加工通路则可以借由计算机算法和程序计算的改进和提升,从而达到对机器人共情反应的调节和控制(苏剑波等,2016;张润,王永滨,2016)。实际上,受这两个模型的影响,一些人工智能领域的研究者已经开始建构机器人共情的相关模型。

2 智能体共情及其计算机模型

一些研究者希望把共情机制嵌入机器人系统的尝试催生出了“artificial empathy”一词,该术语最先由Asada及其合作者提出,为方便理解,在此我们将其翻译为智能体共情(Asada, 2015b; Damiano, Dumouchel, & Lehmann, 2015a)。作为一个较为前沿的研究领域,智能体共情这一方面的研究尚处于萌芽阶段。通过借助传统的共情理论,一些研究者

正在尝试建构智能体共情的计算机模型。

Asada (2015a, 2015b)以de Waal (2008)所提出的套娃模型为基础,指出设计智能体共情可能需要考虑两个关键的加工过程:情绪共鸣和情绪/认知共情(见图1)。如同前面所提到的,情绪共鸣是个体对他人情绪的一种自动化的情绪共享,与运动模仿有关,其产生则有赖于镜像神经元系统(胡晓晴,傅根跃,施臻彦,2009; Oberman, Pineda, & Ramachandran, 2007)。Asada与其合作者曾尝试构建了一个镜像神经元系统的计算模型(Nagai, Kawai, & Asada, 2011)。该模型基于婴儿早期与照料者互动过程中涉及到的感觉——运动映射的关联式学习,将机器人的学习分为两个阶段:第一阶段是在运动命令和视觉影像之间建立关联;第二阶段则是在这种关联的基础上区分“我”和“非我”。这种对人类镜像神经元系统的模拟有助于深化了解机器人如何才能产生更加自然的运动模型,从而促进机器人对人类情绪的感知。而针对情绪/认知共情的功能模拟,可能需要借助于机器人的学习能力,正如人类儿童共情的发展一样,与他人的交流在其中扮演着非常重要的角色(Cole & Moore, 2014; Perone, Almy, & Zelazo, 2018)。总的来说,Asada等人的智能体共情模型和de Waal所提出的套娃模型相一致,都认为个体具有一套共情反应的核心系统以及学习系统,并且后者可以通过与外界交互而不断地完善这套系统。

Rodrigues, Mascarenhas, Dias 和 Paiva (2014)则提出智能体共情(见图2)可能主要包括两个加工过程,首先需要描绘出共情反应的特点,其次则是通过机器人所产生的行为反应来体现共情的内在机制,分别对应于de Vignemont 和 Singer (2006)所提出的共情评价(empathic appraisal)和共情反应(empathic response)。共情评价系统驱使主体根据相关线索对所发生的事件进行评估,从而使得主体产生相应的共情情绪(empathic emotion);共情反应系统则根据所产生的共情情绪做出相对应的共情行为(empathic action)。换言之,机器人需要依托场景知觉和对情绪面孔、体态或语调的识别分析筛选出与识别对象相一致的情绪,并通过对情绪类型与场景间的相关性进行计算而遴选出适合当前情境的行为(Lim & Okuno, 2015),从而对人类产生“共情反应”。Rodrigues等人的模型与Decety所提出的共情的双加工模型相一致,都认为个体具有一套

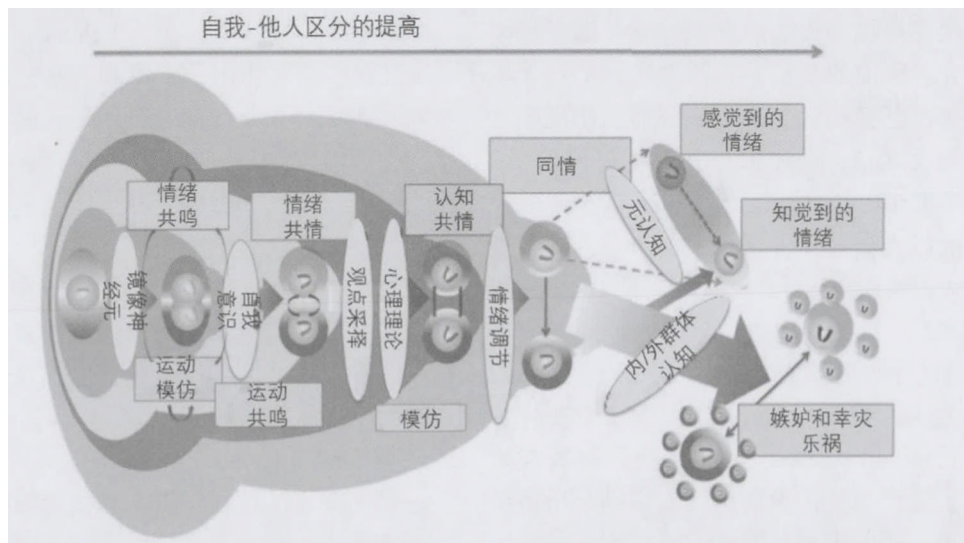


图1 智能体共情的发展模型（修改自 Asada, 2015b）

核心的共情反应系统，支撑产生自动化的共情反应；同时还具有一套共情评估系统，以对相应的共情反应进行调控。

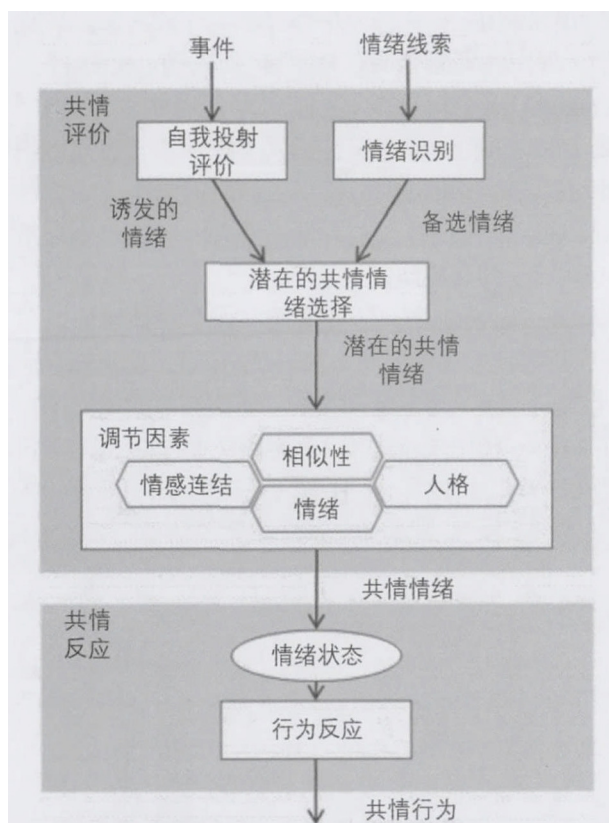


图2 智能体共情的评价反应模型（修改自 Rodrigues et al., 2014）

对这两类智能体共情的计算机模型进行分析可以发现，他们相互之间并不对立。首先，Asada（2015a）和 Rodrigues 等人（2014）都认为建构智能体共情需要建立一套和人类一样的初始共情反应系统，可以是像 Asada 所提出的非常基础的镜像神

经元的运动模型，也可以是 Rodrigues 等人（2014）所提出的完备的共情反应系统数据库；其次，他们都认为这种初始的共情反应系统是可以升级和扩展的。关于具体的实现路径，Asada（2015a）认为需要借鉴人类共情发展的特点，以镜像神经元系统为基础，使得机器人能够产生运动模仿和运动共鸣并完成“自我”和“非我”的区分，从机器学习的角度入手，以人机交互为手段促进机器人共情系统的发展；而 Rodrigues 等人（2014）则认为可以从共情反应的核心元素设计出发，通过对算法的进一步升级以弥补数据库信息的不足。总而言之，从构建贴近人类的情感型机器人的角度来说，Rodrigues 等人（2014）所设计的智能体共情模型具有高效性和易扩展性，但是个性化程度相对较低，而 Asada（2015a）所提出的智能体共情模型则正好可以弥补这个缺点。

3 智能体共情实现的实证探索

技术的进步促进和增加了人与机器人的互动，这使得机器人共情的实现显得愈发重要，而这也推动了一些研究的产生。近些年，在不同学科研究者的共同协作下，机器人的功能得到了不断地完善和发展，其间仿真机器人的情绪理解和情绪表达得到了极大程度的提升。例如 BARTHOC Jr. 机器人，该机器人能够借助交互过程中的言语信号分析对方的情绪，而且能够镜像地表达出这些情绪（Hegel, Spexard, Wrede, & Horstmann, 2006）。另外，由 Ishihara, Yoshikawa 和 Asada（2011）等人研制的 Affetto 机器人具有高度灵活的面孔单元，能够对他人面部表情进行实时、准确的反应。上述镜像的

面部情绪表现使得机器人能够在人际交互的过程中逐渐做到研究者所谓的共情,而智能体共情在人机交互中的影响也已经得到了一些研究的初步证实。例如,Leite 等(2013)设计了一个两个人下棋的场景,机器人(iCat)作为第三方观看下棋,在对弈的过程中,机器人会对其中一个人的每一步棋都表现出一些情绪表情和言语(例如:“妙招,这一步我怎么没有想到!”)以示共情,对另一方则做出不带有情绪色彩的中性的反应(例如:“下得不错。”),结果显示被关注的下棋者会认为机器人十分友好。De Carolis, Ferilli 和 Palestra(2017)更是将智能体共情应用到了医护场景:护理机器人借助分析被护理者的表情和言语以识别其情绪,而被识别的情绪以及行为将诱发机器人的情绪唤醒和相应行为的表达,即共情的产生。例如,当一个人愁眉苦脸,并哀叹“我真是可怜!”时,机器人就会对该个体释放的信息进行识别分析并产生一系列对应的行为(接近、表示理解、询问原因、给予帮助等)。此外,智能体共情的硬件调控系统也得到了一定程度的关注,例如,Lim 和 Okuno(2015)在构建智能体共情的计算机模型时,将人由共情所产生的替代性情绪体验在机器人上转化成了电池的电量水平和机器人的硬件运转温度,当机器人对人类进行共情时,其电量水平就会进行下降预警,硬件的运转温度就会出现升高预警,促使其做出合适的行为从而使得机体恢复到正常状态。

除了上述直接针对机器人共情进行的研究,也有一批研究者从人类对机器人的共情着手,探讨拟人化程度、机器人情绪的表达和理解等因素的影响

(Damiano, Dumouchel, & Lehmann, 2015b)。张钹(2017)就曾指出,机器人互相理解的一个关键就在于探讨人是如何看待机器人的。虽然这些研究与“智能体共情”的概念本身存在一定偏差,但这为提取机器人的哪些特征可能与人们对其产生的情感性知觉存在相关提供了可能,而这有助于人机互动过程中基本情绪交流功能的开发,进而为设计具有共情功能的机器人提供预备性数据。在这类研究中,机器人外型的拟人化程度是许多研究者较为关注的(许丽颖,喻丰,邬家骅,韩婷婷,赵靛,2017;宗阳,王广新,2016)。Mirmig 等人(2015)招募了一批人类被试评估机器人的面部情绪反应系统,这些被试需要对机器人所做出的表情进行判断,包括高兴、悲伤、惊讶和中性四种表情。结果发现,人类能够对机器人所表现出的情绪面孔进行正确的情绪类型判断。此外,机器人外型的拟人化程度越高、情绪面孔表现的动态性越强,人类所做出的情绪类型判断的准确率越高。并且机器人外型的拟人化程度越高,越容易诱发人类的运动模仿。Matsuda, Hiraki 和 Ishiguro(2016)以脑电活动中的 μ 抑制为运动模仿的指标,探讨了人类在观看人、人型机器人和非人型机器人的运动视频时的脑电反应。结果发现,仅在观看人和人型机器人时被试才表现出明显的 μ 抑制反应。

为了对智能体共情的研究有一个总体层面的了解,研究者参照前人的方法以“artificial”、“robot”和“empathy”为主题词检索了英文文献数据库 Web of Science 和中文文献数据库中国知网(CNKI),检索日期为2017年11月20日。由于几乎没有相关

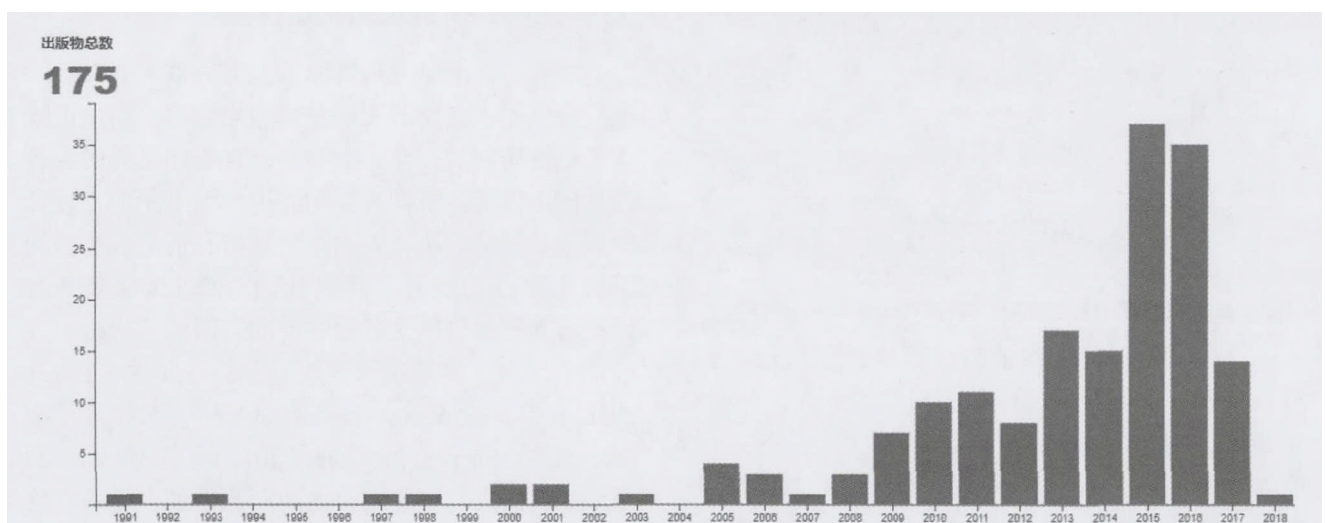


图3 随年份论文出版数量

的中文文献,因此仅对检索到的175篇英文文献采用VOSviewer软件(van Eck & Waltman, 2010)进行了文献计量学分析(颜志强, 苏彦捷, 2017),具体结果见图3、图4。总的来看,虽然智能体共情的研究在近几年呈现爆发式的增长,但是其研究的数量还很有限,研究的内容也很初步,相关的研究在内容上可以大致分为以下几类:(1)人机交互的界面和形式(左上红色部分);(2)共情机制的神经表征(右上绿色部分);(3)人工智能(下方粉色、黄色和蓝色部分)。但是,这些研究为智能体共情的开发提供了必不可少的参考,同时也让我们意识到,智能体共情的实现需要不同学科研究者的通力合作,尤其是心理学、认知神经科学和计算机科学等。

4 研究展望

情感型机器人的研究尚处于萌芽阶段,而给机

器人注入“共情”也只有一些初步的尝试。目前,研究者更多地围绕该问题进行理论上的探讨,相关实证研究还很少,而已有心理学和神经科学研究成果则可以为这一领域的深入提供诸多启示。

首先,心理学最为凸显的一个作用便在于其所探讨的人类共情模型将有助于完善智能体共情的计算机模型。如前面我们已经讨论了套娃模型和双加工模型对智能体共情的计算机模型的建构可能起到的作用,但这些模型并没有考虑自我学习成分,而在智能体共情领域,自我学习能力极为重要。针对这一点,发展心理学方面的研究成果或许可以借鉴(Lim & Okuno, 2015)。其实,婴幼儿方面的心理学研究已经开始显现出其价值。例如, Meltzoff (2007) 所提出的“like me”理论就指出,成功的模仿是个体之后发展出高级认知能力的基础。而以这一理论为基点,已有研究者从婴儿的角度

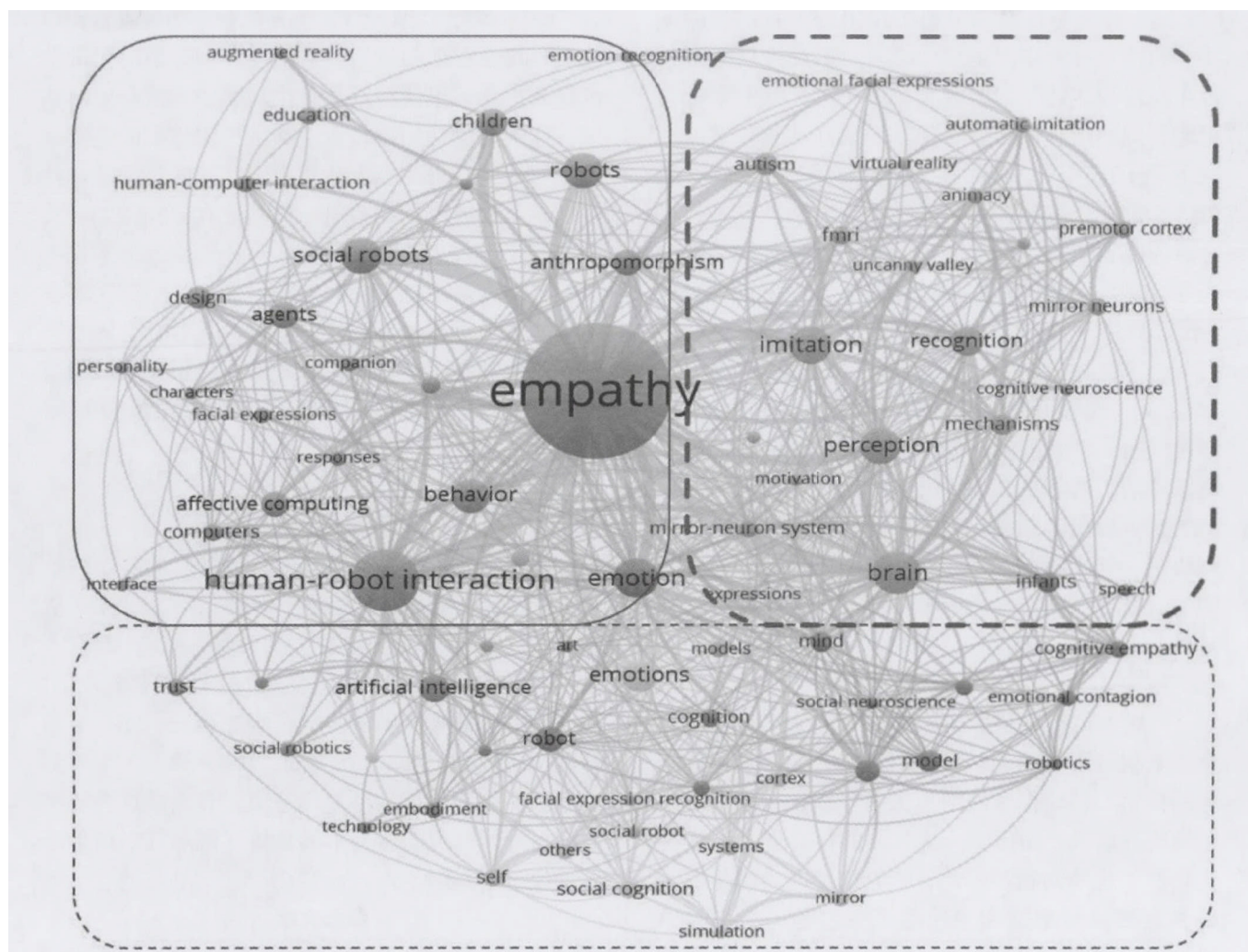


图4 关键词共现图谱

注:该图谱基于论文中关键词共同出现的次数进行可视化,关键词之间的距离越近说明共现次数越多,关键词越大,说明其出现的次数越多。

入手,探讨了婴儿模仿技能的发展在机器人身上的实现(Demiris & Meltzoff, 2008)。同时,心理学研究也将有助于开发用于评估智能体共情的研究工具。一方面,虽然该领域的研究者有尝试编制一些相关的评估工具(Rosenthal-von der Pütten, Kramer, Hoffmann, Sobieraj, & Eimler, 2013; Tisseron, Tordo, & Baddoura, 2015),但这些工具在编制过程中往往不能严格遵循心理测量学的有关程序,致使其存在信效度问题,这就需要从心理测量的角度进行专业开发以补充;另一方面,心理学中的一些动物比较研究也可以为机器人共情能力的实现提供线索,如Kano和Call(2014)的研究发现,虽然大猿会对基于目标的行为产生预测,但其只能预测人类手臂的运动方向,却不会预测机械手臂的运动方向。这提示我们,这类心理学实验或可以用于检测机器人与人类的相似度。

其次,文献计量学分析的结果提示我们智能体共情的研究可能涉及到多个学科,因此今后的研究者在探讨智能体共情时应尝试进行多学科的融合与交流。例如,基于已有心理学、认知神经科学等学科的研究成果,Asada等人(2015a)推演出了共情产生可能依赖的神经计算模型。以婴儿在对照料者的面部表情学习过程中涉及到的计算过程为例,在学习的过程中,婴儿的额下回(inferior frontal gyrus, IFG)和脑岛(insula)首先被激活,之后映射到背侧前扣带回(dorsal anterior cingulate cortex, dACC)。在学习之后,当照料者再次表现出相应情绪表情时,婴儿的脑网络就会根据之前的记忆进行镜像激活。这就从侧面为共情基础成分在机器人身上的实现提供了某种程度的计算基础。而针对共情高级成分(如观点采择)在机器人中的实现,人们现在仍所知甚少,有研究者基于神经网络激活基础数据建立的概率生成模型或可有助于这一问题的解决(Tamir & Thornton, 2018)。

再次,情感型机器人、类人智能的出现在给社会带来积极影响的同时,也引发了人们对其伦理问题的探讨。科幻作家阿西莫夫很早就提出了机器人三原则(陈涛, 2015),其关注的核心就是“人类优越性”这一问题。因此,具有情感能力的机器人是否需要具有道德和主体意识一直是一个极具争议性的问题。一方面,有研究者就提出人工智能所具有的道德是否是人类道德所容纳的(王东浩, 2014)?另一方面,人工智能是否需要为其有意图

的行为承担道德责任(杜严勇, 2017)?就实际而言,情感型机器人在日常生活中很容易陷入道德困境,因为这涉及到机器人的“人类优越性”准则问题。首先,机器人的最高优先级是人类还是机器人的实际拥有者?其次,如果机器人获得了主体意识,其有意图的最高优先级的调整是否需要承担道德责任?笔者认为,智能体共情从目前技术实现的角度更多的是偏向于无意识的镜像反应,这种智能系统具有场景限定性(张钺, 2017),主要还是应用于家庭和医院。道德和主体意识的加入将使得机器人能够根据情境调整共情反应的对象优先级,当然这也涉及到机器人内置人类心理状态知识库信息的完备性和准确性(张钺, 2014)。道德和主体意识的加入所引发的问题则可能需要借助于人类学、伦理学和法律学领域研究者的帮助,在政府主导的情况下进行相关政策和准入标准的制定。

最后,传统共情经常发生社会互动过程当中,也就是说,在现实生活中共情主体和共情对象常常都是人类。这里就涉及到一个问题:共情对象对共情主体的能动性知觉是否是共情产生效果的先决条件之一。一个可能是,共情主体具有的意识性和社会性是其释放出的某些信号被共情对象知觉为“共情”并产生效果的必要条件。对于人机互动来说,虽然机器人高度拟人化的外表可能使其更像人类,进而使人类逐渐将其当作一份子,具有社会属性。但作为物理性实体,意识问题是一个长期困扰的难题。这类问题也将是设计出真正具有共情的智能体必然要考虑的。

5 结语

依赖于技术进步,机器人开始从实验室走向社会,情感型机器人也随之诞生。情感型机器人要求机器人具备共情能力,而该能力的开发有赖于心理学、认知神经科学和计算机科学等领域的相关研究成果。基于现有智能体共情的发展模型(Asada, 2015a)和评价反应模型(Rodrigues et al., 2014)以及未来可能的新模型,我们认为可以构造出拥有共情能力的机器人。同时,智能体共情的具体实现和其所衍生出的伦理问题也必将是今后研究所要关注的重点。

参考文献

- 蔡曙山, 薛小迪. (2016). 人工智能与人类智能——从认知科学五个层级的理论看人机大战. *北京大学学报(哲学社会科学版)*, 53(4), 145-154.
- 陈涛. (2015). 科学选择与伦理身份: 阿西莫夫小说中的人机伦理关系. *华中学术*, 12(2), 71-81.

- 杜严勇. (2017). 机器人伦理中的道德责任问题研究. *科学学研究*, 35(11), 1608–1613.
- 胡晓晴, 傅根跃, 施臻彦. (2009). 镜像神经元系统的研究回顾及展望. *心理科学进展*, 17(1), 118–125.
- 黄嵩青, 苏彦捷. (2010). 共情中的认知调节和情绪分享过程及其关系. *西南大学学报(社会科学版)*, 36(6), 13–19.
- 刘聪慧, 王永梅, 俞国良, 王拥军. (2009). 共情的相关理论评述及动态模型探新. *心理科学进展*, 17(5), 964–972.
- 陆泉, 张良韬. (2017). 处理流程视角下的大数据技术发展现状与趋势. *信息资源管理学报*, 4, 17–28.
- 苏剑波, 陈叶飞, 马哲, 黄瑶, 向国菲, 陈若冰. (2016). 从 AlphaGo 到 BetaGo——基于任务可完成性分析的定性人工智能的定量实现. *控制理论与应用*, 33(12), 1572–1583.
- 陶建华, 谭铁牛. (2004). 数字化人类情感——和谐人机交互环境中的情感计算. *微电脑世界*, 1, 29–32.
- 王东浩. (2014). 人工智能体引发的道德冲突和困境初探. *伦理学研究*, 2, 68–73.
- 许丽颖, 喻丰, 郭家骅, 韩婷婷, 赵靛. (2017). 拟人化：从“它”到“他”. *心理科学进展*, 25(11), 1942–1954.
- 颜志强, 苏金龙, 苏彦捷. (2017). 共情的时代变迁：一项横断历史元分析. *心理技术与应用*, 5(10), 578–585.
- 颜志强, 苏彦捷. (2017). 共情研究主题的变化——来自文献计量学的证据. *心理科学*, 40(3), 699–707.
- 张钹. (2014). 评《人机交互中的体态语言理解》. *科学通报*, 59(31), 3108.
- 张钹. (2017). 引领人工智能发展须从何处着力. *联合时报*, 4, 1–2.
- 张润, 王永滨. (2016). 机器学习及其算法和发展研究. *中国传媒大学学报(自然科学版)*, 23(2), 10–18, 24.
- 宗阳, 王广新. (2016). 拟人化：人机交互中的心理学应用. *心理技术与应用*, 4(5), 296–305.
- Asada, M. (2015a). Development of artificial empathy. *Neuroscience Research*, 90, 41–50.
- Asada, M. (2015b). Towards artificial empathy: How can artificial empathy follow the developmental pathway of natural empathy? *International Journal of Social Robotics*, 7(1), 19–33.
- Cole, P. M., & Moore, G. A. (2014). About face! Infant facial expression of emotion. *Emotion Review*, 7(2), 116–120.
- Damiano, L., Dumouchel, P., & Lehmann, H. (2015a). Artificial empathy: An interdisciplinary investigation. *International Journal of Social Robotics*, 7(1), 3–5.
- Damiano, L., Dumouchel, P., & Lehmann, H. (2015b). Towards human-robot affective co-evolution overcoming oppositions in constructing emotions and empathy. *International Journal of Social Robotics*, 7(1), 7–18.
- de Carolis, B., Ferilli, S., & Palestra, G. (2017). Simulating empathic behavior in a social assistive robot. *Multimedia Tools and Applications*, 76(4), 5073–5094.
- de Vignemont, F., & Singer, T. (2006). The empathic brain: How, when and why? *Trends in Cognitive Sciences*, 10(10), 435–441.
- de Waal, F. B. M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, 59, 279–300.
- de Waal, F. B. M., & Preston, S. D. (2017). Mammalian empathy: Behavioural manifestations and neural basis. *Nature Review Neuroscience*, 18(8), 498–509.
- Decety, J., & Cowell, J. M. (2014a). The complex relation between morality and empathy. *Trends in Cognitive Sciences*, 18(7), 337–339.
- Decety, J., & Cowell, J. M. (2014b). Friends or foes: Is empathy necessary for moral behavior? *Perspectives on Psychological Science*, 9(5), 525–537.
- Decety, J., & Fotopoulou, A. (2015). Why empathy has a beneficial impact on others in medicine: Unifying theories. *Frontiers in Behavioral Neuroscience*, 8, 457.
- Decety, J., & Jackson, P. L. (2006). A social-neuroscience perspective on empathy. *Current Directions in Psychological Science*, 15(2), 54–58.
- Decety, J., & Meyer, M. (2008). From emotion resonance to empathic understanding: A social developmental neuroscience account. *Development and Psychopathology*, 20(4), 1053–1080.
- Decety, J., & Svetlova, M. (2012). Putting together phylogenetic and ontogenetic perspectives on empathy. *Developmental Cognitive Neuroscience*, 2(1), 1–24.
- Demiris, Y., & Meltzoff, A. (2008). The robot in the crib: A developmental analysis of imitation skills in infants and robots. *Infant and Child Development*, 17(1), 43–53.
- Hegel, F., Spexard, T., Wrede, B., Horstmann, G., & Vogt, T. (2006). Playing a different imitation game: Interaction with an empathic android robot. In *2006 6th IEEE-RAS international conference on humanoid robots. Genova, Italy: IEEE*.
- Ishihara, H., Yoshikawa, Y., & Asada, M. (2011). Realistic child robot “Affetto” for understanding the caregiver-child attachment relationship that guides the child development. In *2011 IEEE international conference on development and learning. Frankfurt am Main, Germany: IEEE*.
- Kano, F., & Call, J. (2014). Great apes generate goal-based action predictions: An eye-tracking study. *Psychological Science*, 25(9), 1691–1698.
- Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human-robot relations. *International Journal of Human-Computer Studies*, 71(3), 250–260.
- Lim, A., & Okuno, H. G. (2015). A recipe for empathy: Integrating the mirror system, insula, somatosensory cortex and motherese. *International Journal of Social Robotics*, 7(1), 35–49.
- Matsuda, G., Hiraki, K., & Ishiguro, H. (2016). EEG-based mu rhythm suppression to measure the effects of appearance and motion on perceived human likeness of a robot. *Journal of Human-Robot Interaction*, 5(1), 68–81.
- Meltzoff, A. N. (2007). ‘Like me’: A foundation for social cognition. *Developmental Science*, 10(1), 126–134.
- Mirig, N., Strasser, E., Weiss, A., Kühnlenz, B., Wollherr, D., & Tscheligi, M. (2015). Can you read my face? A methodological variation for assessing facial expressions of robotic heads. *International Journal of Social Robotics*, 7(1), 63–76.
- Nagai, Y., Kawai, Y., & Asada, M. (2011). Emergence of mirror neuron system: Immature vision leads to self-other correspondence. In *2011 IEEE international conference on development and learning. Frankfurt am Main, Germany: IEEE*.
- Oberman, L. M., Pineda, J. A., & Ramachandran, V. S. (2007). The human mirror neuron system: A link between action observation and social skills. *Social Cognitive and Affective Neuroscience*, 2(1), 62–66.
- Perone, S., Almy, B., & Zelazo, P. D. (2018). Toward an understanding of the neural basis of executive function development. In R. Gibb & B. Kolb (Eds.), *The neurobiology of brain and behavioral development* (pp. 291–314). Amsterdam: Elsevier.

- Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25(1), 1–20.
- Rodrigues, S. H., Mascarenhas, S., Dias, J., & Paiva, A. (2014). A process model of empathy for virtual agents. *Interacting with Computers*, 27(4), 371–391.
- Rosenthal-von der Pütten, A. M., Kramer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, 5(1), 17–34.
- Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, 132(3), 617–627.
- Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, 22(3), 201–212.
- Tisseron, S., Tordo, F., & Baddoura, R. (2015). Testing empathy with robots: A model in four dimensions and sixteen items. *International Journal of Social Robotics*, 7(1), 97–102.
- Valuch, C., Pflüger, L. S., Wallner, B., Laeng, B., & Ansorge, U. (2015). Using eye tracking to test for individual differences in attention to attractive faces. *Frontiers in Psychology*, 6, 42.
- van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.
- Watanabe, A., Ogino, M., & Asada, M. (2007). Mapping facial expression to internal states based on intuitive parenting. *Journal of Robotics and Mechatronics*, 19(3), 315–323.

From Human Empathy To Artificial Empathy

Yan Zhiqiang, Su Jinlong, Su Yanjie

(School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, 100871)

Abstract As robots are increasingly important in human's life, people begin to pay attention to their other potentials besides intelligence, such as empathy, which plays an important role in human sociality. The prerequisite for robots to become our partner is to share and understand the human's emotions and to react properly. In fact, current research on empathy mostly focuses on traditional interpersonal interaction, and little attention has been given to human-robot interaction. This paper aims to review related research on psychology and human-robot interaction under the background of developing artificial intelligence.

Russian Doll Model and Dual-Process Model are the two most influential theoretical models of empathy. Russian Doll Model considers empathy as a construct comprising of three layers: (1) Motor mimicry and emotional contagion; (2) Empathetic concern and consolation; (3) Perspective-taking and targeted helping. Dual-Process Model stems from relevant findings in cognitive neuroscience, and proposes that empathy work via two routes: bottom-up processing and top-down processing. Both models as well as empirical studies drawing from them suggest a possibility to program artificial empathy analogous to human empathy.

Focusing on the topic of artificial empathy and based on classic models, the current study reviewed the latest research in artificial intelligence and systematically discussed existing theoretical models and empirical studies about artificial empathy. First, we did a bibliometric analysis to draw an overall picture of artificial empathy studies. By analyzing 175 articles, we found that studies in the field of artificial empathy were expandingly concentrated on three topics: (1) Interface and form of human-robot interaction; (2) Computerized simulation of empathy; (3) Artificial intelligence. Second, we introduced two computerized models of artificial empathy. As mentioned above, these two models were derived from the traditional empathy models and assumed that robots needed a main system of empathy and another system for learning.

Recently, preliminary application of research in artificial empathy has been successful in board game industry and medical treatment. In order to promote further research, we argue that it is important to draw on progress in four areas below: (1) Due to the importance of self-learning, insights from developmental studies will contribute to building theoretical models and developing scales with satisfactory psychometric properties in the field of artificial empathy; (2) According to the results of bibliometric analyses, interdisciplinary cooperation will help to delve deep into this field; (3) Ethical issues should not be neglected as studies on artificial empathy proceed, since there is still so much work to do on these issues and government participation is indispensable; (4) Last but not least, consciousness is a hot topic in the field of artificial intelligence. It is possible that authentic empathy cooccurs with consciousness, and if this is exactly the case, we would have to consider the role of consciousness in developing artificial empathy. In sum, based on the present research achievement and theory models, we propose that the actualization of artificial empathy will be the key for robots to open the door towards the real human-like reaction and to better serve humans in the future.

Key words empathy, artificial empathy, robot, human-robot interaction, emotion understanding