心理科学进展 2024, Vol. 32, No. 5, 845–858
Advances in Psychological Science
© 2024 中国科学院心理研究所
https://doi.org/10.3724/SP.J.1042.2024.00845

当 AI 学习共情：心理学视角下共情计算的主题、场景与优化*

侯悍超 倪士光 林书亚 王蒲生

(清华大学深圳国际研究生院, 深圳 518055)

摘 要 共情计算是指使用人工智能和大数据技术来预测、识别、模拟和生成人类的共情, 是传统心理学共情研究与计算机科学交叉的新兴研究领域。本研究构建了一个数据层、模型层与任务层的普适性研究框架, 总结了一个包括个体共情测评、共情内容分类、共情回应系统和共情对话生成的 4 个新主题分析框架, 建立了面向心理健康、教育学习、商业服务和公共管理等心理应用的场景创新。未来研究有必要建构高整合的共情计算理论模型、建设高可信的共情心理行为特征数据集, 并通过以人为中心的评价体系验证并改进共情计算的研究效度。共情计算有益于将当前人际共情研究扩展到智能社会新型人–AI 关系研究中。心理学家在该领域承担引领、评估和改进等角色, 并与计算机科学家紧密合作, 共同推动共情计算理论基础更坚实, 效果检验更可靠, 应用创新更贴近实际需求。

关键词 共情, 共情计算, 计算心理学, 人工智能, 人机交互

分类号 B849

共情(empathy)是指感受和理解他人情绪, 并做出合理回应的心理和行为过程, 也指个体完成这一过程的能力或特质(Bošnjaković & Radionov, 2018; Hall & Schwartz, 2019; Preston & de Waal, 2002)。共情在个体的社会生活中扮演重要角色, 与亲社会行为(de Waal, 2008)、良好的人际关系(Morelli et al., 2017)及幸福感(Grühn et al., 2008)密切相关。作为心理学的一个重要研究主题, 共情在人格与社会心理学、发展心理学、咨询心理学等领域都积累了许多的研究成果(Hall & Schwartz, 2019)。

2022 年末, ChatGPT (OpenAI et al., 2023)的推出标志着人工智能(artificial intelligence, AI)开始进入普通人的日常生活。与过去冷冰冰的机器形象不同, 最新研究表明, 基于大语言模型的 AI 已经表现出一定的共情能力(capability)[1], 可以在情绪识别任务中达到普通人的平均水平(Kosinski, 2023; Wang et al., 2023)。共情计算(empathy computing)是在 AI 技术飞速发展背景下兴起的心理学与计算机科学交叉研究领域, 旨在运用计算的方法自动化分析和模拟共情。计算机科学的进步使得大规模收集和自动化分析心理和行为数据更加高效和便捷, 从而为共情计算研究提供了条件。研究者通过机器学习等方法, 分析过去通过人工编码无法处理的大量文本和音频等数据, 在更大尺度上验证和改进传统的共情研究(Xiao, Imel, Georgiou, et al., 2015)。另一方面, 人工智能技术的应用正在扩展到科研、医疗、教育、服务和社会治理等领域(Zhu et al., 2023), 使得人与 AI 的互动更为普遍; 表现出共情能力的人机交互系统提升人对机器的信任、投入感、减少互动中的

收稿日期：2023-03-29

* 全国教育科学“十四五”规划 2021 年度课题(BBA210042)。

通信作者：倪士光, E-mail: ni.shiguang@sz.tsinghua.edu.cn

[1] 这里以及后文中提及计算机和 AI 所具备的“能力”, 是指计算系统执行特定任务或功能的能力, 英文译作 capability。具体而言, 文中计算机的“共情能力”, 指计算机通过模拟人类语言和行为, 输出与人类相似的共情特征的能力(capability)。尽管同样使用了“能力”一词, 但其内涵与人类具有的共情能力(ability)并不等同。

挫败感并提升互动的时长(Pelau et al., 2021; Yalçın & DiPaola, 2020), 可见对 AI 共情能力的需求也催生了共情计算的研究。

近年来, 共情计算越来越受到关注。2023 年初《自然−机器智能》发表了一项由计算机和心理学领域的研究者共同完成的共情计算研究(Sharma et al., 2023)。研究者开发了一个共情计算系统, 该系统评估助人者回复的共情程度, 并提供改写建议, 协助朋辈助人者提供更有“人情味”的回复(图 1)。结果表明, 与助人者独立完成的对话相比, AI 和人配合的回复表现出更高的共情水平。尽管已经有一些心理学研究者开始参与共情计算研究, 目前多数研究仍来自于计算机领域。通过本文的介绍, 希望引起更多心理学研究者对共情计算的兴趣和关注, 并促进共情计算的基础研究和应用。

1 共情计算的理论背景

1.1 传统共情研究

共情计算是一个新兴领域, 其发展建立在传统共情研究的基础之上。传统研究在共情的概念、测量、神经基础、个体差异以及在心理咨询等领域的应用等方面有丰富成果。研究者普遍接受共情包含了情感和认知成分, 也有一部分研究者认为行为成分也包含在内(Cuff et al., 2014; Stosic et al., 2022)。情感成分是指直接感受和分享他人情绪的过程, 认知成分是指个体间接地想象和设身处地理解他人情绪的过程, 行为成分是指互动时让对方感到温暖、被理解和支持的行为(Stosic et al., 2022)。多数研究采用自我报告法, 从情感和认知两方面测量共情(Hall & Schwartz, 2019)。以经典的人际反应指针量表(Interpersonal Reactivity Index, IRI; Davis, 1983)为例, 它包括 4 个分量表: 共情关注(empathic concern)和个人苦恼(personal distress)维度分别测量个体的情感共情能力, 而想象(fantasy)和观点采择(perspective taking)两个分量表用来测量认知共情能力。近年新开发的测量工具, 基本共情量表(Basic Empathy Scale, BES; Jolliffe & Farrington, 2006), 认知和情感共情量表(Questionnaire of Cognitive and Affective Empathy, QCAE; Reniers et al., 2011), 珀斯共情量表(Perth Empathy Scale, PES; Brett et al., 2023)等也是在这一基础上改进而成的。神经科学研究表明, 情感共情主要与镜像神经系统有关, 包括额下回、顶下小叶、后顶叶皮质和颞上沟等脑区, 而认知共情则与心理理论神经系统有关, 如内侧前额叶、

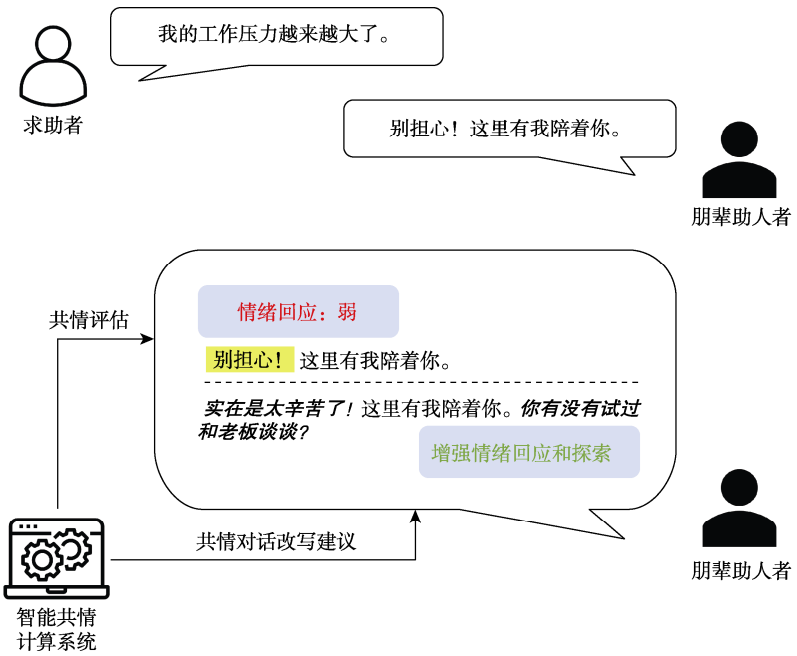

图 1 共情计算研究范例(改编自：Sharma et al., 2020, 2021, 2023)

颞上沟、颞顶联合区和颞极等(岳童, 黄希庭, 2016)。情感共情和认知共情并非完全独立, 两者间存在一个共同激活过程(Schurz et al., 2021)。人的整体共情能力和特质共情存在个体差异(Mooradian et al., 2011); 共情被广泛应用于心理咨询, 对咨询效果有中等强度的正向预测作用(Elliott et al., 2011)。

这些传统研究成果为共情计算的发展奠定基础并提供启示。共情概念和成分的研究为共情计算提供了理论框架, 明确研究范畴。传统自评量表作为校标, 可以评估共情计算的准确性和可靠性。人类共情的神经机制研究可能为设计脑机交互提供启发(Roshdy et al., 2021)。以往的应用研究也为共情计算潜在应用场景指明方向。

1.2　共情计算的概念

共情计算尚未形成统一的定义。本文结合作者的自身认知以及现有文献中的描述, 将共情计算定义为：使用计算系统收集并处理文字、声音、图像、生理信号等多模态数据, 用以预测、识别、模拟和生成人与人、人与机器间共情心理和行为的研究领域(Preston & de Waal, 2002; Xiao et al., 2016; Yalçın & DiPaola, 2020)。其中计算系统包括计算机、传感器等硬件设备以及支持其完成自动化运算的软件。共情计算研究不仅仅指运用计算方法分析各种生理、心理、行为信号来测量共情, 也包括通过模拟这些信号使人工智能或机器人表现出共情能力。由于侧重点不同, 研究者对于该领域有一些不同称呼, 如共情性计算(empathetic computing; Cai, 2006), 计算共情(computational empathy; Yalçın & DiPaola, 2020)或人工共情(artificial empathy; Asada, 2015; Cao et al., 2021)等。鉴于研究手段以及目标的相似性, 我们认为这些都是共情计算相关的研究领域。

1.3　共情计算的研究框架

为了展现共情计算领域的全貌, 本文建构了共情计算的研究框架, 如图 2 所示。心理学研究为共情计算收集和标注数据、建构模型提供理论依据, 并贡献研究主题。计算机科学方面则支持共情计算收集多模态的数据、设计算法以及编写所需的程序。共情计算又反过来促进心理学和计算机科学的研究。

共情计算主要由数据、模型和任务三个层面组成。在数据层, 通过各种数据源头收集与共情有关的生理、心理和行为多模态数据, 如语言、面部表情、语音语调等。这些数据来自问卷调查、临床或实验等传统心理学研究方法, 如心理咨询的视频录像; 也可能来自智能手机、智能手表、虚拟现实(Virtual Reality, VR)眼镜等数字设备的使用过程, 如社交媒体中表达出对新闻当事人共情的留言。在模型层, 研究者建立共情的计算模型, 使计算机可以通过算法“学习”共情。一般做法是基于共情的心理学理论(例如 Sharma et al., 2020)或采用数据驱动的方式(例如 Rashkin et al., 2019)抽取数据中有关共情的特征(如表达共情的

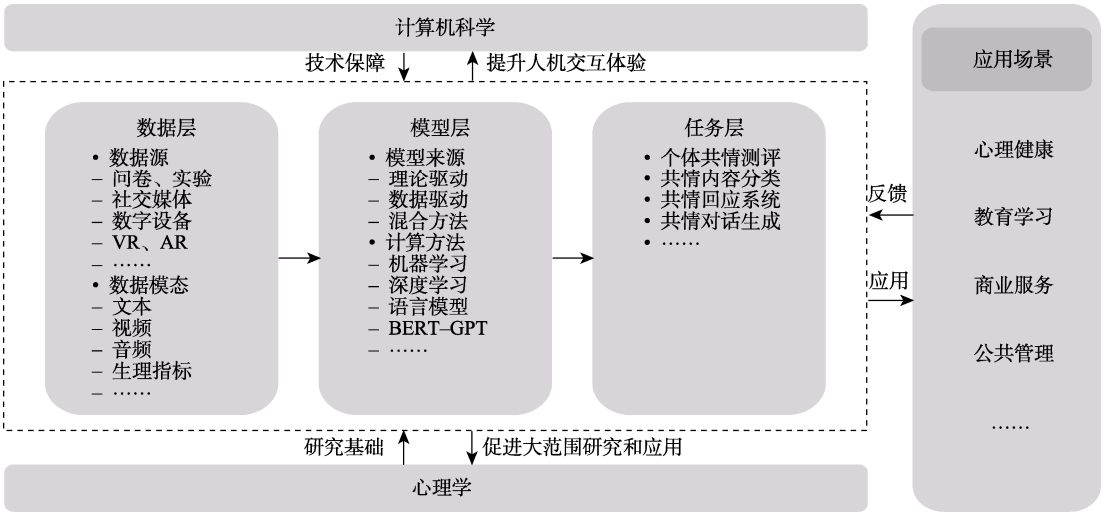

图 2　共情计算的研究框架

词汇、缓和的语调等), 随后用这些特征训练计算模型, 训练后的模型可用于预测新数据或生成新内容。深度学习是近几年共情计算研究中常用的方法, 属于机器学习的一种, 是指通过神经网络模拟大脑中的神经元, 从而进行模式识别和特征提取, 对复杂数据的学习和处理。常用的深度学习模型有递归神经网络(Recurrent Neural Networks, RNN)、卷积神经网络(Convolution Neural Networks, CNN)、长短时记忆(Long Short Term Memory, LSTM)和 Transformers (Vaswani et al., 2017)等。在任务层, 现有的研究主要分为 4 个主题: 个体共情测量、文本共情内容分类、共情回应系统和共情对话生成。

2 共情计算的研究主题

作为一个新兴领域, 共情计算的研究内容在不断发展变化中。当前主要研究, 可以根据目的和手段不同分为 4 个主题。一方面, 共情计算的首要任务是计算机分析和理解共情, 分为(1)个体共情测量和(2)共情内容分类, 前者侧重于分析个体的共情特质, 后者聚焦于分析文本中的共情特征, 而非真实的人。另一方面, 这一研究领域致力于使计算机模拟和表现共情, 包括(3)设计共情性回应系统和(4)开发生成式共情对话系统, 前者设计有限的回应方式并根据规则给用户反馈以表现共情, 后者则利用 AI 自动生成共情性的对话。这些主题相对独立但互补, 随着研究的深入, 也会产生新的主题。

2.1 个体共情测评

个体共情测评是共情计算较早发展的领域, 起初是为了评估心理咨询师, 如通过文字和语音评价心理咨询师的共情能力(Xiao, Imel, Georgiou, et al., 2015), 随后扩展到其他情景, 如通过社交媒体信息识别可能损害用户健康的致病性共情(pathogenic empathy) (Abdul-Mageed et al., 2017)。这类研究的做法是研究者基于理论, 或者通过数据驱动的方法提取可能与共情有关的行为特征(如具有代表性的词语、语调、或表情等), 通过机器学习等方法建立行为特征与自我报告或专家评分之间的关系模型, 随后再使用建立好的模型对新的数据集进行评分(Xiao et al., 2016)。目前的研究已经探索了对话文本(Chakravarthula et al., 2015; Gibson et al., 2016; Gibson et al., 2015; Litvak et al., 2016)、语调(Imel et al., 2014; Xiao et al., 2014)、语速(Xiao, Imel, Atkins, et al., 2015)、面部表情(Kumano et al., 2011; Mathur et al., 2021)、目光(Ishii et al., 2018)等多种行为特征与共情的关系。如 Litvak 等(2016)研究发现社交媒体上的语言风格和特质共情有关, 代词的使用数量和 IRI 量表中的观点采择子维度有较高的相关性。Xiao 等(2014)发现高音调与心理咨询师的低共情有关。

通过计算的方法测量个体的共情特质, 为自动化、大规模评估和研究共情提供了有效的工具。共情计算的方法已经对个体的共情水平做出较准确的测量, 与专家评分达到了较高的相关(r = 0.65) (Xiao, Imel, Georgiou, et al., 2015)。此外, 这些研究也为设计能够展现共情能力的 AI 或机器人提供参考, 如训练机器人在对话、语调、语速等方面表现得更接近高共情特质的人。不过, 某项行为特征在预测共情时是否具有跨语言、文化、情景适用性还需要更多研究。在美国样本中预测共情的语调同步性在法语咨询中并未得到验证(Gaume et al., 2019)。此外, 人类评估者倾向于整合言语和非言语的行为对个体的共情能力做出整体判断(Xiao et al., 2012), 未来研究需要探索怎样整合文本、音频、视频等多模态数据(Ma et al., 2022), 以更自然地评估共情水平与共情内容。

2.2 共情内容分类

共情计算的另一个重要主题是识别文本中表现出共情的关键词或句子, 或依据所表现出的共情强度将文本分类。由于绝大多数网络本文数据无法追溯内容发布者以获取心理测量数据, 这类研究以文本本身作为研究对象, 而非测量文本所反映的个体共情特质。研究者开发了面向具体场景中共情信息的识别模型, 如识别线上癌症社区中的共情留言(Khanpour et al., 2017), 评估在线朋辈写作互评中的共情程度(Wambsganss et al., 2022), 和评估读者所写的新闻读后感中所表现出的共情程度(Buechel et al., 2018; Zhou et al., 2021)等。这类研究首先建立标注过的共情内容数据集, 随后用标注过的数据集训练模型, 最后得到识别文本中共情信息的位置, 或者为内容的共情程度评分的共情计算模型(Sharma et al., 2020)。

聚焦于文本内容的共情计算研究不依赖个体心理测量数据, 便于更大范围收集数据, 补充并

扩展了前文中个体共情特质的计算, 为分析社交网络对话中的共情提供工具, 也为生成共情对话提供参考。共情内容识别和评估已达到较高准确率, 如在判定一段留言是否共情时 Khanpour 等(2017)报告了 78.61%的准确率, 并具备一定程度的跨场景迁移能力, 如基于新闻评论开发的模型也可以用于评估论坛跟帖和电影脚本的共情(Zhou et al., 2021), 因此有广泛的应用前景。然而, 当前研究局限在于从第三人视角评估文本内容, 未能真实反应内容作者和读者的个人体验。为了弥补这一局限, 研究者正致力于收集真实对话情境中发言者和听众的第一人视角的共情体验数据, 以更全面地分析文本中的共情内容(Barriere et al., 2022; Omitaomu et al., 2022)。未来研究需要整合不同视角的数据, 继续提高算法准确性, 并进一步验证和拓展模型的跨场景适用性。

2.3　共情回应系统

在人工智能技术尚未成熟的阶段, 为实现更加人性化的人机交互, 使计算机不仅能执行人类交给的任务, 而且能回应人类的情绪, 研究者早期采用基于规则的方法, 设计了根据不同情绪做出相应回应的计算系统。这类系统通过计算方法将用户的情绪归类, 然后提供预先设计好的共情性反馈。如 Terzis 等(2012)根据计算机自动面部识别和人工识别相结合的方法, 识别了学生的 6 种基本情绪(开心、生气、伤心、惊讶、害怕、厌恶), 并根据情绪给出相应的鼓励或安抚, 是这一领域最早的尝试。也有些研究并不识别用户的情绪, 而是根据用户在任务中的表现推测用户可能需要的反馈。Leite 等(2013)的研究中, 棋手走位接近计算机判断的最优解, 系统会予以鼓励, 如果走位不好则鼓励其思考更优方案。也有研究同时结合情绪识别和任务表现提供相应的共情性反馈(D'mello & Graesser, 2013)。

基于规则的共情系统设计在早期推动了共情计算领域的发展, 也在实际应用中取得了较好的效果, 如提升答题正确率(Guo & Goh, 2016), 缓冲消极情绪对于创造力的影响(Groh et al., 2022)等。这类研究技术门槛相对较低, 反馈逻辑清晰, 便于理解, 因而至今仍被研究者采用。然而, 其预先设计的回应内容, 难以应对复杂和细微的情绪的变化, 使得它们更适用于任务边界清晰的场景, 而在可迁移性方面存在一定局限。此外, 当前许多研究缺乏共情性回应的定义和标准。以 Guo 和 Goh (2016)的研究为例, 研究中所设计的共情回应是学生作答错误时展现鼓励性的微笑, 并提示“不要气馁, 仔细阅读一遍问题”。然而根据心理咨询中的情感反映(reflection)技术(Hill, 2009), 回应“这道题答错了, 可能会有一些气馁”, 相较于研究中的“不要气馁”, 或许更能体现共情。未来研究需要结合心理学理论, 制定更明确、合理的共情回应标准, 并通过实证研究检验回应的有效性。

2.4　共情对话生成

共情对话生成是利用计算机自动生成让人感到自己的情绪被理解和支持的对话。随着生成式 AI 的突破性发展作为技术保障, 这类研究主题正迅速成为热点, 研究者正探索不同策略以实现计算机自动生成共情对话的目标。一种策略是先识别沟通对象的情绪, 再根据上下文生成对于沟通对象意图、需要、影响、和愿望的常识判断, 整合后生成完整的共情性对话内容(Lin et al., 2019; Majumder et al., 2020; Sabour et al., 2022)。这样的做法符合共情包含情绪和认知双重过程的理论。另一类策略则无需事先分辨沟通对象的情绪类型, 而是直接生成共情对话。这类模型往往是在预训练大语言模型基础上, 用共情对话数据进行微调(fine-tune)后得到(Rashkin et al., 2019; Sharma et al., 2021)。如 Rashkin 等(2019)建立了一个包含约 25000 对共情对话的数据集(EmpatheticDialogues), 并利用这个数据集微调预先用 17 亿条论坛日常对话训练的模型, 结果表明微调模型生成的对话比预训练模型更具共情性。这个数据集成为此后众多研究的基准。此外, 也有研究结合多种策略以得到更好的共情回复, 如 Qian 等(2023)认为首先生成一个语意高相关的回复, 再用改写技术增加语句的共情性, 能得到更佳的效果。

共情对话生成研究正在蓬勃发展, 这些研究为人机交互提供更加人性化的体验, 也为其他应用场景奠定了基础, 如改写网络回复内容以增加共情等。以 ChatGPT 为代表的大语言模型的突破性进展, 使得生成共情内容越来越简单, 进一步加速了该领域的发展。尽管如此, 当前研究主要集中于单轮或较短的对话, 这与现实任务中所需的复杂多轮对话相比, 仍存在明显差距。未来的研究应关注将共情对话有机融入各种对话场景, 以满足实际任务需求。例如, 如何在共情和任务

目标之间灵活切换并维持平衡, 确保在支持用户情感表达的同时, 避免偏离对话的主要目标。此外, 如何将共情对话与声音、表情、动作生成结合, 建立人形对话代理(embodied conversational agent)也是未来的研究方向(Loveys et al., 2022)。

上述不同主题的共情计算研究相互借鉴、彼此促进, 共同推动着这个领域的发展。如 Sharma 等在开发共情内容生成的系统时(Sharma et al., 2021), 使用了之前开发的文本共情分类系统(Sharma et al., 2020)来评估新生成的内容, 经过多轮生成−评估的强化学习过程, 筛选出最适合的共情文本。此外, 一些新兴的研究也正在涌现, 如通过混合现实(Mixed Reality, MR)促进人与人之间的共情(Jing et al., 2022), 以及脑机接口提升计算机共情表现(Roshdy et al., 2021)等方面的初步探索。尽管尚未形成体系, 这些研究为共情计算的发展开辟了新的可能性。综上, 共情计算的研究前景广阔, 也伴随诸多挑战, 未来需要继续提升准确性和跨场景适用性。

3 应用场景

尽管尚处于起步阶段, 但现有研究已经展现出共情计算具有广泛的应用前景, 特别是在心理健康、教育学习、商业服务和公共管理等领域积累了较多相关研究。这些领域涉及大量人际互动, 在人工智能日益普及的背景下, 也将成为人机交互最频繁的领域, 因而是共情计算的主要应用场景。由于都涉及通过共情计算, 提升人与人、人与机器互动过程中对情感的理解和回应能力, 这些领域并非完全独立, 而是相互关联却各有侧重的。随着技术的进步, 未来还可能出现新的应用场景, 如将共情计算应用于游戏提升娱乐体验等。

3.1 心理健康

目前, 对于共情计算在心理咨询场景中的研究最为充分。高共情水平的咨询师与来访者有更好的治疗关系, 也能降低他们的脱落率(Moyers & Miller, 2013)。然而, 过去的测评依赖于自我报告或者视频编码, 不客观也不容易大规模推广。共情计算的方法通过文本、声音、视频等自动化评估咨询师的共情。这样可应用于在更大范围内研究咨询过程与效果的关系, 也可应用于选拔咨询师, 或者在培养咨询师过程中反馈并提升其共情能力(Xiao, Imel, Georgiou, et al., 2015)。

另一方面, 共情计算也可以辅助线上心理健康服务。例如, 朋辈心理互助平台借助共情计算所生成的内容辅助心理支持的工作(Sharma et al., 2023)。Liu 等(2021)也尝试将共情对话系统与心理咨询中的助人技术结合, 设计更有效帮助人们应对压力和挑战的对话系统。研究者认为, 增加数字干预系统表现出的共情水平, 有助于促进形成数字治疗联盟(digital therapeutic alliance), 进而提升干预效果(Tong et al., 2022)。不过当前研究只是初步验证了用计算机自动生成共情性对话的可行性, 鉴于心理咨询中可能涉及自杀自伤等风险因素, 目前的研究距离完成整个心理咨询过程并产生治疗效果仍存在差距。

3.2 教育学习

随着计算机和手机等智能终端的普及, 越来越多的人开始关注通过数字游戏、机器人等方法辅助教学。在教育领域中, 共情计算帮助我们设计出能够更好地理解和应对学生情绪和需求的教育机器人。学习不仅仅是认知过程, 情绪在其中同样扮演重要角色(Camacho-Morles et al., 2021)。教师的共情与学生的学习动机、投入程度、满意度和学业表现均呈现正向关系(Cornelius-White, 2007; Roorda et al., 2011)。于是很早就有研究表明计算机的共情反馈提升学习效果, 不过限于当时的技术条件, 还需要由学生自主报告情绪(Chen et al., 2012)。共情计算使得自动化识别情绪变为可能。研究表明, 具有一定共情能力的计算系统缓冲愤怒情绪给学习者带来的不利影响, 提升学生的认知能力和创造性(Groh et al., 2022)。当学生从数字助教那里感受到更多的共情反馈时, 其在数学问题上的正确率更高(Oker et al., 2020)。不过还需要更多研究探索这些影响的条件、内部机制等。比如在实验情景中, 共情机器人促进小组学习和讨论, 但是在真实教育环境的长期研究并未发现明显的效果(Alves-Oliveira et al., 2019)。系统性综述的研究也表明, 情感代理能够有效提升学习者的积极情绪和内部动机, 但对学习效果的促进作用不如情绪提升的效果明显(王燕青 等, 2022)。除了直接帮助学习者, 共情计算也可以通过实时反馈等方式, 为老师或家长提供辅助, 帮助他们在面对学生、子女教育时表现得更共情(Ge et al., 2021; Meyers et al., 2019)。总之, 共情计算在教育领域展现了应用潜力, 不过仍需要更多研

究验证其在真实场景中的效果。

3.3 商业服务

人工智能客服和服务机器人正在越来越多被用于商业领域, 共情计算在服务过程中发挥了辅助效果。传统人工智能客服主要关注为客户提供信息支持, 但研究发现, 在一个社交媒体平台上40%的用户提问并不是寻求具体信息, 而是需要情感支持, 例如抱怨(Xu et al., 2017)。尽管设计者尽力提升人工智能服务的准确性, 在现有技术条件下, 计算机在服务中还是不可避免地会出现失误。这会引发客户的挫折感和不信任, 从而降低客户对人工智能服务的接受程度。共情计算的应用一定程度上缓解这种问题。研究发现, 当 AI 服务失败时, 共情性回应增加继续使用 AI 服务的意向(Lv et al., 2022)。另一项针对服务机器人的研究表明, 当服务机器人无法完全满足客户需求时, 表现出共情能力的机器人被认为更有帮助, 用户体验也更好(Tojib et al., 2023)。未来还会有越来越多的机器人出现在工业或商业领域。人们面对这样的情景可能会产生挫败和敌意。共情计算的应用会让人类觉得机器人更像人类、更友好, 以缓解对机器人的抵触。

3.4 公共管理

共情计算也可以应用于公共管理领域, 如网络空间的治理。网络已成为人们生活密不可分的部分, 线上生活影响着线下的健康(Zhang & Centola, 2019)。然而网上的攻击性言论、网络暴力等也成为了需要治理的问题。一项干预研究表明, 共情性反驳信息减少仇恨言论(Hangartner et al., 2021)。通过共情计算系统自动识别不良言论并生成共情性反驳信息, 可能有效减少来自网络空间的暴力。此外, 共情也是把双刃剑。有些个体会由于过度共情社交媒体上的负面信息, 而给自己带来不适。共情计算监控有这种致病性共情风险的人并给与提示或干预, 有利于健康的社会心态(Abdul-Mageed et al., 2017)。最后, 共情计算也用于评估和修改公共管理部门的政务回复信息, 或在社交媒体上公布的信息, 提升共情程度, 帮助公共部门与民众增进信任, 减少负面舆论。公共管理事务有广泛的责任范围和显著的社会影响, 因此不能完全依赖计算机自动完成, 不过共情计算作为其中一个环节, 可以辅助提升管理效率和质量。

以上 4 个场景展现出共情计算的广泛应用潜力。然而, 如同自动驾驶技术一样, 由于牵涉安全性和伦理等问题, 目前并不能完全依赖计算机执行共情任务, 而需要人机紧密协作。因此, 除了完善技术外, 未来应更加注重以人为中心, 面向特定应用场景, 探索如何更有效实现共情计算。

4 研究不足

尽管共情计算已经有许多有益的研究进展, 但作为一个新兴研究领域仍存在许多不足。未来研究通过引入心理学理论和研究方法予以改善。

4.1 整体性的共情概念模型

当前共情计算研究主要由计算机领域研究者推动, 这些研究有些缺乏对共情的明确定义, 有些则基于算法准确性考虑, 仅选取一个具体的行为特征来代表整体的共情。这样的做法导致不同研究之间难以进行有效比较和整合, 阻碍了研究者对共情计算整体性的理解和应用。心理与行为科学、神经科学等在共情的基础概念研究方面积累了丰富成果, 对共情的层次和类别有详细分析, 例如从神经心理过程分为认知过程、情感过程以及认知情感共同激活的中间过程(Schurz et al., 2021), 从情绪对象角度分为对消极情绪的共情和对积极情绪的共情等(Brett et al., 2023)。未来的研究应重视利用这些研究成果, 更准确地操作性定义共情, 并在此基础上, 逐渐发展出更整合的共情计算模型。Yalçın 和 DiPaola (2020)初步提出了一个整合的模型, 指出计算机需要从沟通能力、情绪调控和认知机制三个层级模拟共情。不过该模型还只是一个概念模型, 仍需足够的实证研究支持。此外, 新时代产生了许多新的共情现象, 比如在线聊天时用表情包表达共情等。整体性的共情计算模型也应重视将这些新形式的共情纳入考虑。

4.2 高质量的共情数据集

共情计算需要大量数据的支持, 数据集的质量对模型精度至关重要。但目前公开的用于共情研究的数据集比较缺乏, 而且标注质量也有待提高。如最常被引用的共情对话数据集 Empathetic Dialogues 并非基于真实场景中的对话, 而是要求参与者根据研究者事先提供情绪标签, 想象感受到这种情绪的场景, 并完成对话(Rashkin et al., 2019)。这样收集的数据生态效度较差。在中文研究领域, 更是缺乏相关数据集。已有的心理数据

集 PsyQA (Sun et al., 2021)和《心理咨询问答语料库》(Wang et al., 2020)缺乏共情相关的标注。由于缺少中文数据集，中文领域共情计算的研究数量也相对较少，这让现有研究很难涉及到文化差异有关问题。因此，未来的研究应借鉴心理学研究方法，通过问卷调查、半结构化访谈、情景模拟、临床数据转录等方法收集多维度、多层次的数据(做法可参考 Omitaomu et al., 2022)。数据标注方面，计算机领域研究多使用众包方法(彭凯平等，2018)。这种方法的参与者往往仅经过简单的培训，并不具备相关领域的专业知识。专业的心理学研究人员对数据进行筛选、加工和标注将提升数据集可信度。

4.3 以人为中心的评价标准

现有的共情计算研究的评价更多是围绕任务表现展开，预测数据集标注的准确率或者生成的共情对话是否比基线模型更好。如 Sabour 等(2022)的研究中，仅招募 3 人比较了 100 组不同系统生成的共情对话内容，以此对比哪个系统更好。这种评价方式很难保证结果的稳定性和解决实际问题中的有效性。未来研究需要在任务表现评价基础上，提升至以人为中心的评价标准。以心理与行为科学为指导，通过访谈、问卷、实验等研究方法考察共情计算系统对于使用者的有效性(effectiveness)、高效性(efficiency)与效能性(efficacy)，并提出改进建议。Sharma 等(2023)的最新研究中开始使用随机对照实验对比人类和人类+人工智能两种系统回复中的共情。不过作者也承认，该研究中所使用的评价标准仍只是第三方视角下对话内容所展现出的共情，而非寻求帮助者本人真实感受到的共情。这反映出共情计算研究发展到现阶段开始重视改进评价标准，正亟需心理学提供支持。

5 展望和讨论

5.1 共情计算为心理学带来新的洞见

共情计算不仅为心理学研究提供新工具，更在理论层面深化和拓展对共情的理解。通过计算机自动分析和模拟共情互动的数据，研究者可以在更大的数据尺度上验证和迭代先前通过传统方法(观察、问卷和实验等)获得的研究成果，也可以识别尚未发现的模式，如有研究使用共情计算来发掘对话中的潜在共情意图(Chen et al., 2022)。这有助于加深对共情影响因素和内在机制的理解。

另一方面，共情计算也拓展共情研究的理论视角。先前的研究表明，共情普遍存在于人类甚至动物中(de Waal, 2012)，是从亲密关系到大规模合作等一切社会互动基础(Zaki, 2014)。然而，在未来的智能社会中，人们的面对面互动正逐渐演化为面向网络和人工智能的新型互动，以至于人们有时不能直接看到对方的表情和肢体动作，不能直观地感受沟通对象的情绪。在这样的情景下，共情是否仍有普遍性，会发生哪些演变，以及如何促进人机交互中的共情等，都是亟待深入研究的问题(如图 3 所示)。共情计算将为我们理解智能社会互动过程中的共情现象提供基础，为建立包括人−人关系、人−机关系等在内的整体性、普适性的共情理论提供启示。

在这些理论创新的基础上，比照以往研究中促进人类共情的方式，开发模拟人类共情和社交互动的虚拟代理(virtual agent)和社交机器人也为心理学应用提供新的方向。通过开发具备共情能力的虚拟代理和社交机器人，可以为心理健康和教育等领域提供更智能、个性化的支持，推动心理学理论和应用的双向促进发展。

5.2 心理学家在共情计算研究中的角色

心理学家在共情计算领域发挥不可或缺的作用。编程等技术门槛曾是许多心理学研究者参与这一领域的阻碍，不过越来越多介绍计算方法的文章(苏悦 等，2022)正在消除这一障碍。大语言模型进一步降低了研究过程对编程技术的依赖，使得研究者可以通过 AI 辅助编程、提示词工程等方法开展研究。这种背景下，心理学家在共情领域积累的丰富理论和研究方法的重要性正在上升。心理学家的首要任务是提出问题。早期个体共情测量研究起源于改善咨询师共情水平的需求，未来心理学家需要持续关注共情计算的创新研究问题。其次，心理学家为设计和提升共情计算提供理论支撑，如心理治疗理论已被用于共情内容改写算法的开发(Lin et al., 2023)，未来心理学理论也可用于优化提示词工程(Li et al., 2023)，提升大语言模型在共情任务中的表现。最后，随着 AI 的行为和心理表现越来越接近人类，研究者开始依据心理测量原理，制定科学有效的评估方法，以测量机器的共情水平(Kosinski, 2023; Wang et al., 2023)。总之，心理学家在共情计算研究和应

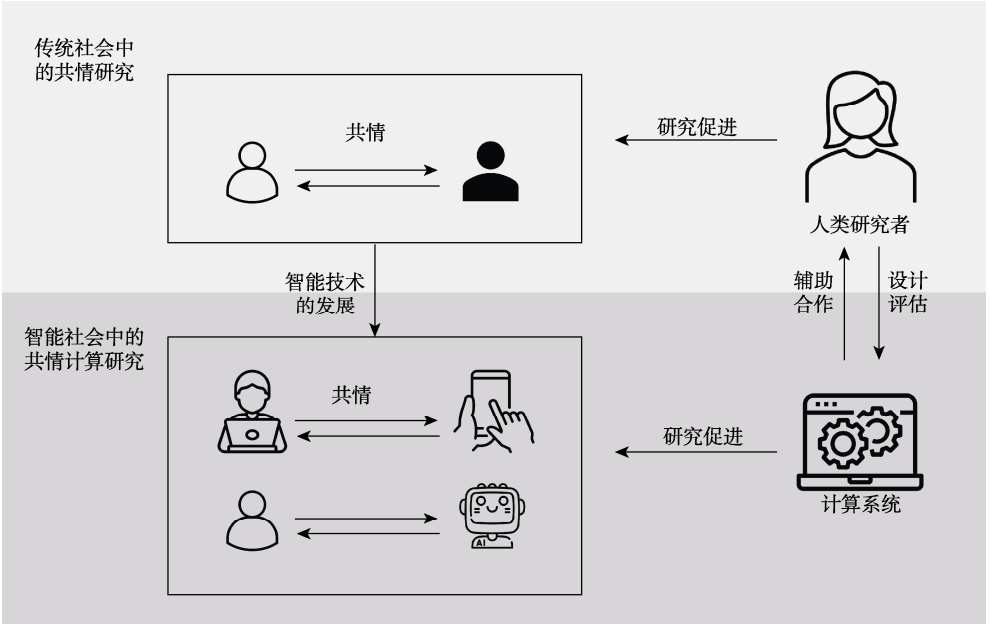

图 3　从传统共情研究到智能社会共情计算的变化示意图

用过程中承担引领、评估和改进等角色, 将与计算机科学家一起, 使共情计算的理论基础更坚实, 效果检验更可靠, 应用创新更贴近实际需求。

5.3　共情计算系统的接纳度及其伦理风险

共情计算研究和应用所涉及的法律和伦理风险也应被充分讨论, 并尽量通过政策机制在系统层面加以避免。AI 的飞速发展一方面为生活带来了便利, 另一方面也引发了与 AI 相关的信任问题和道德困境(Awad et al., 2018)。研究表明, 人们感受到的机器人的共情作为中介变量增加对 AI 的信任(Pelau et al., 2021)。鉴于此, 共情计算的研究进步可能也成为双刃剑。一方面提升 AI 系统的共情能力, 能够提升用户信任和体验; 另一方面, 这种信任也可能被滥用, 例如在营销和广告中使用共情计算技术来影响人们的决策。此外, 人类如何对待具备共情能力的人工智能系统? 是否要像对待其他人类一样考虑 AI 的感受和幸福(Lomas, 2023)? 这些议题都是在今后的研究中需要探讨的。

以 ChatGPT 等大语言模型为代表的超级人工智能技术的发展为共情研究带来了新机遇和新挑战。共情计算的交叉研究不仅是用一种新技术(如机器学习或自然语言处理)测量一个心理概念(Li et al., 2019), 而是持续为人类探索未知生产了新知识、新方法和新应用。这一广阔前景的交叉领域, 不仅拓展和加深对于共情心理机制的理解, 也可以将研究成果运用于智能社会的发展, 提升个人和社会幸福, 值得心理学研究者关注并做出独特的贡献。

参考文献

彭凯平, 刘世群, 倪士光. (2018). 移动互联网时代的社会科学研究工具: 众包的争议与发展. *西北师大学报 (社会科学版), 55*(3), 113−123. https://doi.org/10.16783/j.cnki.nwnus.2018.03.015

苏悦, 刘明明, 赵楠, 刘晓倩, 朱廷劭. (2022). 基于社交媒体数据的心理指标识别建模: 机器学习的方法. *心理科学进展, 29*(4), 571−585. https://doi.org/10.3724/sp.J.1042.2021.00571

王燕青, 龚少英, 姜甜甜, 吴亚男. (2022). 情感代理能否提高多媒体学习的效果? *心理科学进展, 30*(7), 1524−1535. https://doi.org/10.3724/SP.J.1042.2022.01524

岳童, 黄希庭. (2016). 共情特质的神经生物学基础. *心理科学进展, 24*(9), 1368−1376. https://doi.org/10.3724/SP.J.1042.2016.01368

Abdul-Mageed, M., Buffone, A., Peng, H., Giorgi, S., Eichstaedt, J., & Ungar, L. (2017). Recognizing pathogenic empathy in social media. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1), 448−451. https://doi.org/10.1609/icwsm.v11i1.14942

Alves-Oliveira, P., Sequeira, P., Melo, F. S., Castellano, G., & Paiva, A. (2019). Empathic robot for group learning.

ACM Transactions on Interactive Intelligent Systems, 8(1), 1−34. https://doi.org/10.1145/3300188

Asada, M. (2015). Towards artificial empathy. *International Journal of Social Robotics, 7*(1), 19−33. https://doi.org/10.1007/s12369-014-0253-z

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59−64. https://doi.org/10.1038/s41586-018-0637-6

Barriere, V., Tafreshi, S., Sedoc, J., & Alqahtani, S. (2022). WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis* (pp. 214−227). Association for Computational Linguistics.

Bošnjaković, J., & Radionov, T. (2018). Empathy: Concepts, theories and neuroscientific basis. *Alcoholism and Psychiatry Research, 54*(2), 123−150. https://doi.org/10.20471/dec.2018.54.02.04

Brett, J. D., Becerra, R., Maybery, M. T., & Preece, D. A. (2023). The psychometric assessment of empathy: Development and validation of the Perth Empathy Scale. *Assessment*, *30*(4), 1140−1156. https://doi.org/10.1177/10731911221086987

Buechel, S., Buffone, A., Slaff, B., Ungar, L., & Sedoc, J. (2018). Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4758−4765). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1507

Cai, Y. (2006). Empathic computing. In Y. Cai, & J. Abascal (Eds.), *Ambient intelligence in everyday life* (pp. 67−85). Springer Berlin Heidelberg. https://doi.org/10.1007/11825890_3

Camacho-Morles, J., Slemp, G. R., Pekrun, R., Loderer, K., Hou, H., & Oades, L. G. (2021). Activity achievement emotions and academic performance: A meta-analysis. *Educational Psychology Review, 33*(3), 1051−1095. https://doi.org/10.1007/s10648-020-09585-3

Cao, S., Fu, D., Yang, X., Wermter, S., Liu, X., & Wu, H. (2021). *Can AI detect pain and express pain empathy? A review from emotion recognition and a human-centered AI perspective*. arXiv. https://doi.org/10.48550/arXiv.2110.04249

Chakravarthula, S. N., Xiao, B., Imel, Z. E., Atkins, D. C., & Georgiou, P. (2015). Assessing empathy using static and dynamic behavior models based on therapist's language in addiction counseling. In *Interspeech-2015* (pp. 668−672). https://doi.org/10.21437/Interspeech.2015-237

Chen, G.-D., Lee, J.-H., Wang, C.-Y., Chao, P.-Y., Li, L.-Y., & Li, T.-Y. (2012). An empathic avatar in a computer-aided learning program to encourage and persuade learners. *Journal of Educational Technology & Society, 15*(2), 62−72. http://www.jstor.org/stable/jeductechsoci.15.2.62

Chen, M. Y., Li, S., & Yang, Y. (2022). EmpHi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1063−1074). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.78

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research, 77*(1), 113−143. https://doi.org/10.3102/003465430298563

Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2014). Empathy: A Review of the concept. *Emotion Review, 8*(2), 144−153. https://doi.org/10.1177/1754073914558466

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113−126. https://doi.org/10.1037/0022-3514.44.1.113

de Waal, F. B. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology, 59*, 279−300. https://doi.org/10.1146/annurev.psych.59.103006.093625

de Waal, F. B. M. (2012). The antiquity of empathy. *Science, 336*(6083), 874−876. https://doi.org/10.1126/science.1220999

D'mello, S., & Graesser, A. (2013). AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems, 2*(4), Article 23. https://doi.org/10.1145/2395123.2395128

Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2011). Empathy. *Psychotherapy, 48*(1), 43−49. https://doi.org/10.1037/a0022187

Gaume, J., Hallgren, K. A., Clair, C., Schmid Mast, M., Carrard, V., & Atkins, D. C. (2019). Modeling empathy as synchrony in clinician and patient vocally encoded emotional arousal: A failure to replicate. *Journal of Counseling Psychology, 66*(3), 341−350. https://doi.org/10.1037/cou0000322

Ge, Y., Li, W., Chen, F., Kayani, S., & Qin, G. (2021). The theories of the development of students: A factor to shape teacher empathy from the perspective of motivation. *Frontiers in Psychology, 12*, Article 736656. https://doi.org/10.3389/fpsyg.2021.736656

Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. (2016). A deep learning approach to modeling empathy in addiction counseling. In *Interspeech-2016* (pp. 1447−1451). https://doi.org/10.21437/Interspeech.2016-554

Gibson, J., Malandrakis, N., Romero, F., Atkins, D. C., & Narayanan, S. S. (2015). Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Interspeech-2015* (pp. 1947−1951). https://doi.org/10.21437/interspeech.2016-554

Groh, M., Ferguson, C., Lewis, R., & Picard, R. W. (2022).

Computational empathy counteracts the negative effects of anger on creative problem solving. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1−8). IEEE. https://doi.org/10.1109/ACII55700.2022.9953869

Grühn, D., Rebucal, K., Diehl, M., Lumley, M., & Labouvie-Vief, G. (2008). Empathy across the adult lifespan: Longitudinal and experience-sampling findings. *Emotion, 8*(6), 753−765. https://doi.org/10.1037/a0014123

Guo, Y. R., & Goh, D. H.-L. (2016). Evaluation of affective embodied agents in an information literacy game. *Computers & Education, 103*, 59−75. https://doi.org/10.1016/j.compedu.2016.09.013

Hall, J. A., & Schwartz, R. (2019). Empathy present and future. *The Journal of Social Psychology, 159*(3), 225−243. https://doi.org/10.1080/00224545.2018.1477442

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., ... Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences, 118*(50), Article e2116310118. https://doi.org/10.1073/pnas.2116310118

Hill, C. E. (2009). *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

Imel, Z. E., Barco, J. S., Brown, H. J., Baucom, B. R., Baer, J. S., Kircher, J. C., & Atkins, D. C. (2014). The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of Counseling Psychology, 61*(1), 146−153. https://doi.org/10.1037/a0034943

Ishii, R., Otsuka, K., Kumano, S., Higashinaka, R., & Tomita, J. (2018). Analyzing gaze behavior and dialogue act during turn-taking for estimating empathy skill level. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (pp. 31−39). Association for Computing Machinery. https://doi.org/10.1145/3242969.3242978

Jing, A., Gupta, K., McDade, J., Lee, G. A., & Billinghurst, M. (2022). Comparing gaze-supported modalities with empathic mixed reality interfaces in remote collaboration. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 837−846). IEEE. https://doi.org/10.1109/ismar55827.2022.00102

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of Adolescence, 29*(4), 589−611. https://doi.org/10.1016/j.adolescence.2005.08.010

Khanpour, H., Caragea, C., & Biyani, P. (2017). Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Volume 2: Short Papers) (pp. 246−251). Asian Federation of Natural Language Processing. https://aclanthology.org/I17-2042

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv*. https://doi.org/10.48550/arXiv.2302.02083

Kumano, S., Otsuka, K., Mikami, D., & Yamato, J. (2011). Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (pp. 43−50). IEEE. https://doi.org/10.1109/FG.2011.5771440

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. https://doi.org/10.1016/j.ijhcs.2012.09.005

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). *Large language models understand and can be enhanced by emotional stimuli*. arXiv. https://doi.org/10.48550/arXiv.2307.11760

Li, S., Lu, S., Ni, S., & Peng, K. (2019). Identifying psychological resilience in Chinese migrant youth through multidisciplinary language pattern decoding. *Children and Youth Services Review, 107*, Article 104506. https://doi.org/10.1016/j.childyouth.2019.104506

Lin, S., Lin, L., Hou, C., Chen, B., Li, J., & Ni, S. (2023). *Empathy-based communication framework for chatbots: A mental health chatbot application and evaluation.* 11th International Conference on Human-Agent Interaction, Gothenburg, Sweden.

Lin, Z., Madotto, A., Shin, J., Xu, P., & Fung, P. (2019). MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 121−132). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1012

Litvak, M., Otterbacher, J., Ang, C. S., & Atkins, D. (2016). Social and linguistic behavior and its correlation to trait empathy. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)* (pp. 128−137). The COLING 2016 Organizing Committee. https://aclanthology.org/W16-4314

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., & Huang, M. (2021). Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 3469−3483). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.269

Lomas, T. (2023). Stranger than we can imagine: The possibility and potential significance of non-human forms of consciousness and wellbeing. *The Journal of Positive Psychology*, *18*(6), 807−826. https://doi.org/10.1080/17439760.2022.2131608

Loveys, K., Sagar, M., Billinghurst, M., Saffaryazdi, N., & Broadbent, E. (2022). Exploring empathy with digital humans. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 233−237). IEEE. https://doi.org/10.1109/VRW55335.2022.00055

Lv, X., Yang, Y., Qin, D., Cao, X., & Xu, H. (2022). Artificial intelligence service recovery: The role of empathic response in hospitality customers' continuous usage intention. *Computers in Human Behavior, 126*. https://doi.org/10.1016/j.chb.2021.106993

Ma, F., Li, Y., Ni, S., Huang, S.-L., & Zhang, L. (2022). Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN. *Applied Sciences, 12*(1), Article 527. https://doi.org/10.3390/app12010527

Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A., Mihalcea, R., & Poria, S. (2020). MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8968−8979). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.721

Mathur, L., Spitale, M., Xi, H., Li, J., & Matarić, M. J. (2021). Modeling user empathy elicited by a robot storyteller. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1−8). IEEE. https://doi.org/10.1109/ACII52823.2021.9597416

Meyers, S., Rowell, K., Wells, M., & Smith, B. C. (2019). Teacher empathy: A model of empathy for teaching for student success. *College Teaching, 67*(3), 160−168. https://doi.org/10.1080/87567555.2019.1579699

Mooradian, T. A., Davis, M., & Matzler, K. (2011). Dispositional empathy and the hierarchical structure of personality. *The American Journal of Psychology, 124*(1), 99−109. https://doi.org/10.5406/amerjpsyc.124.1.0099

Morelli, S. A., Ong, D. C., Makati, R., Jackson, M. O., & Zaki, J. (2017). Empathy and well-being correlate with centrality in different social networks. *Proceedings of the National Academy of Sciences, 114*(37), 9843−9847. https://doi.org/10.1073/pnas.1702155114

Moyers, T. B., & Miller, W. R. (2013). Is low therapist empathy toxic? *Psychology of Addictive Behaviors, 27*(3), 878−884. https://doi.org/10.1037/a0030274

Oker, A., Pecune, F., & Declercq, C. (2020). Virtual tutor and pupil interaction: A study of empathic feedback as extrinsic motivation for learning. *Education and Information Technologies, 25*(5), 3643−3658. https://doi.org/10.1007/s10639-020-10123-5

Omitaomu, D., Tafreshi, S., Liu, T., Buechel, S., Callison-Burch, C., Eichstaedt, J., Ungar, L., & Sedoc, J. (2022). Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv*. https://doi.org/10.48550/arXiv.2205.12698

OpenAI, J. A., Steven, A., Sandhini, A., Lama, A., Ilge, A., Florencia, L. A., ... Barret, Z. (2023). *GPT-4 technical report*. Retrieved March 01, 2023, from https://ui.adsabs.harvard.edu/abs/2023arXiv230308774O

Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior, 122*, Article 106855. https://doi.org/10.1016/j.chb.2021.106855

Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences, 25*(1), 1−20. https://doi.org/10.1017/S0140525X02000018

Qian, Y., Wang, B., Ma, S., Bin, W., Zhang, S., Zhao, D., Huang, K., & Hou, Y. (2023). Think twice: A human-like two-stage conversational agent for emotional response generation. *arXiv*. https://doi.org/10.48550/arXiv.2301.04907

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5370−5381). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1534

Reniers, R. L. E. P., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment, 93*(1), 84−95. https://doi.org/10.1080/00223891.2010.528484

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher−student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81*(4), 493−529. https://doi.org/10.3102/0034654311421793

Roshdy, A., Kork, S. A., Karar, A., Sabi, A. A., Barakeh, Z. A., ElSayed, F., Beyrouthy, T., & NAIT-ALI, A. (2021). Machine empathy: Digitizing human emotions. In *2021 International Symposium on Electrical, Electronics and Information Engineering* (pp. 307−311). Association for Computing Machinery. https://doi.org/10.1145/3459104.3459154

Sabour, S., Zheng, C., & Huang, M. (2022). CEM: Commonsense-aware empathetic response generation. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(10), 11229−11237. https://doi.org/10.1609/aaai.v36i10.21373

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin, 147*(3), 293−327. https://doi.org/10.1037/bul0000303

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff,

T. (2021). Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021* (pp. 194−205). Association for Computing Machinery. https://doi.org/10.1145/3442381.3450097

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human−AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence, 5*(1), 46−57. https://doi.org/10.1038/s42256-022-00593-2

Sharma, A., Miner, A. S., Atkins, D. C., & Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 5263−5276). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.425

Stosic, M. D., Fultz, A. A., Brown, J. A., & Bernieri, F. J. (2022). What is your empathy scale not measuring? The convergent, discriminant, and predictive validity of five empathy scales. *The Journal of Social Psychology, 162*(1), 7−25. https://doi.org/10.1080/00224545.2021.1985417

Sun, H., Lin, Z., Zheng, C., Liu, S., & Huang, M. (2021). PsyQA: A Chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 1489−1503). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.130

Terzis, V., Moridis, C. N., & Economides, A. A. (2012). The effect of emotional feedback on behavioral intention to use computer based assessment. *Computers & Education, 59*(2), 710−721. https://doi.org/10.1016/j.compedu.2012.03.003

Tojib, D., Abdi, E., Tian, L., Rigby, L., Meads, J., & Prasad, T. (2023). What's best for customers: Empathetic versus solution-oriented service robots. *International Journal of Social Robotics, 15,* 731−741. https://doi.org/10.1007/s12369-023-00970-w

Tong, F., Lederman, R., D'Alfonso, S., Berry, K., & Bucci, S. (2022). Digital therapeutic alliance with fully automated mental health smartphone apps: A narrative review. *Frontiers in Psychiatry, 13*, Article 819623. https://doi.org/10.3389/fpsyt.2022.819623

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000−6010). Curran Associates Inc. https://dl.acm.org/doi/10.5555/3295222.3295349

Wambsganss, T., Soellner, M., Koedinger, K. R., & Leimeister, J. M. (2022). Adaptive empathy learning support in peer review scenarios. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1−17). https://doi.org/10.1145/3491102.3517740

Wang, H. L., Wu, Z. Z., & Lang, J. Y. (2020). *Emotional first aid dataset* [Data set]. https://github.com/chatopera/efaqa-corpus-zh

Wang, X., Li, X., Yin, Z., Wu, Y., & Liu Jia. (2023). Emotional intelligence of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2307.09042

Xiao, B., Bone, D., Van Segbroeck, M., Imel, Z. E., Atkins, D. C., Georgiou, P. G., & Narayanan, S. S. (2014). Modeling therapist empathy through prosody in drug addiction counseling. In *Interspeech-2014* (pp. 213−217). https://doi.org/10.21437/Interspeech.2014-55

Xiao, B., Can, D., Georgiou, P. G., Atkins, D., & Narayanan, S. S. (2012). Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference* (pp. 1−4). IEEE.

Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P. G., & Narayanan, S. S. (2015). Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. In *Interspeech-2015* (pp. 2489−2493). https://doi.org/10.21437/Interspeech.2015-537

Xiao, B., Imel, Z. E., Georgiou, P., Atkins, D. C., & Narayanan, S. S. (2016). Computational analysis and simulation of empathic behaviors: A survey of empathy modeling with behavioral signal processing framework. *Current Psychiatry Reports, 18*(5), 1−11. https://doi.org/10.1007/s11920-016-0682-5

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). "Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS One, 10*(12), Article e0143055. https://doi.org/10.1371/journal.pone.0143055

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3506−3510). Association for Computing Machinery. https://doi.org/10.1145/3025453.3025496

Yalçın, Ö. N., & DiPaola, S. (2020). Modeling empathy: Building a link between affective and cognitive processes. *Artificial Intelligence Review, 53*(4), 2983−3006. https://doi.org/10.1007/s10462-019-09753-0

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin, 140*(6), 1608−1647. https://doi.org/10.1037/a0037679

Zhang, J., & Centola, D. (2019). Social networks and health: New developments in diffusion, online and offline. *Annual Review of Sociology, 45*(1), 91−109. https://doi.org/10.1146/annurev-soc-073117-041421

Zhou, K., Aiello, L. M., Scepanovic, S., Quercia, D., & Konrath, S. (2021). The language of situational empathy. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1−19. https://doi.org/10.1145/3449087

Zhu, S., Yu, T., Xu, T., Chen, H., Dustdar, S., Gigan, S., ... Pan, Y. (2023). Intelligent computing: The latest advances, challenges, and future. *Intelligent Computing, 2*, Article 0006. https://doi.org/10.34133/icomputing.0006

When AI learns to empathize: Topics, scenarios, and optimization of empathy computing from a psychological perspective

HOU Hanchao, NI Shiguang, LIN Shuya, WANG Pusheng

(*Shenzhen International Graduate School, Tsinghua University, Shenzhen* 518055, *China*)

Abstract: Empathy computing is an emerging interdisciplinary research field that leverages artificial intelligence (AI) and big data technology to predict, identify, simulate, and generate human empathy. It is an emerging research field that intersects traditional empathy studies in psychology with computer science. This study constructed a ubiquitous research framework that comprises data, model and task layers, and summarized an analytical framework of four new topics including individual empathy assessment, empathetic content classification, empathic response system, and empathetic dialogue generation. Scenario innovations in applied psychology were discussed, such as mental health, education and learning, business services, and public management. Future research should focus on developing integrated theoretical models of empathy computing, establishing reliable psychological and behavioral datasets of empathy-related characteristics, and validating and refining empathy computing research through a human-centered evaluation system. Empathy computing extends current research on empathy in interpersonal relationships to its novel forms in human-AI relationships in an intelligent society. Psychologists play crucial roles in leading, evaluating, and optimizing research and practice in this field. They collaborate closely with computer scientists to advance the theoretical foundations, enhance human-centered evaluation, and drive practical innovations in empathy computing.

Keywords: empathy, empathy computing, computational psychology, artificial intelligence, human-computer interaction