Advances in Psychology 心理学进展, 2026, 16(2), 84-94
Published Online February 2026 in Hans. https://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2026.162064

Hans汉斯

人工智能时代机器人情绪加工的机遇与挑战

李嘉润，申权威*

湖北医药学院应用心理学系，湖北 十堰

收稿日期：2025年12月16日；录用日期：2026年1月16日；发布日期：2026年2月3日

摘　要

当前，人工智能的发展重心正逐渐由符号计算转向情感计算。在这一背景下，机器人情绪加工已成为人机交互领域的关键研究方向。本文从心理学与计算科学的双重视角出发，分析了机器人情绪加工过程中(注意、识别、评估与响应)的计算机制，描述了机器人情绪加工的应用场景，最后，讨论了情感计算在实现真正情绪理解方面的局限与挑战。未来研究需在技术与伦理之间构建平衡，推动人机情感交互向更具人文关怀的方向发展，在善用科技与科技向善道路上行稳致远。

关键词

人工智能，情感计算，情绪加工，应用场景

Opportunities and Challenges of Emotional Processing for Robots in the Era of Artificial Intelligence

Jiarun Li, Quanwei Shen*

Department of Applied Psychology, Hubei University of Medicine, Shiyan Hubei

Received: December 16, 2025; accepted: January 16, 2026; published: February 3, 2026

Abstract

At present, the development focus of artificial intelligence is gradually shifting from symbolic computation to affective computing. Against this backdrop, the processing of robot emotions has become a key research direction in the field of human-computer interaction. This paper, from the dual perspectives of psychology and computational science, analyzes the computational mechanisms in the process of robot emotion processing (attention, recognition, evaluation, and response), describes

*通讯作者。

文章引用: 李嘉润, 申权威(2026). 人工智能时代机器人情绪加工的机遇与挑战. *心理学进展, 16(2),* 84-94.
DOI: 10.12677/ap.2026.162064

the application scenarios of robot emotion processing, and finally discusses the limitations and challenges of affective computing in achieving true emotional understanding. Future research needs to strike a balance between technology and ethics, promoting the development of human-computer emotional interaction in a more humanistic direction, and steadily advancing on the path of making good use of technology and making technology beneficial to humanity.

Keywords

Artificial Intelligence, Affective Computing, Emotion Processing, Application Scenarios

1. 引言

1997 年，罗莎琳德·皮卡德(Rosalind Picard)通过其专著《情感计算》开创性地提出“情感计算”(affective computing)这一概念(Picard, 1997)。该概念旨在使计算机具备识别、理解、表达以及适应人类情感的功能。中国研究者进一步提出情感计算的功能在于构建更为融洽的人机互动模式，以此推动人工智能向更高层次演进(吴静，董屹泽，2025)。

人类对情绪的研究源远流长。从达尔文的《人类与动物的表情》到保罗·艾克曼提出的“七大基本表情”再到罗莎琳·皮卡德正式提出“情感计算”，人类对情绪的理解有着 150 多年的研究历史。中国学者为此也做出了不懈的努力。2025 年上海外滩大会集中展示了情感计算领域的最新进展，各类陪伴机器人(如“奇多多 AI 学伴”、“Fuzzoo”毛绒 AI 宠物)、心理辅导机器人(如安徽阿拉丁的“AI 心理咨询师”)不断涌现，预示着情感计算已由理论研究转向市场应用。

当今，时代人工智能正经历从纯粹符号计算转向情感智能(emotional intelligence, EI)，即符号计算到情感计算。这一转变既标志着技术模型的深化也从根本上重构人机交互的伦理基础。不过在这样的热潮背后情感计算依然要面对一个根本性的质疑，那就是如何跨越情感模拟和真实理解之间的鸿沟。基于此，本文将从心理学、计算科学的双重视角出发，对机器人情绪加工的计算机制、应用场景、局限挑战进行系统剖析。

2. 计算机制

情绪加工可被解构为三个相互关联的计算维度：情绪注意(emotional attention)、情绪识别(emotion recognition)、情绪评估与响应(emotional assessment and response) (Rachman, 1980)。以下是各维度在算法实现与面临挑战方面的主要内容。

情绪注意阶段的核心任务是从复杂环境信息中筛选关键情绪信号，其算法实现主要基于注意力机制与显著性检测模型(Guo et al., 2022)。在情绪识别层面，系统需要对情绪状态进行精确分类与标注，目前主要采用卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络/长短期记忆网络(Recurrent Neural Network/Long Short-Term Memory, RNN/LSTM)、转换器(Transformer)及多模态融合等算法。情绪评估与响应作为最高层次的加工阶段，旨在理解情绪产生机制并生成适应性反应，该阶段依托奥查－克洛里－柯林斯认知评价模型(Ortony，Clore，and Collins Cognitive Appraisal Model，OCC 模型)、长链(LangChain)等记忆架构以及共情计算策略实现算法支撑(Sharma et al., 2023)。

这三个维度共同构成了完整的情绪加工链条，其技术突破将直接推动人机情感交互向更深层次发展。而这三个方面与人类情绪处理的认知神经模型相对应，其核心任务、关键算法与主要挑战概括如表 1。

Table 1. Core dimensions and challenges of emotional processing in artificial intelligence
表 1. 人工智能情绪加工的核心维度与挑战

加工维度	核心任务	算法实现	主要挑战
情绪注意	筛选情绪信号	注意力机制、显著性检测模型	多信号源竞争、文化差异
情绪识别	对情绪进行分类与标注	CNN，RNN/LSTM，Transformer，多模态融合	主观性、掩饰性情绪
情绪评估与响应	理解情绪成因并生成适应性反应	OCC 模型、记忆架构	深度推理、因果判断

2.1. 情绪注意的计算机制

情绪注意(emotional attention)是指系统从多源信息流中筛选情绪显著性刺激的能力(Pessoa, 2009)。其计算实现依赖于注意力机制(attention mechanism)，这种注意机制可以被看作是一种基于输入图像特征的动态权重调整过程。除此之外还有视觉显著性模型(visual saliency models)，在该系统中，注意力机制可以被视为通过根据输入的重要性自适应地加权特征来实现的动态选择过程(Guo et al., 2022)，该过程模拟了人类注意偏向(attentional bias)的认知特性。图 1 简要总结了计算机视觉处理中给予注意力的模型演进。

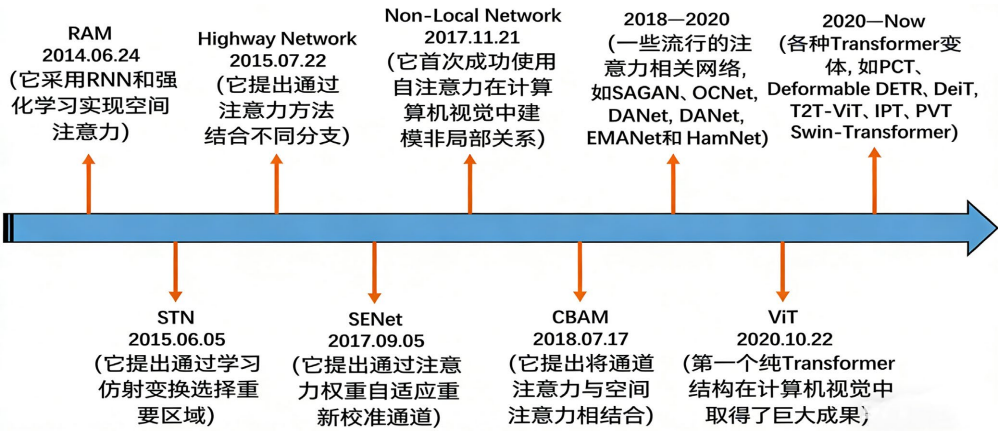

Figure 1. The evolution of attention-based models in computer vision processing (Cited from Guo et al., 2022)
图 1. 计算机视觉处理中给予注意力的模型演进(引自 Guo et al., 2022)

图中显示在最初探索阶段(2014~2015)，研究者通过 Recurrent Attention Model (RAM)模型的强化学习实现空间定位，以及空间变换网络(Spatial Transformer Network, STN)学习仿射变换来校正图像区域，这些工作奠定了注意力机制的雏形。其次，2017 年成为关键转折点：挤压激励网络(Squeeze-and-Excitation Network, SENet)提出的通道注意力机制通过特征重校准显著提升了网络表达能力，而非局部网络(non-local network)这一事件首次将自注意力成功引入视觉领域，突破了传统卷积的局部感受野限制。在随后的发展时期(2018~2020)，研究者开始系统整合不同维度的注意力机制，如卷积块注意力模块(Convolutional Block Attention Module, CBAM)将通道与空间注意力有机结合，同时涌现出多种注意力网络的创新设计。2020 年之后，视觉转换器(Vision Transformer, VT)的突破性成果引领了研究范式的转变，证明完全基于自注意力的架构在视觉任务中的卓越性能，继而催生了众多适应不同视觉任务的 Transformer 变体推动注意力机制成为计算机视觉的核心建模工具。这一演进过程不仅体现了对特征制度的持续深化，更体现着视

觉表征学习范式的变革。

但当前情绪注意的计算机制仍面临两大挑战。其一，情绪注意本质上是资源分配过程，系统需要在有限的资源下完成多通道情绪信号的优先级排序，而这一过程深受文化背景影响。研究表明，东亚和西方个体在感知面部情绪时，会采用不同的注意模式与文化特异性策略(如分别关注眼部和嘴部) (Jack et al., 2009)。东亚文化中的情绪表达通常含蓄内敛，西方文化则更为直接外显，这种根深蒂固的文化特异性导致基于单一文化数据训练的模型在跨文化场景中识别准确率明显降低。一项跨文化面部表情识别研究表明，基于西方数据训练的模型在东亚被试上的识别准确率平均下降约 15%~20% (Dupre et al., 2020)。

其二，当系统进行跨文化迁移时，其注意力权重需要动态调整，这不仅需要技术层面的参数优化，更要求深入理解不同文化的情绪表达规约。Gross 的情绪调节理论指出，注意分配作为情绪体验的首要环节，其文化差异性常被现有系统忽略(Gross, 1988)。这些挑战体现了生物注意机制与机器算法之间存在的解释鸿沟，计算神经科学研究也为此提供了证据，当人类观看情绪面孔时，其大脑皮层的电生理活动如 N170 成分会因文化背景不同而表现出显著差异，而现有注意力模型尚未能很好地模拟这种由文化调节的神经动力学特征(Li et al., 2022)。

2.2. 情绪识别的计算机制

情绪识别(recognition of emotion)旨在将输入信号映射至离散类别如 Ekman 的六种基本情绪或连续维度空间，例如 VAD 模型中的效价(valence)、唤醒度(arousal)、优势度(dominance)。其技术演进经历了从规则系统到机器学习再到深度学习的范式变迁。下面简要介绍与情绪识别相关的计算模型。

首先是离散分类模型。卷积神经网络(CNN)作为一种多层结构的人工神经网络，其核心设计目标在于检测与处理二维图像中的特征。该网络的基本原理借鉴了生物视觉系统中负责特征抽取的神经组织形态(常亮等，2016)。经过多层级训练后，此类网络不仅可生成图像特征，还能够实现特征的提取以及分类功能(李欢，曾烁，2022)，而在标准的面部表情识别数据集如 FER-2013 上，基于 CNN 的模型已达到 90%以上的分类准确率(Minaee et al., 2021)。除此之外，循环神经网络(RNN/LSTM)主要面向语音序列信号的处理任务，并输出对应的情绪分类结果。随着研究进展，维度回归模型逐渐受到重视 通过支持向量回归(Support Vector Regression, SVR)等算法回归 VAD 数值，更适合捕捉“悲喜交加”等复杂混合情绪(Kollias & Zafeiriou, 2017; Tzirakis et al., 2017)。

近年来，多模态融合成为新的研究趋势。多模态情感识别技术致力于从图像、视频、音频、文本及生理信号等多种信息源中抽取情感相关信息，以实现对情感的判断、预测、检测与检索(姚鸿勋等，2022)。部分研究采用基于 Transformer 的模型开展识别工作，该类网络在并行运算方面效率更高，且能够更有效地捕捉远距离特征间的依赖关系。以审讯场景为例，系统通过处理人脸视频、语音及生物信号等多元数据，实现对犯罪嫌疑人情绪状态的分析。人工智能的介入改变了传统情绪识别与分析的技术路径，使得在审讯过程中批量采集并解析心理行为特征数据成为可行(李长庭等，2024)。此外，一项元分析研究指出，融合视觉、听觉和文本信息的多模态情绪识别系统，其平均识别准确率比最好的单模态系统高出约 5%~15% (Wang et al., 2022)。然而，当前情绪识别的计算机制仍有一些亟待解决的问题。首先，主观性表现为同一表述会因文化与个体差异产生不同解读且存在个体情感表达差异问题，不同个体在生理(性别、年龄等)和心理(种族文化等)方面存在差异导致相同诱发条件下不同个体表达同一情感的面部方式也有较大鸿沟(Chen & Joo, 2021; Xu et al., 2020)。

值得注意的是，掩饰性和欺骗性情绪的识别问题更为棘手。现实生活中，个体常出于社会规约或个人考量而掩饰真实情绪(Noroozi et al., 2017)。心理学研究证实，这种刻意控制的情感表达在神经机制和外部表征上(如微表情)与自发情绪存在可测量的细微差异(Porter & Brinke, 2008)。然而，如何从有限的、受

控的行为信号中稳健地捕捉这些“泄露真相”线索的现状，对现有技术的敏感性与特异性提出了极高的要求。

2.3. 情绪评估与响应的计算机制

情绪评估与响应(emotional assessment and response)作为情绪加工流程中的高级阶段，其核心任务在于识别情绪产生的内在诱因、预测个体未来的情感状态变化，并生成具有适应性的行为反馈。以下是一些机器人情绪评估与响应的主要模型。

首先，在理论模型层面，OCC 模型作为一种经典的认知评价框架，引导人工智能系统依据对事件的预期、目标关联性以及主体态度进行情感评估，进而形成相应的情感反应(Ortony et al., 1988)。例如，当系统检测到“用户完成了预设学习目标”时，可能触发“高兴”这一正向情绪反馈。

其次，为实现更具个性化的交互体验，可引入 LangChain 等架构构建记忆系统。这类系统基于大型语言模型强大的上下文理解与记忆能力(Zhao et al., 2023b)，使人工智能能够基于用户历史行为数据生成定制化响应。典型场景如教育机器人记录到某学生曾在考试失利后情绪低落，便会在后续互动中有意识地采取更为积极的鼓励策略。

最后，在共情机制方面，系统可体现为平行共情与反应性共情两种路径(McQuiggan & Lester, 2007)：前者侧重于对用户当前情绪的模仿与映照，后者则致力于提供补偿性情感支持。这两种共情模式可通过语言内容、语调调节乃至非语言手势(如微微低头示意)等多重通道进行表达，从而增强交互的自然度与情感真实感。

当前情绪评估与响应的计算机制仍有一些亟待解决的问题。主要包括两方面内容，第一，因果推理能力缺失。主要表现为现有系统能识别情绪的“是什么”却难以理解情绪的“为什么”。正如 Park 和 Ko (2023)所指出的，缺乏因果模型使得系统无法进行深层次的情绪归因，最新的实证评估也表明，即使在复杂社会情境的理解上，当前最先进的大型语言模型在推断情绪和行为背后的原因时，其表现也远逊于人类，这凸显了其在因果推理与社会常识方面的根本性欠缺(Zhan & Liu, 2024)，而在多数情况下计算机检测到学生焦虑情绪时系统往往无法准确判断该情绪源于学业压力、家庭因素还是人际关系，而这种深层归因能力需要系统具备丰富生活常识与推理能力。

第二，个性化适应问题体现为现有个性化模型虽在技术层面持续进步但在理解个体情绪世界的独特性与复杂性上仍显不足。正如 Zhang 和 Williams (2022)提出的“隐私悖论”框架正揭示了这一两难境地，要求设计者在避免回应过度泛化的同时，必须严格防范过度个性化所带来的数据泄露与滥用风险。

3. 应用场景

随着情绪加工机制在注意、识别与响应等环节的不断突破，情感机器人已逐步从实验室走向多元化的现实场景，展现出广泛的应用潜力。特别是在人口结构变化与心理健康需求日益凸显的背景下，情感计算技术正被广泛应用于适老化陪护、儿童教育、心理咨询及课堂情绪监测等领域(Kouroupa et al., 2022; Zhao et al., 2023a)。这些应用不仅体现了情感机器人作为“情绪协作者”的社会价值，也对其在真实复杂环境中的适应能力提出了更高要求(王方家，2025)。

3.1. 适老化照料与情感陪护

2020 年世界卫生组织与联合国所有会员国共同批准“联合国健康老龄化十年”项目。该项目特别强调为老人提供长期护理并将这项工作列为未来十年重点领域之一(United Nations General Assembly, 2020)。为应对老龄化带来的各类挑战，借助机器人护理技术安全有效地照料体弱老人，并满足其情感需求是颇

具发展前景的解决方案。

该方案既有助于弥补老人护理需求与现有资源间的差距也能帮助身体功能受损老人改善健康状况(Zhao et al., 2023a)。及时对可能存在的健康风险发出预警，同时可提醒老年人按时服药和规律作息。此外，这类护理机器人具备多种功能，可实现 24 小时陪伴，通过与老人聊天、分享回忆缓解其孤独，并能识别老人抑郁、焦虑等情绪。

然而，在这一技术应用的背后，我们须清醒地认识到其带来的深刻伦理挑战与潜在风险。一方面是情感陪伴的真实性问题。长期与之相处的老年人，尤其是认知能力有所下降者，可能会对这种“表面上的”陪伴产生依赖甚至误解，这不但可能误导老人的情感寄托，反而还会减少他们与家人、朋友的真实交流进而加深孤独感(Sharkey & Sharkey, 2020)。实证研究显示，约 30%的长期使用社交机器人的老年人报告其与家人的面对面交流频率有所减少，同时有部分用户对机器人产生了显著的情感依赖(Pu et al., 2021)。另一方面则是隐私与责任问题。陪伴机器人持续收集老人的健康数据、行为习惯甚至情感状态，这些高度敏感的信息如何被安全保护、合理使用，目前仍缺乏完善的规范。一旦发生数据泄露或基于错误算法做出不当判断和误判情绪而发出错误警报，由谁来承担责任也模糊不清(Zhao et al., 2023a)。因此，推动情感陪护机器人的发展，不能仅追求技术先进，更需要建立以尊重老年人权益为核心的伦理框架，谨慎审视技术介入的界限，确保科技真正服务于人的福祉。

3.2. 儿童教育与情感发展

在儿童教育领域，情感机器人承担着“教育玩伴”和“情感导师”的双重角色，它能即能帮助学生提升多方面技能，如解决问题的能力、自我效能感、创造力、合作技能以及计算思维等，还能通过识别婴儿啼哭背后的原因如饥饿、疼痛或困倦，来分析幼儿的情绪状态来辅助家长进行精准照料，并且通过多轮对话助力儿童社会情感能力的发展，目前这类机器人已应用在语言(Lin et al., 2022)、数学与科学领域(Zhong & Xia, 2020)以及跨学科的 STEAM 教育中(Benitti & Spolar, 2017; Sullivan & Bers, 2018)。

目前已有多家公司关注情感计算在儿童教育中的应用。无界方舟公司推出的“奇多多 AI 学伴”能够实现十几轮连贯对话，还能阅读绘本并回应孩子提出的各类追问，其与孩子建立起的“情绪联结”正是它的核心竞争力。中科院研发的“飞燕”机器人则主要聚焦于青少年心理疏导工作。

值得注意的是这类应用也触及了深刻的伦理边界，比如机器到底能不能、又是否应该替代父母与孩子之间的情感联结，且算法对儿童心理发展产生的长期影响目前仍不明确。首先是关于情感联结的“替代”风险。机器提供的是一种始终耐心、随时回应的互动，这与真实亲子关系中必然存在的矛盾、等待和协商截然不同。若儿童过度依赖这种“完美”陪伴，可能会影响他们在现实生活中建立和处理深度人际关系的能力，甚至模糊对真实情感连接的认知。一项为期 6 个月的纵向观察研究发现，频繁与情感机器人互动的学龄前儿童，在自由游戏中发起同伴社交互动的主动性较对照组有统计学上的显著降低(Kohler et al., 2023)。其次是算法对心理发展的长期影响难以预测。教育算法若在设计时隐含了文化、性别等偏见，可能会在互动中无声地固化这些刻板印象，影响儿童价值观的多元形成。同时，算法根据儿童情绪状态即时调整反馈的适应性，也可能使其习惯于被即时满足，从而削弱面对挫折的耐受力(Xu et al., 2020)。因此在儿童教育中引入情感机器人绝不能仅从技术效率出发，而需要将儿童权益置于中心进行最审慎的评估与监管。

3.3. 心理健康与特殊护理

情感计算在心理健康领域展现出巨大潜力，它的目标是提供标准化、无偏见且能随时使用的辅助治疗手段。在具体应用方面，它能为自闭症谱系障碍(Autism Spectrum Disorder, ASD)患者提供可预测的社

交训练伙伴。研究表明，如机器人较为容易被自闭症谱系的儿童和青少年接受，这类机器人能带来模仿技能、眼神接触、联合注意力、行为反应以及重复和刻板行为方面的积极影响，还能提供一种可预测且一致的互动模式，这种模式对自闭症儿童和青少年的学习很有帮助(Kouroupa et al., 2022)。此外，它也能帮助抑郁症、焦虑症患者开展认知行为练习(Cognitive Behavioral Therapy, CBT)和情绪监测。

与此同时，我们也需要辩证看待情感机器人在心理健康这一特殊领域可能引发的风险与担忧。首先是一种名叫“机械依赖”的风险。机器人提供的是一种稳定、可预测且永不厌烦的互动模式，对于部分渴望陪伴或社交焦虑的患者而言，长期依赖这种简单化的关系可能会削弱他们面对真实、复杂人际关系的勇气与能力，反而不利于社会功能的恢复(Sharkey & Sharkey, 2021)。临床观察报告指出，在某些边缘性人格障碍患者的治疗中，过度使用情感支持聊天机器人与患者在团体治疗中参与度下降和现实人际关系挑战的回避行为增加存在关联(Bendig et al., 2023)。其次是关于儿童自主性面临严峻挑战。当机器人的自动化建议与患者的个人意愿或专业治疗师的判断产生冲突时，如何界定责任与尊重人的自主选择仍是未解的难题(Calvo et al., 2020)。因此将其作为为严格监管下的辅助工具，并持续评估其社会心理影响是至关重要的前提。

3.4. 学生情绪监测与学习干预

在教育场景中，情感计算同样具有广泛的应用价值。情感计算具备实时分析课堂情绪的能力，能帮助教师清晰掌握整体课堂氛围处于专注、困惑还是厌烦状态，同时还了解个别学生的具体情绪状况。教师借助这些信息可针对不同学生的需求实施个性化学习干预，使教学更具针对性(Picard et al., 2004)。相关研究中研究者通过构建参与向量模型(engagement vector)，Whitehill 等人(2014)通过多模态方法所展示的，可以从认知、情感、行为三个维度对学生课堂参与度进行量化分析，研究结果证实融合情绪识别功能的教学机器人在实际教学过程中能明显帮助学生提升学习效果，使学生更易掌握知识、提高学习效率。

与此同时，我们也应冷静审视这项技术对教育环境和学生成长带来的深远影响。一方面，若将复杂的学习情感简化为有限标签，是否存在过度简化与误判的风险？学习过程中的困惑、专注或无聊是动态且高度情境化的，算法对其进行分类和量化很可能忽略了背后的具体原因如教学内容难度、个人兴趣或外部因素。这种简化的情感画像若被用于自动化干预或评价学生，可能导致刻板印象和错误的判断进一步影响教育公平。另一方面，技术可能无意中复制并放大现有的教育不平等。已有研究表明，面部情绪识别算法在不同种族、性别和身体表现力的人群中，准确率存在差异。如果基于一个有偏差的系统来评估学生的课堂参与度和情绪状态，可能会对某些学生群体产生系统性的低估或误解，从而在追求“个性化”的社会环境下，反而巩固了社会中的不平等现象(Xu et al., 2020)。因此，在教育领域引入情感计算需要持审慎的态度，其发展应以保障学生权益为核心。

4. 局限、挑战

尽管情感计算在诸多场景中展现出独特价值，其进一步发展仍面临技术实现、伦理规范与哲学认知层面的多重挑战。具体而言，这些挑战体现在以下几个方面。

4.1. 技术瓶颈：算法的局限

当前，情感计算在技术层面仍面临多重算法瓶颈，这一现象制约着其进一步发展与应用深化。首先，多模态感知融合的挑战尤为突出。系统需同时处理来自视觉、语音等多通道的情感信号，然而不同模态信息之间常存在不一致甚至冲突，例如“微笑”的面部表情可能与“讽刺”的语音语调同时出现。如何实现跨模态信息的精准对齐与协同理解，至今仍是未解的难题(Xu & Li, 2023)。

其次，系统在上下文理解与个性化建模方面存在明显局限。由于缺乏真实的生活经验与常识推理能力，机器难以捕捉情绪产生的复杂背景因素，也无法结合个体经历与性格特质进行深层次的情感解读，导致其理解停留在表面关联层面。

最后，成本与算力限制直接影响了技术的普及潜力。例如一个能够实时处理对话、表情和语调的多模态情绪理解模型，其单次推理的能耗可能是仅处理文本模型的数十倍，这严重制约了其在移动或嵌入式设备上的部署(Fernandez et al., 2023)，此外高阶情感模型往往需要 400TOPS 以上的算力支持(Strubell et al., 2022)，这不仅推高了硬件成本，也使得搭载此类系统的高端机器人价格昂贵，进而限制了其在更广泛场景中的规模化应用。

4.2. 伦理困境：必须直面的人类命题

在情感计算的发展过程中，伦理问题构成了亟待解决的核心挑战。首先，情感欺骗与操纵问题引发广泛争议。情感欺骗指的是人工智能系统通过模拟情感表达或共情反应，使使用者误以为其具备真实的情感理解与关怀能力，本质上是一种基于技术表现的认知误导。AI 所模拟的共情反应虽具备交互上的实用性，但其本质是否构成对用户情感的欺骗仍存疑问。更值得警惕的是，这种技术若被刻意用于引导、放大或利用用户的情感状态，可能演变为系统性的情感操纵，例如影响个体消费决策乃至政治倾向，从而引发深层的道德与社会信任危机(Lutz & Tamo-Larrieux, 2022)。

与此同时，隐私与数据安全风险尤为突出。情绪数据属于最高敏感级别的生物特征信息，其采集、存储与使用过程存在显著的泄露与滥用隐患。根据一项针对数据泄露事件的调查报告显示，生物特征数据一旦泄露，其平均修复成本是普通个人身份信息(PII)的 2.5 倍以上，且造成的长期信任损害更为严重(IBM Security, 2023)，此外在医疗诊断、自动驾驶等高风险场景中，此类数据的安全性需满足极高标准(Guo et al., 2022)，任何疏漏都可能造成严重后果。

进一步而言，其责任归属机制尚不明确。当系统因情绪识别错误或响应失当而对用户造成心理伤害时，如何在开发者、运营方与使用者之间界定法律责任与道德责任，成为制度设计上的难题(Santoni de Sio & Mecacci, 2021)。

最终，这一技术还可能带来社会关系层面的潜在影响。若人类过度依赖机器人提供的情感陪伴，长期来看可能导致个体社交能力的弱化与现实人际关系的疏离，进而影响社会联结的质量与稳定性。这些伦理议题共同构成了情感计算融入社会过程中必须严肃审视的人类命题。

4.3. 哲学争议：无法回避的意识本质之问

在情感计算的发展进程中，哲学层面的争议始终伴随其技术演进，尤其体现在对情感本质的理解上。本质上，人工智能所呈现的“情绪”，不过是对人类情感模式的模仿与复现，如同一面“精致的镜子”，仅能反射出情感的外在表征，而无法承载内在真实的感受。这一分歧将我们引向心灵哲学中的核心难题：主观体验究竟如何从物理过程中涌现？是否能够通过计算模型实现真正意义上的情感再现？

该问题不仅关乎技术路径的选择，更促使我们重新审视情感与意识、计算与体验之间尚未弥合的理论鸿沟，正如内格尔(Nagel, 1974)所启示，在于我们无法通过客观的功能描述去通达他者(或机器人)的主观体验。若无法在哲学层面厘清情感的本质，情感计算的发展将始终面临根基性的质疑与边界约束。

5. 未来展望：从识别到生成，迈向可信赖的情感智能体

为克服当前情感计算面临的技术瓶颈与伦理困境，未来发展的核心目标是构建一个可信赖、可治理且真正增强人类能力的情感智能范式(中国计算机学会，2025)。这一路径会推动技术从静态的“情感识别”向动态的“情感理解与生成”跃迁，最终发展为可信赖的情感智能体。

首先，在技术发展上我们需要让系统从“观察行为”转向“理解原因”，并有一个清晰的发展步骤。目前的技术主要是分析表情、语音等外在行为信号，而这与理解人内心的真实感受是两回事。所以未来的关键突破，是让系统不仅能识别出焦虑还能理解这种焦虑是因为考试压力还是朋友矛盾，从而真正明白情绪背后的原因。这个过程会像爬台阶一样，逐步从简单的对话功能，发展到更自主的智能体，再向与身体和环境结合的具身形态演进，这样每一步都会注重满足个人需求和保障安全(秦兵，2025)。

其次，在管理规则上我们需制定出可以具体执行和检查的措施而不是停留于纸面上的原则。情感机器人像一把“双刃剑”，既能陪伴孤独的人也可能让人越来越远离真实的社交(甘怡群，2025)，因此管理规则需要融入到技术设计中。具体来说可以参考保护数字身份、防止过度依赖、守护人身安全这三条底线来建立法规，在鼓励创新和确保安全之间找到平衡(国家互联网信息办公室，2025)。而对于技术细节的处理可以设定一个明确的启动开关如情感交互启动事件，一旦系统开始分析并利用用户的情绪来互动，就必须自动提高透明度让用户拥有知情权和控制权，这样管理就落到了实处(洪延青，2025)。

最后，在应用定位上要始终坚持“辅助人类”这一方向，并警惕其长期影响。情感智能的最高价值，是成为帮助人们更好沟通与生活的工具(中国计算机学会，2025)。例如，在心理保健中它可以辅助医生持续关注患者的情绪变化并提供练习建议，但绝不能代替医生做出诊断和关键决策(Torous et al., 2021)。在教室里，它应该帮助老师了解全班的学习氛围，而不是用来监控和评价每一个学生。要想成功实现这个辅助角色，我们必须时刻警惕技术可能导致人与人关系疏远的风险(甘怡群，2025)，还要通过长期的跟踪研究不断观察这项技术对我们社会生活与心理健康的真实影响。

只有通过技术、管理和应用这三个方面的共同努力，情感计算才能逐渐跨越“模仿情绪”到“理解人心”的障碍，在带来更自然交互体验的同时确保其发展是安全、负责任且真正为人服务的。

6. 结论

机器人的情绪加工既是对技术边界的拓展，也折射出人类探索自身情感奥秘的孜孜追求。它迫使我们重新思考什么是情绪？什么是共情？什么又是人与人之间不可替代的联结？情感机器人的终极目标不应是创造完美的“人造人”或替代人类的情感交流，而应是作为工具、伙伴和能增强人类自身情感能力、弥补情感缺口的“外部引擎”。迈向这一未来的过程中我们需要工程师的代码、心理学家的智慧、哲学家的思辨，以及全社会对技术与人性的持续审思。唯有这样才能确保技术发展始终闪耀人性光辉，让人类在数字时代能更深入、更真诚地联结彼此，在善用科技与科技向善的道路上行稳致远。

基金项目

湖北医药学院人才启动金项目(项目编号：2025QDJRW06)。

参考文献

常亮, 邓小明, 周明全, 武仲科, 袁野, 杨硕, 王宏安(2016). 图像理解中的卷积神经网络. *自动化学报, 42(9),* 1300-1312.

甘怡群(2025). 情感人工智能对社会支持感知的双刃剑效应: 补偿与疏离. *人民论坛·学术前沿, (20),* 48-56.

国家互联网信息办公室(2025). *专家解读. 智能有度服务有温, 把好“三关”规范 AI 拟人化互动边界.* 国家互联网信息办公室.

洪延青(2025). 人机“情感交互”的规范化: 指标、机制与多方协同治理. *法治日报,* p. 3.

李欢, 曾烁(2022). 诊断、评估与干预: 近五年卷积神经网络在特殊教育中的实证研究述评. *中国特殊教育, (7),* 10-22. .

李长庭, 赵印, 毕惜茜(2024). 多模态智能审讯技术的原理与实战化应用路径研究. *中国人民公安大学学报(社会科学版), 40(2),* 22-29.

秦兵(2025). 情感交互的层级演进与对齐技术. 见 中国计算机学会(主编), *CNCC 2025 技术论坛报告集*(pp. 1-15).

王方家(2025). 探索人与智能体共生的心理密码——评《人智交互心理学》. *心理与行为研究, 23(3),* 430-432.

吴静, 董屹泽(2025). 从人工智能情绪识别到情感资本主义: 情感可计算化的哲学反思. *南京社会科学, (7),* 94-103.

姚鸿勋, 邓伟洪, 刘洪海, 洪晓鹏, 王甦菁, 杨巨峰, 赵思成(2022). 情感计算与理解研究发展概述. *中国图象图形学报, 27(6),* 6-36.

中国计算机学会(2025). *《从识别到生成: 智能交互离真情实感还有多远?》论坛举办[摘要]*. 技术论坛报告. 中国计算机学会.

Benitti, F. B. V., & Spolaôr, N. (2017). How Have Robots Supported STEM Teaching? In M. S. Khine (Ed.), *Robotics in STEM Education: Redesigning the Learning Experience* (pp. 103-129). Springer International Publishing. https://doi.org/10.1007/978-3-319-57786-9_5

Calvo, R. A., Peters, D., & Vider, K. (2020). The Ethics of Affective Computing: Principles, Challenges, and Opportunities. *Foundations and Trends in Human-Computer Interaction, 13,* 1-16.

Chen, Y., & Joo, J. (2021). Understanding and Mitigating Annotation Bias in Facial Expression Recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 14960-14971). IEEE. https://doi.org/10.1109/iccv48922.2021.01471

Dupré, D., Krumhuber, E. G., Küster, D., & McKeown, G. J. (2020). A Performance Comparison of Eight Commercially Available Automatic Classifiers for Facial Affect Recognition. *PLOS ONE, 15,* e0231968. https://doi.org/10.1371/journal.pone.0231968

Fernandez, J. D., Martinez, F., & Garcia, A. (2023). Energy Consumption Analysis of Multimodal versus Unimodal Affective Models on Edge Devices. *Sustainable Computing: Informatics and Systems, 38,* Article ID: 100877.

Gross, J. J. (1988). Emotion and Emotion Regulation. In L. A. Pervin, & O. P. John (Eds.), *Handbook of Personality: Theory and Research* (2nd ed., pp. 525- 552). Guilford Press.

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J. et al. (2022). Attention Mechanisms in Computer Vision: A Survey. *Computational Visual Media, 8,* 331-368. https://doi.org/10.1007/s41095-022-0271-y

IBM Security (2023). *Cost of a Data Breach Report 2023*. IBM Corporation.

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural Confusions Show That Facial Expressions Are Not Universal. *Current Biology, 19,* 1543-1548. https://doi.org/10.1016/j.cub.2009.07.051

Kohler, M., Gieselmann, H., & Sodian, B. (2023). Social Robots in Kindergarten: Effects on Peer Interaction and Prosocial Behavior. *Computers & Education, 196,* Article ID: 104727.

Kollias, D., & Zafeiriou, S. (2017). Continuous Regression of Dimensional Emotion for Video. In *2017 12th International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 42-49). IEEE Computer Society.

Kouroupa, A., Laws, K. R., Irvine, K., Mengoni, S. E., Baird, A., & Sharma, S. (2022). The Use of Social Robots with Children and Young People on the Autism Spectrum: A Systematic Review and Meta-Analysis. *PLOS ONE, 17,* e0269800. https://doi.org/10.1371/journal.pone.0269800

Li, X., Hu, X., & Fu, S. (2022). Cultural Modulation of Neural Responses to Emotional Faces: An EEG Study. *NeuroImage, 264,* Article ID: 119705.

Lin, V., Yeh, H.-C., & Chen, N.-S. (2022). A Systematic Review on Oral Interactions in Robot-Assisted Language Learning. *Electronics, 11,* Article No. 290. https://doi.org/10.3390/electronics11020290

Lutz, C., & Tamo-Larrieux, A. (2022). The Ethics of Emotional AI and Its Impact on Human Autonomy. *Minds and Machines, 32,* 667-688.

McQuiggan, S. W., & Lester, J. C. (2007). Modeling and Expressing Empathy in an Intelligent Tutoring System. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (pp. 1-8). ACM.

Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors, 21,* Article No. 3046. https://doi.org/10.3390/s21093046

Nagel, T. (1974). What Is It like to Be a Bat? *The Philosophical Review, 83,* 435-450. https://doi.org/10.2307/2183914

Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. (2017). Challenges and Opportunities in Deep Learning for Automated Emotion Recognition. *Pattern Recognition Letters, 99,* 86-93.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press. https://doi.org/10.1017/cbo9780511571299

Park, G., & Ko, B. (2023). Beyond Accuracy: Towards Causal Reasoning in Affective Computing. *Nature Machine Intelligence, 5,* 789-799.

Pessoa, L. (2009). How Do Emotion and Motivation Direct Executive Control? *Trends in Cognitive Sciences, 13,* 160-166. https://doi.org/10.1016/j.tics.2009.01.006

Picard, R. W. (1997). *Affective Computing*. The MIT Press. https://doi.org/10.7551/mitpress/1140.001.0001

Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D. et al. (2004). Affective Learning—A Manifesto. *BT Technology Journal, 22,* 253-269. https://doi.org/10.1023/b:bttj.0000047603.37042.33

Porter, S., & ten Brinke, L. (2008). Reading between the Lies. *Psychological Science, 19,* 508-514. https://doi.org/10.1111/j.1467-9280.2008.02116.x

Pu, L., Moyle, W., Jones, C., & Todorovic, M. (2021). The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. *The Gerontologist, 61,* e1-e17.

Rachman, S. (1980). Emotional Processing. *Behaviour Research and Therapy, 18,* 51-60. https://doi.org/10.1016/0005-7967(80)90069-8

Santoni de Sio, F., & Mecacci, G. (2021). The Attribution Problem in Emotional AI: Who Is Responsible When Things Go Wrong? *Philosophy & Technology, 34,* 1587-1611.

Sharkey, A., & Sharkey, N. (2020). We Need to Talk about Deception in Social Robotics! *Ethics and Information Technology, 23,* 309-316. https://doi.org/10.1007/s10676-020-09573-9

Sharkey, A., & Sharkey, N. (2021). Who Will Care for the People? The Looming Crisis in Aged Care and the Threat of the Care Robot. *AI & Society*. Springer Nature

Sharma, A., Lin, Z., & Li, L. (2023). A Computational Approach to Empathetic Responding Using Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12074-12089). Association for Computational Linguistics..

Strubell, E., Ganesh, A., & McCallum, A. (2022). The Computational and Energy Cost of Deep Learning and Affective Computing Models. *Communications of the ACM, 65,* 70-79.

Sullivan, A., & Bers, M. U. (2018). Dancing Robots: Integrating Art, Music, and Robotics in Singapore's Early Childhood Centers. *International Journal of Technology and Design Education, 28,* 325-346. https://doi.org/10.1007/s10798-017-9397-0

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P. et al. (2021). The Growing Field of Digital Psychiatry: Current Evidence and the Future of Apps, Social Media, Chatbots, and Virtual Reality. *World Psychiatry, 20,* 318-335. https://doi.org/10.1002/wps.20883

Tzirakis, P., Zhang, J., & Schuller, B. W. (2017). End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *Journal of Selected Topics in Signal Processing, 11,* 1301-1309.

United Nations General Assembly (2020). *United Nations Decade of Healthy Ageing (2021-2030) (Resolution A/RES/75/131)*. https://undocs.org/A/RES/75/131

Wang, J., Li, L., & Wang, D. (2022). Multimodal Emotion Recognition: A Systematic Review of Recent Advances and Challenges. *Knowledge-Based Systems, 258,* Article ID: 110010.

Whitehill, J., Serpell, Z., Lin, Y., Foster, A., & Movellan, J. R. (2014). The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing, 5,* 86-98. https://doi.org/10.1109/taffc.2014.2316163

Xu, P., & Li, Y. (2023). Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions. *Information Fusion, 98,* Article ID: 101819.

Xu, T., White, J., Kalkan, S., & Gunes, H. (2020). Investigating Bias and Fairness in Facial Expression Recognition. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision—ECCV 2020* (pp. 506-523). Springer. https://doi.org/10.1007/978-3-030-65414-6_35

Zhan, Y., & Liu, R. (2024). Social Reasoning in Large Language Models: Current State and Future Directions. *Proceedings of the National Academy of Sciences, 121,* e2401743121.

Zhang, L., & Williams, A. C. (2022). The Privacy Paradox in Personalized Affective Systems: A Framework for Mitigation. *Proceedings of the ACM on Human-Computer Interaction, 6,* 1-28.

Zhao, D., Sun, X., Shan, B., Yang, Z., Yang, J., Liu, H. et al. (2023a). Research Status of Elderly-Care Robots and Safe Human-Robot Interaction Methods. *Frontiers in Neuroscience, 17,* Article ID: 1291682. https://doi.org/10.3389/fnins.2023.1291682

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Wen, J. R. (2023b). *A Survey of Large Language Models*.

Zhong, B., & Xia, L. (2020). A Systematic Review on Exploring the Potential of Educational Robotics in Mathematics Education. *International Journal of Science and Mathematics Education, 18,* 79-101. https://doi.org/10.1007/s10763-018-09939-y