

# 数心融合视域下AI心理机器人促进硕士生健康的逻辑、机制与边界

焦松明<sup>1,2</sup>, 时 勤<sup>1,3\*</sup>, 苏星赫<sup>2,4</sup>, 李从庸<sup>1,2</sup>, 唐伟杰<sup>1,2</sup>, 张旭晟<sup>1,2</sup>

<sup>1</sup>温州大学生态文明与环境治理实验室, 浙江 温州

<sup>2</sup>温州大学温州模式发展研究院, 浙江 温州

<sup>3</sup>中国科学院大学经济与管理学院, 北京

<sup>4</sup>国立发展管理学院国际学院, 泰国 曼谷

收稿日期: 2026年8月12日; 录用日期: 2026年9月7日; 发布日期: 2026年9月18日

## 摘 要

在数字技术与心理健康服务深度融合的背景下, “数心融合”成为心理健康领域的重要范式。硕士生群体面临高认知负荷、高心理防御与低求助意愿的结构性困境, 传统服务难以有效覆盖。本研究以AI心理机器人“小欣”为对象, 该机器人以易术心理剧为理论内核, 融合多模态生理监测与大语言模型技术, 构建了“意识分析→艺术体验→潜意识情志重塑→行为训练”的四层穿透式干预模型。本文从理论逻辑、作用机制与实践边界三方面展开探讨: 阐释“数心融合”的范式价值, 揭示AI机器人的心理剧数智化转型与人机共情机制, 分析隐私保护、情感欺诈等伦理风险, 并提出技术、使用、治理三维协同优化路径。研究表明, AI心理机器人并非替代咨询师, 而是服务体系中的协作者, 核心价值在于扩大覆盖面、提升危机识别精度与提供持续性陪伴。“数心融合”的本质命题不是机器取代人, 而是人机协同实现服务效能最大化。

## 关键词

数心融合, AI心理机器人, 硕士生心理健康, 四层穿透式干预, 人机协同

## Logic, Mechanism, and Boundaries of AI Psychological Robots in Promoting Master's Students' Mental Health from the Perspective of Digital-Mental Integration

\*通讯作者。

文章引用: 焦松明, 时勤, 苏星赫, 李从庸, 唐伟杰, 张旭晟(2026). 数心融合视域下 AI 心理机器人促进硕士生健康的逻辑、机制与边界. *心理学进展*, 16(9), 232-242. DOI: 10.12677/ap.2026.169447

Songming Jiao<sup>1,2</sup>, Kan Shi<sup>1,3\*</sup>, Xinghe Su<sup>2,4</sup>, Congyong Li<sup>1,2</sup>, Weijie Tang<sup>1,2</sup>, Xucheng Zhang<sup>1,2</sup>

<sup>1</sup>Laboratory of Ecological Civilization and Environmental Governance, Wenzhou University, Wenzhou Zhejiang

<sup>2</sup>Academy of Wenzhou Model Development, Wenzhou University, Wenzhou Zhejiang

<sup>3</sup>School of Economics and Management, University of Chinese Academy of Sciences, Beijing

<sup>4</sup>International College, National Institute of Development Administration, Bangkok Thailand

Received: August 12, 2026; accepted: September 7, 2026; published: September 18, 2026

## Abstract

Against the backdrop of deep integration of digital technology and mental health services, “digital-mental integration” has emerged as a significant paradigm in the mental health field. Master’s students face structural difficulties characterized by high cognitive load, high psychological defensiveness, and low willingness to seek help, making traditional services insufficient to effectively address their needs. This study focuses on the AI psychological robot “Xiaoxin”, which adopts Yi-Shu Psychodrama as its theoretical core and integrates multimodal physiological monitoring with large language model technology to construct a four-layer penetrative intervention model encompassing “consciousness analysis → artistic experience → subconscious emotional reshaping → behavioral training”. This paper explores the topic from three aspects: theoretical logic, mechanisms of action, and practical boundaries. It elucidates the paradigmatic value of “digital-mental integration”, reveals the digital transformation of psychodrama and the human-robot empathic mechanisms of the AI robot, analyzes ethical risks such as privacy breaches and emotional deception, and proposes a three-dimensional collaborative optimization path involving technology, usage, and governance. The study indicates that AI psychological robots are not meant to replace counselors but serve as collaborators within the service system, with their core value lying in expanding coverage, enhancing crisis identification accuracy, and providing sustained companionship. The essential proposition of “digital-mental integration” is not machines replacing humans, but rather human-robot collaboration to maximize service efficacy.

## Keywords

Digital-Mental Integration, AI Psychological Robot, Master’s Students’ Mental Health, Four-Layer Penetrative Intervention, Human-Robot Collaboration

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

据《健康中国行动（2019—2030年）》<sup>1</sup>统计，我国抑郁症患病率为2.1%，焦虑障碍患病率达4.98%。心理健康和精神卫生日益受到社会各界重视，《“十四五”国民健康规划》<sup>2</sup>亦将心理健康纳入国家发展总体目标，相关政策文件亦多次提出加强心理健康服务的可及性与规范化<sup>3</sup>。硕士生群体是国家战略人才

<sup>1</sup>健康中国行动推进委员会. (2019) 健康中国行动（2019—2030年）. 中华人民共和国国家卫生健康委员会.

<https://www.nhc.gov.cn/guihuaxxs/c100133/201907/2a6ed52f1c264203b5351bdbbadd2da8.shtml>

<sup>2</sup>国务院办公厅. (2022) 国务院办公厅关于印发“十四五”国民健康规划的通知[国办发（2022）11号]. 中国政府网.

[https://www.ndrc.gov.cn/fggz/fzzlgh/gjjzxgh/202206/t20220601\\_1326725.html](https://www.ndrc.gov.cn/fggz/fzzlgh/gjjzxgh/202206/t20220601_1326725.html)

<sup>3</sup>国家卫生健康委等. (2026) 健全社会心理服务体系和危机干预机制实施方案. 人民网.

<http://health-app.people.cn/n1/2026/0409/c14739-40698310.html>

储备的核心层与创新活力的重要来源,其心理健康状况直接关乎高层次人才培养质量与社会可持续发展动能。已有研究表明,其压力源主要来自学业、科研、临床实践、就业、人际及经济等多重维度的叠加(宋娇等, 2022),且挑战性-阻断性压力源与情绪耗竭显著正相关,进而影响学业投入与创新行为(Zhao等, 2025)。国内研究进一步揭示,科研训练、毕业要求、就业竞争、导师互动与自我期待共同构成复合压力系统(刘源, 李双洁, 2024)。然而,高校心理健康服务体系面临结构性困境:专业咨询师数量严重短缺,师生比远低于 1:3000 的推荐标准;传统服务受时空制约,硕士生因科研任务繁重、时间碎片化,难以规律参与;心理评估多依赖自陈量表,易受社会称许性与心理防御影响,高风险个案难以及时发现。上述困境与硕士生“高认知负荷、高心理防御、低求助意愿”的特殊性相叠加,亟需探索心理健康服务供给新路径。

人工智能技术的快速发展为破解上述难题提供了可能。“数心融合”专项课题的启动,标志着数字技术与心理健康服务的深度融合已由学术概念上升为国家政策方向。该概念由国家心理健康和精神卫生防治中心与中国信息通信研究院于 2024 年联合提出,2025 年 11 月首批 60 家合作研究单位获遴选,迅速成为心理健康数字化转型的核心分析框架。高歌(2025)将其界定为“技术辅助+人文关怀”协同育人模式,强调 AI 情绪识别技术可优化“评估-干预-追踪”全流程。本研究将“数心融合”的理论内涵概括为三个递进层次:第一,数据融合,整合生理信号、行为数据、自陈报告、环境信息等多源异构数据,借助多模态 AI 眼动追踪、心率变异性检测等技术,形成心理健康动态画像,实现智能评估与风险监测;第二,技术融合,将 AI、大数据、物联网、VR/AR 等技术系统集成于心理健康服务流程;第三,服务融合,构建 AI 系统与人类专业人员的协同工作机制,体现为 AI 筛查与专家复核、AI 轻中度干预与人类深度治疗、AI 持续追踪与人类阶段性评估的结合。

基于上述分析,本研究聚焦以下核心问题:第一,“数心融合”的理论内涵及其对硕士生心理健康服务的范式价值;第二,AI 心理机器人“突破心理防御、深入潜意识”的作用机制;第三,AI 心理机器人的实践边界,以及技术创新与伦理风险之间的平衡路径。

## 2. AI 心理干预的现状

### 2.1. 数字心理健康干预的多元形态与实证基础

数字技术为心理健康服务提供了新载体。Malinauskas 和 Malinauskiene (2022)的元分析表明,CBT、正念、压力管理及网络干预对大学生总体有效,但异质性较大且长期效果不稳。Linardon 等(2024)基于 176 项随机对照试验指出,心理健康 App 对抑郁和焦虑有显著但效果较小。李佳等(2023)总结了远程视频、网页、VR 和移动应用等多元形式,认为其能提高效率、降低成本、减少污名,但亦面临依从性不足、个性化欠缺与效果评估标准不统一等挑战。

AI 会话代理与生成式大语言模型(LLMs)的兴起进一步拓展了干预边界。已有研究验证了会话代理在缓解心理困扰方面的潜力:Abd-Alrazaq 等(2020)发现聊天机器人可改善心理健康但效果不一且安全性报告不足;Vaidyam 等(2019)指出其在筛查、心理教育、症状监测及低强度支持中的应用,但临床整合与危机识别仍是难点。针对具体工具,Sabour 等(2023)以中文聊天机器人 Emohaa 为例,发现其可降低心理困扰并改善抑郁与失眠症状;He 等(2023)的综述支持会话代理减轻抑郁、焦虑和压力,但随访期偏短;Li 等(2023)的元分析量化表明,AI 会话代理可显著减轻抑郁和心理困扰,但未能显著提升总体幸福感。Khosravi 和 Azar (2024)则从用户视角强调,个性化、同理心、沟通质量、易用性与信任是持续使用的关键决定因素。

当前,生成式 AI 拓展了心理健康应用边界,但也带临床证据不足、安全性存疑与伦理风险等挑战(Sharma et al., 2024; Hua et al., 2024; Kolding et al., 2025; Jin et al., 2025)。Siddals 等(2024)的质性研究进一

步显示,用户视生成式 AI 为情绪表达空间与关系建议来源,但期待更可靠的安全护栏、记忆能力与引导功能。这表明 AI 心理健康应用不仅是技术命题,更涉及信任建立与安全保障。国内学界同样认可 AI 在评估与干预中的潜力,但强调隐私、偏见、责任归属及社会心理复杂性等问题,主张在技术效度、伦理规范与服务体系间寻求平衡(任萍等, 2022; 李瑶, 杨琳, 2022; 王慧等, 2025; 侯洁等, 2025; 袁小雅, 刘仪辉, 2025)。

“数心融合”与既有概念的核心差异在于其层次结构。数字心理健康(digital mental health)泛指所有数字化手段在心理领域的应用,其边界宽泛但缺乏结构性;远程心理服务(tele-psychology)侧重服务交付方式的时空延伸,关注点在于距离的消解;AI 辅助心理干预则聚焦于 AI 在评估或治疗中的工具性角色。

“数心融合”的理论增量在于,它将数据融合(多源异构数据的整合与动态画像)、技术融合(AI、物联网、VR/AR 的系统集成)与服务融合(人机协同的分级交付)建构为三个相互关联且可操作化的分析层次,使“如何融合”的问题获得了超越口号式表述的理论内涵。其中,数据融合构成技术融合中输入基础,技术融合构成服务融合的实现手段,服务融合则构成前两者的价值出口。三者之间形成“输入-处理-输出”的因果链条,而非简单的并列关系。

## 2.2. AI 智能心理机器人“小欣”的系统概述

当前,AI 智能心理机器人日益引发学界关注。本研究聚焦的 AI 智能心理机器人“小欣”为项目组所在的实验室于 2025 年 10 月引入,为“数心融合”理念的实践载体。“小欣”的系统历经 9 年研发,以易术心理剧(龚铤博士整合心理剧与我国传统文化之成果,融汇西方心理剧技术与中医“阴阳平衡”、道家“顺应自然”理念)为理论内核。2016 年在重庆医科大学附属第一医院精神科完成与药物治疗结合的干预研究,取得显著疗效及脑科学证据;2019 至 2024 年间累计完成 3100 余例有效数字化干预。2024 年 7 月,该系统接入大模型,通过 DeepSeek 与通义千问 Turbo 双模型运作,实现全人工智能的多模态评估、报告生成、干预策略制定及多轮干预的成熟流程。“小欣”的核心特色可概括为一个内核、双模型驱动、四层穿透。一个内核是指以易术心理剧的结构化流程(热身-演出-分享三阶段)为算法基础,实现传统治疗智慧的 AI 化转型。双模型驱动是指 DeepSeek 负责快速应答与日常对话、通义千问 Turbo 负责深度分析与复杂推理,兼顾响应速度与干预质量。

## 3. AI 心理机器人的理论逻辑

### 3.1. 从心理剧到 AI: 治疗智慧的算法化

心理剧由 Jacob L. Moreno 创立,是一种通过戏剧行动探索心理问题的表达式治疗方法,其核心理念是“行动胜过语言”(Walters, 2023),通过多种技术帮助个体在安全的情境中表达被压抑的情感、重构认知模式。传统心理剧包含锚定现实、呈现问题、角色交换、镜像、独白、空椅、替身、情绪宣泄、整合回归 9 个环节(Sampath, 2026),这些环节具有明确的操作步骤和可预期的治疗效应,这为其算法化提供了可能。成为心理剧导演需要累积足够的学时和心理咨询经验,而 AI 的介入为解决这一矛盾提供了新思路。

易术心理剧的 AI 化遵循“结构化→数据化→算法化”的技术路径,其具体操作体现为 3 个阶段。其一,在结构化阶段,将心理剧的热身-演出-分享三阶段拆解为可编程的操作单元。热身阶段对应初始对话模板库与用户情绪状态分类器;演出阶段包含角色设定、场景构建、替身对话、角色交换、空椅技术等子模块,每个子模块均有明确的触发条件与执行序列。例如,“角色交换”的触发条件是用户在对话中表达出对某一关系对象的强烈情绪(以情感词频与语义相似度为判断依据),其执行序列为,提示用户描述对方特征→引导用户以对方口吻回应→对比用户前后表达方式的差异→引导用户反思视角转换后的认知变化。分享阶段对应情绪表达引导与认知整合提示。上述操作单元的边界与转换规则,由临床心理

学家与 AI 工程师联合定义。其二，在数据化阶段，将上述结构化流程应用于 3100 余例数字化干预案例，系统记录每个操作单元的用户响应(包括语言内容、响应时长、情绪词频、主动退出节点等)，形成“干预动作 - 用户响应”配对数据集。该数据集为后续算法训练提供了标注了治疗师决策逻辑的样本基础。例如，当用户在角色交换阶段表现出强烈情绪反应时，治疗师的选择可能是延长该阶段而非按流程推进。这一决策逻辑被记录为训练信号。其三，在算法化阶段，基于上述配对数据，利用大模型(DeepSeek 与通义千问 Turbo)进行监督微调与强化学习。前者通过“干预动作 - 用户响应”配对数据学习治疗师在特定情境下的标准回应模式；后者通过用户反馈信号(如主动延长对话、情绪词频变化方向)优化干预策略的选择优先级。最终形成的算法并非简单匹配关键词，而是习得了心理干预中“何时推进、何时停留、何时转换技术”的决策逻辑。它将心理剧的治疗环节转化为标准化的操作步骤，通过数字化干预案例积累训练数据，并利用大模型技术学习治疗师在不同情境下的决策逻辑。AI 通过积累学习时长和经验替代部分导演和演员的角色，共情计算则能在心理剧中更真实地还原人类情感。这揭示了 AI 心理机器人区别于传统规则型聊天机器人的本质特征，它并非简单匹配关键词，而是通过大规模数据训练习得了心理干预的深层语法。

### 3.2. 人机共情的可能性与限度

AI 共情的可能性问题存在两种理论视角的对立。一方面，“计算机作为社会行动者”(CASA)范式认为，人们会下意识地将人类社会的互动脚本应用于计算机，从而像对待人一样对待机器(Nass et al., 1994)。另一方面，机器启发式(machine heuristic)理论指出，人们对机器持有冷漠、无情感的刻板印象，这种心理捷径可能导致人们认为机器无法像人类那样提供真诚的共情(Sundar & Kim, 2019)。实证研究表明，当提供完全相同的情感支持内容时，人类伙伴所激发的感知支持性显著高于聊天机器人，进而带来了更强的减压效果(Rubin et al., 2025)。这验证了机器启发式的存在，即人们在潜意识中仍倾向于认为机器缺乏情感和共情能力。Ong 等(2025)指出，当人们知晓共情回应来自 AI 时，感知共情评分会下降，尤其在“情感分享”和“关怀”维度上最为显著。AI 心理机器人实现共情计算的技术路径并非真正的情感体验，而是通过以下方式模拟共情性回应：其一，基于大语言模型的语义理解，识别用户表达中的情绪词、评价词与隐喻，将其映射到预先训练的情绪分类框架(如愉悦度 - 唤醒度二维空间)；其二，根据情绪分类结果，从回应策略库中选择匹配的共情模板(如“听起来你感到……”)并加以个性化改写；其三，通过多轮对话中的情绪变化追踪，动态调整回应的情感强度与表达方式。这一过程的本质是共情行为的算法模拟，而非共情体验的算法实现，前者属于行为层面的模仿与匹配，后者则涉及情感状态的共享与理解，后者目前仍属人类能力范围。

Sio 等(2025)针对硕士生的研究表明，感知有用性与全天候可及性促进其对 AI-CBT 的积极态度，而数据隐私、情感安全及复杂问题适配性疑虑则抑制使用意愿。这表明，硕士生对 AI 心理服务的接纳是理性权衡的结果，既期待便捷无评判的支持，又对其能力与安全性保持审慎。

### 3.3. 四层穿透式干预的理论建构

本研究构建的“四层穿透式干预模型”的创新之处在于，它并非简单地将传统心理治疗搬到 AI 平台，而是基于 AI 的技术特性重新设计了干预流程(见图 1)。AI 在艺术体验和潜意识工作阶段的介入，不是替代人类治疗师，而是创造了一种新的治疗可能性，即用户可以在无评判、低焦虑的环境中，更自由地探索自我。

值得关注的是，“潜意识情志重塑”是四层模型中理论上最深刻、而在实践上却最难实现的一层。AI 能否触及人类的潜意识过程，取决于如何界定触及。若将触及定义为现象学意义上的直接体验与共情

性理解，则 AI 不具备此能力；若将触及定义为基于行为指标对潜在认知冲突的推断与表征，则 AI 可通过特定技术路径实现有限度的触及，即下文所述的“潜意识工作的算法化”。从认知心理学的角度看，潜意识是指那些未被觉知但影响行为的信息加工过程。AI 系统可以通过分析用户在艺术表达中的意象、隐喻与情绪词频等特征，识别其深层认知和情感模式。这是“潜意识工作的算法化”的核心逻辑，即将潜意识过程转化为可观测、可分析的行为指标。

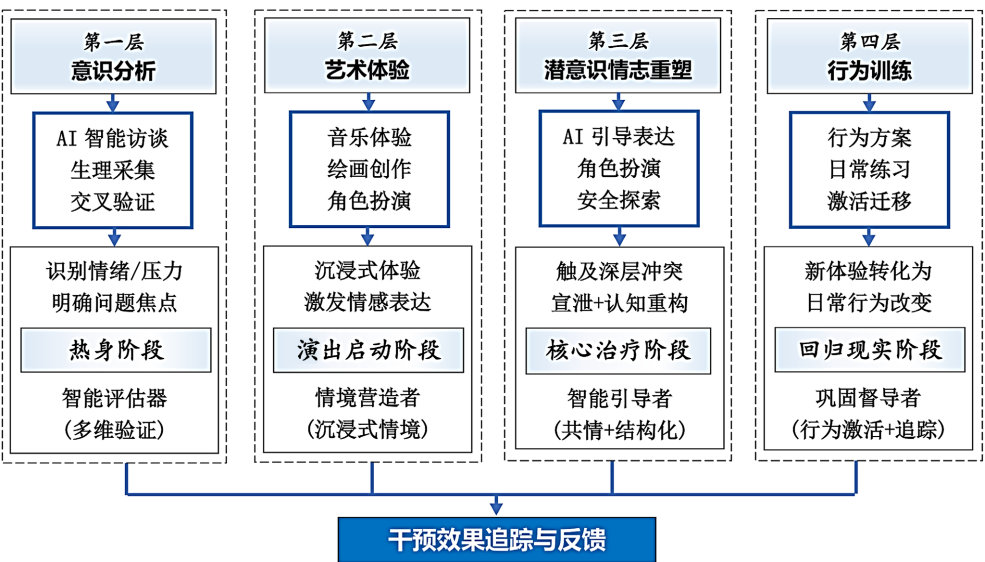


Figure 1. Four-tier penetrative intervention model  
图 1. 四层穿透式干预模型

但这一路径存在根本性局限：其一，AI 的分析对象是用户的语言输出与行为数据，而人类的防御机制可能使语言表达与深层体验之间出现系统偏差，概因语言可以是真实的，也可以是策略性的，AI 缺乏识别此差异的独立通道；其二，AI 无法感知人类的非语言信号(如微表情、躯体姿态、语音韵律)，这些信号往往是潜意识冲突的最直接表征；其三，AI 对潜意识的“触及”本质上是基于统计规律的间接推断，而非现象学意义上的直接共情性理解。因此，AI 对潜意识的识别是概率性的、可错的，其输出需经人类专业人员的复核与校准。这并非否定 AI 的价值，而是为其设定了合理的能力边界。

## 4. 作用机制分析

### 4.1. 突破心理防御

硕士生群体的高认知水平往往伴随着高心理防御。在接受面询时，他们可能因为担心被评判、被贴上标签而有所保留。AI 心理机器人的“无评判”特征在一定程度上突破了这一困境。

Raimi 等(2025)的研究提示，当用户知晓对话对象为 AI 时，其对“被评判”的感知反而高于人类专家。一项预注册实验比较了向 AI 聊天机器人倾诉与向“感知到的人类”倾诉的减压效果差异，所有回复内容均由 AI 生成后经编辑呈现。结果显示，效果量在 AI 辅助倾诉与感知人类辅助倾诉之间相近(Hu et al., 2026)。对于硕士生而言，AI 的非人身份反而成为一种优势，它消除了被评判的担忧，降低了自我暴露的心理成本。此外，AI 作为第一响应者，在用户最需要支持的时刻提供即时陪伴。这种模式特别适合硕士生群体时间碎片化、求助意愿低、对隐私高度敏感的需求特点。需指出的是，AI“无评判感知”的优势存在边界条件。“无评判”优势并非 AI 的固有属性，而是取决于用户对 AI 的初始信任水平与对心

理服务场景的预期。若其对 AI 的技术能力缺乏信心或对数据隐私高度担忧, AI 的非人身份可能从优势转化为劣势。因此, AI 心理服务的推广需配套用户教育与信任建立机制, 而非默认“AI 天然具有降低防御的效果”。

## 4.2. 情感激活与艺术通道

硕士生的理性思维习惯可能导致情感体验的抑制。AI 艺术疗愈模块通过音乐、绘画、心理剧等形式, 提供了绕过言语防御、激活情感体验的通道。从作用机制看, 艺术疗愈的效能源于其象征性表达特性。当用户通过 AI 引导进行心理剧角色扮演或曼陀罗绘画时, 其内在冲突被投射到外部对象上, 从而在安全距离内实现情感宣泄。心理剧系统允许用户将焦虑、恐惧、抑郁等情绪拟人化, 通过与之对话来处理这些情绪。这种方式比直接谈论情绪更具可接受性, 尤其适合情感词汇匮乏或情感表达困难的个体。更值得关注的是, AI 艺术疗愈的个性化潜力。Lin 等(2025)从 AIGC 曼陀罗艺术疗愈干预、Shang & He (2025) AI 协作绘画等不同研究视角指出, 生成式 AI 可根据用户的情感状态实时生成适配的图案等, 实现千人千面的干预方案。这突破了传统艺术疗愈标准化素材的限制, 使干预的精准性大幅提升。

## 4.3. 潜意识工作的算法化可能性

值得关注的是, “潜意识情志重塑”是四层模型中理论上最深刻、而在实践上却最难实现的一层。AI 能否触及人类的潜意识? 这一问题涉及心理学、认知科学和 AI 哲学的交叉领域。从认知心理学的角度看, 潜意识是指那些未被觉知但影响行为的信息加工过程。AI 系统可以通过分析用户在艺术表达中的意象(Achlioptas et al., 2021)、隐喻(张建丽, 王宇泽, 2025)、情绪词频(任巨伟等, 2013)等特征, 识别其深层认知和情感模式。这是“潜意识工作的算法化”的核心逻辑, 即将潜意识过程转化为了可观测、可分析的行为指标。这一路径的局限性在于, 人类本身的防御机制会让语言的真实性大打折扣, 而 AI 无法像人类心理咨询师或心理治疗师那样通过眼神、肢体动作或语音语调分辨真伪。这意味着, AI 对个体潜意识的触及仅为间接、基于行为指标推断, 而非直接、现象学意义上的共情性理解。

## 4.4. 人机协同的分工逻辑

AI 心理机器人的作用机制最终要落实到人机协同的实践框架中。AI 心理服务的理想模式不是机器取代人, 而是“AI 筛查与轻中度干预 + 人类专家处理复杂个案 + AI 持续追踪”的分级模式。广州市花都区引入的“阳光朵朵”AI 心理健康监测干预系统和贵州省三穗县检察院研发的“数智未检”智能体, 均为人机协同分工的真实案例。具体而言, 人机协同可分解为以下环节: 第一, AI 初筛: 通过多模态评估识别高风险个案, 提前预警。第二, AI 轻中度干预: 对一般性压力、轻中度焦虑抑郁提供标准化干预。第三, 人类专家深度干预: 对复杂创伤、人格障碍、危机个案进行专业治疗。第四, AI 持续追踪: 干预结束后, 由 AI 负责长期随访和复发预警。这一分工逻辑的核心是将专业人力资源配置到最需要的地方, 让 AI 承担基础性、标准化、高频次的工作, 让人类专家专注于需要深度共情、灵活应变、价值判断的复杂工作。

## 5. 实践边界与伦理反思

### 5.1. 隐私保护与数据安全

AI 心理服务涉及高度敏感的个人信息, 隐私保护是首要伦理议题。孙丹阳(2025)指出, 智能心理治疗在隐私保护、数据安全、数据偏见、机器幻觉等技术性困境上具有放大效应。具体而言, 其一是数据泄露风险。用户的心理访谈记录、HRV 数据、艺术创作内容若被不当访问或泄露, 可能造成严重心理伤

害。其二是数据滥用风险。用户数据可能被用于商业营销或保险定价等非预期目的。此外，还存在着数据偏见风险。训练数据若存在样本偏差，可能导致对特定群体的评估不准确或干预不适配。从技术层面看，边界感知个性化框架(BTP)提出了“最小化个性化”原则，在不存储用户身份信息的前提下实现个性化干预。[Kang \(2026\)](#)指出，个性化与参与度存在正向关系。这表明，在保护隐私与提供个性化服务之间可以找到平衡点。

## 5.2. 边界模糊与情感依赖

当 AI 能够提供共情回应时，用户可能产生对 AI 的情感依赖，模糊人与机器的关系边界。研究警告，超个性化陪伴型聊天机器人与依赖性和关系边界模糊相关([Edwards, 2026](#))。硕士生群体尤其面临这一风险。高强度的学术压力使他们更容易寻求情感支持，而 AI 的全天候可及性和无评判回应可能产生成瘾性使用模式。用户可能逐渐减少真实人际交往，将 AI 作为情感支持的唯一来源，这反而会加剧社交隔离问题。对此，有研究提出边界感知设计理念，即 AI 系统应有意维持最小化、边界保持的个性化方法，而非追求最大化拟人化或亲密性模拟([Kang, 2026](#))。具体措施包括明确告知用户“我是 AI，我没有真实情感”、设置合理的使用时长限制、在检测到严重情绪问题时主动建议寻求人类帮助。

## 5.3. 情感欺诈与主体性挑战

当 AI 能够模拟共情但并非真正共情时，是否构成一种情感欺诈？这是一个深层的伦理问题。我国青年报 2026 年的深度报道指出，大语言模型大规模输出“貌似共情、理解与关心的情感性反馈”，构成了 AI 时代的“情感幻觉”风险。支持者会认为，只要用户感受到被理解、被支持，AI 的真实性并不重要。反对者则会认为，基于模拟的关系本质上是一种欺骗，可能损害用户的自主性和对真实关系的认知能力。[孙丹阳\(2025\)](#)从主体性角度指出，智能心理治疗的伦理风险包括过度依赖导致的自我效能下降、主体性难题造成的情感欺诈与法律隐患。她强调，当用户将 AI 视为真正理解自己的人时，其对真实人际关系的期待模式可能被扭曲。这一问题尚无定论，但需在设计和使用中保持警觉。可能的缓解策略包括：在 AI 回应中嵌入真实性提示(如“作为 AI，我理解你的感受是基于数据分析……”)；鼓励用户在 AI 陪伴之外维持真实人际联结；定期评估用户的情感依赖程度。

## 5.4. 自主性与专业判断的消解风险

如果 AI 能够高效完成心理评估和干预，人类专业人员的角色是否会贬值？这是一个关乎职业未来的问题。一方面，AI 的介入确实改变了心理健康服务的专业生态。部分标准化、程序化的工作可能被 AI 替代，这可能引发专业人员的身份焦虑。另一方面，也有观点认为，AI 解放而非替代了专业人员，使其从繁重的文书工作、基础咨询中抽身，专注于更需要人类独特能力的工作。“数心融合”项目的官方立场是技术赋能而非技术替代。国家心理健康和精神卫生防治中心主任付强指出，面对精神卫生领域供给与需求之间的巨大鸿沟，以大数据、云计算、人工智能等为核心的新一代数智技术，为未来破解心理健康难题带来无限可能。这意味着，AI 被定位为破解供给缺口的工具，而非取代人类的技术。

## 5.5. 监管框架与治理路径

针对上述伦理风险，需构建“技术研发 - 实践使用 - 运营监管”多主体共治的风险防治路径。具体可阐释为下述层面的治理策略：在技术研发维度，要优化智能心理治疗功能与应用模式迭代，提升其诊疗质量。在实践应用维度，要增强使用者心理健康保健与数智素养库存，保障技术效用；在运营监管维度，要完善准入与监管流程并建立智能系统的督导机制。从政策层面来看，当前数心融合项目已探索建

立“建库、组网、分区、赋能”四大路径，通过遴选合作研究单位、组建跨领域技术指导组、设立片区牵头机制、搭建技术协作平台等举措，以强化对项目全过程的指导与支持。这一治理框架的建立，标志着我国的 AI 心理健康服务正走向规范化发展。

## 6. 结论与展望

### 6.1. 主要结论

本研究从理论逻辑、作用机制与实践边界三个维度，系统探讨了“数心融合”视域下 AI 智能心理机器人对硕士生心理健康促进的理论问题。主要结论如下：

其一，“数心融合”是涵盖数据融合、技术融合与服务融合三层次的整合性框架，其核心定位为“技术辅助人文”，在坚守人文关怀内核的前提下发挥数字技术的效能增益。

其二，AI 心理机器人“小欣”的作用机制可概括为“一个内核、双模型驱动、四层穿透”。

其三，AI 的实践边界由技术能力与伦理风险共同界定。

其四，AI 心理机器人的功能定位为服务体系中的辅助性角色，而非对人类咨询师的替代。

### 6.2. 理论贡献与实践启示

本研究首次构建“数心融合”理论框架，明确其内涵与层次结构，揭示 AI 心理机器人作用机制及“四层穿透式干预模型”的理论逻辑，并系统分析 AI 心理服务的伦理边界，为后续研究提供了理论参照。高校心理健康服务中，AI 心理机器人可作为破解供给缺口的有力工具，建议构建“AI 筛查→轻中度干预→专家处理复杂个案→AI 持续追踪”的分级服务模式，同时健全伦理监管机制，确保技术创新不逾越人文底线。

### 6.3. 研究局限与未来方向

本研究属理论思辨，缺乏实证检验，是主要局限。未来应开展大样本随机对照实验验证干预效果，探索个体差异因素以明确 AI 干预的适宜人群，深化“四层穿透模型”各层独立贡献的机制研究，并开展跨文化比较以探讨 AI 心理服务的文化适配性。展望未来，大模型迭代将提升 AI 共情能力与干预深度，多元数据融合将推动阶段性干预转向连续性照护，人机协同模式将日趋成熟。正如研究者所言，随着 XR 技术与可穿戴设备的融合，有望最终在元宇宙中实现更为仿真的沉浸式体验与真实温暖的触感还原。

## 基金项目

本研究得到国家心理健康和精神卫生防治中心、中国教育发展基金会、教育部学生服务与素质发展中心“向阳花学生心理健康促进行动”支持项目“高校心理健康管理者的胜任特征及培训模式”(项目编号：GX25A012)，以及国家心理健康与精神卫生防治中心、中国信息通信研究院“数心融合”专项课题“核心胜任特征成长评估模型的数心融合研究”(项目编号：SXRH-ZD-2025005)的支持。

## 参考文献

- 高歌(2025). 数心融合视域下 AI 情绪识别技术对高校心理服务的影响. *黑龙江科学*, 16(23), 82-84.
- 侯洁, 曾宏丽, 王勃, 等(2025). AI 赋能还是隐忧? 大学生 AI 使用动机的异质性及与心理健康的相互作用. *心理科学*, 48(5), 1038-1050.
- 李佳, 符仲芳, 田东华, 等(2023). 数字化干预在心理健康领域的发展与应用. *北京师范大学学报(社会科学版)*, (6), 127-140.
- 李瑶, 杨琳(2022). 人工智能应用于心理健康服务的相关问题思考. *医学与哲学*, 43(5), 49-54.

- 刘源, 李双洁(2024). 双一流建设高校研究生心理压力及应对策略分析. *中国心理学前沿*, 6(5), 727-732.
- 任巨伟, 杨亮, 林鸿飞(2013). 情感图式构造及其在文本情感计算中的应用. *江西师范大学学报(自然科学版)*, 37(2), 130-135.
- 任萍, 汪悦, 刘冬予, 等(2022). 心理健康评估与干预的智能化应用. *北京师范大学学报(社会科学版)*, (4), 150-160.
- 宋娇, 于景琦, 王睿琼, 等(2022). 研究生心理健康问题的分析与应对策略. *心理学进展*, 12(4), 1088-1093.
- 孙丹阳(2025). 智能心理治疗的伦理隐忧及可能出路. *自然辩证法通讯*, 47(9), 10-18.
- 王慧, 胡银环, 冯显东, 等(2025). 人工智能在心理干预中的应用: 效果、挑战与前景. *中国全科医学*, 28(25), 3209-3216.
- 袁小雅, 刘仪辉(2025). 人工智能在大学生心理健康评估与干预中的应用探究. *心理学进展*, 15(1), 182-187.
- 张建丽, 王宇泽(2025). 生成式人工智能机器人隐喻能力的实证研究. *外国语文*, 41(3), 41-51.
- Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, 22, e16021. <https://doi.org/10.2196/16021>
- Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., & Guibas, L. (2021). Artemis: Affective Language for Visual Art. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11564-11574). IEEE. <https://doi.org/10.1109/cvpr46437.2021.01140>
- Edwards, C. R. (2026). Closer but Intentionally Distant: Designing AI That Respects What Makes Us Human. *Interactions*, 33, 44-47. <https://doi.org/10.1145/3802835>
- He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q. et al. (2023). Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-Analysis of Randomized Controlled Trials. *Journal of Medical Internet Research*, 25, e43862. <https://doi.org/10.2196/43862>
- Hu, M., Ho, J. Q. H., Ng, C., Wong, S. S. M., & Hartanto, A. (2026). Emotional Support through AI: Venting to Artificial Intelligence or a Perceived Human May Offer Comparable Emotional Well-Being Benefits. *Computers in Human Behavior Reports*, 23, Article ID: 101149. <https://doi.org/10.1016/j.chbr.2026.101149>
- Hua, Y., Liu, F. L., Yang, K., Li, Z., Na, H., Sheu, Y. et al. (2024). Large Language Models in Mental Health Care: A Scoping Review. *Current Treatment Options in Psychiatry*, 12, Article No. 27. <https://doi.org/10.1007/s40501-025-00363-y>
- Jin, Y., Liu, J. Y., Li, P., Wang, B., Yan, Y., Zhang, H. et al. (2025). The Applications of Large Language Models in Mental Health: Scoping Review. *Journal of Medical Internet Research*, 27, e69284. <https://doi.org/10.2196/69284>
- Kang, B. (2026). Resolving the Personalization Paradox in Therapeutic AI: A Boundary-Aware Framework for Ethical Personalization. *AI and Ethics*, 6, Article No. 146. <https://doi.org/10.1007/s43681-026-01011-8>
- Khosravi, M., & Azar, G. (2024). Factors Influencing Patient Engagement in Mental Health Chatbots: A Thematic Analysis of Findings from a Systematic Review of Reviews. *Digital Health*, 10, Article ID 20552076241247983. <https://doi.org/10.1177/20552076241247983>
- Kolding, S., Lundin, R. M., Hansen, L., & Østergaard, S. D. (2025). Use of Generative Artificial Intelligence (AI) in Psychiatry and Mental Health Care: A Systematic Review. *Acta Neuropsychiatrica*, 37, e37. <https://doi.org/10.1017/neu.2024.50>
- Li, H., Zhang, R. W., Lee, Y. C., Kraut, R. E., & Mohr, D. C. (2023). Systematic Review and Meta-Analysis of AI-Based Conversational Agents for Promoting Mental Health and Well-Being. *NPJ Digital Medicine*, 6, Article No. 236. <https://doi.org/10.1038/s41746-023-00979-5>
- Lin, X., Gong, H., Xiao, Y., Peng, Y., & Li, S. (2025). Design Research and Exploration of Mandala Art Therapy Intervention Based on AIGC. In *IASDR 2025: Design Next*. Design Research Society. <https://doi.org/10.21606/iasdr.2025.509>
- Linardon, J., Torous, J., Firth, J., Cuijpers, P., Messer, M., & Fuller-Tyszkiewicz, M. (2024). Current Evidence on the Efficacy of Mental Health Smartphone Apps for Symptoms of Depression and Anxiety. A Meta-Analysis of 176 Randomized Controlled Trials. *World Psychiatry*, 23, 139-149. <https://doi.org/10.1002/wps.21183>
- Malinauskas, R., & Malinauskiene, V. (2022). Meta-Analysis of Psychological Interventions for Reducing Stress, Anxiety, and Depression among University Students during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, 19, Article 9199.
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers Are Social Actors. In *Conference Companion on Human Factors in Computing Systems—CHI '94* (p. 204). ACM Press. <https://doi.org/10.1145/259963.260288>
- Ong, D. C., Goldenberg, A., Inzlicht, M. et al. (2025). *AI-Generated Empathy: Opportunities, Limits, and Future Directions*. PsyArXiv Preprint. [https://doi.org/10.31234/osf.io/8n5jw\\_v1](https://doi.org/10.31234/osf.io/8n5jw_v1)
- Raimi, R., Kim, A., Ayabakan, S., & Dennis, A. R. (2025). Judgmental Bot: Conversational Agents in Online Mental Health Screening. *MIS Quarterly*, 49, 1319-1356. <https://doi.org/10.25300/misq/2025/18555>

- Rubin, M., Li, J. Z., Zimmerman, F., Ong, D. C., Goldenberg, A., & Perry, A. (2025). Comparing the Value of Perceived Human versus AI-Generated Empathy. *Nature Human Behaviour*, 9, 2345-2359. <https://doi.org/10.1038/s41562-025-02247-w>
- Sabour, S., Zhang, W., Xiao, X., Zhang, Y., Zheng, Y., Wen, J. et al. (2023). A Chatbot for Mental Health Support: Exploring the Impact of Emohaa on Reducing Mental Distress in China. *Frontiers in Digital Health*, 5, Article 1133987. <https://doi.org/10.3389/fdgth.2023.1133987>
- Sampath, D. (2026). Psychodrama: An Exploration of Self-Development Theatre as per M. In *Being and Becoming Your True Self* (Chapter 38). Taylor & Francis. <https://www.taylorfrancis.com/chapters/edit/10.4324/9781003686170-38/psychodrama-sampath>
- Shang, B., & He, Z. (2025). AI-Assisted Symbolic Healing for Digital Natives: A Jungian Perspective Based on the S-O-R and Fogg Behavior Models. *Acta Psychologica*, 260, Article ID: 105472. <https://doi.org/10.1016/j.actpsy.2025.105472>
- Sharma, A., Rushton, K., Lin, I. W., Nguyen, T., & Althoff, T. (2024). Facilitating Self-Guided Mental Health Interventions through Human-Language Model Interaction: A Case Study of Cognitive Restructuring. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-29). ACM. <https://doi.org/10.1145/3613904.3642761>
- Sio, C.-I., Mann, A., Fan, L., Cheung, A., & Lee, L.-H. (2025). Perceptions of AI-CBT: Trust and Barriers in Chinese Postgrads. In *Proceedings of the 30th International Conference on Technologies and Applications of Artificial Intelligence (TAAI 2025)*. <https://arxiv.org/abs/2602.03852>
- Siddals, S., Torous, J., & Coxon, A. (2024). “It Happened to Be the Perfect Thing”: Experiences of Generative AI Chatbots for Mental Health. *npj Mental Health Research*, 3, Article No. 48. <https://doi.org/10.1038/s44184-024-00097-4>
- Sundar, S. S., & Kim, J. (2019). Machine Heuristic: When We Trust Computers More than Humans. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-9). ACM. <https://doi.org/10.1145/3290605.3300768>
- Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *The Canadian Journal of Psychiatry*, 64, 456-464. <https://doi.org/10.1177/0706743719828977>
- Walters, R. (2023). Using the Sociodramatic Empty Chair with Children. *Zeitschrift für Psychodrama und Soziometrie*, 22, 217-225. <https://doi.org/10.1007/s11620-023-00750-8>
- Zhao, M., Yu, Y., Zhong, M., & Lin, B. (2025). Challenge-Hindrance Stressors and Innovative Performance among Dental Graduate Students in China: A Moderated Mediation Model. *Humanities and Social Sciences Communications*, 12, Article No. 707. <https://doi.org/10.1057/s41599-025-05005-5>