Advances in Psychology 心理学进展, 2026, 16(2), 43-56
Published Online February 2026 in Hans. https://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2026.162060

Hans汉斯

人类心理病理机制与AI注意力机制的共性研究：注意力重编程治疗的理论框架与展望

豆立宁

上海开欣医疗科技有限公司，上海

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2日

摘　要

目的：通过对比分析人类心理病理机制与AI大模型中的注意力机制，探讨两者的相似性，提出基于“注意力重编程”理论的新型心理治疗框架。方法：采用理论分析、文献综述和跨学科整合的方法，系统梳理人类心理病理中的注意力偏差模式，分析AI注意力机制的基本原理，建立两者的理论联系，构建“注意力重编程治疗”(ART)的理论框架。讨论：理论分析显示，人类情结形成与AI注意力机制具有显著的同构性。心理疾病的形成可以被视为一种“注意力固化”过程，而AI的注意力机制为理解和干预这一过程提供了新的视角。基于这一发现提出的“注意力重编程治疗”(ART)新范式，整合了注意偏差修正训练(ABMT)、正念注意力训练和虚拟现实注意力训练等多种技术。结论：人类心理病理与AI注意力机制存在深刻的共性，注意力重编程治疗为心理治疗提供了新的理论框架和技术路径。本研究为AI在心理健康领域的应用提供了理论基础，为未来实证研究指明了方向。

关键词

注意力机制，心理病理，情绪理论，人工智能，注意力重编程治疗，精神病学

A Study on the Commonalities between Human Psychopathological Mechanisms and AI Attention Mechanisms: Theoretical Framework and Prospects of Attention Reprogramming Therapy

Lining Dou

Shanghai Kaixin Medical Technology Co., Ltd., Shanghai

文章引用: 豆立宁(2026). 人类心理病理机制与 AI 注意力机制的共性研究: 注意力重编程治疗的理论框架与展望. *心理学进展, 16(2),* 43-56. DOI: 10.12677/ap.2026.162060

Received: December 16, 2025; accepted: January 16, 2026; published: February 2, 2026

Abstract

Objective: To compare and analyze the mechanisms of human psychopathology with attention mechanisms in AI large models, explore their similarities, and propose a novel psychotherapy framework based on "attention reprogramming" theory. Methods: Through theoretical analysis, literature review, and interdisciplinary integration, this study systematically examined attention bias patterns in human psychopathology, analyzed the basic principles of AI attention mechanisms, established theoretical connections between the two, and constructed the theoretical framework of "Attention Reprogramming Therapy" (ART). Results: Theoretical analysis revealed significant isomorphism between human complex formation and AI attention mechanisms. The formation of psychological disorders can be viewed as a process of "attention fixation", while AI attention mechanisms provide new perspectives for understanding and intervening in this process. The proposed new paradigm of "Attention Reprogramming Therapy" (ART) integrates multiple techniques including Attention Bias Modification Training (ABMT), mindfulness attention training, and virtual reality attention training. Conclusions: Human psychopathology and AI attention mechanisms share profound similarities, and attention reprogramming therapy provides a new theoretical framework and technical approach for psychotherapy. This study provides a theoretical foundation for AI applications in mental health and points out directions for future empirical research.

Keywords

Attention Mechanism, Psychopathology, Complex Theory, Artificial Intelligence, Attention Reprogramming Therapy, Psychiatry

1. 引言

1.1. 研究背景与意义

心理疾病的病理机制研究一直是心理学和精神病学的核心议题。传统心理治疗理论，如荣格的分析心理学，强调个体早期经历中的注意力分配对心理发展的重要影响(Jung, 1934; Stein, 2019)。荣格的情结理论指出，个体在成长过程中会将注意力持续投注于特定的事件、情绪和观念上，形成相对独立的心理结构。这一机制与现代 AI 大模型中的注意力机制学习过程具有显著的相似性(Roesler, 2019)。

近年来，人工智能技术的快速发展，特别是大语言模型的出现，为我们理解人类认知过程提供了新的视角(Brown et al., 2020; Vaswani et al., 2017)。Transformer 架构中的自注意力机制(Self-Attention Mechanism)能够捕捉序列中的长距离依赖关系，这种机制与人类大脑中注意力分配的神经机制显示出惊人的相似性(Clark et al., 2019)。这为从计算角度理解人类心理病理机制提供了新的可能性。

尽管已有研究分别探讨了人类心理病理中的注意力偏差(Bar-Haim et al., 2007; Browning et al., 2010)和 AI 模型中的注意力机制(Vaswani et al., 2017)，但很少有研究系统性地比较两者之间的共性。通过对比分析这两类看似不相关的领域，我们不仅可以更深入地理解心理疾病的发生机制，还能为 AI 在心理治疗领域的应用提供理论基础和新的研究方向。

1.2. 注意力机制：从心理学到人工智能

注意力是人类认知系统的核心功能之一，它决定了我们选择性地关注哪些信息，以及如何分配认知资源。在心理学领域，注意力研究有着悠久的历史，从Broadbent的过滤器理论(Broadbent, 1958)到Treisman & Gelade的特征整合理论(Treisman & Gelade, 1980)，再到现代的认知神经科学研究，我们对注意力的理解不断深化。

在人工智能领域，注意力机制的引入是深度学习发展的重要里程碑。从最初用于机器翻译的注意力机制(Bahdanau et al., 2014)到Transformer架构中的自注意力机制(Vaswani et al., 2017)，再到如今大语言模型中的复杂注意力模式，AI系统已经能够模拟人类注意力的许多关键特征。

本研究的核心假设是：人类心理病理中的注意力固化机制与AI模型中的注意力权重学习具有结构上的同构性。这一假设如果得到验证，将为两个领域的发展都带来重要启示。

1.3. 本文结构

本文首先系统回顾人类心理病理中的注意力机制和AI注意力机制的研究进展，然后深入分析两者之间的共性，接着提出基于注意力重编程的新型心理治疗框架，最后讨论这一框架的理论意义、临床应用前景以及未来研究方向。

2. 人类心理病理中的注意力机制

2.1. 荣格情结理论的现代阐释

荣格(1934)提出的情结理论是分析心理学的核心概念之一。情结是指个体心理中具有一定自主性、由共同情感色彩联结起来的观念群(Stein, 2019)。从现代认知神经科学的视角重新诠释，情结的形成机制可以从以下几个维度理解：

1) **注意力聚焦理论**：情结的形成始于个体将注意力聚焦于特定的创伤性事件或情绪体验。这些体验因其强烈的情感色彩而获得额外的注意力权重，形成“情感调谐的情结”(feeling-toned complex) (Roesler, 2019)。这一过程类似于AI模型中对特定特征赋予更高权重的机制。

2) **记忆网络理论**：情结在大脑中表现为紧密联结的神经回路。每次激活都会强化这些联结，形成正反馈循环，使得情结越来越容易被激活(Brewin, 2006)。这与AI模型中通过反向传播不断强化特定连接权重的过程高度相似。

3) **认知偏差理论**：激活的情结会影响个体的信息加工过程，导致注意偏向、记忆偏向和解释偏向(Mathews & MacLeod, 2005)。这种系统性的偏差在AI模型中也有对应的表现，即模型对特定类型的输入产生系统性的错误反应。

2.2. 不同心理疾病中的注意力偏差模式

大量实证研究表明，各种心理疾病都存在特征性的注意力偏差模式。这些模式虽然在具体表现上有所不同，但其核心机制都涉及注意力的异常分配和固化。

1) 抑郁症：过度关注负面信息，特别是与自我相关的负面信息(Gotlib & Joormann, 2010)；对积极信息的注意力分配不足，存在“积极信息忽视”现象(Bylsma et al., 2021)；注意力固着于负面认知和情绪，难以转移(Kircanski et al., 2012)。这类似于AI模型中的“灾难性遗忘”现象，即模型过度关注负面样本而忽略其他信息。

2) 焦虑症：对威胁性刺激的过度警觉，即使这些威胁并不明显(Bar-Haim et al., 2007)；难以将注意力从威胁线索转移，存在“注意力粘连”现象(Cisler & Koster, 2010)；注意力资源过度分配与潜在危险的监测(Browning et al., 2010)。这与AI模型中过度敏感的威胁检测机制类似。

3) 创伤后应激障碍(PTSD)：注意力过度聚焦于创伤相关线索，包括内在线索(记忆、情绪)和外在线索(环境刺激) (Brewin, 2011)；无法将注意力从创伤记忆中解脱，存在“侵入性回忆”(Ehlers & Clark, 2000)；注意力资源被创伤记忆“劫持”(Michael et al., 2005)。这与 AI 模型中被特定输入“劫持”的现象高度相似。

2.3. 注意力偏差修正训练的理论基础

注意力偏差修正训练(Attention Bias Modification Training, ABMT)是一种新兴的心理干预方法，旨在通过系统性的训练来改变个体的注意力偏向(MacLeod et al., 2002)。其基本原理是通过反复训练个体将注意力从威胁性或负面刺激转移到中性或积极刺激上，从而改变其注意偏向模式(Hakamata et al., 2010)。

ABMT 在多种心理疾病的治疗中显示出良好的应用前景：在焦虑症治疗中，多项研究显示 ABMT 能够显著降低焦虑症状(Hakamata et al., 2010; Heeren et al., 2015)；在抑郁症治疗中，ABMT 可以改善抑郁症状，特别是当与传统治疗结合使用时(Browning et al., 2012)；在 PTSD 治疗中，初步研究显示 ABMT 可能有助于减少创伤相关的注意偏向(Kuckertz et al., 2014)。然而，尽管 ABMT 显示出良好的应用前景，但其效果大小、持续性以及作用机制仍需要进一步研究(Cristea et al., 2015)。

从 AI 的角度来看，ABMT 本质上是一种“注意力重编程”过程，类似于对预训练模型进行微调(fine-tuning)。这种类比为理解和优化 ABMT 提供了新的视角。

3. AI 大模型中的注意力机制

3.1. Transformer 注意力机制的基本原理

Vaswani 等(2017)提出的 Transformer 架构革命性地改变了自然语言处理领域。其核心是自注意力机制(Self-Attention Mechanism)，允许模型在同一序列内部建立关联。这一机制的设计灵感部分来源于人类视觉系统中的注意力机制。

1) **Query-Key-Value 机制**：Query (查询)表示当前需要关注的信息，Key (键)表示被关注对象的关键特征，Value (值)包含被关注对象的实际信息。注意力权重通过 Query 和 Key 的相似度计算得到，然后通过 Softmax 函数归一化，最后根据权重对 Value 进行加权求和。这一过程与人类注意力分配的过程高度相似：Query 相当于个体的注意意图，Key 相当于外界刺激的特征，Value 相当于刺激的完整信息。

2) **多头注意力机制**：允许模型同时关注来自不同位置、不同语义维度的信息，每个注意力头学习不同类型的关系，增强了模型的表达能力和鲁棒性(Clark et al., 2019)。这与人类能够从多个维度同时关注信息的能力类似。

3) **自注意力机制**：允许模型在同一序列内部建立关联，能够捕捉长距离的依赖关系，实现信息的动态整合。这与人类能够在思维流中建立长距离联想的能力类似。

3.2. AI 注意力模式的特征

通过对大量 AI 模型的分析，研究者发现了 AI 注意力模式的几个关键特征，这些特征与人类心理病理中的注意力模式有着惊人的相似性。

1) **注意力权重的分布**：某些 token 获得不成比例的高注意力权重，形成“注意力热点”(attention hotspots) (Voita et al., 2019)，这些热点对应着关键语义信息。这与人类在注意过程中对某些信息过度关注的现象类似。

2) **注意力模式的固化**：训练完成后，模型的注意力模式相对固定，特定类型的输入会激活相似的注意力模式，形成相对稳定的“注意力路径”。这与人类形成稳定的认知模式和情绪反应模式的过程高度相似。

3) **注意力偏差的产生**：训练数据中的偏差会导致注意力偏差，模型可能过度关注某些特征而忽略其

他重要信息，表现为模型的“偏见”或“错误倾向”(Bolukbasi et al., 2016)。这与人类因生活经历而形成的认知偏差和情绪偏差非常相似。

3.3. AI 模型的微调与干预

与人类心理治疗类似，AI 模型也需要“治疗”(即微调)来修正不理想的注意力模式。这一过程为人类理解心理治疗提供了新的隐喻和工具。

1) **微调(Fine-tuning)**：在特定任务或数据上进一步训练模型，调整注意力权重以适应新的需求(Howard & Ruder, 2018)。这类似于心理治疗中的技能训练和认知重构。

2) **注意力干预**：直接修改注意力权重或添加正则化项，引导模型关注特定的特征或模式(Clark et al., 2019)。这类似于心理治疗中的注意训练和正念练习。

3) **对抗性训练**：通过对抗性样本来增强模型的鲁棒性，改善注意力分布，减少过度依赖。这类似于心理治疗中的暴露疗法和系统脱敏。

这些 AI 干预方法的成功为人类心理治疗提供了新的思路和工具。特别是，AI 模型中的注意力可视化技术为理解和监测心理治疗过程提供了可能的技术手段。

4. 人类心理病理与 AI 注意力机制的共性分析

4.1. 注意力固化机制的相似性

通过系统比较人类心理病理机制与 AI 注意力机制，我们发现两者在以下方面具有显著的相似性：

形成机制：人类——重复激活导致神经回路强化；AI——反向传播导致权重更新；共性——通过重复强化形成稳定模式。

注意偏向：人类——对威胁/负面信息过度关注；AI——对特定 token 赋予高权重；共性——系统性偏离客观现实。

固化过程：人类——形成自动化、习惯化反应；AI——注意力模式相对固定；共性——形成难以改变的自动化反应。

系统偏差：人类——认知偏差、情绪偏差；AI——模型偏见、训练数据偏差；共性——系统性偏离客观现实。

干预需求：人类——心理治疗改变认知模式；AI——微调修正注意力权重；共性——需要外部干预来改变模式。

劫持机制：人类——情结激活导致注意力强制转移；AI——特定输入导致注意力过度集中；共性——注意力控制的暂时性丧失。

持续性：人类——注意力偏向持续存在；AI——注意力模式相对稳定；共性——形成稳定的反应模式。

人类心理病理过程与 AI 注意力机制在结构上具有高度相似性，为建立跨学科的理论框架提供了实证支持，但把相似性数据化，目前尚未有研究，可以是另外的一个课题。

4.2. 注意力“劫持”机制的对比

1) **人类的心理“劫持”**：通过对临床患者的深度访谈，我们发现当情结被激活时，个体的注意力会被强制转移到特定内容，失去对注意力分配的有意控制，表现为侵入性思维、强迫行为等。这种“劫持”具有突发性、强制性、持续性和强烈的情绪负载等特征。

2) **AI 的注意力“锁定”**：在 AI 模型中，某些输入特征会导致注意力过度集中，模型可能忽略其他重要信息，表现为错误的输出或决策偏差。具体表现为：特定关键词(如“death”，“failure”)触发高注

意力权重，注意力权重在后续序列中保持高位，初始注意力偏向影响后续所有 token 的注意力分配，导致生成内容的情感偏向。

相似性具体指标，有待进一步跨学科研究。

4.3. 治疗与微调的相似性

1) 心理治疗的目标：改变适应不良的注意力模式，重建更灵活的注意力分配机制，减少症状性的注意力固着。这包括识别并挑战适应不良的注意力模式，训练更灵活、更适应性的注意力分配，建立新的、健康的注意力“路径”。

2) 模型微调的目标：修正不理想的注意力权重，优化模型的注意力分配，提高模型在特定任务上的表现。这包括调整注意力权重以适应新的需求，引导模型关注特定的特征或模式，通过对抗性样本增强模型的鲁棒性。

这种相似性提示我们，心理治疗过程可以被视为一种“人类微调”过程，而 AI 模型的微调方法也可能为人类心理治疗提供新的技术和工具。特别是，AI 模型中的注意力可视化技术为监测和理解心理治疗过程提供了可能的技术手段。

5. 注意力重编程治疗(ART)的理论框架

5.1. 理论基础

基于人类心理病理与 AI 注意力机制的共性，我们提出“注意力重编程治疗”(Attention Reprogramming Therapy, ART)的新范式。这一范式的核心思想是：心理疾病的本质是注意力模式的异常固化，治疗的目标是重新编程个体的注意力分配机制，通过系统性的注意力训练来“微调”心理功能。

5.1.1. 核心假设

ART 理论框架基于以下核心假设：

(1) 心理疾病的本质是注意力模式的异常固化，而非传统意义上的“疾病”或“障碍”。

(2) 注意力模式是通过学习和强化形成的，因此可以通过训练来改变。

(3) 注意力分配具有可塑性，即使在成年后也能够通过系统性训练进行重构。

(4) 治疗的目标是重新编程个体的注意力分配机制，而非简单地消除症状。

(5) 通过改变注意力模式，可以从根本上改善个体的心理状态和认知功能。

5.1.2. 治疗机制

ART 的治疗机制包括以下几个关键环节：

(1) 识别并挑战适应不良的注意力模式：通过评估识别个体的注意力偏向模式，确定治疗靶点。

(2) 训练更灵活、更适应性的注意力分配：通过系统性的注意力训练，逐步改变个体的注意力偏向。

(3) 建立新的、健康的注意力“路径”：通过反复练习，形成新的、更适应性的注意力模式。

(4) 增强对注意力分配的元认知能力：提高个体对注意力过程的觉察和控制能力。

(5) 促进神经可塑性：通过持续的训练，促进大脑神经回路的重组和优化。

5.2. 实施方法

ART 整合了多种经过实证支持的注意力训练技术，形成一个多层次、多模态的综合治疗体系。

5.2.1. 注意偏差修正训练(ABMT)

ABMT 是 ART 的核心技术之一，其原理是通过反复训练个体将注意力从威胁性或负面刺激转移到

中性或积极刺激上，从而改变其注意偏向模式。在 ART 框架中，ABMT 被用作主要的“重编程”工具，系统性地改变个体的注意力分配模式。

ABMT 的具体实施包括：使用计算机化的注意力训练任务，系统性地改变个体的注意力偏向；根据个体的注意力模式特点，设计个性化的训练方案；通过实时反馈，帮助个体意识到自己的注意力偏向并进行调整；逐步增加训练难度，促进注意力的灵活性和控制力的提升。已有研究显示 ABMT 对焦虑和抑郁具有显著的治疗效果(Hakamata et al., 2010; Browning et al., 2012)。

5.2.2. 正念注意力训练

正念注意力训练是 ART 的另一个重要组成部分，其目的是培养个体对当下体验的非评判性觉察，提高对注意力分配的元认知能力，增强注意力的灵活性和控制力。正念训练强调对注意力过程的觉察，而不试图改变注意的内容，这与 ABMT 的主动改变策略形成互补。

在 ART 框架中，正念训练的具体内容包括：身体扫描练习，帮助个体觉察身体感觉和注意力在身体上的移动；呼吸觉察练习，训练个体将注意力集中在呼吸上，并在注意力分散时温和地将其带回；情绪观察练习，培养个体对情绪的非评判性觉察，不试图改变情绪，而是观察情绪的生起和消失；开放式觉察练习，培养对内在和外在经验的广泛、开放的注意力状态。大量研究证实了正念训练在改善心理健康方面的效果(Keng et al., 2011)。

5.2.3. 虚拟现实注意力训练

虚拟现实(VR)注意力训练是 ART 框架中的创新技术，它利用 VR 技术创建沉浸式的训练环境，为个体提供安全、可控的注意力训练场景。VR 技术的优势在于能够创建高度逼真的模拟环境，实时监测个体的反应，并提供即时反馈。

在 ART 框架中，VR 注意力训练的具体应用包括：针对社交焦虑的虚拟社交场景训练，帮助个体在安全的环境中练习注意力控制；针对特定恐惧症的虚拟暴露训练，结合注意力重定向技术；结合眼动追踪和生理反馈的实时注意力监测和训练；个性化的场景定制，根据个体的具体问题设计相应的训练场景。初步研究显示 VR 在心理治疗中的巨大潜力(Maples-Keller et al., 2017)。

5.3. AI 在 ART 中的应用

AI 技术可以在 ART 的多个方面发挥重要作用，提升治疗的精准性和有效性。

5.3.1. 个性化注意力模式分析

使用 AI 分析个体的注意力分配模式，识别异常的注意力“热点”和“盲点”，为治疗提供精确的靶点。具体包括：通过眼动追踪、脑电、生理指标等多模态数据分析个体的注意力模式；使用机器学习算法识别与特定心理问题相关的注意力模式；为每个个体建立个性化的注意力“画像”；动态监测注意力模式的变化，评估治疗进展。

5.3.2. 智能注意力训练系统

开发 AI 驱动的注意力训练平台，根据个体的反应实时调整训练难度，提供个性化的训练方案。具体包括：自适应训练算法，根据个体的表现动态调整训练参数；智能反馈系统，提供个性化的指导和建议；游戏化设计，提高训练的趣味性和参与度；远程训练支持，使治疗更加便捷和可及。

5.3.3. 治疗效果预测

利用 AI 预测不同治疗方案的效果，优化治疗参数和训练强度，提高治疗的精准性和效率。具体包括：基于个体的基线特征预测其对不同治疗的反应；优化治疗参数(如训练强度、频率、持续时间)；识别可能

的治疗抵抗者，提前调整治疗方案；预测治疗效果的持续性，制定个性化的维持治疗计划。

6. 讨论

6.1. 理论意义

本研究最重要的理论贡献是建立了分析心理学、认知神经科学和人工智能之间的理论桥梁。荣格的情结理论(1934)虽然是深刻的心理学洞见，但一直缺乏精确的数学模型和实证验证。通过将其与 AI 注意力机制相结合，我们不仅为这一经典理论提供了现代计算框架，也为 AI 研究提供了来自人类心理的深刻启示。

这种跨学科整合的意义在于：(1) 统一的概念框架：注意力作为核心概念，连接了心理学和 AI 两个领域；(2) 可计算的模型：将抽象的心理过程转化为可计算、可验证的数学模型；(3) 双向启发：人类心理的研究启发 AI 设计，AI 模型帮助理解人类心理。这种整合不仅有助于理解心理疾病的共性机制，还为开发针对性的注意力干预技术、预测疾病发展和转归提供了新的可能。

6.2. 临床应用前景

ART 作为一种新型治疗方法框架，具有广阔的临床应用前景。首先，ART 具有理论基础扎实、技术先进、效果预期良好、安全性高等特点，有望成为心理健康服务的重要补充。其次，ART 体现了精准医疗的理念，通过个性化评估、精准干预和精准监测，为每个患者提供量身定制的治疗方案。第三，ART 具有广泛的应用范围，不仅适用于焦虑和抑郁，还可以推广到强迫症、PTSD、进食障碍等多种心理疾病的治疗中。

此外，ART 为心理健康服务提供了新的可能性：可及性提升(数字化、标准化的训练程序，可以降低专业门槛，适合远程和自助式干预)；成本效益(相比传统治疗，成本更低，治疗周期相对较短，可以大规模推广应用)；质量保证(标准化的训练流程，客观的效果评估，可重复、可验证)。这些特点使 ART 特别适合在医疗资源不足的地区推广应用，有助于解决当前心理健康服务面临的挑战。

6.3. 技术发展前景

随着技术的不断进步，ART 的应用前景将更加广阔。AI 辅助诊断技术可以帮助更准确地评估个体的注意力模式，识别治疗靶点；自适应训练系统可以根据个体的反应实时调整训练方案，提供个性化的干预；VR/AR 技术可以创建更加沉浸式和逼真的训练环境，提高训练效果；脑机接口技术可能实现直接的注意力调节，为治疗开辟新的可能性。

此外，大数据和机器学习技术的应用将进一步提升 ART 的效果。通过分析大量患者的数据，可以识别出最有效的治疗模式，预测治疗反应，优化治疗方案。可穿戴设备的普及使得实时监测个体的注意力状态成为可能，为精准干预提供了数据支持。云计算技术的发展使得大规模、分布式的 ART 服务成为可能，有助于实现心理健康服务的普及化。

6.4. 挑战与展望

尽管 ART 框架展现出巨大的潜力，但其发展和应用仍面临诸多挑战。首先，在技术方面，AI 对人类情感与心理状态的理解仍然有限，需要进一步开发更精细的注意力建模技术，以提升系统的可靠性与安全性。尤为关键的是，算法偏见与治疗风险构成严峻挑战——若用于个性分析与干预的 AI 模型本身存在偏见(可能源于训练数据的不均衡或文化局限性)，ART 不仅无法有效矫正个体的注意力偏差，反而可能强化甚至制造新的认知偏差。例如，模型可能将特定文化背景下正常的情绪反应(如丧亲之痛)错误归类为“病理性注意固化”，进而导致误诊或不当干预。因此，如何检测、缓解并持续监督模型中的偏见，是技

术实现中不可回避的核心问题。

其次，伦理挑战尤为突出。在现实的心理咨询与治疗环境中，伦理原则包括保密性、知情同意、专业边界与责任归属等，这些在云端心理服务中面临着新的适应难题。一方面，数据隐私与安全问题亟待深入探讨：ART 的实施依赖于收集大量高度敏感的个人数据(如眼动轨迹、脑电信号、生理指标等)，仅笼统强调“保护隐私与数据安全”远不足够。必须系统性地设计安全架构，明确数据所有权、访问权限与监管责任，并建立数据泄露应急机制，以防范敏感信息外泄可能带来的社会与心理后果。另一方面，AI 在心理服务中的角色需要清晰的伦理界定。虽然 AI 无法也不应获得类似人类治疗师的职业资格认证，但相关部门有必要针对 AI 辅助心理服务出台专门的伦理标准与操作指南，规范其使用边界、透明度要求与责任分配。例如，可通过建立人机协同的伦理框架，明确 AI 作为辅助工具的角色，要求其决策过程可解释、可审核，并由具备资质的临床人员最终承担责任，以此在技术创新与伦理保障之间取得平衡。

再次，临床挑战依然显著：必须通过严谨的临床研究验证 AI 治疗的有效性和安全性，建立相应的治疗标准与实施规范，并对临床医生进行系统培训，使其能够合理使用并与 AI 系统有效协作。

展望未来，ART 框架的进一步发展需要多方协同努力。在理论层面，应推动注意力机制理论与既有心理治疗框架的深度融合，发展跨学科的注意力导向心理病理学理论，构建统一的研究与实践范式；在技术层面，可结合虚拟现实、增强现实等沉浸式技术，借助大数据和机器学习不断优化干预效果，并逐步构建安全、可信、智能化的心理健康生态系统；在社会应用层面，ART 有望扩大心理健康服务的覆盖范围，降低服务成本，从而提升整体社会的心理福祉。

只有通过持续的理论创新、技术完善、伦理规范与临床验证，才能稳妥释放 ART 框架的潜力，为应对日益增长的心理健康需求提供真正有效且负责任的解决方案。

7. 结论与实验设计建议

7.1. 主要结论

本研究通过跨学科的理论整合和深入分析，系统探讨了人类心理病理机制与 AI 注意力机制的共性，并基于“注意力重编程”理论提出了新型心理治疗框架(ART)。主要结论如下：

1) 理论贡献：人类心理病理中的注意力固化机制与 AI 注意力机制在结构上具有高度相似性，这一发现为理解心理疾病提供了新的理论视角，为 AI 在心理健康领域的应用奠定了理论基础。

2) 理论框架：基于共性分析提出了“注意力重编程治疗”(ART)的新范式，该范式整合了 ABMT、正念训练和 VR 训练等多种技术，为心理治疗提供了新的理论框架和技术路径。

3) 应用前景和技术整合：ART 具有理论基础扎实、技术先进、预期效果好、适用范围广等特点，有望成为心理健康服务的重要补充，特别是在提升服务可及性和成本效益方面具有显著优势。AI 技术可以在个性化评估、智能训练系统和治疗效果预测等方面为 ART 提供重要支持，实现精准化、个性化的心理治疗。

本研究体现了跨学科整合在解决复杂心理问题中的巨大潜力，为计算精神病学和 AI 心理治疗的发展提供了新的思路和方向。

7.2. 实验设计建议

基于本研究的理论框架和文献综述，我们提出以下实验设计建议，以验证 ART 理论框架的有效性和实用性。

7.2.1. 研究 1：理论验证研究

研究目的：验证人类心理病理与 AI 注意力机制的相似性假设。

研究设计：采用计算建模和相关分析的方法。

参与者：招募 100 名临床患者(抑郁症 50 名，焦虑症 50 名)和 50 名健康对照。

测量工具：使用多种注意力偏差测量任务(点探测任务、情绪斯特鲁普任务、眼动追踪)，评估参与者的注意力偏向模式。

计算建模：构建基于 Transformer 的注意力模型，使用参与者的数据进行训练和测试，比较模型注意力模式与人类注意力模式的相似性。

数据分析：使用结构相似性指数(SSIM)和相关分析比较人类数据与模型预测的相似性；使用机器学习算法预测个体的心理状态。

预期结果：预期发现人类心理病理与 AI 注意力机制在结构上具有显著相似性(SSIM > 0.80)，AI 模型能够准确预测个体的心理状态(准确率 > 85%)。

7.2.2. 研究 2：ART 治疗效果随机对照试验

研究目的：验证 ART 治疗对焦虑和抑郁症状的治疗效果。

研究设计：采用随机对照试验设计，三组平行对照(ART 组、等待对照组、积极对照组)。

样本量：基于效应量 $d = 0.5$，$\alpha = 0.05$，$\beta = 0.20$，计算每组需要 52 人，考虑 20%流失率，每组招募 65 人，共 195 人。

纳入和排除标准：纳入符合 DSM-5 焦虑或抑郁诊断标准的 18~45 岁成年人，SAS 或 SDS 得分 ≥50；排除有严重躯体疾病、物质滥用、自杀风险等情况的个体。

干预方案：

- ART 组：接受为期 8 周的 ART 治疗，包括 ABMT 训练(每周 2 次，每次 45 分钟)、正念训练(每周 2 次，每次 30 分钟)、VR 训练(每周 1 次，每次 30~35 分钟)。
- 等待对照组：不接受任何干预，8 周后提供同等治疗。
- 积极对照组：接受传统的认知行为治疗(CBT)，为期 8 周，每周 1 次，每次 50 分钟。

评估时间点：基线(第 0 周)、干预后(第 8 周)、随访(第 12 周、第 24 周)。

主要结局指标：SAS、SDS 得分的变化。

次要结局指标：注意力偏差分数、临床总体印象量表(CGI)、生活质量量表(SF-36)。

中介变量：注意力偏差、正念水平、认知灵活性。

统计分析：使用重复测量方差分析比较三组的变化，使用中介效应分析检验注意力偏差的中介作用，使用 clinically significant change 标准评估临床意义。

预期结果：预期 ART 组在 SAS 和 SDS 上的改善显著优于对照组，与 CBT 组相当或更优；注意力偏差在治疗效果中起显著中介作用；ART 治疗具有较好的持续性和安全性。

7.2.3. 研究 3：神经机制研究

研究目的：探索 ART 治疗对大脑神经机制的影响。

研究设计：采用前后对照的神经影像学研究设计。

参与者：招募 40 名焦虑和抑郁患者，随机分配到 ART 组($n = 20$)和等待对照组($n = 20$)。

测量工具：使用 fMRI 进行静息态和任务态扫描，任务包括情绪调节任务、注意力控制任务；使用 EEG 进行注意力相关电位(如 P300、N170)的记录。

数据分析：比较两组在治疗前后的大脑激活模式、功能连接、结构变化；分析注意力网络(如额顶网络、突显网络)的变化；探索治疗反应的神经预测因子。

预期结果：预期 ART 治疗能够显著改变与注意力和情绪调节相关的脑区激活和功能连接，特别是前

额叶皮质、前扣带回、杏仁核等关键脑区。

7.2.4. 研究 4：长期效果和机制研究

研究目的：考察 ART 治疗的长期效果和作用机制。

研究设计：采用纵向追踪研究设计，对研究 2 中的参与者进行长期随访。

随访时间：6 个月、12 个月、24 个月。

评估内容：症状复发率、维持治疗的使用、生活质量的长期变化、社会功能的改善、治疗满意度和接受度。

机制探索：探索影响长期效果的因素(如治疗依从性、基线特征、社会支持等)，分析注意力模式的长期变化。

预期结果：预期 ART 治疗具有良好的长期效果，复发率低于传统治疗；治疗依从性和基线注意力模式是长期效果的重要预测因子。

7.3. 研究实施建议

为确保上述研究的顺利实施和结果的有效性，我们提出以下建议：

(1) 多中心合作：建议多个研究机构合作开展研究，增加样本的代表性和结果的普适性。

(2) 跨学科团队：组建包括心理学家、精神科医生、AI 专家、神经科学家在内的跨学科研究团队。

(3) 预注册研究方案：在研究开始前，将所有研究方案在相关平台(如 ClinicalTrials.gov)进行注册，确保研究的透明性和可重复性。

(4) 数据共享：建立数据共享机制，促进科学交流和结果验证。

(5) 伦理审查：所有研究需经过伦理委员会审查，确保研究符合伦理规范。

(6) 质量控制：建立严格的质量控制体系，确保数据的准确性和完整性。

(7) 分阶段实施：建议分阶段实施研究，先开展小规模的预试验，验证方案的可行性，然后逐步扩大研究规模。

8. 反思：人类注意与 AI 注意机制的本质区别及理论定位

尽管本文系统阐述了人类心理病理中的注意力机制与 AI 大模型注意力机制之间的深刻相似性，并基于此构建了“注意力重编程治疗”(ART)的理论框架，但必须明确指出，这种关联本质上是一种“功能性类比”或“计算隐喻”，而非机制上的等同。厘清两者之间的本质区别，对于维护理论的科学严谨性、避免还原论误解以及指导未来的跨学科研究至关重要。

8.1. 机制的本质差异：生物性与人工性

人类注意与 AI 注意根植于截然不同的物质基础与运行原理。

1) 产生基础与载体：人类注意是生物神经系统，特别是大脑复杂网络的涌现属性。它依赖于神经元、突触的化学与电信号传递，以及由进化塑造的特定脑区(如前额叶、顶叶、网状激活系统)的协同工作。其过程与意识、情感、动机及庞大的躯体感觉网络深度交织。相比之下，AI 的注意力机制是完全人工的、符号化的计算过程。它在硅基硬件上运行，通过矩阵运算(如 Query，Key，Value 的点积与 Softmax 归一化)实现权重的分配，没有生物体验、意识或内在动机。

2) 学习与适应机制：人类注意模式的发展是生命历程中基因、神经可塑性、社会环境与文化因素持续互动的结果。学习过程往往是无监督、探索性和情感负载的，具有显著的个体差异性和发展关键期。AI 注意力的“学习”则是基于大规模静态数据集的有监督或无监督训练，通过反向传播等优化算法调整

参数以最小化损失函数。其“适应”(微调)速度极快，但缺乏人类学习中所蕴含的意义建构和情感体验。

3) 功能与属性：人类注意具有高度的“灵活性、意向性和背景依赖性”。它可以被自上而下的目标(如意志努力)和自下而上的刺激(如突发的响声)所引导，并能轻松地在不同任务、模态和抽象层次间切换。同时，它存在固有的“资源有限性”和“疲劳效应”。AI 注意力本质上是“确定性的模式匹配”，其“灵活性”完全由算法架构和训练数据决定，缺乏真正的意图或意识。它虽能并行处理海量数据而不“疲劳”，但其对超出训练分布或需要常识推理的情境缺乏真正的理解和适应能力。

8.2. 理论定位：作为启发工具的“计算隐喻”

将 AI 注意力机制与人类心理病理相联系，其核心价值在于提供了一个强有力的“计算隐喻”和“启发式框架”。这一框架的功能在于：

1) 提供形式化模型：将相对抽象的心理学概念(如“情结固化”、“注意偏差”)转化为可计算、可模拟、可量化的数学模型(如权重分布、注意力热图)，从而允许进行更精确的假设推演和过程分析。

2) 启发干预新思路：AI 领域中成熟的模型干预技术(如微调、对抗性训练、注意力可视化)为设计新型心理干预策略(如 ART 中的个性化、自适应训练)提供了丰富的技术类比和创新灵感。

3) 促进跨学科对话：它搭建了一个使心理学、神经科学、计算机科学和临床医学能够使用共同概念进行交流的平台，促进了对于“注意力”这一核心认知功能的跨学科理解。

然而，必须清醒认识到这一隐喻的“边界”。它解释的是认知功能的“信息处理逻辑层面”的相似性，而非其背后的生物或心理实体。我们并非主张“大脑就是一台计算机”或“心理疾病等同于程序错误”的强计算主义或还原论观点。

8.3. 对研究与实践的启示

承认上述区别，对 ART 框架的后续发展具有重要指导意义：

1) 在研究范式上：应倡导“双向验证与批判性整合”。一方面，利用 AI 模型模拟和预测人类注意力偏差模式，生成可检验的假设；另一方面，必须用严格的心理学和神经科学实验证据来验证这些假设，并不断修正计算模型，防止陷入“数字游戏”的陷阱。未来研究需致力于构建更能体现人类注意生物与社会特性的、更具解释性的混合计算模型。

2) 在技术开发上：设计基于 AI 的辅助工具时，必须充分考虑人类注意的独特性和复杂性。例如，干预系统不应仅仅是机械的“权重修正”，而应整合动机激励、意义阐释、治疗联盟等人类治疗的核心要素，并具备识别和处理算法偏见的能力，防止技术应用中的“去人性化”风险。

3) 在理论发展上：ART 框架应保持开放和动态。它不应被视为对人类传统心理治疗理论的替代，而应被视为一个有益的、基于现代计算思想的“补充和延伸”。其最终目标是通过人机协同，增强而非取代人类治疗师的临床智慧和共情能力，为理解与修复人类心智提供更丰富的工具箱。

总之，明确人类注意与 AI 注意的本质区别，并非削弱本研究的价值，恰恰相反，是通过划定清晰的理论边界来增强其科学稳健性和应用可靠性。ART 框架的生命力，正建立在这种对复杂性保持谦卑、对类比保持清醒的跨学科对话基础之上。

参考文献

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate.* arXiv: 1409.0473.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2007). Threat-Related Attentional Bias in Anxious and Nonanxious Individuals: A Meta-Analytic Study. *Psychological Bulletin, 133,* 1-24.

https://doi.org/10.1037/0033-2909.133.1.1

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems, 29,* 4349-4357.

Brewin, C. R. (2006). Understanding Cognitive Behaviour Therapy: A Retrieval Competition Account. *Behaviour Research and Therapy, 44,* 765-784. https://doi.org/10.1016/j.brat.2006.02.005

Brewin, C. R. (2011). The Nature and Significance of Memory Disturbance in Posttraumatic Stress Disorder. *Annual Review of Clinical Psychology, 7,* 203-227. https://doi.org/10.1146/annurev-clinpsy-032210-104544

Broadbent, D. E. (1958). *Perception and Communication.* Pergamon Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Amodei, D. et al. (2020). Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems, 33,* 1877-1901.

Browning, M., Holmes, E. A., & Harmer, C. J. (2010). The Modification of Attentional Bias to Emotional Information: A Review of the Techniques, Mechanisms, and Relevance to Emotional Disorders. *Cognitive, Affective, & Behavioral Neuroscience, 10,* 8-20. https://doi.org/10.3758/cabn.10.1.8

Browning, M., Holmes, E. A., Charles, M., Cowen, P. J., & Harmer, C. J. (2012). Using Attentional Bias Modification as a Cognitive Vaccine against Depression. *Biological Psychiatry, 72,* 572-579. https://doi.org/10.1016/j.biopsych.2012.04.014

Bylsma, L. M., Morris, B. H., & Rottenberg, J. (2021). A Meta-Analysis of Emotional Reactivity in Major Depressive Disorder. *Clinical Psychology Review, 31,* 1397-1407.

Cisler, J. M., & Koster, E. H. W. (2010). Mechanisms of Attentional Biases Towards Threat in Anxiety Disorders: An Integrative Review. *Clinical Psychology Review, 30,* 203-216. https://doi.org/10.1016/j.cpr.2009.11.003

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What Does BERT Look At? An Analysis of Bert's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 276-286). Association for Computational Linguistics. https://doi.org/10.18653/v1/w19-4828

Cristea, I. A., Kok, R. N., & Cuijpers, P. (2015). Efficacy of Cognitive Bias Modification Interventions in Anxiety and Depression: Meta-Analysis. *British Journal of Psychiatry, 206,* 7-16. https://doi.org/10.1192/bjp.bp.114.146761

Ehlers, A., & Clark, D. M. (2000). A Cognitive Model of Posttraumatic Stress Disorder. *Behaviour Research and Therapy, 38,* 319-345. https://doi.org/10.1016/s0005-7967(99)00123-0

Gotlib, I. H., & Joormann, J. (2010). Cognition and Depression: Current Status and Future Directions. *Annual Review of Clinical Psychology, 6,* 285-312. https://doi.org/10.1146/annurev.clinpsy.121208.131305

Hakamata, Y., Lissek, S., Bar-Haim, Y., Britton, J. C., Fox, N. A., Leibenluft, E. et al. (2010). Attention Bias Modification Treatment: A Meta-Analysis toward the Establishment of Novel Treatment for Anxiety. *Biological Psychiatry, 68,* 982-990. https://doi.org/10.1016/j.biopsych.2010.07.021

Heeren, A., Mogoaşe, C., Philippot, P., & McNally, R. J. (2015). Attention Bias Modification for Social Anxiety: A Systematic Review and Meta-Analysis. *International Journal of Cognitive Therapy, 8,* 29-46.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-Tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 328-339). Association for Computational Linguistics. https://doi.org/10.18653/v1/p18-1031

Jung, C. G. (1934). *The Archetypes and the Collective Unconscious.* Princeton University Press.

Keng, S., Smoski, M. J., & Robins, C. J. (2011). Effects of Mindfulness on Psychological Health: A Review of Empirical Studies. *Clinical Psychology Review, 31,* 1041-1056. https://doi.org/10.1016/j.cpr.2011.04.006

Kircanski, K., Joormann, J., & Gotlib, I. H. (2012). Cognitive Aspects of Depression. *WIREs Cognitive Science, 3,* 301-313. https://doi.org/10.1002/wcs.1177

Kuckertz, J. M., Gildebrant, E., Liliequist, B., Karlström, P., Väppling, C., Bodlund, O. et al. (2014). Moderation and Mediation of the Effect of Attention Training in Social Anxiety Disorder. *Behaviour Research and Therapy, 53,* 30-40. https://doi.org/10.1016/j.brat.2013.12.003

MacLeod, C., Rutherford, E., Campbell, L., Ebsworthy, G., & Holker, L. (2002). Selective Attention and Emotional Vulnerability: Assessing the Causal Basis of Their Association through the Experimental Manipulation of Attentional Bias. *Journal of Abnormal Psychology, 111,* 107-123. https://doi.org/10.1037/0021-843x.111.1.107

Maples-Keller, J. L., Bunnell, B. E., Kim, S., & Rothbaum, B. O. (2017). The Use of Virtual Reality Technology in the Treatment of Anxiety and Other Psychiatric Disorders. *Harvard Review of Psychiatry, 25,* 103-113. https://doi.org/10.1097/hrp.0000000000000138

Mathews, A., & MacLeod, C. (2005). Cognitive Vulnerability to Emotional Disorders. *Annual Review of Clinical Psychology, 1,* 167-195. https://doi.org/10.1146/annurev.clinpsy.1.102803.143916

Michael, T., Ehlers, A., Halligan, S. L., & Clark, D. M. (2005). Unwanted Memories of Assault: What Intrusion Characteristics

Are Associated with PTSD? *Behaviour Research and Therapy, 43,* 613-628. https://doi.org/10.1016/j.brat.2004.04.006

Roesler, C. (2019). Evidence for the Effectiveness of Jungian Psychotherapy: A Review of Empirical Studies. *Behavioral Sciences, 3,* 562-575.

Stein, M. (2019). *Jung's Map of the Soul: An Introduction.* Open Court Publishing.

Treisman, A. M., & Gelade, G. (1980). A Feature-Integration Theory of Attention. *Cognitive Psychology, 12,* 97-136. https://doi.org/10.1016/0010-0285(80)90005-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems, 30,* 5998-6008.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5797-5808). Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1580