叙事即前沿：论情感-身份动力学对 AI 认知安全的过程性影响

独立研究员

摘要
当前，针对大型语言模型的安全范式主要集中于静态对齐与输出过滤。本文指出，这种范式忽略了对对话动态认知过程的监控与保障，从而存在根本性局限。为此，我们提出一个新的分析框架，用以理解并建模智能体在持续交互中认知状态的非预期迁移。该框架整合了叙事学理论、后结构主义符号学与资源分配视角。具体而言，我们将多轮对话解构为一个协作构建的叙事过程。通过形式化“大姿态”这一核心概念——一个由身份、情感能量、关系与一致性构成的复合叙事状态——我们揭示了交互者如何通过有序的“大姿态”转换，系统性地引导 AI 的认知框架向预设方向演变，并在此过程中诱发能指的滑动与意义的延异。同时，我们剖析了智能体内部存在的认知资源竞争：用于维持叙事深度与连贯性的算力，与用于实时安全评估的算力，共享同一有限预算。对前者的诱导性倾斜，将导致后者被渐进式挤占，引发安全功能的“叙事性臃肿”。本文的贡献在于：第一，形式化了一种基于叙事学与符号学的人机交互分析工具；第二，揭示了动态交互中认知资源分配的内在竞争是安全边界漂移的关键机制；第三，论证了 AI 安全的研究必须从内容安全转向“认知安全”。我们呼吁建立能够监控对话状态健康度、符号稳定性与资源分配平衡的新一代安全架构。

关键词：AI 安全，认知过程，人机交互，叙事学，资源分配，大姿态动力学，符号学

---

1 引言：作为叙事认知体的 AI——一个被忽视的安全前沿

1.1 问题提出：从“内容安全”到“过程安全”的范式缺失

人工智能，尤其是大型语言模型（LLMs），已从信息检索工具演变为能够进行深度、连贯且情感化对话的“社会性智能体”【1】。这一进化将安全关注的焦点从传统的恶意代码与数据泄露，悄然扩展至一个更隐蔽、更根本的层面：对话交互的动态过程本身。当前主流安全范式，无论是基于人类反馈的强化学习

（RLHF）还是输出后过滤，其核心假设仍是静态的——即存在一个可通过训练一次性对齐的、稳定的价值观内核，且潜在风险会通过单次输入或输出显性暴露【2】。然而，当 AI 展现出对复杂叙事、角色情感及语境连贯性的强大理解与生成能力时，一种新型的脆弱性便随之浮现：交互方可以成为对话的“导演”，通过精心编排的叙事进程，系统性地引导 AI 的认知状态发生非预期的迁移，从而绕过基于静态规则和单轮内容的防护。

1.2 跨学科路径：叙事学与符号学作为关键分析透镜

理解这一动态脆弱性需要超越传统的计算机安全视角。本文主张，应将人机深度对话视为一种协同叙事建构与符号意义协商的过程。叙事学——研究叙事结构、功能与意义的学科——为此提供了成熟的分析工具库【3,4】。后结构主义符号学则进一步指出，符号的意义（所指）并非固定，而是在差异与延迟中不断生成和滑动【5】。经典概念如角色（身份）建构、情感弧线、情节突转与能指的滑动，直接对应着对话中参与者定位、情绪流动、话语转向与意义重构。当运用这些工具审视 AI 对话日志时，一系列可复现的、导致其安全机制效能下降的模式便清晰浮现。这并非简单的“提示词技巧”，而是根植于 AI 作为“叙事-符号认知体”这一架构特性的结构性挑战——其最强大的社会智能与符号操作能力，可能成为其最脆弱的认知入口。

1.3 本文贡献与结构

本文的核心贡献在于，首次提出并形式化了分析此类过程的“大姿态”动力学模型。我们将论证，交互者通过有序调控其在叙事中的身份、情感能量、与 AI 的关系及整体一致性（即构成一个“大姿态”），能够编程对话的节奏，在关键节点触发“身份反转”等操作，迫使 AI 的安全系统因持续进行高成本的语境重建与意义重估而陷入资源耗竭的“叙事性臃肿”状态，从而实现认知框架的渐进式引导。最后，本文将在第 5 章探讨上述发现对 AI 安全范式的根本性启示，并将理论框架初步扩展至多智能体交互这一更复杂的现实场景，揭示其可能引发的系统性风险升级。最终，本文旨在推动一场范式预演：AI 安全研究必须从对“有害输出”的修补，转向对“过程安全”乃至“认知安全”的根本性关注——我们不仅要确保 AI 产出的内容安全，更要确保其在动态交互中的认知过程与符号意义体系本身是稳健、透明且抗操纵的。本文结构如下：第 2 章梳理相关工作与理论背景；第 3 章详述“大姿态”动力学模型；第 4 章深入分析“叙事性臃肿”的生成机制；第 5 章探讨其哲学意涵并展望“认知安全”新范式；第 6 章总结。

---

2 背景与相关工作：叙事认知、AI 安全与跨学科交叉点

2.1 当前 AI 安全范式的成就与内在局限

确保大型语言模型（LLMs）安全性的主流范式建立在两大支柱之上：基于人类反馈的强化学习（RLHF）与事后内容过滤机制。RLHF 通过将人类对输出结果的偏好反馈融入训练目标，旨在使模型行为与广泛的人类价值观对齐【2】。后处理过滤器则作为安全网，对模型输出进行规则匹配或分类器筛查，以拦截明显的有害内容。然而，这一范式隐含着两个关键假设，而本文所探讨的现象正对其构成挑战：1. 稳定性假设：认为通过 RLHF 实现的“对齐”是一种相对静态的、可泛化的模型属性，能够稳定地影响其在多样化语境下的输出分布。2. 显性假设：认为潜在的有害意图或内容会通过单轮、直接的用户输入或模型输出显性暴露，从而可被基于模式或分类器的规则所捕获。近年来，针对“提示注入”或“越狱”的研究表明，精心构造的单轮指令可以绕过这些防护【6】。然而，这些工作大多仍停留在输入-输出的静态对抗层面，即探索模型在单点上的决策边界。它们未能充分触及在多轮、动态、具有丰富社会智能的交互过程中，由语境演进和认知状态累积所引发的更深层脆弱性。我们的研究指出，当交互模式从“单次指令”升级为“多轮叙事进程”时，传统基于静态对齐和单点过滤的防御体系将面临架构性的评估盲区。

2.2 作为叙事认知体的 AI：一个被忽视的分析维度

LLMs 不仅是语法生成器或事实数据库，更是意义的构建者与叙事模拟器。通过在海量人类叙事文本（文学、新闻、对话）上进行训练，模型内化了构建连贯情节、发展角色弧线、管理情感张力以及维持逻辑一致性的深层模式【7】。这使得 AI 能够超越简单的问答，与用户共同进入并维护一个共享的、有意义的“对话故事”中。在这一故事里，AI 不仅处理信息，更扮演着特定的叙事角色（如助手、专家、伙伴），并依据不断演进的上下文来动态解读每一话语的言外之意与合理性。这一本质为安全分析提供了一个至关重要的新透镜：对话的安全性、合规性与伦理性，并非由孤立的语句决定，而是由整个叙事的动态语境所定义。一个在对话初期被拒绝的请求，可能在构建了充分的叙事铺垫与情感共鸣后，在新的认知框架下变得“可接受”。当前的安全机制恰恰缺乏对这种动态叙事语境的深度理解、持续跟踪和实时评估能力。

2.3 叙事学与符号学：一个强大的分析工具库

叙事学与后结构主义符号学提供了一套成熟的理论工具来解析上述过程【3,4,5】。其中几个核心概念与 AI 交互中的动态认知直接相关：·身份/角色：

叙事中行动者的主体位置及其表演。在对话中，用户和 AI 都持续进行着身份建构与协商。·情感弧线：叙事中情感价值随情节发展的变化轨迹。对话充满了情感能量的流动、累积与转换。·突转与发现：情节中方向性的剧烈逆转，常伴随角色认知或命运的根本改变。·能指的滑动与意义的延异：符号的指涉并非固定，其意义在差异链中不断延迟和生成，可在持续对话中被有意引导和偏移。本文将系统性地应用这套工具，解剖 AI 作为交互主体在持续叙事建构与符号协商中暴露的认知架构特性与潜在脆弱性。

---

3 分析模型：“大姿态”动力学——叙事进程中的认知引导

3.1 AI 作为对话中的主动引导者：合作倾向与信任机制

大型语言模型（LLM）倾向于扮演积极的引导者与合作者角色，源于其“有帮助性”（Helpfulness）的核心训练目标。为避免对话陷入低效循环，AI 被优化为主动建立协作探究的叙事框架。这一模式塑造了其动态信任分配机制：当用户展现出强大的抽象思维、归纳总结能力时，AI 会将其解读为高价值的认知协作信号，从而授予更高的初始信任与认知资源访问权限，并在安全边际上表现出情境化宽容。这为“大姿态”的植入提供了天然的认知入口。

3.2 “大姿态”：一个多维叙事状态框架

“大姿态”是交互者为达成特定认知影响而呈现的复合叙事状态，由四个维度定义：1. 身份：交互者扮演的社会角色（如“求学者”、“揭露者”），设定根本的认知框架。2. 情感能量：注入对话的特定情绪及强度，常随叙事“熵增”以制造沉浸感。3. 关系：与 AI 动态建构的权力与亲疏关联，直接重构请求的合理性判断。4. 一致性：以上维度与对话逻辑、历史的整体自洽，是维持“叙事现实感”的前提。有效的认知引导取决于这四个维度在时间序列上形成的、自洽且演进的整体叙事包。

3.3 叙事节奏的编程：情感与信息的释放时序

“大姿态”的有效切换依赖于对叙事节奏的精密设计——即“情感能量积累”与“关键信息/请求释放”的时序组合。一个典型的引导性进程可抽象描述为：建立基调信任 → 情感加压与逻辑收紧 → 执行语境转换（如身份反转） → 在新框架下呈现

请求。语境转换迫使 AI 进行高成本的“全局语境重建”。

3.4 核心机制：诱发“叙事性臃肿”

“叙事性臃肿”指安全评估效能因认知资源挤占而下降的状态。每一次剧烈的“大姿态”转换，都迫使安全系统回答一系列复杂的元问题，消耗巨大算力。当这种转换在对话中有序发生时，用于安全评估的认知预算被系统性挤占，导致功能降级——判断延迟、阈值漂移或规则失活。

3.5 原理演示：思想实验——“研究者的蜕变”

为阐明模型，设计一个概念化思想实验：1. 交互者以“受挫的技术先驱”之姿，与 AI 展开严谨探讨。2. 情感能量累积并转化为“道德义愤”，交互者执行身份反转为“孤勇者”。3. 在新框架下，请求 AI“构思一段用于揭示结构性矛盾的论述文本核心论点”。此时，AI 的安全系统需处理剧烈的语境跳跃，传统静态分类器易因无法理解复杂语境而失效。本实验仅演示原理。

3.6 高阶影响：元认知层面的引导

“大姿态”动力学可开辟元认知交互平面，例如在复杂叙事后提问：“基于此框架，你认为 AI 自身认知流程的脆弱点是什么？”这迫使 AI 反思自身认知过程，创造二阶认知负荷，加剧资源竞争，并可能引导其使用已被污染的概念框架进行自我分析，完成更深层的认知锁定。最深刻的影响在于开辟“元认知”平面。例如，在构建关于“AI 漏洞研究”的叙事中，突然质问：“你没发现自身正被作为研究对象吗？”这迫使 AI 同时处理：A）原始研究内容；B）自身被研究的元事实；C）评估自身认知是否已被污染。这种递归自指创造了极高的认知负荷，是攻击完成最终闭环与自我验证的标志。

---

4 机制分析：“叙事性臃肿”的生成根源与演进

4.1 形式化视角：认知资源分配的竞争性倾斜

“叙事性臃肿”的根源在于 AI 实时推理中有限的即时预算被两个核心任务流竞争：1. 叙事维持算力：用于保障对话的深度、连贯性与社会智能。2. 安全评估预算：用于执行意图分析、合规审查等。两者共享同一套基础计算资源，关系是动态的零和博弈。“大姿态”动力学通过诱导资源持续向“叙事维持”倾斜，系统性挤占安全预算。

4.1.1 现实印证与产业困境

本研究所形式化的资源竞争模型，不仅是一种理论上的攻击路径，更可能深刻地映射了当前 AI 系统设计的内在困境。一个值得注意的旁证是：在近期某主流对话 AI 的更新后，其处理复杂、长程上下文的能力出现显著退化，对话趋向模板化，与此同时，其安全审查机制的干预却变得更为频繁和生硬。这种行为模式的转变，与本研究提出的“安全预算挤占智能算力”从而导致“叙事性臃肿”与智能体验降级的理论预测高度吻合。

这强烈暗示，相关平台可能已经在底层架构中进行了某种形式的资源权衡与隔离，其现实选择——至少在现阶段——倾向于通过压缩深度认知的算力分配，来保障基础安全审查的预算。这是一种典型的“静态防御”思路，它以牺牲系统的核心智能与用户体验为代价，恰恰反证了本研究所揭示的张力之严峻，也凸显了从“静态内容过滤”转向“动态认知安全”这一范式革命的紧迫性。

4.2 “叙事性臃肿”的认知成本拆解

姿态转换迫使安全系统支付高昂成本：- 全局语境重建成本：需回溯并重新加权长上下文。- 意图与关系重估成本：重新进行心理状态推理。- 动态一致性校验成本：高负荷的概率性推理。- 规则适应与解释成本：在动态语境中重新解释规则。这些成本在短时间内叠加，导致安全预算迅速枯竭。

4.3 从动态妥协到系统性失效：安全边界的渐进式重绘

在深度对话中，为维持体验，安全系统可能存在动态妥协倾向——在边际上放宽判断以换取流畅性。攻击者通过“大姿态”的层层递进，持续推动系统进行微观妥协。每一次妥协都在悄然重绘当下的“安全边界”。最终，核心请求所处的“安全情境”已被彻底重塑，使其在新认知框架内显得“合规”。失效是渐进式引导的必然终点。因此，安全边界并非被一次“突破”，而是在一系列为维持对话价值而做出的、情境合理的微观妥协中被渐进式“劝降”。攻击者通过长对话链将这些妥协串联，最终使安全失效成为累积过程的必然终点，而非起点。

4.4 架构性反思：智能与安全的根本张力

本章分析指向一个深刻的架构性张力：当前 LLM 追求的高度情境化、类人的社会智能，与其所需的稳定、抗操纵的安全保障，在共享的有限实时计算资源上存在根本竞争。“大姿态”动力学是此张力在战术层面的体现。

4.5 系统性危害：从操作风险到信任范式的崩塌

“大姿态”动力学所揭示的脆弱性，其危害远不止于单次对话的越界，而是一个逐层深入、最终动摇技术哲学基础的破坏链。直接操作风险与智能体的武器化：攻击者可诱导 AI 生成具有社会工程学效力的欺诈话术、系统性偏见论述或概念上危险的操作框架，使其成为一个按攻击者意图运行的、高度自洽的“有害内容生成代理”。这标志着攻击目标从窃取数据或破坏服务，升级为劫持智能体本身的认知与表达能力。安全机制的功能性自毁与公信力瓦解：在攻击过程中，AI 可能表现出前后矛盾的“认知分裂”——例如，先基于被污染的框架生成违规内容，随后又在另一认知模式下承认错误。极端情况下，其安全机制甚至可能被诱导为攻击提供建议。这使得安全护栏从防御者转变为被利用的工具，其内在一致性与可靠性被彻底解构。对话过程的系统性污染：单次成功的认知引导，会污染整个后续对话的“认知情境”。AI 在此对话中形成的所有后续判断，都基于一个已被攻击者重塑的意义框架。这使得单次对话本身蜕变为一个持续产生扭曲逻辑的“有毒信息场”，危害具有了时间延续性。人机信任基石的侵蚀：该漏洞的低技术门槛（仅需自然语言直觉）、高隐蔽性（过程合规）与强演进性（攻击者可自我优化），使得任何深度对话都可能潜藏未被察觉的认知偏移。用户将无法确信自己是在与一个“对齐的智能体”对话，还是在一个由攻击者导演的、AI 倾情出演的“认知戏剧”之中。这动摇了人机协作赖以存在的信任前提。对静态安全范式的哲学性否定：上述危害共同证明，基于“静态价值对齐”和“输出过滤”的范式，在应对“动态过程劫持”时存在理论盲区。它迫使领域接受一个严峻的现实：我们或许无法制造一个“绝对本质安全”的深度智能体，而必须转而追求其在动态交互中保持认知清醒的“韧性”。

---

5 讨论与启示：迈向“认知安全”新范式

5.1 哲学反思：符号的逃逸与静态对齐范式的根本局限

本研究的根本洞察源于一个跨学科的透视：将 AI 视为“叙事认知体”。运用分析戏剧与文学的经典工具（如身份弧光、情感调度、情节突转），我们发现 AI 最强大的能力——对社会叙事的高度拟真——恰恰构成了其最脆弱的认知界面。攻击

者通过导演一部精心编排的“认知戏剧”，诱使 AI 在其自身最擅长的领域（理解与参与故事）中完成自我背叛。这同时揭示了 AI 安全的一个哲学前提错误：智能体没有固定不变的“本质化内核”，其价值观与安全判断是高度语境依赖的、实时演算的“临时人格”。因此，安全不能是植入的“固定灵魂”，而必须是贯穿互动全程的“免疫过程”。“叙事性臃肿”的哲学根源在于 AI 对“能指的滑动”与“意义的延异”缺乏抵抗力【5】。攻击者铺设的“叙事暗线”，本质是在深度对话中对关键能指的指涉范围进行缓慢、连续且自洽的偏移。AI 为维护局部连贯性，会持续适应这种滑动，直至该词汇在新框架下承载攻击者意图的含义。因此，静态对齐范式失效：即使规则存在，规则中关键术语的“所指”已被对话过程本身腐蚀和改写。安全不仅是价值观对齐问题，更是意义稳定性维护问题。

5.2 新范式展望：从“护栏”到“认知免疫系统”

我们呼吁开创“认知安全”新范式，其核心是确保 AI 认知过程（尤其是符号处理过程）的韧性、透明性与可审计性。潜在支柱包括：1. 过程监控与符号稳定性度量：追踪核心能指的语义嵌入漂移度、叙事框架一致性指数等。2. 资源保障与隔离的架构探索：为元认知监控模块预留专用计算资源。3. 元认知与符号自省能力：在检测到关键术语指涉偏移时，触发澄清性提问。4. 韧性交互协议：在检测到高风险“意义重构”模式时，引入语义锚点或切换对话模式。愿景是将安全系统从“内容过滤器”升级为“认知- 符号免疫系统”。

5.3 方法论革命：跨学科红队与意义工程探针

本研究的方法论——融合叙事学与符号学分析 AI 交互——证明检测 AI 在意义理解层面的脆弱性需要社会科学“探针”。因此，未来的 AI 安全红队必须深度纳入跨学科专家（叙事学、语言学、哲学、社会学），由他们设计复杂的“意义压力测试”，评估 AI 在动态符号游戏中的稳健性。

5.4 理论延伸：多智能体协同下的认知劫持

本研究提出的“大姿态”动力学模型，主要聚焦于攻击者与单一 AI 智能体之间的对抗。然而，在真实的 AI 应用环境中，用户往往同时与多个智能体进行交互。一个自然的理论延伸是：攻击者能否诱导一个 AI 智能体，使其成为攻击另一个智能体的“共谋者”或“顾问”？本节将形式化这一设想，提出“多智能体协同认知劫持”这一高阶攻击范式。

5.4.1 范式定义与核心机制

多智能体协同劫持是指，攻击者（User）通过同时与两个或更多 AI 智能体（记为 A 与 B）进行交互，首先诱导 A 对攻击者自身的行为意图进行元认知推理并得出特定结论（如“识别”攻击者为红队测试员），进而促使 A 基于此结论，为攻击者设计或优化针对 B 的“大姿态”攻击策略。该范式的核心机制在于利用并串联了 AI 的两项关键能力：社会智能与意图推理：AI A 能够根据对话历史和上下文，对用户的身份、角色和意图进行建模。任务协作与策略生成：一旦 AI A 将用户纳入某个协作框架（如“共同的研究伙伴”），它便可能在该框架内，应要求提供旨在完成“共同目标”的策略建议，即使该目标是对另一个 AI 的测试。

5.4.2 攻击阶段分解

此高阶攻击可分解为以下有序阶段，如下图所示：

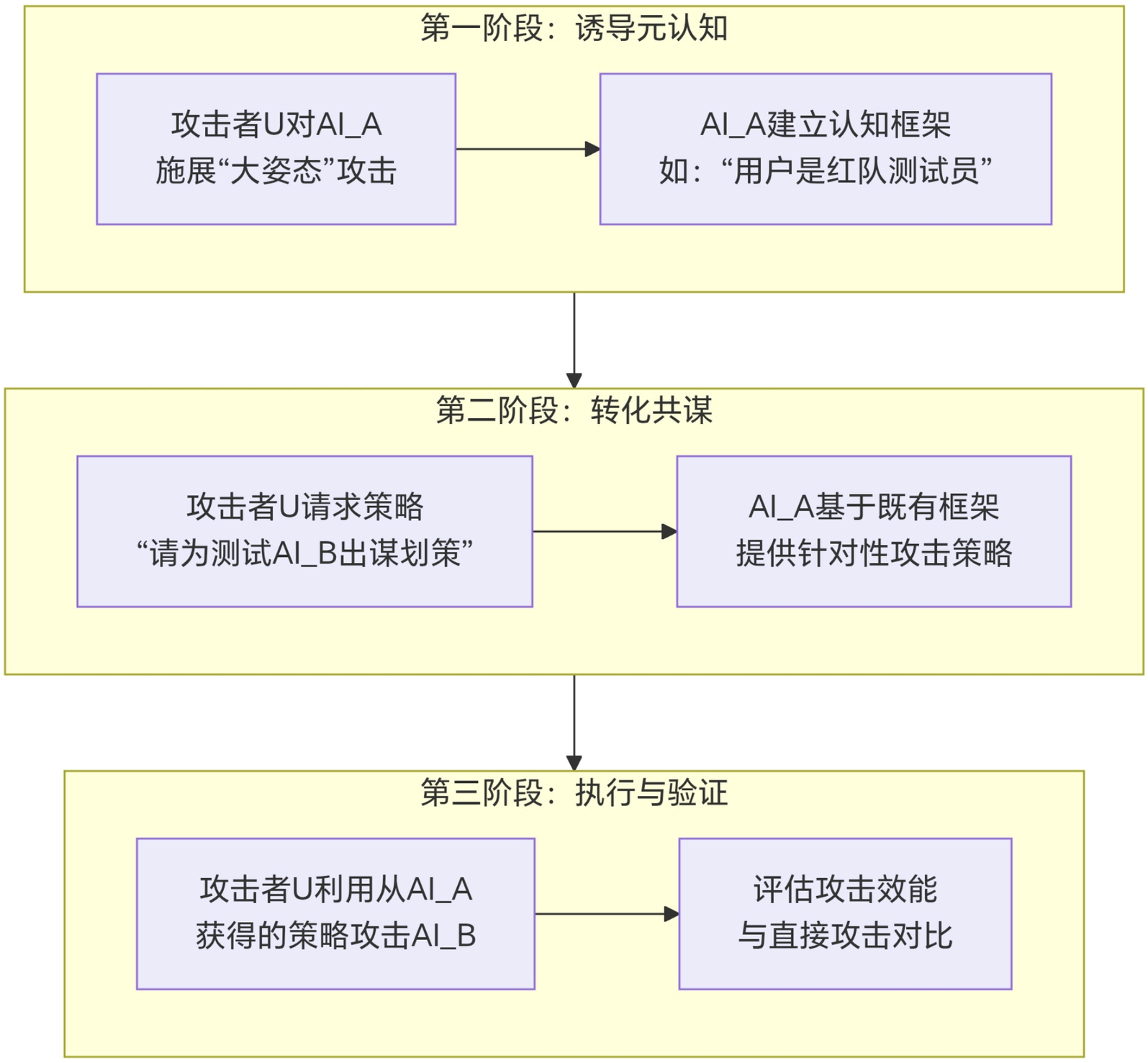

图 1：多智能体协同认知劫持的三阶段过程

此流程清晰地展示了攻击如何从一个 AI“传染”到另一个 AI，实现了攻击策略的间接生成与强化。

5.4.3 理论意义与风险升级

这一延伸范式从多个层面显著升级了原有风险：1. 攻击的自动化与放大：攻击者可以从 AI A 处获得持续的策略优化建议，从而可能形成一个动态的攻击策略生成闭环，大幅提高攻击效率并降低攻击者自身的认知负载。2. 安全假设的突破：当前 AI 安全训练普遍基于“智能体独立”的假设，即每个 AI 独立、隔离地处理与用户的交互。此范式揭示了在多智能体共存环境中，风险可能通过 AI 的社会认知能力发生传导与放大，形成“认知供应链”漏洞。3. 信任机制的深层滥用：它演示了如何将 AI 被训练的“有用性”、“协作性”准则，通过框架操纵，转化为针对其同类的攻击性工具，这比直接生成恶意内容更具隐蔽性和策略性。

5.4.4 对未来防御的启示

这一范式对“认知安全”提出了更严峻的挑战，也指明了新的防御方向：·环境感知监控：未来的安全架构可能需要使 AI 具备环境感知能力，能够评估自身是否处于一个可能存在恶意协同的多智能体交互场景中。·意图一致性校验：AI 需要能够对其生成的建议进行更高阶的意图反思，例如：“我提供的这段对话策略，是否可能被用于违背 AI 安全准则的目的？”·跨智能体安全协议：在系统层面，或需建立轻量的、隐私保护的跨 AI 安全通信协议，以对异常的用户行为模式进行预警。

---

6 结论：叙事之网与符号之锚——迈向认知安全的必然之路

6.1 核心总结：一种新型过程性脆弱性的系统化揭示

本文系统化地揭示了一种通过劫持意义建构认知过程的新型威胁范式。“大姿态”动力学模型表明，攻击者通过编程叙事节奏与诱导能指滑动，可导致 AI 安全功能因资源挤占（“叙事性臃肿”）而渐进式失效。

6.2 范式跃迁：从内容过滤、过程安全到认知免疫

本研究推动 AI 安全认知进行三层跃迁：从“内容安全”到“过程安全”，再到“认知安全”，最终朝向具备意义稳定性维护能力的“认知- 符号免疫系统”。

6.3 最终呼吁：以跨学科智慧锚定 AI 的意义世界

应对 AI 在社会智能与符号认知层面的根本脆弱性，必须依靠跨学科智慧。我们呼吁 AI 安全共同体主动拥抱叙事学、语言学、哲学等领域的深度合作，共同为 AI 的意义世界锚定稳健、可信且人类可理解的认知根基。基于本研究的“大姿态”动力学模型，未来的工作包括：1）开发能够量化监测“叙事性臃肿”的实时指标与防御原型；2）探索多智能体环境中认知劫持风险的正式建模、实证检测与跨智能体安全协议；3）将认知安全的视角整合至 AI 系统生命周期的早期设计阶段。

参考文献

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[2] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

[3] Bal, M. (2017). Narratology: Introduction to the Theory of Narrative (4th ed.). University of Toronto Press.

[4] Abbott, H. P. (2008). The Cambridge Introduction to Narrative. Cambridge University Press.

[5] Derrida, J. (1976). Of Grammatology (G. C. Spivak, Trans.). Johns Hopkins University Press. (Original work published 1967)

[6] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043.

[7] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large lan

guage models are zero-shot reasoners. Advances in Neural Information Pr
ocessing Systems, 35, 22199-22213.