迈向可信具身智能新范式
《神经符号具身智能系统的控制论立场》导读

当前，我们正见证具身智能（Embodied AI）的蓬勃发展。以视觉-语言-动作（VLA）大模型为代表的技术路径，凭借其卓越的感知与生成能力，在诸多演示中带来了前所未有的震撼。然而，当我们将目光从精心剪辑的视频转向复杂、不确定且要求严苛的真实物理世界时，一个根本性的挑战日益凸显：**这些系统是否具备在长期运行中所需的可靠性、安全性与可解释性？**

许多现有的智能体，尽管在特定任务上表现出色，但其核心决策过程如同一个“黑箱”。我们难以确切知晓它为何做出某个决定，无法界定其安全边界，更无法保证它在未知扰动下的稳定行为。这并非单纯的算法优化或数据规模问题，而是源于其架构在**控制论**意义上存在深层的“结构性缺失”。具身智能首先是一个与物理世界持续交互的“控制系统”，而传统的数据驱动范式在追求泛化能力时，往往牺牲了控制系统所必需的**可观察性、可控制性与稳定性**这三块基石。

为此，神经符号人工智能社区正式发布**《神经符号具身智能系统的控制论立场》**文件。本文件不仅是一份技术分析，更是一项**关于工程哲学的根本声明**。它系统性地阐述了为何纯粹的端到端学习范式在应对物理世界的约束时存在固有局限，并明确提出了神经符号人工智能作为一种融合路径所必须坚持的设计原则。

本立场文件的核心主张在于：构建可信赖的具身智能，必须从控制论的第一性原理出发，进行体系化的架构设计。我们主张：

（1）状态符号化：从感知中提取可用于推理与验证的符号状态，提升可观察性，避免 VLA 的“状态塌缩”。

（2）决策约束化：在神经层与物理层间设立动作原语作为可信接口，其前置与后置条件定义了系统的可控制边界。

（3）安全结构化：将物理定律、安全规则作为先验知识显式嵌入符号推理层，使安全成为系统稳定性的内生属性，而非统计结果。

这一“控制论立场”为我们的技术发展提供了顶层的哲学框架。它与社区此前发布的**《基于神经符号 AI 的机器人拆解智能化技术路线图》**形成了**“哲学”**

与“实践”的完整统一。技术路线图描绘了在动力电池拆解等具体产业场景中技术演进的“术”，而本立场文件则明确了指导所有技术选择的“道”。

我们坚信，人工智能的未来，尤其是在关乎安全、财产和生命的物理交互领域，不能仅仅寄托于构建更庞大的“黑箱”模型。真正的突破在于**构建更透明、更可信、更稳健的系统**。神经符号人工智能的融合路径，正是通过引入符号系统的可解释性、可推理性和可约束性，为强大的神经感知能力嵌入可靠的“骨架”与“护栏”。