神经符号人工智能社区立场文件
——论神经符号具身智能系统的控制论立场

摘要

本文件系统阐述了神经符号人工智能社区关于构建可信具身智能的核心工程哲学——**控制论立场**。我们认为，当前主流的视觉—语言—动作（VLA）模型因其在控制论意义上的**结构性缺失**，难以满足在物理世界中长期、安全、可靠运行的根本要求。本立场文件明确了神经符号具身智能系统应遵循的设计原则，即通过**分层融合架构**，将神经网络的感知能力置于符号系统的推理与约束框架之内，从而系统性保障智能体的**可观察性、可控制性与稳定性**。

一、核心问题：VLA 模型的结构性缺失

当前基于端到端学习的视觉—语言—动作（VLA）模型，其根本矛盾在于作为一个强大的“统计关联引擎”，其架构目标与具身智能体在物理世界中的根本需求存在内在不一致。具体表现为四大结构性缺失：

1、可观察性塌缩：其潜在表示倾向于压缩对控制至关重要的物理量（如质量、摩擦因数、装配间隙等），导致基于失真、不完整状态估计进行决策。

2、安全边界模糊：基于奖励函数的概率性学习，无法回答“什么绝对不该做”，缺乏对能力边界的认知和数学上的稳定性证明。

3、层级控制分离失效：层间传递的 token 或 embedding 接口，构成的是隐式输出反馈，而非真正的职责分离，无法满足实时物理交互的稳定性要求。

4、物理先验缺失：缺乏嵌入已知物理定律的归纳偏置，需以极低的数据效率重新发现基础规律。

二、理论基础：控制论第一性原则

任何与物理世界持续交互的智能体，首先必须是一个**控制系统**。其设计必须满足控制论的三个第一性原则：

- **可观察性**：能否获取对决策真正关键的状态信息。
- **可控制性**：是否清楚自身的能力与安全边界。
- **稳定性**：在扰动下行为是否仍可预测、可收敛。

神经符号具身智能系统的设计，正是以重建这三个基石为根本出发点。

三、神经符号具身智能系统的控制论立场

神经符号具身智能（NSAI）系统的控制论立场，在于系统性地弥补主流 VLA 模型在控制理论基础上的结构性缺陷。其核心在于通过体系化的架构设计，重构智能体的**可观察性、可控制性与稳定性**，具体体现为以下三个环环相扣的原则：

首先，为从根本上解决 VLA 模型的“**可观察性塌缩**”问题，NSAI 坚持**“状态必须符号化”**。系统从高维、连续的感知数据中，提取并生成可被明确推理、验证的符号状态（例如“已稳固抓握”、“目标位于禁区”），而非依赖不可解析的潜在表示（如 Token）。这一过程将物理世界中对控制至关重要的弱可观测变量，转化为清晰、可形式化处理的符号命题，从而为决策提供了可靠且可解释的事实基础。

第二，为确保行为的边界明确且安全，NSAI 确立了**“动作必须经由原语接口”**的原则，以此实现可靠的**可控制性**。系统在神经感知与物理执行之间，引入“动作原语”作为唯一的可信接口。每一个动作原语（如“对准目标”、“摸索套接”、“力控旋松”）都封装了明确的前置条件、可预测的状态转移效果以及内在的安全假设，并可被更高层的符号系统所审核、拒绝或中断。这标志着从 VLA 模型中“**token 流→动作输出**”的隐式、统计式映射，向“**符号状态→动作原语**”的显式、约束式规划的范式转变，从而清晰地定义了系统的能力与安全边界。

第三，为实现长期可靠运行，NSAI 将**“安全定义为结构属性”**，以保障系统的内在**稳定性**。通过将物理定律、操作规范等先验知识**显式嵌入**符号推理层，系统能够在任务规划源头禁止物理上不可能或规范上危险的行为。这使得安全性不再依赖于数据分布的统计相关性和训练后的参数调优，而是成为由系统架构和知

识库所保障的、可验证的**结构性约束**。这一立场从根本上将智能体的可靠性，从一种难以保证的“统计结果”提升为一种可设计、可检验的“工程属性”。

因此，上述神经符号具身智能系统的控制论立场，代表了一条以可解释的符号框架约束和引导神经网络感知能力的融合路径，从而在系统设计之初就嵌入可信赖的基石。

核心范式融合：在这一立场下，系统应遵循“**感知—推理—控制**”的清晰分层架构：

1、神经系统（如 VLA）作为“感知大脑”：负责提升**可观察性**，理解“世界是什么样”，实现泛化。

2、符号系统作为“推理与约束中枢”：负责保障**可控制性**，明确“为何与不可怎样”，进行可验证的规划。

3、经典控制层作为“可靠小脑”：负责**稳定**执行，解决“如何做”，实现高频、精准的物理交互。

四、工程实践印证：从自动驾驶到动力电池拆解

神经符号具身智能系统的控制论立场，不仅被英伟达 Alpamayo 等前沿探索所验证，更在复杂、高危的工业场景中得到了系统性实践与深化。例如，英伟达在自动驾驶模型 Alpamayo 中引入独立的“推理层”，生成可解释的“思维链”，正是上述控制论立场的工程体现。神经符号人工智能社区在“动力电池自主拆解”这一典型非结构化任务上的开创性工作，为 NSAI 的控制论原则提供了切实可行的工程范本。

动力电池自主拆解：构建 NSAI 可信拆解智能体

退役动力电池的拆解，是具身智能控制难题的集中体现：**对象（电池型号、使用状态）不确定、环境（螺钉锈蚀、连接件变形）动态变化、任务（拆解顺序、拆卸操作）需实时规划**，且对**安全性、可靠性**要求极高。传统的预编程或纯感知驱动的 VLA 式方法，在此类场景下面临着前文所述的全部结构性缺失风险。

针对以上挑战，社区开展的研究与实践完整贯彻了 NSAI 的控制论立场，研发了如 BEAM-1 自主拆解移动操作机器人及系列工作站，其系统架构清晰体现了分层融合思想：

- **符号化状态感知（解决可观察性塌缩）**：系统并非直接处理原始 RGB 图像，而是通过“神经谓词”，将视觉信息**转化为符号状态描述**，如“螺钉 A—已定位—未对准”、“连接件 B—处于卡紧状态”。这明确提升了对非结构化场景关键特征的**可观察性**，为决策提供了可推理的明确事实。
- **知识与原语驱动的规划（解决可控制性缺失）**：规划核心是一个**符号化的任务与运动规划器**。它接收符号状态，并调用预定义的**动作原语库**（如“力控套接螺钉”、“柔顺分离连接件”）。每个原语都封装了**前置条件**（如“末端执行器已对准”）、**成功条件**及**安全约束**（如最大许可扭矩）。这确保了系统行为始终在**可控制**的、物理可行的边界内。
- **感知-控制闭环与先验嵌入（保障稳定性）**：在执行层，系统通过**基于力感知的柔顺控制模型**，将符号规划输出的动作原语转化为实时控制指令。这一过程**显式嵌入了运动学与动力学先验知识**，使机器人能动态适应拆解中的阻抗变化，从而保障了操作过程的物理**稳定性**与安全性。
- **数字孪生与持续验证（实现可验证性）**：社区同步构建了高保真的**机器人智能拆解工作站数字孪生系统**。所有算法和策略可在虚拟环境中进行“仿真—验证—优化”的闭环迭代，为系统的稳定性和决策逻辑提供了可重复、可溯源的**验证平台**，将安全从统计结果转变为可测试的结构属性。

这一系列实践的成功（相关成果荣获“2025 英特尔人工智能创新应用大赛”特等奖），标志着 NSAI 控制论立场在解决真实产业痛点上的有效性。它证明，通过**神经**（感知/适应）与**符号**（推理/约束）的深度融合，能够构建出应对动态未知环境、决策过程可解释、行为边界可保障的**可信具身智能系统**。

因此，工程实践的印证是双向的：英伟达 Alpamayo 展现了**分层推理架构**在开放道路场景中的必要性；而动力电池自主拆解工业项目则证明，将这一架构与**领域知识（物理定律、工艺规范）深度结合**，是解决高价值、高风险具身任务的必由之路。两者共同印证了本立场文件的核心论断：未来的可靠具身智能，必然是神经与符号在控制论框架下深度融合的产物。

五、结束语

神经符号人工智能社区认为，具身智能的未来不在于构建更大的“黑箱”模型，而在于设计更**稳健**的系统。我们倡导的**控制论立场**，其本质是推动一场工程范式变革：从纯粹依赖“数据驱动”的关联智能，迈向由“**架构保障**”的信赖智能。

未来的可信具身智能，必将是神经与符号深度融合的产物——**利用大模型（神经）理解世界，借助符号系统与控制理论（符号）可靠地丈量并改变世界**。