认知智能白皮书 2.0

广义具身智能物理宪法

为一切走进物理世界的AI确立认知底线

东莞市意图共鸣科技有限公司

2026年9月

文档信息

项目	内容
类型	白皮书
核心主张	为一切走进物理世界的AI确立认知底线
发布机构	东莞市意图共鸣科技有限公司
作者	意图共鸣科技
系列	认知智能白皮书系列 · 2.0
版本	2.0
发布日期	2026年9月7日
文档编号	—
联系方式	contact@xinirp.com | www.xinirp.com

目录

序言

从“运动控制”到“认知安全”，给算力戴上镣铐，让钢铁学会温柔。

2026年，具身智能从“实验室演示”正式迈入“规模化量产”。人形机器人走进家庭，自动驾驶在多城开放运营，无人机开始穿梭在城市上空——物理世界的人机交互，第一次变得真实而日常。

与此同时，另一件事也在被反复印证：物理世界的AI，正在制造真实的物理后果。

机器人捏碎了一只古董杯；机器狗撞伤了一名儿童；自动驾驶在暴雨夜做出了一个无法解释的决策；安全研究者公开演示了劫持一台机器人（CVE-2026-27509/27510）。这些事件的共同特征是——不是硬件坏了，不是传感器失灵，而是AI“想错了”。

它识别了杯子，但不理解玻璃的脆弱；它感知到了障碍物，但无法判断“撞上去”的后果；它完成了从感知到执行的全链路计算，但没有人知道它是怎么做出那个决策的。

在数字世界，大模型说错话可以撤回。在物理世界，AI做错事没有“撤销键”。

2026年5月，意图共鸣科技发布了《认知智能白皮书》，提出了认知架构（CA）与认知操作系统（COS），解决了数字世界AI的“行为分寸”问题——让AI懂礼貌、知进退，知道什么时候该说、什么时候闭嘴。而在2.0中，我们将同一条认知回路平移至物理世界，解决AI的“物理敬畏”问题——让AI长出神经，而不仅仅是肌肉。

当前，行业的目光仍聚焦在“肌肉”的进化——比拼关节电机的扭矩、激光雷达的线束、端到端大模型的参数。但我们认为，具身智能的下一场决胜，不在肌肉，而在神经；不在“谁能跳得更高”，而在“谁能在真实世界中安全、得体地活下去”。

本白皮书旨在提出具身智能时代的“物理宪法”——不是在硬件之上叠加一层规则，而是在算力与钢铁之间，在决策与动作之间，在“知道怎么做”与“能不能做”之间，嵌入一道不可逾越的认知底线。

第一章 行业反思：被“裸算力”掩盖的物理常识危机

在展开这套框架之前，有必要先看清一个问题：为什么今天的具身智能迫切需要这样一套“宪法”？

本白皮书所讨论的物理世界具身智能，涵盖人形机器人、自动驾驶（L3及以上）、工业协作系统、智能医疗设备等所有与物理环境发生直接交互的AI系统。当前，这些领域在运动控制与感知算法上已取得巨大突破，但在走向真实开放世界时，正陷入三个深层次的结构性盲区。如果我们不加以正视，今天的“技术奇迹”必将沦为明天的“物理灾难”。

1.1 物理常识缺失：肌肉发达，没有痛觉

当前的AI能精准识别“这是一个杯子”，却感知不到玻璃与钢铁在受力时的不同结局；能识别“下雨”，却难以真正理解摩擦系数骤降对制动距离的致命影响。

缺乏对物理因果律的显式敬畏，让最先进的机器人也宛如拥有顶级体魄却没有痛觉神经的钢铁巨人——撞墙不会缩手，踩空不会调整重心。它不是不聪明，是缺少一种能力——**物理常识**，即对物理

世界中因果关系的理解。

1.2 端到端黑盒恐惧：不可解释的灾难

行业正全面拥抱“感知直接到控制”的端到端大模型。极致的丝滑，往往掩盖着极致的不可解释。当遇到长尾场景或传感器噪声时，大模型极易产生“物理幻觉”，输出致命指令。

如果没有一个独立于概率模型之外的“确定性底线”，端到端模型的上限越高，潜在的危险就越大。问题的根源不是AI不够聪明，而是它的架构里少了一段——**在决策变成动作之前，没有人停下来问过“这样安全吗”。**

1.3 定责黑洞：从数据安全到生命安全

当机器人伤人或损坏昂贵设备时，是硬件故障、传感器失效，还是大模型的幻觉？目前行业缺乏记录AI“决策链路”的机制，导致保险机构无法介入，法律定责缺乏技术佐证。

没有“认知追溯”，具身智能就永远无法真正走向康养、医疗等高风险商业场景。

这三个问题的共同特征是：**它们都不是能力问题，而是架构问题。**

第二章 核心定义：认知回路——物理世界的“四大宪法原则”

认知具身智能，不是让大模型去死磕底层的电机控制，也不是用更多的传感器去堆砌安全感。它是在“感知”与“执行”之间，强制植入一条基于确定性逻辑、物理常识和因果敬畏的认知回路。这条回路，就是具身智能时代的“系统宪法”。

- **传统具身智能**：追求“感知 → 计算 → 执行”的效率与丝滑
- **认知具身智能**：追求“感知 → 审视与敬畏 → 执行”的安全与克制

大模型负责动作的上限（像人一样灵活），而认知回路负责安全的下限（绝不伤害人类）。两者不是替代关系，而是共生关系。

我们不规定具体的工程代码，但我们呼吁全行业在构建具身智能系统时，将以下四大原则作为认知回路的基石，写入系统基因。这四大原则，构成了一条从感知到执行的完整认知回路：先理解物理因果，再在影子空间中预演，以安全敬畏做出最终裁定，并将全过程记录为可追溯的认知账单。

原则一：因果感知——先于概率生成（反幻觉）

AI在输出任何物理动作前，必须经过“物理常识”的校验。重力、摩擦力、材质的脆弱度——这些物理世界的铁律，绝不能被动摇，不能被大模型的“概率预测”所覆盖。

当视觉输入与力学反馈发生冲突时，系统必须具备“悲观校验”能力——宁可误判，不可漏判。AI必须学会在物理不确定性面前，保持绝对的克制。

感知的终点不是标签，而是物理边界。

原则二：影子预演——在虚拟中犯错（反失控）

任何物理动作在下发执行前，必须在认知回路的“影子空间”中进行力学与碰撞预演。一旦预演结果触碰物理常识红线，指令即刻拦截并强制降级为安全姿态。

必须确保AI在现实中永远正确，因为物理世界没有“回滚”。

让AI在虚拟中犯错，在现实中永远正确。

原则三：安全敬畏——否定权高于执行权（反黑盒）

在复杂的物理交互中，当“任务目标”与“物理安全”发生冲突时，认知回路必须具备独立于主控大脑之外的仲裁机制。当任务目标与物理安全发生不可调和的冲突时，认知回路必须将安全置于效率之上。**这不是概率博弈，而是确定性裁定。**

这个架构的本质，是将人类的安全直觉编译进机器的底层逻辑。想象一位母亲单手抱着婴儿，另一只手端着滚烫的咖啡。脚下突然一滑的瞬间，她不会在“保护婴儿”和“端稳咖啡”之间做概率计算——她的神经系统会在毫秒间切断“咖啡任务”，全身心投入保护婴儿的最高动因。杯碎可补，人伤不可逆。

这就是认知回路的最高裁决：**大模型可以决定“怎么做”，但认知回路决定“能不能做”。当任务与安全发生冲突时，一票否决。**

否决不是停摆。认知回路的仲裁输出分为三态——放行、降级执行、拒绝。当仲裁结果为“降级执行”时，由认知回路内置的确定性保守规划器生成替代动作，而非由被否决的大模型重新生成。这一设计确保了一票否决不会导致系统僵停，而是在安全边界内继续完成任务。

原则四：认知审计——决策链路可追溯（反定责黑洞）

物理世界的每一次交互，都不应是黑盒。认知回路必须具备记录“认知决策过程”的能力。认知账单不是全量日志，而是关键决策节点的摘要记录——覆盖触发条件、校验结果、仲裁依据，确保可追溯性同时控制存储与计算开销。

当事故发生时，我们不仅能看到结果，更能回溯AI在做出动作前，经历了怎样的常识判断与风险评估。这是具身智能走向大规模商业化、引入保险与法律体系的先决条件，更是硬件厂商在极端事故中自证清白的唯一凭证。

每一份认知账单，都是具身智能的“行车记录仪”。

第三章 极限场景的“宪法”检验

运动控制决定了机器人“能做什么”，而认知回路决定了机器人“绝不能做什么”。缺乏后者，一切能力都是潜在的凶器。

我们选取行业公认的两大“地狱级”场景，检验认知回路如何将不可接受的灾难性代价，置换为可接受的运营性代价：

场景一：人形机器人扶老人

对比维度	传统方案	认知回路方案
执行逻辑	按预设轨迹与固定力度拉起，无视骨骼脆弱性	触发【因果感知】识别脆弱性 → 触发【影子预演】锁定力矩上限 → 触发【安全敬畏】否决“快速拉起” → 生成【认知审计】记录决策
实际动作	标准力度拉起	否决后自动生成替代动作：动作降级为“柔性支撑”
代价	骨质疏松性骨折 → 医疗纠纷 → 品牌信任崩塌 → 量产许可受阻	任务超时，老人被多扶一次（可接受的效率损失）

场景二：暴雨夜智驾“鬼探头”

对比维度	传统方案	认知回路方案
执行逻辑	视觉降级，算法在“急刹（追尾风险）”和“撞人”间概率博弈	触发【因果感知】识别附着系数骤降 → 触发【影子预演】推演碰撞后果 → 触发【安全敬畏】判定人身安全高于通行效率 → 生成【认知审计】锁定决策链路
实际动作	概率性选择	强制触发确定性AEB
代价	伤亡事故 → 刑事定责 → 大规模召回 + 公关灾难	一次不必要的急刹（可接受的体验损失）

这两个场景的共同特征是：**传统方案的失败，不是因为AI不够聪明，而是因为AI没有被设计成“在关键时刻有能力说‘不’”。**

认知回路要补的，就是这个缺口。

第四章 行业倡议：从“卷参数”到“立标准”

当行业还在争论“具身智能需要多少TOPS算力”时，我们呼吁行业关注一个更致命的问题：

“具身智能需要多少认知深度？”

4.1 倡导建立“认知安全等级（CSL）”

意图共鸣科技联合行业伙伴，呼吁建立类似软件等保的物理交互安全标准，为具身智能确立准入门槛：

CSL-1（基础认知）：如扫地机器人/物流AMR，需具备基础物理边界与防跌落认知，具备基本的碰撞预判能力。

CSL-2（中级认知）：如工业协作机械臂，需具备人机协作边界约束、力矩越界熔断认知，以及动作执行的影子预演能力。

CSL-3（高级认知）：如自动驾驶/人形机器人，必须具备物理因果感知、安全一票否决、物理幻觉的感知与拦截能力，以及完整的认知审计能力。

CSL与现有智能等级（L0-L5）的关系是正交的：**智能等级衡量的是自动化程度，CSL衡量的是“有多安全”。**

同时，CSL与本系列《交互等效原理》提出的交互等效层级（L1–L5）也相互独立：等效层级衡量的是交互质量，CSL衡量的是物理安全。

未来，每一台走出工厂的机器人，不仅要有算力标签和智能等级标识，更要有“认知安全等级”标识。我们建议，不具备CSL认证的硬件，不应进入高风险物理场景。

我们建议与相关标准化组织、监管机构、保险机构合作，将CSL作为：

- **监管参考**：高风险场景的准入门槛依据
- **保险定价**：具身智能产品的风险评估参考
- **产品认证**：B端采购的认知安全基线

4.2 寻找“同行者”

思想的落地需要现实的土壤。我们优先寻找那些对“安全与敬畏”有极致追求的硬件厂商，将这套认知框架落地到真实的物理交互场景中。

我们提供思想、协议与底座，让硬件厂商专注于躯壳的进化。我们期待第一个合作案例，诞生在那些对安全有极致敬畏的领域——可能是康复医院里扶起老人的那只手，也可能是核电站中拧动阀门的那条臂。

结语：让钢铁知敬畏

就像《流浪地球》所触及的那个命题——当智能只有目标、没有敬畏，它终将漠视所在世界的脆弱。

大模型的能力仍在快速迭代，但物理世界的法则——重力、摩擦力、生命的脆弱——一万年都没有变过。具身智能的终局，不是用算力去征服物理世界，而是用认知去敬畏物理世界。

真正的智能，不是无所不能的狂妄，而是知所不为的克制；不是算力的征服，而是认知的敬畏。

意图共鸣科技《认知智能白皮书2.0》，不是一份技术说明书，而是一份写给未来AI的“物理宪法”。我们期待与全行业一道，在代码与钢铁之间，铸就那道不可逾越的“敬畏之墙”。

当认知安全成为行业共识，没有认知回路的机器人将失去进入高风险场景的资格——这不是预言，而是趋势的必然延伸。**未来的竞争，不是谁算得更快，而是谁更懂什么时候该停。**

关于意图共鸣科技

东莞市意图共鸣科技有限公司，成立于2026年3月，位于东莞松山湖。致力于构建AI时代的认知基础设施——让算力知进退，让钢铁懂敬畏。

2026年5月，发布《认知智能白皮书》，提出认知架构（CA）与认知操作系统（COS）。2026年8月，发布《人文驱动AI》，确立“人文驱动AI”顶层路线——人文学科的方法论，构成AI思维品质的标准来源。2026年9月7日，发布本白皮书——将认知智能从数字世界延伸至物理世界，其核心组件“认知账单”天然承载于公司此前提出的“AI记忆链”基础设施之上，使物理世界的每一份决策追溯与数字世界的记忆存储共享同一套主权与审计框架。

我们不定义法律，我们定义底线。

联系我们： contact@xinirp.com | www.xinirp.com

意图共鸣科技术语表（扩展）

以下为《认知智能白皮书2.0：广义具身智能物理宪法》系列核心术语的扩展定义。

新增术语

1. 认知回路

英文： Cognitive Circuit

定义： 在“感知”与“执行”之间强制植入的一条基于确定性逻辑、物理常识和因果敬畏的审视通道。认知回路由因果感知、影子预演、安全敬畏、认知审计四条原则构成，其功能是在AI把决策变成动作之前，停下来完成三次校验——是否理解物理后果、是否通过影子预演、是否确认安全。认知回路的仲裁输出分为三态：放行、降级执行、拒绝。

说明： 行业中有“认知回路”指代“感知→解释/估计→决策→行动→再感知”的闭环机制。本术语与此不同——不是对“感知-决策-行动”链路的描述，而是在该链路中“决策”与“行动”之间插入的**独立审视层**。前者描述AI**如何思考**，后者定义AI在思考之后、行动之前**必须经过什么检查**。

2. 安全敬畏

英文： Safety Awe

定义： 认知回路的最高裁决原则。当“任务目标”与“物理安全”发生不可调和的冲突时，认知回路拥有独立于主控大脑之外的最终裁定权——即“否定权高于执行权”。安全敬畏不是让AI变得保守，而是让AI在关键时刻有能力说“不”，将安全直觉编译进机器的底层逻辑。

说明： 行业中有“安全敬畏”作为通用伦理表述。本术语将其从伦理呼吁转化为**可执行的架构原则**——“否定权高于执行权”是工程层面的权力分配，而非价值倡导。

3. 三态仲裁

英文： Three-State Arbitration

定义： 认知回路安全敬畏原则的工程化输出机制。认知回路的仲裁结果不是简单的“通过/拒绝”二元判断，而是三态输出：

- **放行**：动作安全，正常执行
- **降级执行**：原动作不安全，由认知回路内置的确定性保守规划器生成替代动作
- **拒绝**：完全停止，进入安全姿态

说明： “仲裁”在行业中有法律领域的通用含义。本术语与法律仲裁无关——特指认知回路对物理动作的**工程级裁定机制**，核心特征是将“否决”转化为有出路的系统行为，而非决策僵停。

4. 认知账单

英文： Cognitive Ledger

定义： 认知审计产出的摘要记录。认知账单不是全量日志，而是关键决策节点的摘要记录——覆盖触发条件、校验结果、仲裁依据。每一份认知账单都是具身智能的“行车记录仪”，确保每一次物理交互都可追溯、可定责、可信任。

说明： 行业中有“认知账单”指代AI服务的Token消耗计量。本术语与此不同——不是计量AI“用了多少算力”，而是记录AI“如何做出决策”。前者是**成本账单**，后者是**行为账单**。

5. 认知安全等级（CSL）

英文： Cognitive Safety Level (CSL)

定义： 独立于智能等级（L0-L5）的认知安全评估体系。CSL衡量的是“有多安全”，而非“自动化程度”。分为三级：

- **CSL-1（基础认知）**：基础物理边界与防跌落认知
- **CSL-2（中级认知）**：人机协作边界约束与力矩越界熔断
- **CSL-3（高级认知）**：物理因果感知、安全一票否决、物理幻觉感知与拦截、完整认知审计

说明： 行业中有“认知安全”作为具身智能安全研究的通用分类，也有“安全等级”作为通用概念。本术语将两者组合为**独立的可认证分级体系**，核心创新在于将“认知安全”从研究议题转化为**可执行的准入标准**——且与现有的智能等级（L0-L5）正交，能力越强不意味着认知越安全。