安全右脑的技术定位

02122025SQ

目录

1，前提

本文讨论安全右脑技术定位，不讨论基本概念和基本原理，是对已有架构的重新解释。

2，技术定位

安全右脑使命：

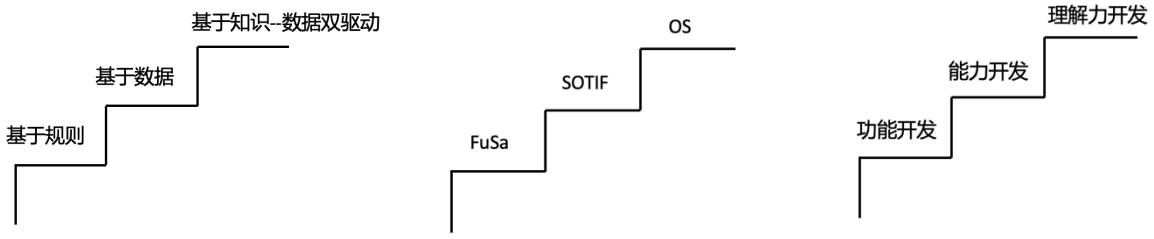
- 突破 E2E 的技术瓶颈
- 开辟 OS 研究平台

从三个技术发展阶段视角出发，安全右脑都是处在第三阶上：

a）技术路线定位：基于规则 → 基于数据 → 基于双驱

b）能力定位：功能开发 → 能力开发 → 理解力开发（行为责任开发）

c）安全领域定位：FuSa → SOTIF → OS



(a) 技术路线定位　　(b) 安全领域定位　　(c) 系统定义定位

图 1 立足点三阶段定义法

3，技术核心

技术核心指得是技术出发原点位置，而不是指“核心技术”。

知识图谱（KG）是安全右脑技术的硬核部分。利用 KG 给 VLM 提供硬边界，结果就是双驱路线：“黑盒子+白盒子+先进 AI”。

KG 是对 OS 风险的三层解析结果，风险分析是 OS 要解决的核心。三层风险世界模型是对三层风险的理解力。这个三层结构，则是根据人的思维过程而产生的，模仿人类自己开车经历了怎样一个风险管控过程。

技术外延的过程是：OS 风险分解 → 风险世界模型（KG） → 模型增强 （KG+LM）→ 模型融合

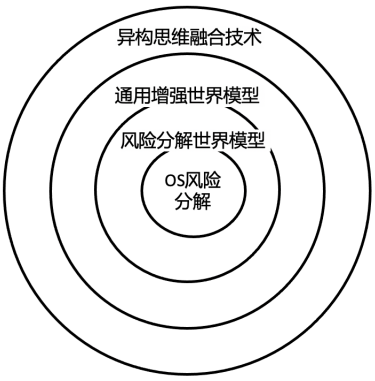


图 2 核心技术定位

其中“异构思维融合” 是感知融合原理的自然延伸，技术内涵包括：

- 左右脑融合
- 知识-数据融合
- 多层世界模型融合
- 端到端与大模型的融合

4，技术路线主干

AI 日新月异，是否有一不变应万变的架构？答案是不存在一劳永逸的制胜绝招，但是存在相对保持长久稳定的架构内核。

对风险的透明解析方式是一切外围技术的起源。安全右脑有相对不变的部分还有动态变化的部分。不变的是 KG 内核（四个风险专用世界模型），动态变化部分是各种最新 AI 通用世界模型和持续学习。

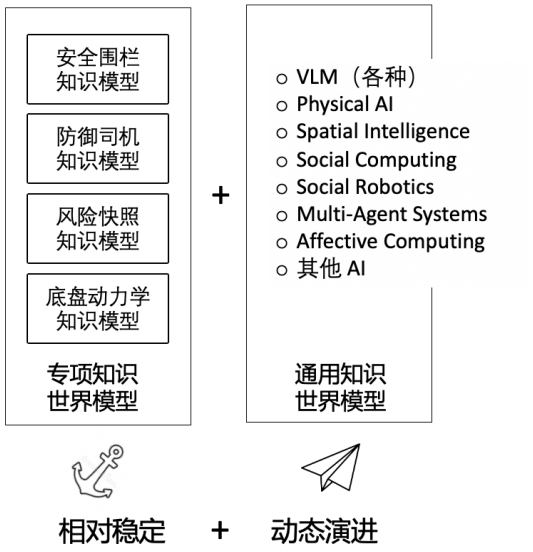


图 3， 技术路线的“anchor 部分”与“进化部分”

有了 KG 内核，安全右脑就有了主干，至于 KG augmented LM 或者 VLM augmented KG 都无所谓，可以在探索过程中决定。

主干固定不等同于固守观念。主干定位有两个结局：不断被行业最新认知加持加固；或者由于立足不稳而改弦更张。安全右脑属于前者。

除非本源问题已经得到解决，否则主题不会轻易改变。

成长：枝叶部分是工具层，随时更新使用最新工具

5，安全右脑提供的解决方案

端到端技术的最大优点在于能够实现系统整体优化和极高的自适应能力，使复杂任务的处理更加高效和智能。其他优点包括：

1）系统整体优化（Global Optimization）

- 打破模块化限制，这种设计避免了不同模块和功能之间信息传递的误差累积和信息丢失，能够在全局范围内寻找最优解。
- 极大提高处理效率，直接学习输入与输出之间的复杂映射关系，减少了中间环节，提高了响应速度和处理效率。

2）高度自适应能力（High Adaptability）

- 从数据中自动学习特征
- 易于扩展和迁移学习

3）复杂任务处理能力强（Handling Complex Tasks）

- 处理非线性和高维数据
- 自动发现隐含关系

4）快速迭代与持续优化（Rapid Iteration and Continuous Improvement）

- 数据驱动的快速迭代
- 适合在线学习和实时更新

5）降低系统复杂度（Reduced System Complexity）

- 简化架构设计，将所有功能整合到一个模型中，减少调试和维护成本，相对于传统智驾系统团队规模小，所以端到端系统的开发效率更高，资源消耗也更少。
- 提高系统一致性：统一的网络结构和优化目标，确保系统在不同场景下的行为一致性，避免模块间的不协调导致的决策冲突

但是，E2E 仍然面临很多挑战和发展瓶颈

5.1，E2E 挑战和安全右脑解决方案

如何突破端到端的技术瓶颈。

	E2E 技术瓶颈	E2E 问题描述	安全右脑解决方案
1	黑盒子的不可解释性	E2E 决策时，外部观察者（包括工程师、监管机构，甚至系统开发者本人）很难理解其具体的决策逻辑和依据，主要表现在以下几个方面： - 模型内部的特征难以解释； - 无法明确追溯决策原因； - 难以验证与监管（监管机构需要理解系统的决策逻辑）； - 模型的鲁棒性难以评估（提前预判风险需要解释模型在特定场景下的行为）	提供右脑白盒子解决方案： - 用知识图谱解释决策逻辑解释； - 知识图谱按照功能划分成可以独立解释和调试 - 模块化世界模型； - 引入可解释人工智能模型可视化与日志追踪
2	行为不可设计	行为不可设计和行为不可解释是同一个问题，主要原因是： - 黑箱模型特性； - 缺乏人类可理解的中间语义层； - 数据驱动的局限性，无法明确指出问题的根源：	安全右脑有多种行为设计手段，包括（不限于）： - 修改知识图谱 ontology 架构的基本语义网络 - 调整 arbiter 模块的舒适度阈值、安全度阈值、设定不同的舒适度与安全度 metrics - 调整多重异构世界模型交叉检验的安全级别
3	缺乏安全价值目标	不是一个“目标-能力-功能”的分解过程，所以不清楚自己的最终目标是什么，也不清楚自己的能力界限在什么地方	设计架构允许人工正向能力分解，能根据安全目标分解出能力，再定义出功能。可以通过以下方法为系统注入目标： - 在 arbiter 里融入价值目标优先原则 - 在 KG 里也编写价值冲突处理原则

			- 也可以设置专门的安全目标/价值判断模块处理价值冲突（比如使用 Rulebook 方法）
4	长尾问题	长尾场景指的是那些在数据中出现频率极低，但对系统性能和安全性至关重要的复杂或极端情况。主要挑战是： - 数据稀缺性：长尾场景的数据非常少，即便通过数据增强或仿真生成数据，也很难完全覆盖所有潜在的复杂场景； - 泛化能力不足； - 安全风险放大：系统可能在关键时刻失效； - 调试和验证困难：“黑箱”特性使得难以进行有针对性的优化和验证	- 增加三重异构的风险模型，从多个角度暴露潜在风险 - 用安全右脑的推理能力预见训练数据集里没出现过的风险模式 - 根据安全目标价值，利用 AI 技术自动分解陌生任务，组合出整体解决方案
5	跷跷板效应（可扩展性）	在优化深度学习模型时，当提升一个指标时，另一个可能会下降，导致系统在某些关键方面表现不佳。具体表现： - 性能与安全的跷跷板：比如更流畅、更高效的路径规划可能会牺牲安全性； - 提高安全上限的同时会牺牲安全下限（详见下面 6 项） - 特定场景优化 vs. 泛化能力的跷跷板：特定场景或数据集上表现出色，但在长尾场景中泛化能力较差； - 可扩展性与复杂性的跷跷板：为了提升模型的复杂性和处理能力，可能会增加计算资源需求，降低系统的实时性和可扩展性； - 数据量与模型鲁棒性的跷跷板：依赖大规模数据集可以提升模型准确性，但过度依赖数据可能导致模型对数据噪声或异常值敏感，降低鲁棒性。	- 知识图谱是解决泛化能力的重要手段，能避免对数据量的无限需求，又能避免单纯“基于规则”开发路线的“规则爆炸” - KG 在理论上可以在不破坏原有规则和知识的基础上添加新的规则与知识 - 安全右脑提供用户泛化界面 - 具有多种性能冲突调节的手段和途径（见本表第 3 项）
6	安全下限的问题	“做得好可以达到很好的效果，做的不好比传统更差”，也就是在数据充足、场景常规的理想条件下能够达到非常优秀的性能，但是在面临陌生场景、数据噪声等环境突变时，表现会大幅下降 其实这也是“跷跷板效应”里的一种	安全右脑的左右脑可以独立发展，用知识架构（比如安全围栏功能）保证安全底线。左脑不断训练（端到端架构）提高上限的同时，不会降低右脑的下限
7	不清楚自己的能力边界在哪里	主要原因是： - 黑箱特性：系统在做出决策时无法解释或量化其自信度，难以明确知道自己是否能做出正确决策； - 泛化能力的不确定性：自己无法判断自己是否进入了“能力边界之外”的领域； - 自信度评估机制缺失：端到端系统的所有功能被合并为一个整体，无法在每个环节设置置信度阈值进行自信度评估，导致系统可能在低置信度情况下仍继续运行； - 缺乏冗余和异常检测：端到端系统通常依赖单一的深度学习模型，缺乏独立的冗余模块来交叉验证结果，难以及时检测到自身失效。在面对异常情况时，系统可能无法及时察觉并调整策略，增加了失控风险	- 风险分解量化，可以进行模块化置信度评估； - 双脑驾驶系统是感性思维与理性思维的综合体，当策略不符合基本常识常理的时候，KG 可以提供明确的规则和边界，系统会改变策略（自适应决策）； - 安全右脑提供四重风险世界模型，用异构冗余方法对环境风险进行交叉检验 - 四个风险评估单元可以实时评估当前环境的风险水平，如果风险超过预设阈值，系统可以降低自动化等级
8	数据与算力壁垒	以特斯拉为例（截止 2024 年末） - 数据量：每年收集数据量超过 10 亿英里（约 16 亿公里）的行驶数据，累积的数据量已超过数十亿小时的视频数据，每辆车每天上传的行驶数据大约为 100MB 到数 GB 不等，每天的数据总上传量可能达到数 PB（Petabytes，千万 GB）级别 - 数据标注	- 安全驾驶知识图谱（KG）提供先验规则和决策约束，减少系统对大规模长尾场景数据的依赖。 - 利用实际死亡事故数据，重点训练系统在高风险场景下的表现，提升系统的泛化能力，减少海量数据需求 - 利用四重风险知识单元，通过实时风险评估，在低风险环境中启用轻量化模

		- 训练算力：一个完整 Dojo 机柜包含 3000 多个 D1 芯片，每个 D1 芯片拥有 362 TFLOPS 运算能力，总算力超过 1.1 ExaFLOPS（每秒百亿亿次浮点运算） - 推理算力（车载算力）：FSD4.0 算力 250 TOPS	型，降低推理算力需求；在高风险环境中启用复杂模型，平衡车载算力与能效
9	资源投入对比	- 后期投入成指数上升，主要用在提高安全上限上 - 蛮力投入是唯一出路 - 对基础理论研究要求不高，所以不用配备基础研究团队	- 安全右脑是系统工程，需要从理论基础开始研究，所以需要配备 E-I-P 三层研究梯队；智力投入比蛮力投入更重要 - 前期工程量较大，需要投入更多的团队人力资源，但是一旦建立起基础 KG 以后，整体的的开发费用会有数量级的下降
10	同质化竞争	大家都拥挤在端到端赛道上。有人开始实践“端到端+大模型”，但是本质上还是“黑盒子+黑盒子 = 黑盒子”，上述端到端解决方案存在的挑战依然存在	- 另辟“黑盒子+白盒子 = 灰盒子”赛道
11	建立 trust 体系	Trust 体系由“能力”和“资质”两个要素构成。端到端能提供能力证明，但是很难提供“资质”证明，类似于会开车但是没有驾照。原因是“资质”是理性体系的产物，比如驾照、毕业证书、职称证书都是系统解析与论证的结果。资质的目的是证明行为结果的概率。	- 右脑的白盒子可以提供决策逻辑依据； - 双脑系统能保证安全下限； - 双脑系统能保证行为的可设计性，行为设计方法见本表第 3 项； - 双脑系统可以证明行为的稳定性和可再现、可重复性，能证明在某些环境下的大概率表现 - 双脑系统清楚自己的能力界限 - 双脑系统通过多重异构化风险世界模型来解释环境，因此具有环境理解力 - 双脑系统可以依靠“自身能力界限认知”、“环境理解力”和“确定性行为概率”提供高级自动驾驶的准入“资质”证明
			-

另外，参考文献[1]里提出了类似的 8 个端到端挑战，基本与上表总结的内容类似：

1）可解释性；2）泛化能力；3）因果混淆；4）安全性和鲁棒性；5）实时性和计算资源；6）伦理和法律问题；7）多任务学习；8）长尾分布

安全右脑的对应价值分析见[2]。

右脑是给端到端左脑大的补丁。

5.2，SOTIF 挑战

SOTIF 技术瓶颈	SOTIF 问题描述	安全右脑解决方案
Known unsafe 问题	挑战主要在于 - 可解释性不足；已知的不安全场景，系统难以明确检测和标记这些风险，即便开发人员知道某些场景存在风险，也难以在模型中直接注入安全规则或确保模型在这些场景下的正确行为，	已知（known）风险知识系统化： - 用安全围栏保证牢固的安全下限，不会因为左脑端到端提高上限而丢失下限

	- 缺乏明确的安全边界控制：缺乏能力边界和自信度评估 - 数据驱动模型无法完全覆盖已知不安全场景 - 模型更新与验证困难 - 端到端的知识不系统，无法保证安全下限	
Unknow 能 unsafe 问题	- 系统无法预测和识别未知风险 - 缺乏自适应与推理能力 - 可解释性差导致难以对模型进行诊断和修正	未知（unknown）的隐式风险显式化 - 不确定性量化（Uncertainty Quantification）：通过贝叶斯神经网络或蒙特卡罗 Dropout 等技术，评估系统决策的不确定性。在模型自信度低于阈值时，系统可以自动降低自动化等级或提醒驾驶员接管。 - 异常检测模型：三大风险世界模型
21448 无法产生能力定义	因为 21448 没有明确规定行为要求，能力是对行为的分解	- 可以针对未来安全行为标准进行正向针对性设计 - 可以从行为要求分解出能力要求，比如对安全目标的分解能力，对安全-舒适-效率-availability 的平衡能力
21448 无法产生功能要求	目前系统功能全部诞生于用户体验，而不是为满足车辆行为能力要求而诞生的	可以从能力要求分解出功能定义，比如语义感知内容、分层世界模型、风险识别功能、风险规避功能、响应时间、稳定性要求等等

5.3，OS 挑战

目前 OS 还没有系统性的研究平台。安全右脑以风险分解定义为入手点，开辟以下研究领域：

- 交互风险分类与度量方法
- 交互风险模型建立方法
- 交互行为安全评价方法
- 交互行为正向设计方法

OS 挑战	OS 问题描述	安全右脑解决方案
交互行为安全仍是空白领域	OS 安全需要以安全行为研究为基础，包括 - 行为理解与预测 - 行为风险的度量 - 安全行为定义 - 行为正向设计 但是目前以上研究均以暗箱为基础，无法实现策略调整	- 采用 KG 进行白盒研究 - 随时采用 social AI 领域的最新成果 - 实现行为正向设计（5.1）
交互风险定义的挑战	风险三角形已经存在于行业多年，但是风险分解和量化存在很大困难	- 风险分级：一共分成三级，按照 planning horizon 时序分别为：安全围栏、防御司机和风险快照 - 风险量化：每级风险独立输出风险强度、风险概率、风险类别；arbiter 提供最终的综合度量指标

环境语义的理解力的局限		采用“知识+数据双驱动”推理架构提供环境理解力。

6，AI 应用

AI 方法是一个贯穿全部开发、验证与应用的系统工程。目前没有某一种石破天惊的算法会全部解决我们所提出来的问题，而是在开发全过程中无处不在。

AI 不是目的，只是手段。

目的是三个立足点，是**本源问题**，AI 人才并不能替代我们思考本源问题，这也正是公司的价值所在。在图 2 中，本源问题是稳定不变的，AI 工具是随时动态应变的。

	安全右脑开发领域	AI 需求
1	专项风险模型建模	四个模型的 ML 训练，Physical AI，Social AI，KG 工程，ChatGPT 辅助 KG 生成
2	推理	KG 推理，大模型推理，KD3 推理
3	大模型应用	KG 增强技术、轻量化技术（大模型的蒸馏、量化、剪枝）、车载优化技术
4	Data Engine	
5	世界模型进化	持续学习技术
6		

6.1，知识图谱与大模型的关系：数据-知识双驱动（KD3）模型的开发

“左脑+大模型”是现在很多人正在尝试的技术路线，这条路线的主要问题是：

黑盒子 + 黑盒子 = 黑盒子

所以端到端黑盒子特性相关的天生问题，依旧存在。知识图谱与大模型结合的推理引擎是我们的独有技术，双脑系统透明度为：

黑盒子 + 白盒子 = 灰盒子

大模型可以有很多选择，与 KG 融合的工作量可以参考“附录 I 大模型与 KG 融合 KD3”。

附录 I 是以英伟达的 Omniverse AI 为例，主要环节见下图，其他大模型同理。

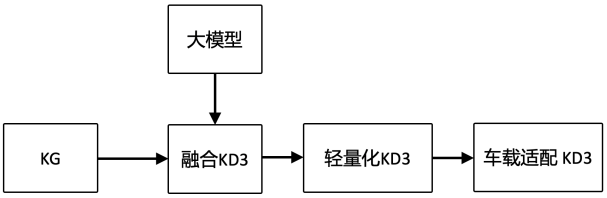


图 4 KD3 构件

6.2，任务分解

功能的从属关系见图。

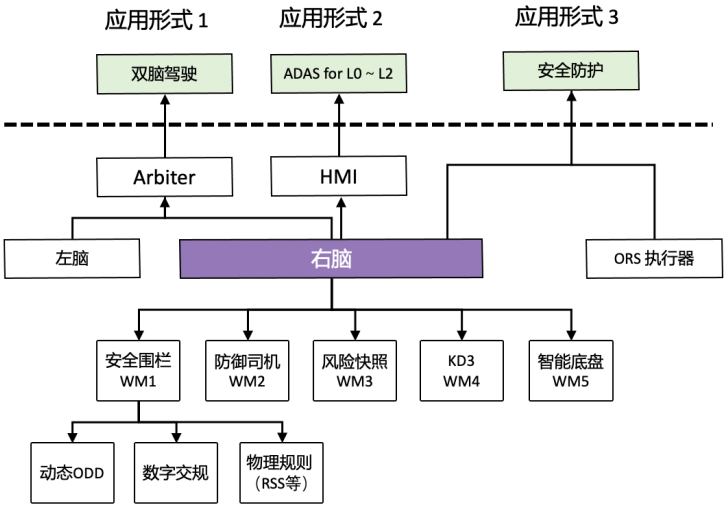


图 5 （这里 KD3 特指被 KG 增强以后的大模型）

起步内容应当包括 WM1～3 中的一个或多个、Arbiter、KD3 大模型。

每个 WM 都要求按照统一协议输出风险评分，风险的定义按照 IEC 61058。另外，知识模型单独要求如下。

6.2.1 安全围栏

6.2.1.1 动态 ODD

6.2.1.2 数字交规

6.2.1.3 物理规则

6.2.2 防御驾驶

6.2.3 风险快照

1）能根据场景要素的组合预测当前场景的事故类型、强度和概率，并提出应对措施

2）在满足最低预测精度的前提下要求输入颗粒度最粗化，以便于输入量的传感获取

3）将已经完成的 NAIS 模型、FARS（25 万）模型、CRSS（60 万）模型、NHTSA（260 万）统一到一个模型（“统一模型”）

4）对“统一模型”进行输入量敏感性分析，删除次敏感量，以便减少传感量，提高传感可实施性

5）将 GIDAS、ITARDA 等数据库纳入“统一模型”，形成 “Global 模型”，最好能找到适合于各个国家/地区的一致性输入参量，以便于场景传感的标准化

6）形成标准化传感变量列表

6.2.4 Arbiter 仲裁单元

7 产品应用形式

7.1 Worst Cases 分析

1）全配置方案

假设四个风险模块都开发成功，生成 4 个风险世界模型，最复杂的应用场景见图 6：

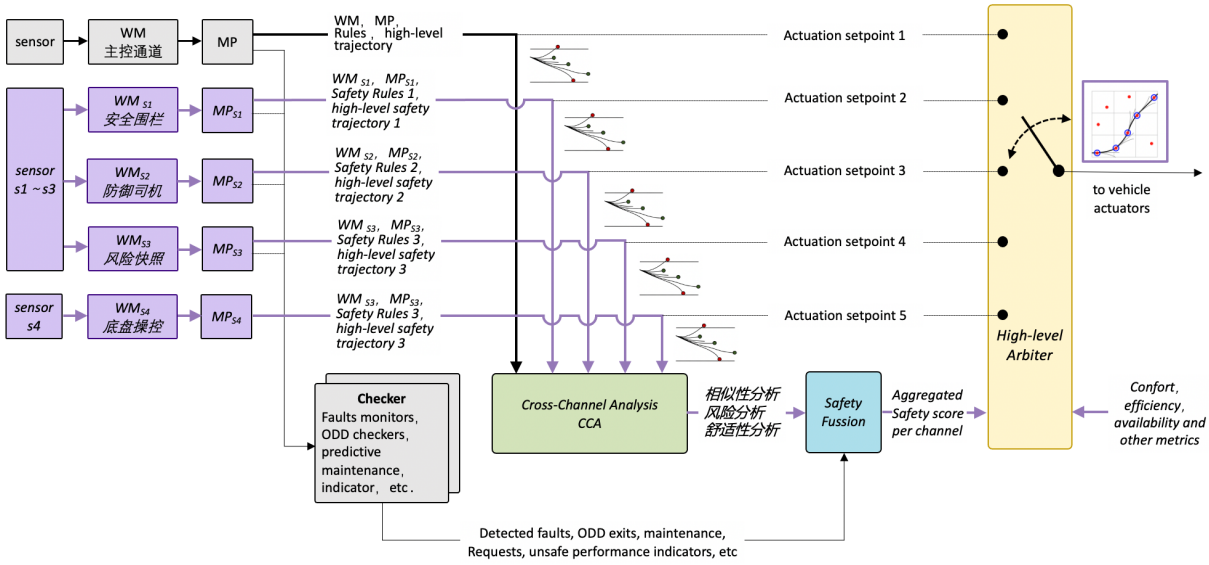


图 6 风险模型×4 结构

这种架构中每个风险模型都有投票权，会直接影响到结果输出。

2）融入 KD3 方案

如果前期已经完成 KD3 开发（开发方法见附录 I_用 omniverse 开发 KD3 02132025SQ），则 KD3 可以单独构成一个通道，见图 7。

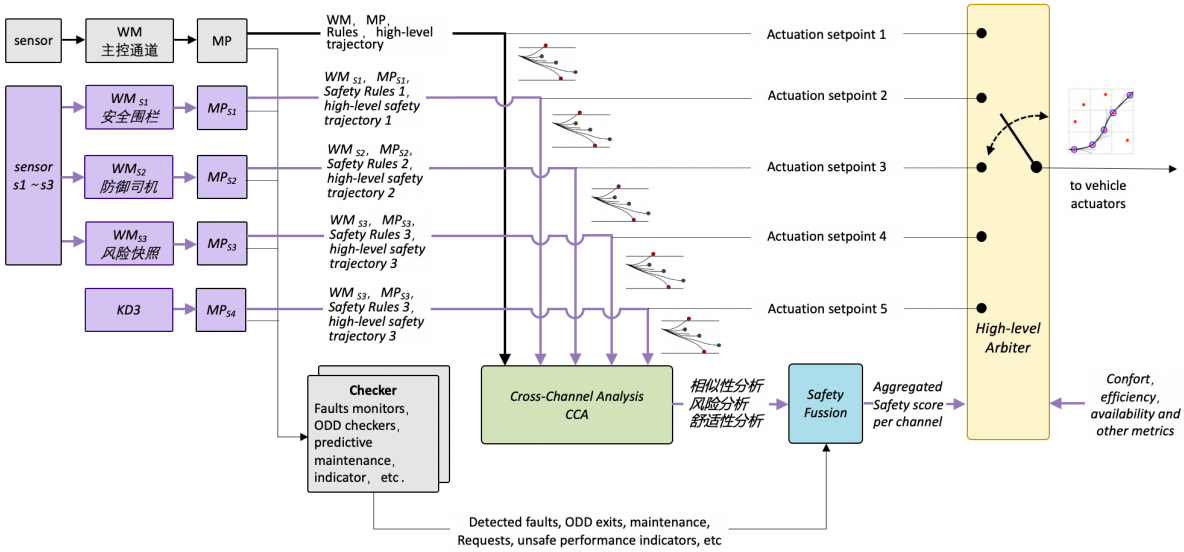


图 7 风险模型×3+KD3 结构

3）左脑+单模型方案

如果三个风险模型里任何一个开发成功，都可以单独与左脑构成异构冗余架构，都随时有可能产生双脑响应，比如，风险快照就比较成熟，现在就具备条件尝试单独插入：

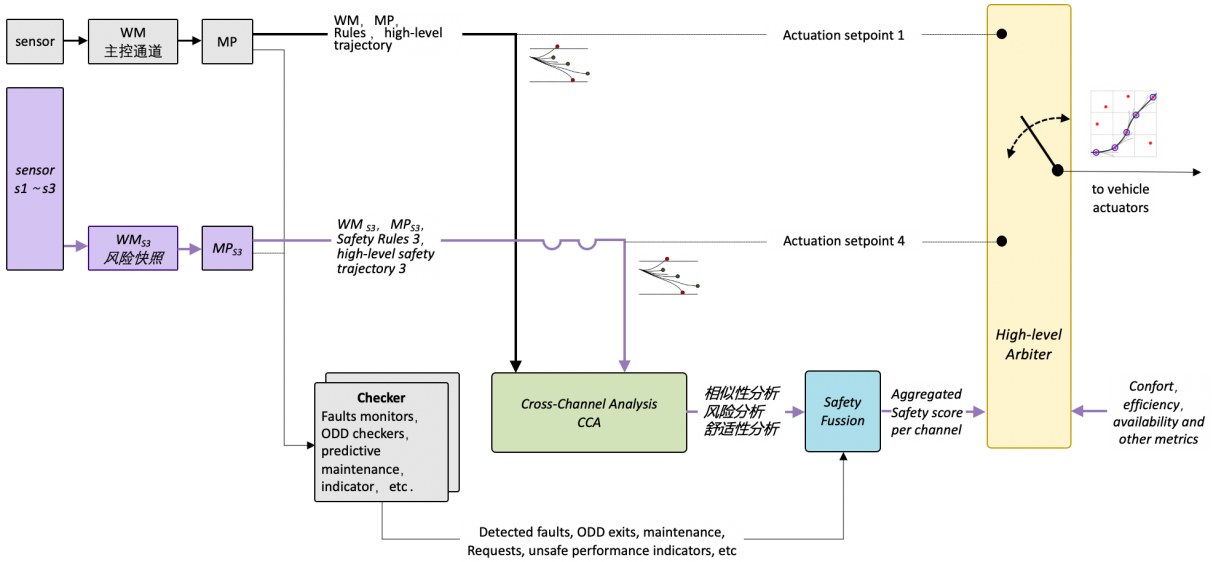


图 8 左脑+单模型

4）左脑+KD3

先将 n 个风险模型与 LLM 集成到一个 KD3（方法见附录 I），然后将 KD3 与左脑形成交叉互检异构冗余，见图 9。

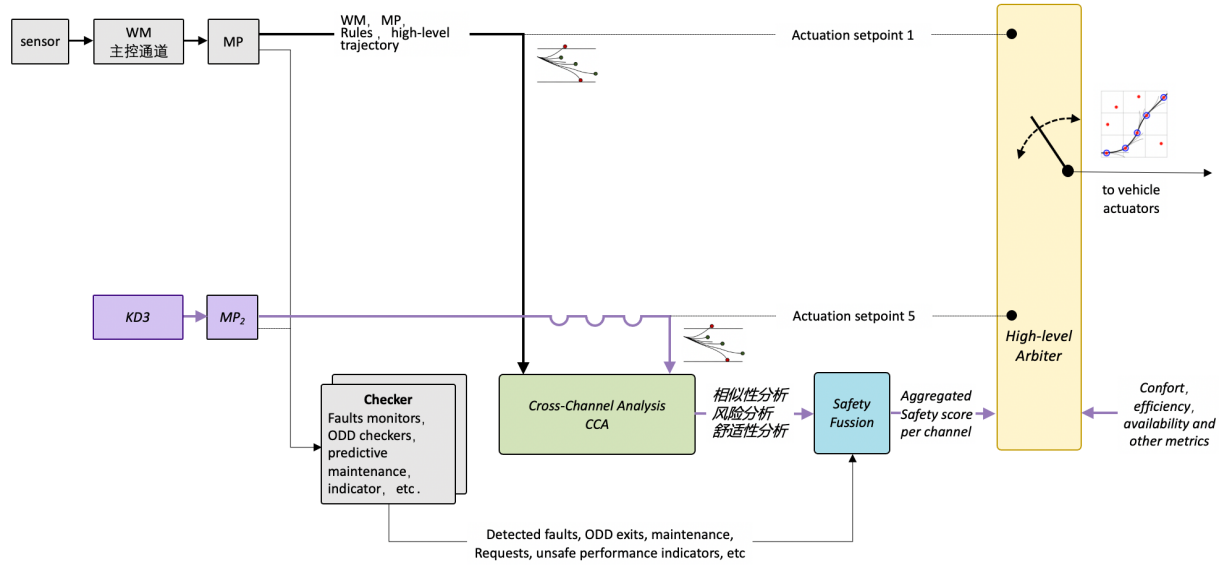


图 9 左脑+KD3 架构

4）左脑+右脑

将三个风险模型融合成一个右脑，然后与左脑冗余，见图 10。

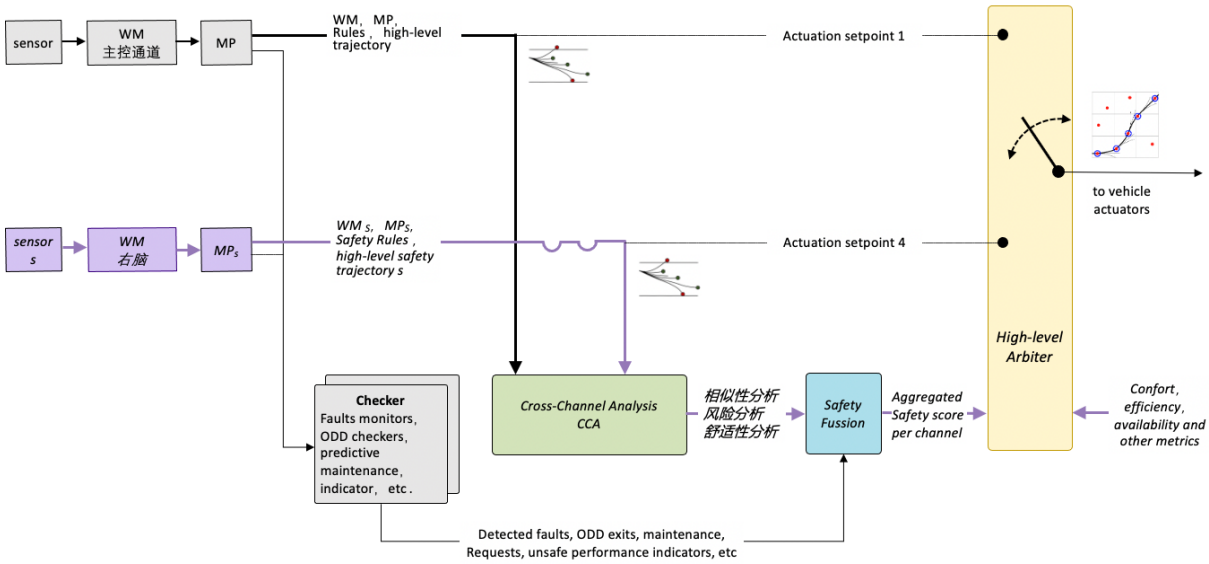


图 10 左脑+右脑架构

7.2 形式互检

很多理论符合三段式论证“潜规则”，代表了某种自然的暗规律，可以用来检验论断的稳定性。

多信息源之间的概念对照如下，映射对齐关系可以验证结论的合理性。多领域的概念投影对齐相当于科学方法里的交叉验证。

	对应关系		
产品重心	功能产品	能力产品	理解力产品
安全类别	FuSa	SOTIF	OS
技术路线	基于规则	基于数据	基于知识-数据双驱动
安全效果	动作安全	Scene 安全	Scenario 安全
认知域（抽象空间）	物理域	环境域	社会域
规划范围	执行层	战术层	战略层
先进 AI 工具	Physical AI	空间 AI	社会 AI；心理 AI

思维方式	左脑	右脑	双脑
安全架构	同构 Homo 冗余	版本 Multi-version 冗余	异构 Hetero-冗余
生物等效	爬行动物	哺乳动物	人类
……			

可以看到，跨领域对齐在形式上比较工整。

参考文献：
[1] End-to-end Autonomous Driving: Challenges and Frontiers
[2] CGX, 自动驾驶人工智能模型的整理.xlsx