安全大脑 AI 基座建议

Oct 20/2024

SQ

1 双脑智驾目标

功能目标：利用风险管控提高自动驾驶安全的可预测性。

智力水平要求：有直觉，有常识，能推理，有长期记忆，有安全目标

2 杨立昆演讲的启发

Yann LeCun 于 10 月 15 日 在 Hudson Forum 发表了题为 Human-Level AI 的演讲，演讲 ppt[1]为 YouTube 视频截图，文字来自于现场发言，网址见[1]

2.1 演讲摘要

p1

推理、规划、持久记忆和理解物理世界，是四个 AI 不具备的人类智能特征（p1）。

p2

Objective-Driven AI 事建立在一个世界模型之上，这时一个关于世界如何运作的心理模型，具有长期记忆，是 LLM 所不具备的。

p3

AI 的成功，包括 LLM，依赖于自监督学习技术。LLM 利用自回归预测，按照理想目标补齐缺陷，比如用来预测文本中的单词。这个概念 50 年代就有，只不过现在算力足够。

p4

但是自回归预测主要的一个局限性是没有真正的推理。还有一个限制，就是这只适用于以离散对象、符号、标记、单词、其他可以以离散化形式出现的数据。

p5

Moravec 悖论：对人类越简单的事，对机器就越难，反之亦然。

p6

4 岁儿童的信息输入量要大于 LLM。

p7

Objective-Driven AI 架构与 LLM（前馈神经网络）有很大的不同。前馈过程看到一个观察结果，产生一个输出。有很多情况下，对于一个感知，有多个可能的输出解释。人类的感知系统就是这样做的，大脑会自发地循环遍历这些解释。

绿色框是世界模型，是你关于世界如何运作的心理模型。可以想象你将要采取一系列动作，你的世界模型会预测这一系列动作对世界的影响，所以绿框会预测世界的最终状态是什么，或者预测世界中将要发生的事情的整个轨迹。你将它馈送到一堆目标函数，比如测量目标实现的程度，任务完成的程度，或者其他目标比如安全护栏，等等。

p9

找到使这些目标最小化的动作序列，这就是推理过程，所以它不仅仅是前馈。使用世界模型进行推理的好处，是可以完成新任务而不需要任何学习。也可以将大多数形式的推理简化为优化。这里需要的新事物是：如何用合适的抽象表达世界。

p13

Objective-Driven AI 规划的层级化特征，从纽约到巴黎，不可能规划到肌肉在 10 秒以内的动作。但是从哪里开始规划呢？AI 完全没有主意。层级化对 AI 是一个很艰难的任务，到现在没人作抽象分层这件事，是很大的挑战。只有抽象分层，才能实现目标驱动 objective- driven AI。

p14

人类直觉物理 （intuitive physics，physical intuitions）的培养，远在语言发展以前，比如对稳定性，重力，惯力等等的感觉。

p15

用 LLM 这种方法是不能学习出常识的，上个 10 年一直这样干，结果没学到任何东西。LLM 能生成图像，但无法用内部关联部反映现实世界（不是 objective-driven），所以不能将其付诸实践。在理解世界的运作结构 structure of the world 方面, AI 事完全的失败。

p16

Generative Architecture （生成式结构）无法预测视频。YL 用 JEPA 解决这个问题。主要特点是放弃预测像素，而是要学习 abstract representation ，然后在这个抽象空间进行预测。不同层次的表达实际上是科学的基本问题。

p18

预测的问题就是如何能找到正确的表达，很多其他参数都是无用参数。那么怎样训练这样一个 JEPA 系统能呢？关键是找出 cost function。

p19

Cost function 的能量表达。

p21

建议：放弃当今的 AI 四大支柱：generative model，probabilistic model，contrastive model，和 Reinforcement Learning。

p22，p23

VICReg 原理。

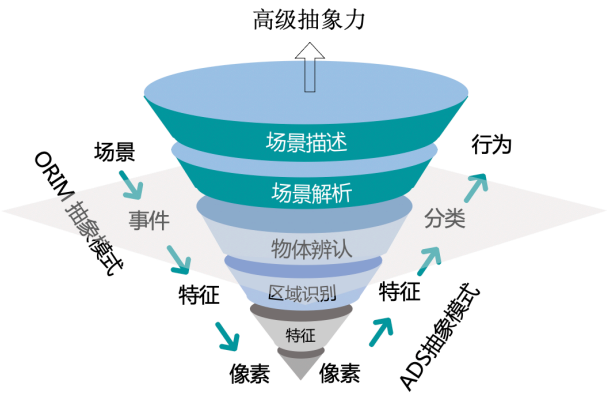
p25～27

JEPA 效果。

2.2 Human-Level AI 讨论

1）HL-AI 理念与双脑 Orim 理念很类似，其中包括：

- HL-AI 的目标是推理、规划、持久记忆和物理直觉；安全大脑的智力目标是能推理，有持久记忆，有常识，有社会交往直觉、有安全目标
- HL-AI 和 Orim 都主张规划分层，认为分层抽象是理解世界的前提
- 不认为 LLM 是推理引擎的核心，LLM 不能学出常识导致理性
- JEPA 放弃预测像素，学习 abstract representation ，然后在抽象空间进行预测；Orim 主张风险管控一定建立在高阶抽象的基础之上（见下图）



2）二者的不同之处

- Human-Level AI 将推理表达为优化问题，Orim 将推理归结为 knowledge-based agent 的问题
- JEPA 从环境中学习抽象表达；Orim 从核心领域知识里定义抽象分层
- 人类靠物理直觉和社会交往直觉处理日常事务。HL-AI 着重建立物理直觉，Orim 着重建立社会直觉，但是可以参考 HL-AI 的 JEPA 方法

3）讨论

杨立昆与辛顿对 AI 技术路线、内涵、未来前景的理解非常不同。本文倾向于杨立昆的基本观点。

苹果公司 10 月 7 日发表的一片论文也质疑了 LLM 的推理能力[2]。LLM 虽然能以语言为载体进行“信息填空”式的推理，但是 LLM 并不理解世界的运作原理，也就是杨立昆所说的世界模型。Marcus 对此表示，“我在 2001 年出版的 The Algebraic Mind 一书中提出的观点依然有效：符号操作必须是其中的一部分。在符号操作中，知识被真正地抽象为变量和对这些变量的运算，就像我们在代数和传统计算机编程中看到的那样。神经符号人工智能将这种机制与神经网络相结合——很可能是继续发展的必要条件。”

Ppt 演讲稿 p7 里面的绿色世界模型是 Objective-Driven AI 的重要组成部分。这个世界模型不同于目前用于 ADS 力的世界模型。ADS 世界模型知识描述环境，但是 OD-AI 具有长期记忆，对外界响应是有心理预期的，这是非常重要的区别。Orim 的世界模型更接近于 OD-AI 的描述。

关于直觉问题。在物理直觉和社交直觉当中，婴儿优先发展物理直觉，甚至早于语言发展。杨立昆的一个目标是通过看录像学习直觉、推理和预测能力，但是难点在于，智能体没有物理反馈是形不成物理直觉的（婴儿有大量的物理传感器）。关于物理世界理解，代表思想是李飞飞的空间智能公司 World Labs。“我们现在看到的大语言模型和多模态语言

模型，它们是底层表达其实是一种一维表达”。空间智能就是让人工智能直接绕过一切中间障碍，直接地感受、理解所身处的三位世界，然后采取行动。

与物理直觉不同的是，Orim 优先追求社会直觉，并且能通过传感系统在交互过程中获得社交反馈，前提是确定交往安全性的衡量目标。所以 Orim 培养起社交直觉要比培养物理直觉更容易。况且，即使没有物理直觉，只靠物理原理，控车就足够目前用的了。

3 流行安全技术与端到端（E2E）解决方案

3.1 E2E 的流行

安全监管目前处于真空期，国家法律法规及标准层面都没有对 FuSa 和 SOTIF 有强制性要求，所以安全显得并不那么紧急。但是，安全确实关乎企业的长久生存问题，又是一个重要议题，所以安全成了一个“重要但不紧急”的事情。

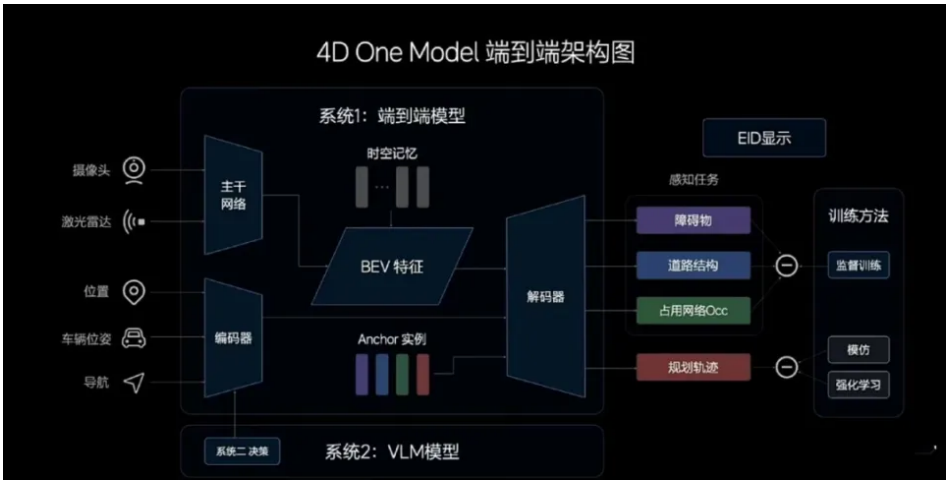
为了解决几乎无穷的“Corner case”，行业一股脑地投入到“端到端”热潮。就算 E2E 的安全性无法充分证明，也没有人因此而停止 E2E 的研发节奏。目前看来，E2E 是解决 corner case 的最佳途径。与安全相比，E2E 是一个“重要且紧急”的任务，大多数企业会把开发 E2E 放在优先于安全的地位。既然没有法规和监管限制，那就先把 E2E 开发出来形成实力，然后再慢慢优化和提升安全。目前企业对 E2E 的宣传，主要还是以市场卖车为目标。E2E 是在特斯拉公布以后才火起来的。受慑于特斯拉的强大，也出于没有能力寻求其他技术出路，今天的企业言必称 E2E，否则就会觉得要出局，使 E2E 成为了一种情绪载体，而非出于理性选择。

但是，E2E 并不是一个学术概念。

3.2 典型方案

3.2.1 理想

理想采用 One Model 的模式。系统 1 是一个 E2E 模型，具备快速处理信息能力，用来应对日常驾驶 95%的场景，系统 2 是一个 VLM 模型，具备逻辑思考能力，用来应对剩余的 5%困难场景。在系统 1 面对自己处理不了的复杂情况时，就会求助系统 2，系统 2 通过自己的大模型推理能力，持续向系统 1 输出环境理解，驾驶决策建议、驾驶参考轨迹等驾驶策略。



3.2.2 华为

华为的 ADS 3.0 E2E 方案采用的是 Two Model 模式，提出一个 “本能安全网络”的概念，作为智驾行为的下限兜底策略，确保输出结果具有安全底线。 “本能安全网络”目前在公开资料里还没有关于它的详细论述。

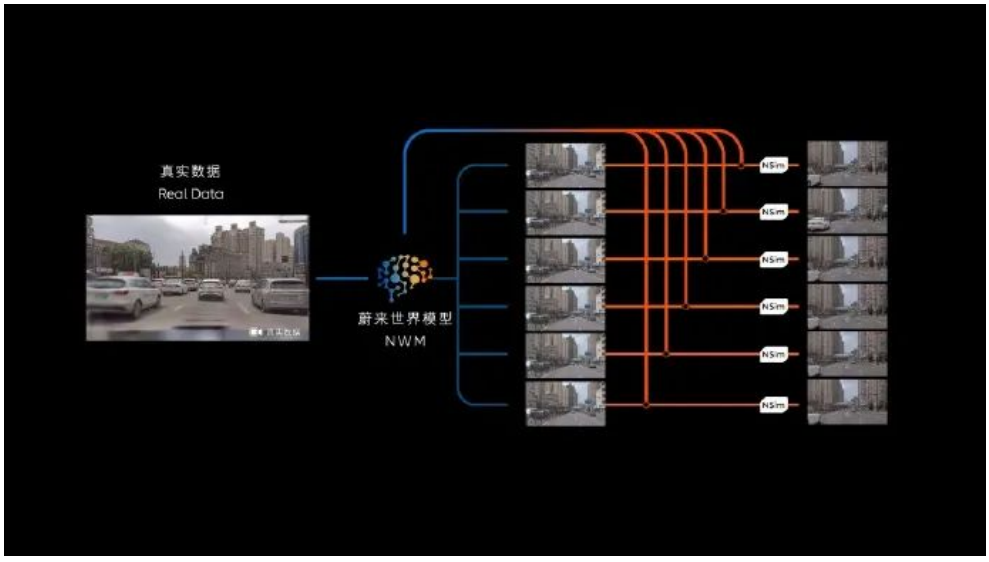
3.2.3 小鹏

小鹏汽车的 E2E 具体是 One Model 的模式还是 Two Model 不是很明确。小鹏的 E2E 模型包括了神经网络 XNet+规控大模型 XPlanner+大语言模型 XBrain，应该偏于 Two Model 模式。

3.2.4 世界模型

蔚来、理想等车企都发布了自研的“世界模型 World Model”，供应商中有地平线、小马智行等。

所谓世界模型，一般理解为对真实世界的仿真与建模，可以真实准确地还原比如十字路口等场景的变化。同时，世界模型还是一个评分体系，对自动驾驶系统的表现做出评价，能够得知 A 系统和 B 系统相比谁更好。



蔚来自动驾驶副总裁任少卿：“相比于常规的端到端的模型，新的世界模型有三个我们认为主要的优势。第一个是在空间理解上，通过生成式模型，从重构传感器的方式，更加泛化地抽取了信息。第二个，通过自回归模型，自动建模长时序环境。第三个，万千世界需要更多数据，通过自监督的方式，无须人工标注，它是一个多元自回归生成模型结构，让我们学得更好。”

地平线在感知上引入了“World Model”的概念，通过 World Model 的算法训练可以解决场景的泛化、功能的连续性以及体验的一致性的问题。在规控算法上，保留了 Rule-based 的链路。

Rule-based（基于规则）或者说 Principle-based 还是没有放弃，完全依靠 E2E“黑盒子”来解决问题，包括特斯拉、华为、小鹏等企业，结果还有待观察。

3.3 目前流行技术路线存在的安全问题

经常面临的疑惑包括：

1）依靠 E2E L2+能进化到 L4 吗？

2）怎样平衡 E2E 的安全上下限？

3）如何解决 E2E 不透明、不可解释行与认证监管之间的矛盾？

4）LLM 的角色和作用

5）E2E 需要对硬件资源和大数据的依赖问题，会消耗大量的成本。有没有其他能走入寻常百姓家的亲民技术路线？

3.4 讨论

3.4.1 安全是一个概率过程

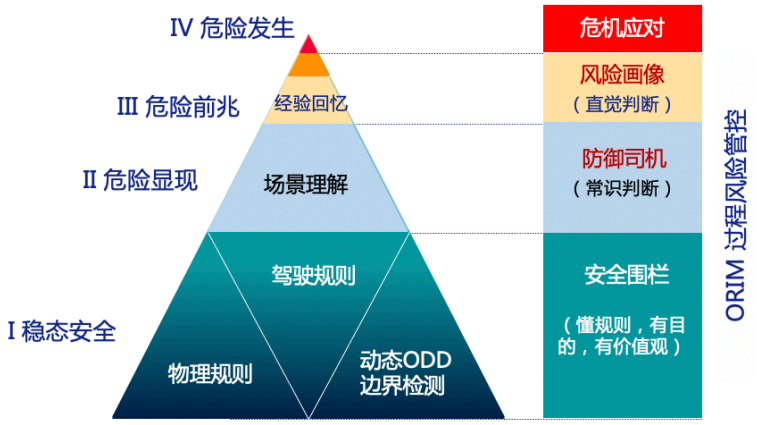
世界上不存在绝对的安全。安全的定义是“不存在不可接受的风险”，而不是不发生事故。

和疾病问题一样，安全是一个概率过程，某些健康问题和交通事故都是永远无法避免的。安全设计的目标是把安全变成可控，控制在“可接受范围之内”。如果安全系统的设计理念是“避免一切事故”，那么结果一定是失败，因为，安全是一个概率过程，是概率驱动，而不是一个确定过程，不可用微分方程来描述。把概率过程当成确定过程来管理，必然无法阻止小概率事件（长尾/corner 事件）的发生。这是当前安全理论的一个主要局限性。所以需要寻找新的评估方法来论证与评价，这便是安全大脑的方向：降低危害发生的概率。

目前 ADS 安全技术主要是寻求必然性解决途径，把安全当成确定过程来处理，而忽略了概率性。稳健的安全策略应当将安全划分为应对危机（确定性过程）和风险管理（概率性过程）两阶段分阶段管理，也就是先期过滤用右脑，临近反应用左脑（而不是像理想那样采用反过程）。

安全设计的大目标应是追求大概率，采用什么技术路线都可以。IPM 次数、透明性、可解释性本身并不是目的，而只是通向“大概率安全”的手段。

我们需要相信概率的力量。其中有两层含义：一是相信在数据统计层面上，概率一定会呈现相应的趋势规律，在大数据面前，事物一定会按照相应的概率分布，这意味着首先要相信风险金字塔是一个事实。二是我们努力要做的是提高有利事物的概率，降低有害事物的概率。这就是安全金字塔的内涵。



3.4.2 对 E2E 可解释性的要求

特斯拉认为有证据表明，FSD 的安全性比人类驾驶高 100 倍，但为什么仍然不能获取用户和政府监管的完全信任？为什么 NHTSA 宁可接受安全性低 100 倍的人类驾驶？

一个完全透明的 rule-based 系统目前在监管体制面前，心理上是被接受的，但是实际效果并不一定令人满意，比如不能覆盖 corner-case，能力很低，经常会发生异常导致安全问题，那么我们也是不能接受的。

特斯拉试图用结果说话，监管部门试图用可解释的过程说话，但是都指向了同一个最终目标，那就是想证明安全的可预测性。但是到目前为止双方并不互相买账。

怎样证明安全的“大概率的可预测性”？让我们从日常出行举例说明。

例 1: 你搭乘出行的时候，是否对司机放心？

为此我们通常要做两件事。

第一件事是考察驾驶员的背景。人们搭乘的理由首先是信任。比如，某个熟人是个谨慎的人，我可以搭他的车；我的朋友是个非常靠谱的人，这次他派他的儿子来机场接我，应当没问题。出租司机第一次接触，怎样建立信任？这时我们会使用信任链，比如这个网约公司的运行安全记录良好，整个行业的安全记录良好，这就相当于间接的背景人格确认。

第二件事是可能还要考察一下驾驶员的即时行为，比如操作是否滑顺，开车是否玩手机，是否打瞌睡，说话是不是不靠谱。

如果以上两条都得到确认（这个过程会以潜意识完成，会非常快），你才有可能在高速公路上放心休息。应当注意到，即使上述两条的确认结果都很满意，我们也不会相信会百分之百不出事故，但是剩下的危险概率是处在“可接受范围”，这就是我们能接受的“安全”预期。

从这个例子可以看到，保证“大概率可预测性”需要有两件事：司机人格背景调查和实际操作效果考察，也就是要经过“人格+能力”的双重认证，缺一不可。

FSD 目前只有操作效果证据，只展示了能力，面临的挑战是不能出示人格证据，所以难以获得用户的信赖和政府认证。想象一下，人怎样才能证明自己的人格？第一要有作人的原则（信仰、信念、道德准则、礼仪规范等等），第二是遵循原则办事，二者缺一不可，这和例 1 中的“两件事”是一个道理。第一次接触的人，无论他在你面前的表现多么完美无瑕，因为你不了他的人格背景，所以你也不可能把他当成一个可信赖的人。这就是 FSD 面临的挑战，这个挑战并不能随着里程的积累而自然得到解决，和里程积累无关。人格证明是 FSD 和政府监管之间最大的博弈点。

可解释性并不是 E2E 问题的根本障碍。在例 1 中，我们完成了两项核实任务之后，还需要司机随时向我们解释每个动作的动机吗？根本就不需要，90%以上的动作都是靠直觉的潜意识就直接完成了。当然，如果 FSD 有了可解释性，在应对监管时会更加主动。但是透明和可解释并不是被认可的唯一原因。

也许，可解释性与直觉性存在一个最佳平衡点。杨立昆在演讲的 p13 触及了这个问题：从纽约到巴黎不需要肌肉动作具有可解释性。可解释性是有一个平衡点的，这个平衡点取决于我们在世界模型将环境抽象成多少层，所以这里有回到了认知抽象层的问题。

人格是个人行为方式的概率保障，而不是行为确保机制。可解释性不是安全概率保障系统，ADS 的人格才是概率保证机制。所以，如何进行人格建设和人格认证是 L4 推广面临的挑战。

人格可以通过可解释性获得，也可以通过其他途径获得，比如信任链。安全大脑可以提供人格基础。

如果 FSD 不能提供人格证明，那只能等待 FSD 和 NHTSA 双方达到博弈的纳什均衡点。

3.4.4 对影响人物观点的讨论

1）马斯克相信，L2 能进化成 L4

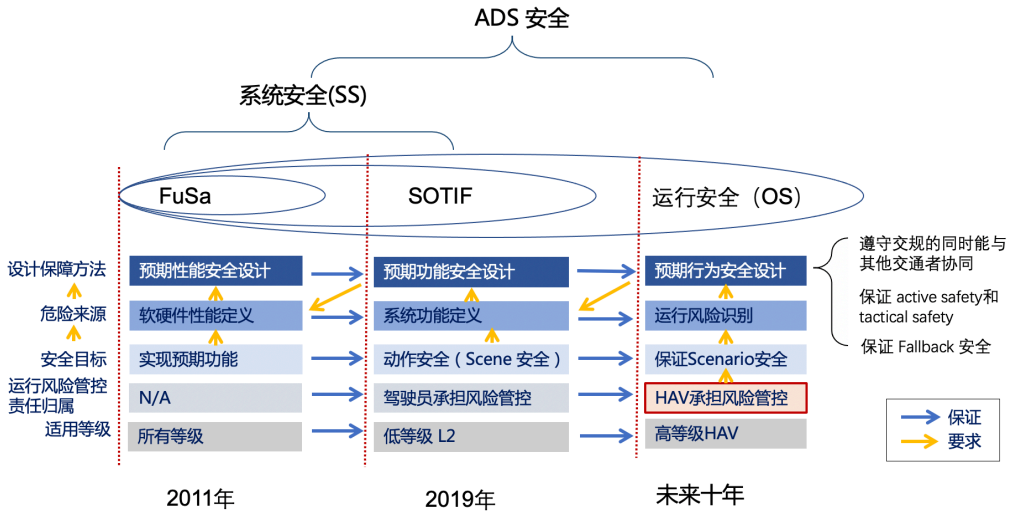
特斯拉做 Robotaxi 的逻辑是通过 L2 辅助驾驶的数据不断积累实现 L4。

本文对此持有不同意见。绿皮书总结出了一个 L2 和 L4 的根本区别，就是风险责任的归属问题。L2 的风险归驾驶员，L4 的风险责任完全归机器（见下图），这是绿皮书得出的最重要结论。

	确定过程管理 （应激响应）	概率过程管理 （风险管控）
L2	机器	人
L4	机器	机器

上表也可以表达为：L2 只要通过能力认证就可以，但是 L4 必须通过能力和人格双重认证。

风险责任的转移无法通过自然进化而完成，因为二者的概率过程管理机制不一样，不可能从人进化到机器。所以本文不认可自然进化论。



2）地平线市场总监刘文尧：

“E2E 端到端的技术路线由于是数据驱动的，在上限上的表现会更高，它能明显在更复杂的场景当中有更好的体验。但是，由于它是一个不可解释的黑盒的模式，完全放弃规则驱动（Rule-based），意味着它的下限就会不可控，很可能出现人都无法解释的诡异的驾驶安全行为。而且在这个行为出现的情况下，你很难对它做一个快速的 Bad Case 的迭代，因为你自己都解释不出来它为什么这么做，根因是很难找到的。”

本文赞同此观点。

3）BotAuto 侯晓迪：“L4 必须有可解释性，不能仅依赖说不清楚的数据黑盒子。 E2E 是个黑盒子，充满不可解释性”

本文意见：同意“L4 不能仅依赖说不清楚的数据黑盒子”，但不同意“L4 必须有可解释性”，因为可解释性不是目标，黑盒子也不是障碍，通向“安全可控性”终极目标的途径也不是只有可解释性一条路。详细理由见 3.4.2 的讨论。

4）小马智行 CTO 楼天城：“L2 做得越厉害，它离 L4 越远。反之也是如此。一个越好的 L4 公司，它离 L2 越远”

本文偏向于支持这个观点。主要理由还是风险归属问题（见上面 FuSa-SOTIF-OS 分析图）：

L4 的驾驶主体是机器，所有事情都要机器端解决，不能交给别人，没有人给系统兜底，车辆的驾驶权属于系统，责任属于企业，所以 L4 更关注的是安全。L2 的驾驶主体是人，智能辅助驾驶背后有人类司机兜底，关注的核心是成本、覆盖范围和体验，所以，L2 和 L4 的产品设计出发点不同。

5）小马智行楼天城：“World Model”是目前最佳最重要的东西，将其理解为通往自动驾驶的唯一解。

本文认为世界模型很重要，但是应当采用杨立昆 human-level AI 的世界模型，包括了对世界运作原理的心理预期，不是光停留在描述阶段。有了这个功能，世界模型就成了推理和预测机制的一个组成部分。

具有多层抽象、立体认知特征的，带有“世界运作机理心理预期”（杨立昆语）的世界模型是自动规划的基本前提，是 human-level AI 的基础。

4 安全大脑的 AI 架构搭原则

从杨立昆的 Human-Level AI 模型，行业 E2E 现状讨论，结合安全大脑的安全理念， AI 架构要满足以下要求：

- 安全目标是提高危险事物的可接受性、可预测性，提高安全概率
- 承认安全本质是概率驱动这一事实，用右脑是从概率的角度去解决安全问题
- 区分“过程驱动”和“概率驱动”两类安全任务
- 双脑系统要同时具备安全“过程控制”与“概率控制”两种能力
- 右脑负责建立 ADV 的人格体系（概率保证）
- 解决“概率安全”和“过程安全”的双向奔赴性
- 寻找可解释性与与 E2E 条件反射的平衡点，在正确的抽象层上有适度的可解释性
- 建立 objective-driven AI 的世界模型
- 体现双向避险特征：双脑系统可以设想用端到端方向从正向角度提高左脑应激反应能力，用右脑概率系统从反向降低危害 exposure rate

L4 的核心是如何完成一个稳定的系统，尤其是用不稳定的模块去完成一个稳定系统。L4 需要的是智慧，而不仅仅是资本和算力、数据、智商，只能通过架构层创新来保证系统的安全性。

5 要点结论

1）安全大脑的智力水平目标：有直觉，有常识，能推理，有长期记忆，有安全目标、能分层抽象

2）杨立昆 Human-level AI 的智力预期：推理、规划、持久记忆和理解物理世界

3）杨立昆-辛顿的分歧可以归结为“人是智力生物进化的中间一环，还是人是智力进化的终点”的问题，这是哲学“有神论-无神论”、“一元论-二元论”之争的延续，所以永远会同时存在两条健康的技术路线。安全大脑的 AI 架构应当具备兼容性。

4）安全大脑与杨立昆的智力模型思路更接近。

5）Marvin Minsky [3] 认为“分层抽象”是人脑智慧的独有特征。受其启发，安全大脑将分层抽象的世界模型当作认知基础。分级抽象也是 Human-Level AI 分级规划的前提。

6）安全是概率驱动，是可接受度（可控性），而不是零事故

7）安全可控性应当从“确定过程”管理和“概率过程”管理两方面的机制同时入手。概率过程控制是对右脑需求的原点

8）L2 只要通过能力认证就可以，但是 L4 必须通过“能力+人格”双重认证。人格建设和人格认证是推广 L4 面临的最大挑战

9）AI 可解释性不是目的而只是手段，不是安全概率的充要条件。不透明性不是 E2E 的主要障碍

10）安全大脑的主要任务是建立 HAV 人格体系，不光依靠可解释性，同时还要建设信任链

参考文献

[1] 杨立昆演讲 https://www.youtube.com/watch?v=4DsCtgtQlZU，15 Oct 2025

[2] Iman Mirzadeh et al，GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models，7 Oct 2024

[3] Marvin Minsky, The Emotion Machine, 2006