白盒子 vs 黑盒子
— 再议 E2E/LLM

20240608

1 白盒子的需求

安全大脑的目标是根据第一性原理而设计的：AI 思维应以人脑模型为基础。安全大脑认为汽车 AI 必须具备“长期记忆”、“常识推理”和“经验规则”三方面知识，使命是为汽车 AI 提供系统性的透明知识架构。

端到端（以下简称为 E2E）的根本缺点是很明显的：针对各种场合的表现，AV 的行为具有不可预测的不确定性，可以相信这就是 FSD v.12 只拿到 1%订阅的原因。即使发现车辆行为不尽人意，E2E 的黑盒子系统也很难做针对性局部设计修改。白盒子知识架构使自动驾驶的“行为设计”成为可能。行为是决定用户接受度的根本。

从目标上来看，安全大脑针对实际需求而提出，而不是先进技术的堆积或者技术路线模仿。举一个简单的例子，实际发生过的千万次死亡事故场景 AV 需不需要知道？E2E 是否有能力全面获取人类的领域经验知识？只要这个知识需求是实际存在的，那么安全大脑就是必要的。

如果这个需求不存在，那么就应退回到上级命题讨论（包括不限于）：AV 是否缺失右脑？右脑的结构是什么？AV 需要什么杨的知识结构？目前的安全大脑结构是假设上级命题讨论已经完成。

2. 技术比较

白盒子知识（安全右脑）与黑盒子知识（E2E）技术路线之间互有长短（见网会材料-王红老师部分），目前看来不能完全互相取代。E2E 的优势是，如果算力和数据足够，自动化的训练过程比较高效，但是也面临诸多挑战，比如：

- 不可解释性
- 行为不可设计性
- 缺乏统一的价值与目标体系，无法中和多种需求的冲突
- 长尾问题
- 跷跷板效应
- 不清楚自己的能力边界在哪里
- 数据壁垒（可以假设没有）
- 算力壁垒（可以假设没有）

- 行业同质化竞争
- Trust 体系建设
- 准入所需要的“理解力”证明

白盒子与黑盒子的目标是不一样的。白盒子的输出是知识，黑盒子的输出是行为。从目标先进性上看没有可比性。

另外，工具/方法层面的先进性可以用是否能高效达成目标来评价。以 SpaceX 星舰壳体材料 301 不锈钢为例，301 远没有铝合金、碳纤维高大上，但是综合评价高低温性能、成本、重量却是最好的，第四飞的表现充分证明，301 材料本身很“落后”，但是应用却是十分先进。白盒知识技术不需要 100 E-TOPs 这类先进技术的支持，但是在应用方面并不落后，甚至有更高的目标。

3. 安全右脑不排斥 E2E，不是 E2E 的竞争方案。

安全大脑很乐于和 E2E 的战术层结合，不关心战术层是 rule-based 还是 E2E-based。

在安全大脑内部结构里，推理模块里也可能包括 E2E/LLM 方法。

4. E2E/LLM 并不是完美的解决方案

杨立昆等人强调 AGI 的思维完整性，不认为 LLM 是突破 AGI 的终极方案。

5. 为什么大家都在走 E2E 或者大模型

E2E 的工具开发已经成熟，尤其是 transformer，有特斯拉等先驱趟路，在应用上已经没有技术障碍，剩下的只是如何获取高质量有效数据和足够的算力。对于基础好的厂家，E2E 只是投入规模问题，能提供全栈解决方案，是个不错的选择。

大模型是近期新技术，主要作用是提供通用知识和自动推理，并不能全面系统地提供 AV 所需要的全部知识，目前还处于探索开发阶段，Wayve 在尝试提供全栈解决方案，投入是 10 亿美元数量级。很多其他人也在探索常识推理的应用。目前还没有成为市场应用的主流。

虽然欧洲、日本一直在探索白盒知识技术，但是由于没有整体知识架构目标定义，开发工具零散，没有全栈通用工具，需要各个击破，开发速度较慢，所以在市场应用的速度上落后于 E2E 的迅猛发展。E2E/LLM 的发展得益于近期 AI 芯片性能的大幅度提升，但是这并不意味着对白盒知识的需求已经不存在了。

白盒子知识架构开发比较麻烦，但却是一条可解释、可持续成长、低算力门槛的实用化路线。鉴于白盒子的天生固有优势，知识-数据双驱动 EPTech 右脑也会形成对 E2E 的吸引力。