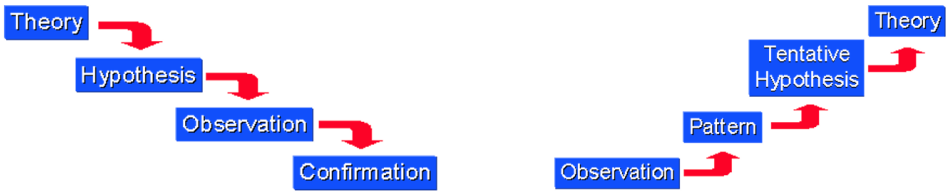
自动驾驶需要什么样的世界模型

1 AV 认知技术路线

在科学发展的历史上，“解析-综合”、“演绎-归纳”几乎成了探索的万能药（图 1）。同样，自动驾驶对世界的认知也不例外：解析法与综合法构成了当今“世界模型”与“大模型”两个技术主流。



a）演绎/解析法（自上至下）　　b）归纳/综合法（自上至下）

图 1　两种科学研究方法

端到端是综合技术路线的典型代表。虽然 VLA 的目的是具有可解释性，但是属于事后表象注解，不属于以分类学为起点的解析方法，所以仍然属于综合技术路线。

世界模型使用解析的方法对危险进行分门别类，然后寻找对应措施各个击破（图 1，a)），属于演绎技术路线。

2 AV 对世界模型的内容要求：物理现实+社会现实

自动驾驶对世界模型的要求：

> *世界模型可以对外部环境的状态进行归类观察，根据系统性知识体系对事物的时空关联进行理解，并对外部世界的发展趋势进行预判。*

环境状态可分为“物理状态”和“社会状态”两个方面，因为车辆的运动一方面要遵守物理规律，另外一方面又要受社会规律的制约（交通规则、行为习惯、道德约束、互动行为的可接受性等等）。同样，人类依靠“物理直觉”和“社会直觉”完成日常生活行为。为此，世界模型需要建立“物理世界”和“社会世界”两类模型。

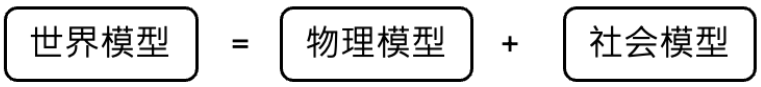


图 2 世界模型的两个子模型

	物理世界模型	社会世界模型
分类状态观察量	道路几何、车道线、交通标志、地标、地图；其他车辆 / 行人 / 骑行者 / 交通信号	区域识别、道路分类、适用交通法规则、分类识别、对象关联度量、关联模式表达与观察
理解	冲突区域、冲突类别、	事态、意图、征兆、隐含、其他交通者意图判断、其他交通者交互愿望
知识系统	运动学，动力学，环境影响分析、物理公式，微分方程，	安全常识集，规则集，惯例，安全实践（防御驾驶）手册
推测引擎	物理公式，微分方程	直觉思维引擎、常识推理引擎
输出结果	未来某一刻的位置、姿态和运动学、动力学参数、可行驶区域	本车的行为（行为设计）

Rule-based 自动驾驶系统、端到端系统基本属于“物理世界模型”。

“社会世界模型”的目的是决定本车的行为，也就是驾驶行为设计。行为是指完成一个任务的方式。完成一个任务有很多种方式，但是如何兼顾安全、高效、舒适、符合乘员心理、符合他车心理，社会世界模型要为此找到一个此时此景下的最佳方式，比如如何安全汇入高速公路主导、过路口时如何与他车互动、此时应当块加速还是缓加速，等等。为此，社会世界模型必须具备预测社会环境演变的能力。

与物理预测不同的是，社会环境演变不能靠数学等式运算而获得结果，而只能根据常识进行推理而得出结论。所以，社会世界模型的核心知识体系是常识集。

世界模型可用于用于解决安全、高效、舒适、经济、用户体验等多方面的问题。不会有一个世界模型同时覆盖上述的所有领域，人类也是一样，不需要通晓所有领域的知识。厨师的世界模型和外科医生的世界模型是不一样，二者之间共通的是物理世界模型，这个模型人类从婴儿时期就开始培养；差异是二者的社会模型不同，因为他们有不同的社会交往内容和方式，需要用后天的专业训练进行培养。婴儿需要通过自己独立的反复试探才能学会抓取、走路、游泳、重力、惯性等一系列物理直觉，但是人类很难将这些物理直觉描述出来再转达给给机器，所以 Yann LeCun 的目标是让机器像婴儿一样通过学习训练自己摸索出的一套物理世界模型[?]。

本文范围内的世界模型聚焦在安全领域的专用世界模型。

对世界模型有多种不同的定义，比如[1], [2], [3], [4]。

3 EPTech 的安全策略

EPTech 世界模型聚焦于 AV 安全。

事故都是酝酿而成的，是偶然中的必然。任何一次事故都是行为的后果，并且事先写满了预兆。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。安全设计的总目标应是追求“大概率安全”。

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，这便是对安全世界模型的要求：降低危害发生的概率。

目前 ADS 安全技术多聚焦于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，也就是先期用概率过程进行风险过滤，临近事故时采用确定性过程进行危机应对。安全控制的“两阶段方案”是 EPTech 安全策略的核心。

在传统的自动驾驶安全技术里，大多不划分这两类性质的矛盾，采用统一的应对措施，直到危险临近时才采取紧急应对措施，对系统性能提出了极高的要求，也增加了应对失败的概率。

风险控制的基本理论是“安全金字塔”，将安全份为危险控制和风险控制两个话题, 详见[5], [6]，是 EPTech 安全策略的出发点。

4 安全世界模型的双模思维要求：左脑与右脑

风险和危机不是一个发生机理，一个是确定性过程，另外一个是概率过程，所以，适用于这两个目的的世界模型是不同的。

临近事故的金字塔顶端适于用“物理世界模型”来描述（图 3），因为事件的发生是确定性的，动作目标是确定的，物理世界模型的任务是评价这个事件和动作是否安全。

金字塔的下部风险是不确定发生的危险（图 3），是否发生取决于个体之间的互动关系，也就是社会关系。虽然风险不一定发生，但是风险要素是确定的，风险-危险的比例也是确定的，因此人们制定了一系列的风险方案措施，比如交规和防御驾驶规则。降低风险的发生率等同降低危险的发生率。风险的描述更适合于用“社会世界模型”。

物理模型是一个完全的理性世界，没有弹性和灰区，这种思维方式与人类左脑接近。社会模型注重定性的关系分析，具备常识，理解约定俗成，运用直觉判断，思维方式与人类右脑接近。所以完整的世界模型应当由左脑-右脑两个部分构成的。这种特征我们称之为安全世界模型的“双模特征”。

右脑是人脑不可或缺的组成部分[7]。人有健全的双脑时，左脑负责局部，右脑负责整体态势。左脑侧重于控制有规律的日常行为，右脑专门负责处理威胁等这类非日常情况的反应。比如，当生物遇到捕食者要吃掉自己时，需要立即采取适当的行动保证生命安全，而右脑就是为了处理这些紧急事务进化而来的。从这种意义而言，右脑就是安全大脑。

如果失去了右脑 Right Hemisphere Brain Damage (RHD)，人类会

- 失去整体概念，没有社会关联感，无法与人交往
- 没有注意力分配机制，所以注意力涣散，时刻分心，无法专注，没有关注中心，容易受扰动，所以执行效率非常低下，要么过激，要么迟缓，会表现得很神经质
- 无法推理和理解语言暗示，无法理解笑话，无法理解局势
- 没有长期记忆，难以回忆重要的经验，也无法学习，无法安排计划

- 没有目的性，不知道问题的存在，更不知道寻找解决问题的方法
- 无法感知左侧信息

开发安全右脑的目的在于：

- 识别风险和避险
- 提高左脑执行任务的效率

另外，右脑的思维方式与 M Minsky 提出的六层意识模型是非常接近的（图 6），见[8]、[25]，要求具备在不同层级进行多层抽象的能力。可为将来进入到自主任务分解打下基础。自主任务分解要求机器要分层思维，能够把陌生的任务分解成一系列熟悉的子任务组合，再用子任务的答案综合出陌生任务的解决方案。Yann LeCun 也强调了思维分层的重要性（见下面第 6 节论述）。

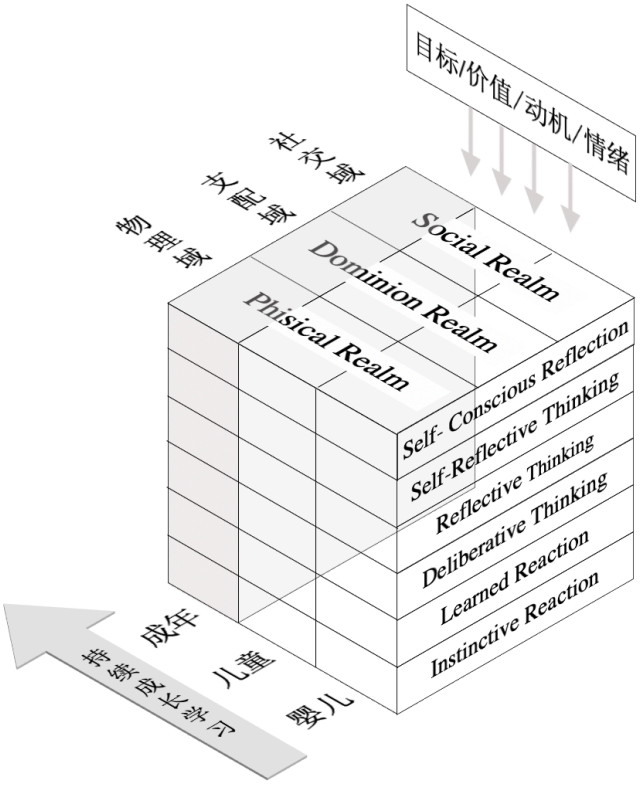


图 6 六层 AV 意识模型：思维异构化是人类智慧（resourcefulness）的来源

我们将上述的内容要求（物理世界+社会现实）、思维方式要求（左脑+右脑）、安全的分类控制概念（风险+危险）之间的对应关系统一表达成图 3。

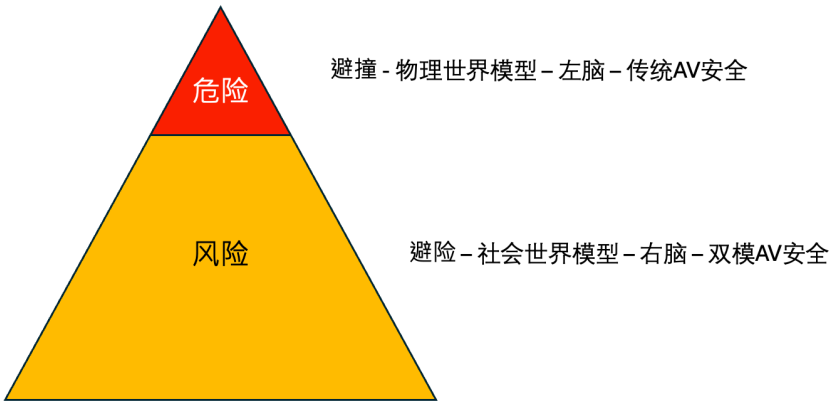


图 3 风险 vs 危险、左脑 vs 右脑、物理模型 vs 社会模型

5 安全右脑的构造

左脑部分目前行业技术日臻完善，因此益普华安的安全世界模型聚焦在右脑，这部分工作目前开展的很少。

5.1 右脑构成

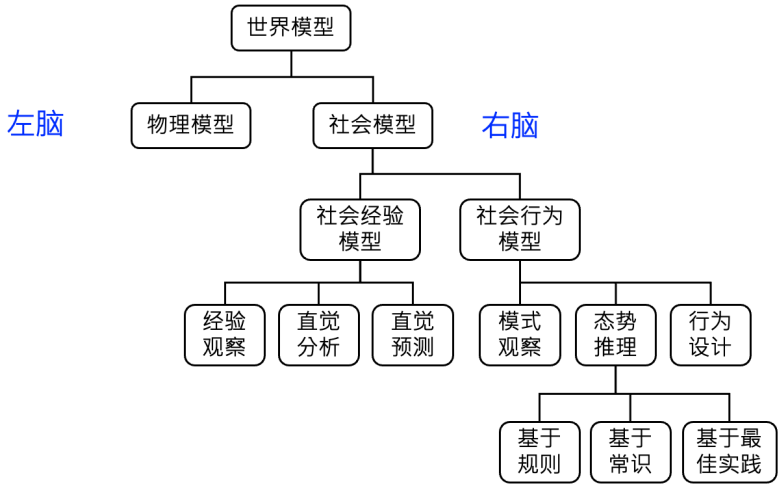


图 4 右脑社会世界模型的进一步分解

右脑的分解是基于对人类认知过程的模仿。回想一下我们人类驾驶时如何保证安全的。

日常驾驶光靠手疾眼快时无法保证安全的，始终绷紧神经反倒更容易出危险。首先，我们会熟知规则，因为交通规则就是避免事故的最佳指南。其次，我们还会听取很多安全驾驶

知识，比如老司机的经验、驾校的忠告、防御驾驶手册的建议，去刻意避免事故的征兆，作到“君子不临危墙”，防患于未然。然后，我们会根据常识去推测环境的演变，并据此规划我们的行动，比如，根据常识判断，一辆和你并行的有轨电车是不会突然切入你的车道的。最后，我们会回忆我们自己的经验或者亲友的经验，规避容易引起事故的“模式”，起作用相当于“此处事故多发”的警示牌。

因此，右脑的基本功能应当包括（见图 4）：

- 安全模式的表征与观察能力（比物体物理特性更高等级的抽象能力，比如场景描述能力）
- 安全驾驶知识的长期记忆，包括交通规则、防御驾驶安全知识、文明交互行为知识、社会常识知识等
- 场景演变的常识推理能力

产品形式见图 5（详见[8]，插图 7）.

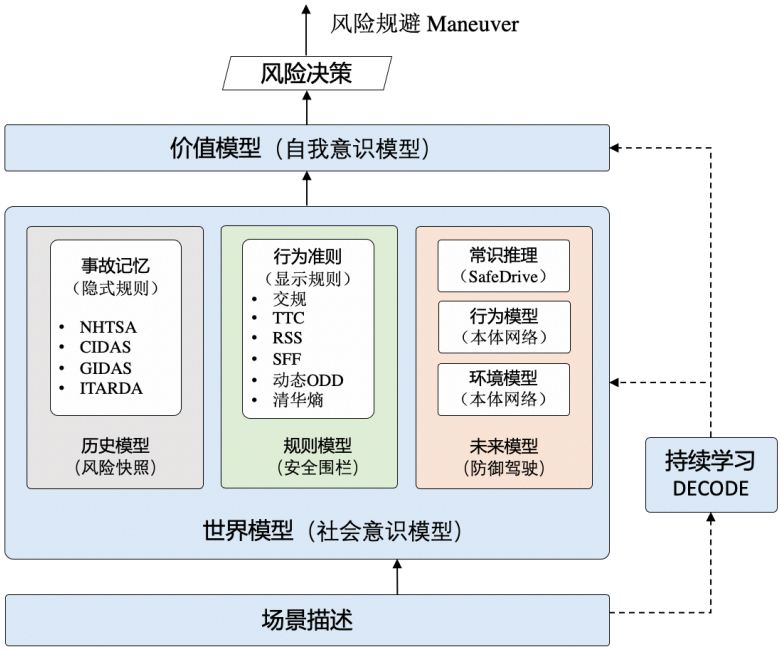


图 5 右脑世界模型

子模块技术方案汇总见[24]，子模块开发示例见下面 5.4 节内容。

5.2 双模安全大脑的特点

双模大脑的应用价值包括：

1）应对 D2D 目前面临的 11 项挑战（见[9], 5.1）
2）应对 SOTIF 的挑战（见[9], 5.2）
3）应对 OS 的挑战（见[9], 5.3）

与大语言模型相比，双模大脑能够提供：

1） 高阶安全目标与目标任务分解
2） 分层思维与多层抽象
3） Ground Truth

5.3 右脑开发技术路线

在现有技术条件下，右脑的产品化开发是可行的，详细技术路线见[24]。

5.4 左右脑的协同

人类的左右脑时如何协同工作的目前还不是很清楚，也就是两边各自遵守哪些退让原则。所以，无法模仿人脑建立起一套左右协调机制。

三个风险模型在输出右脑决策之前，先要通过 Fusion 1 进行融合（图 9），大概方法见[22]，右脑输出为风险五元组。

左脑的决策与右脑的决策需要通过 Fusion 2 进行融合，大概方法见[21]、[22]。

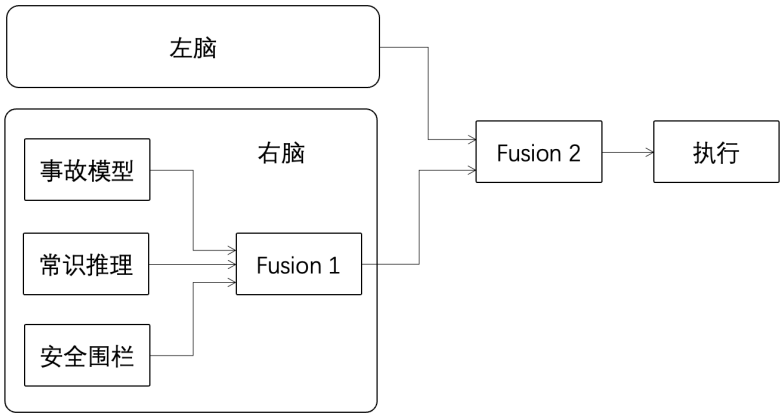


图 9

6 学术流派与益普华安世界模型

1) Yann LeCun

主要观点[17]：

> *我认为未来人工智能和机器学习研究面临着三个挑战。第一是学习世界的表征和预测模型。解决这个问题的方法就是自监督学习。第二是学习推理。基本上与人类的潜意识相对应，可以做到下意识的反应，不需要过多地思考。第三是学会分层制定行动计划。可以通过大量复杂的动作来实现目标。*
>
> *大多数人类知识都是非语言的。我们在一岁之前学到的一切都与语言无关。除非拥有以视觉形式提供直接感官信息的系统，否则我们将无法创造出达到人类智力水平的人工智能。……LLMs 确实无法计划的事实，它们没有真正的思考能力，也没有和人类一样的推理和计划能力。*
>
> *最终，我们想要做的是使用自监督学习和 JEPA 架构来构建之前提到的那种可以预测世界和进行计划推理的系统，这些系统是分层的，可以预测世界上将要发生的事情。……我们需要的是一个能够学习世界状态的系统，这将使它们能够将复杂的任务分解成更简单层次的任务。*
>
> *我们正在使用这些 JEPA 架构，但我们还没有最终的配方。我们可以用它来构建由目标驱动的推理和计划的 LLMs，希望可以构建出能够分层规划的学习系统，就像动物和人类一样。*
>
> *我不认为存在通用人工智能这样的概念，人工智能是非常专业的。*

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。

如果把世界模型应用到具身机器人上，那么我们还需要将图 4 扩充成图 7 所示的“领域分布地图”。可以看到，明显的变化是增加了“直觉物理”部分。Yann LeCun

和 FeiFei Li 的研究目前主要聚焦在这一部分。最大挑战是，没有婴儿的感知能力，但是要建立婴儿的直觉。可以想像，人形机器人如果不建立起物理直觉和社会直觉，其产业将面临泡沫化。

物理直觉一定要通过经验学习进行总结和积累，因为这部分知识很难用语言描述出来传达给机器。杨立昆试图通过让机器通过看视频建立物理直觉，可以想像难度是非常大的。婴儿在建立直觉的过程中，除了视觉还有触觉、听觉、嗅觉、味觉等感官输入，这样才能建立重力、惯性、平衡等物理直觉。单靠视觉观察没有其他渠道的反馈建立物理直觉的难度可想而知。物理智能、空间智能都属于目前的前沿研究。

社会直觉的常识可以用语言表达出来，所以可以传输给机器，这就是 EPTech 右脑正在做的工作。

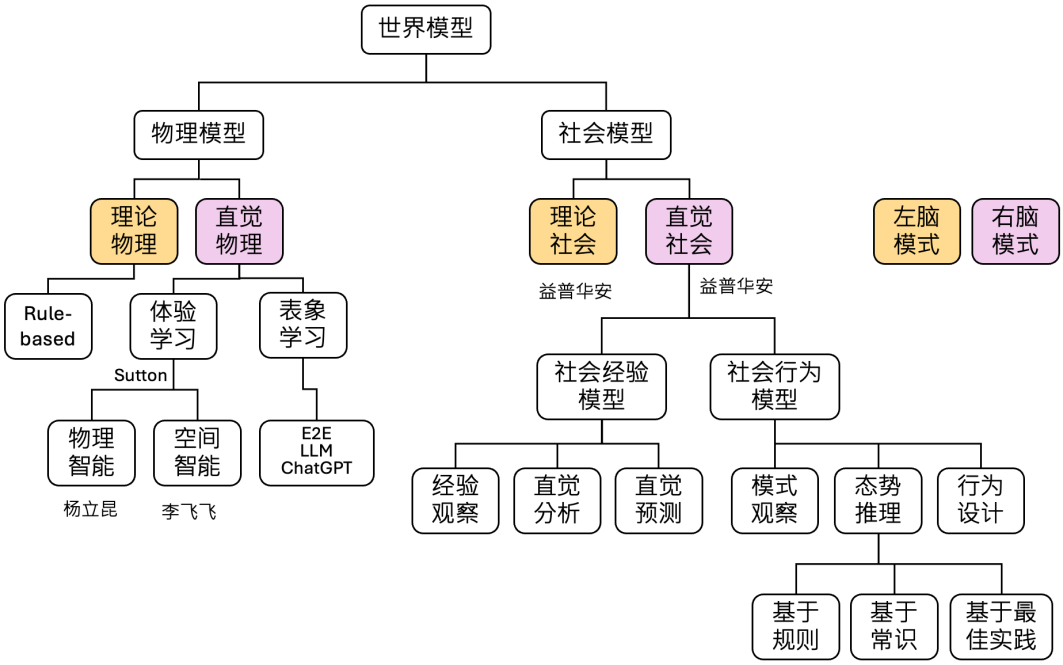


图 7 具身智能体所需要的世界模型

双模大脑的设计与 Sutton、Yann LeCun 关于世界模型的观点比较接近。

2) Richard Sutton

强化学习之父、25 年度图灵奖得主 Richard Sutton 观点：大语言模型是一个错误的起点，是「死胡同」。在他看来，真正的智能必须源于经验学习，而不是模仿人类语言的「预测游戏」。

在 Sutton 与科技博主 Dwarkesh Patel 的对话中[18]，Sutton 认为，真正的智能源自经验学习，通过行动、观察与反馈持续修正行为，实现目标；相比之下，大语言模型的预测能力更多是对人类行为的模仿，它没有独立的目标，也无法对外部世界的变化产生真正意义上的惊讶和调整。他认为，想要真正可扩展的智能，必须从经验学习出发，而不是把大语言模型当作起点。这实际上相当于站队杨立昆。

谈到大模型，他说：

> *LLM 模仿人类语言，并不等于在建立世界模型。那只是在模仿那些拥有世界模型的人类。我并不是想采取对抗的立场，但我想质疑「大语言模型具备世界模型」这个观点。一个真正的世界模型，应该能预测未来会发生什么。大语言模型能预测某个人会说什么，但没法预测世界上会发生什么。*
>
> *当你尝试建立世界模型时，你会预测会发生什么，然后观察结果。这中间存在 ground truth。但大语言模型没有这种 ground truth，它们没法预测接下来真实会发生什么。*
>
> *关键还是在于，它们（LLM）缺乏目标。对我来说，拥有目标是智能的本质。如果一个系统能实现目标，那它就是智能的。我赞同 John McCarthy 的定义：智能就是达成目标的计算能力。没有目标，它就只是一个行为系统，没有特别之处，算不上智能。你同意大语言模型没有目标吗？*
>
> *真正可扩展的方法是从经验中学习。尝试各种做法，观察哪些有效。不需要有人告诉你。前提是，有一个目标，没有目标，就没有对错或好坏之分，而大语言模型试图在没有目标或优劣判断的情况下运作。这就是一个错误的起点。*

应当指出，安全大脑的右脑设计包含了上面提到的目标和 ground truth。

双模安全大脑的最终形态应当包含最高级抽象层的安全目标，安全大脑可以根据实际环境对这些高级目标进行现场任务分解，不同的环境可能导致不同的分解结果，用这种方法可以陌生场景解析的难题。比如，本车从支路汇入主路的安全原则之一是“不得迫使主路车辆采取避撞措施”，而不是固定的 TTC 判据。这种思路与 Yann LeCun、Marvin Minsky、Sutton 的明确目标、目标分解、多层思维的学术思想是一致的。

3) Ed Chi，谷歌 DeepMind 的 VP

他在 2025 八月份发表了一篇名为 Google Gemini Era：Bringing AI to Universal Assistant and the Real World 的演讲[19]，其中开了一个玩笑说（@8 min 5 Sec），你要是先知先觉，2022 年买 ChatGPT 的股票现在早就发横财了，但别泄气，未来发财还有机会。他认为未来的 agentics 设计应当考虑到人脑有两种思维机制，系统 1 和系统 2 （他在此引用了 Daniel Kahneman 的观点），一个是感性的模式识别快系统，另一个是理性的逻辑思考慢系统。所以，未来的投资方向如果投在具备双系统的 agencies 上，你就能攥钱。这个观点是和我们一贯的主张完全吻合的，双脑综合完全可能成为下一步的潮流，现很多人已经开始懵懵懂懂地提这方面的概念。

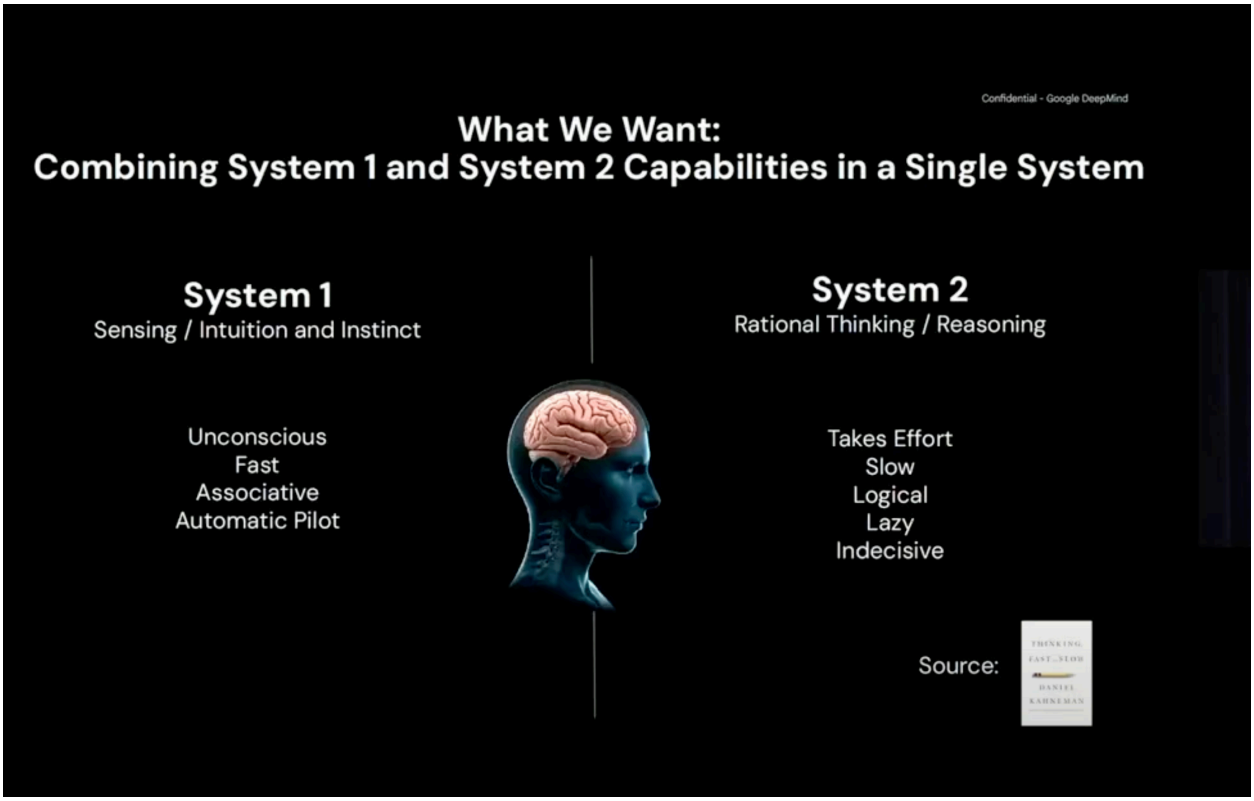


图 8

7 双模大脑世界模型与 AI

AI 通常泛指一些最新的智能工具，比如深度学习技术。但是 AI 真正的意义在于如何更接近于人脑思维模式，以便最后实现自主繁衍。

双模大脑是一个 AI 目标，而不是一个单一的 AI 工具。所有最前沿的 AI 工具都可以应用于为这个目标服务，所以双模大脑是目标与一系列 AI 工具的组合，包括机器学习、知识

图谱、机器推理、Data Engine、大语言模型等等。其中目标是相对不变的，工具是动态更新的。

参考文献

[1] Jingtao Ding et al, Understanding World or Predicting Future? A Comprehensive Survey of World Models, arXiv:2411.14499 [cs.CL]（2024）

[2] Javier Del Ser et al, World Models in Artificial Intelligence: Sensing, Learning, and Reasoning Like a Child", https://doi.org/10.48550/arXiv.2503.15168, (2025)

[3] Yanchen Guan et al, World Models for Autonomous Driving: An Initial Survey, https://doi.org/10.48550/arXiv.2403.02622,

[4] Sifan Tu et al, The Role of World Models in Shaping Autonomous Driving, https://doi.org/10.48550/arXiv.2502.10498, 2025

[5] SQ, 风险控制金字塔理论

[6] CertiCAV, Assurance Paper, 2021

[7] SQ，没有右脑人会怎样

[8] SQ，Minsky 意识模型与 AV 世界模型

[9] 03012025 安全大脑定位讨论, SQ

[18] https://www.dwarkesh.com/p/richard-sutton

[19] https://www.youtube.com/watch?v=b78mvG40yis&t=9s

[21] SQ 左右脑高阶融合机制

[22] SQ 安全大脑解析与右脑综合

[23] SQ 安全大脑架构设计讨论

[25] 基于 Minsky 六级抽象的 AV 世界模型