世界模型与法律

Claudio Novelli[a]

[a] 耶鲁大学数字伦理中心，美国康涅狄格州纽黑文

摘要。当前的人工智能或许具备关于我们如何使用人类语言的模型，但这并不意味着它能可靠地表征物理现实。世界模型有望弥补这一差距：它们通过学习环境随时间变化的方式，并预测某个动作可能造成的后果来实现这一点；目前，它们已被用于操控机器人和生成可供探索的世界。本文探讨欧盟、加利福尼亚州和中国这三大主要法律环境中的人工智能监管如何应对世界模型带来的独特挑战。我提出了一个定义，并依据谁或什么主体会根据其预测采取动作，对世界模型进行分类，同时指出对其产生不当依赖的风险。我认为，现行规则往往涵盖世界模型，但仍留下三个问题有待解决，即行动溯源、轨迹有效性以及模拟证据的权重；本文并建议了填补这些空白的办法。

1. 引言

大语言模型（LLM）并非为构建世界模型而设计：它们或许具备关于我们如何使用人类语言的模型，也可能掌握零星的世界知识，但这并不意味着它们能可靠地理解物理定律或生物学。因此，世界模型有望成为一种替代方案，或至少是一种互补且突破性的技术范式，并且正日益获得市场认可。

从法律角度来看，毫无疑问，世界模型可能会根据其使用方式而落入现有人工智能监管框架的范围，但这些监管框架是否已准备好应对世界模型，或者相反，是否需要进行类似于大语言模型所引发的调整（例如《人工智能法案》中的调整），则不那么明确。[1] 在本文中，我将探讨一些最重要的人工智能监管模式是否足以应对世界模型带来的挑战。

我试图说明，从法律角度看，世界模型最重要的问题，主要在于：由谁、在多大程度上，将其生成的内容视为某种证据。我强调世界模型在工作流程中所发挥作用的重要性，以及以三种方式处理其生成的证据的重要性：(a) 建立一种分类法，根据由谁或什么对其输出采取动作来对世界模型进行分类（第3节）；(b) 确定三类风险，说明这种依赖可能如何被错误地置于其上（第4节）；以及 (c) 为监管机构提出七项建议（第6节）。这些建议都受到同一原则的启发：赋予世界模型的权重，应当根据支持它的证据进行校准。其中一些建议源于一项总体性的法律关切：世界模型的输出，是关于某个动作将导致何种结果的主张，而这些主张会在一系列复杂且相互关联的事件中逐步展开。因此，应当在一段时间跨度内（第6.2节）以及在现实世界条件（包络）下（第6.3节）检验其可靠性，

[1] Natali Helberger 和 Nicholas Diakopoulos，《ChatGPT 与〈人工智能法案〉》（2023 年）12(1)《互联网政策评论》<https://doi.org/10.14763/2023.1.1682>，文中指出，《人工智能法案》的基于风险的方法是针对具有预期目的的系统制定的，因此并不适合生成式人工智能。

在这些条件下，系统依赖于它们。其他建议则涉及世界模型学习所依据的数据，尤其是其中的动作是如何被选择和识别的（第6.1节）。

2. 什么是世界模型

认知科学家、强化学习研究人员、视频生成系统的设计者，以及销售三维内容工具的公司的市场部门，都使用过“世界模型”这个术语。不同的人使用这个术语时，所指的并不完全相同。让我们试着缩小讨论范围，以聚焦于我们感兴趣的技术。

2.1. 起源

心理学家 Kenneth Craik 于1943年首次提出这样一个观点：现实中的实体（例如生物）会在头脑中携带一份外部现实的“微型模型”，并在行动之前用它来尝试各种备选方案。[2] 此后，这一观点及其类似变体被应用于不同领域；在一些更规范或语言学的研究领域中，甚至也能发现与之相似之处（例如，在认知语言学中，“单身汉”这一概念预设了一种理想化的婚姻模型）。这一概念后来进入早期人工智能领域，其主要用途是为机器提供对周围环境的显式手工编码模型，但这种做法只适用于范围狭窄、由规则支配的微观世界。[3]

世界模型这一术语更现代的含义，也就是我在这里具体讨论的含义，可以追溯到2018年，当时 David Ha 和 Jürgen Schmidhuber 发表了一篇题为《World Models》的短文。他们采用了启发式方法文献中相当熟悉的棒球击球手这一案例：击球手必须在几毫秒内决定如何挥棒，这比视觉信号抵达大脑所需的时间还短。他们强调，击球之所以成为可能，是因为击球手的大脑会预测球将出现在哪里，而不是进行精确计算（不过，归根结底仍然依赖一种启发式方法）。在此基础上，他们开发了一个系统，将智能体所看到的内容压缩成一种小型内部表征，并学习预测这一小型表征将如何响应智能体的动作而演变；随后，他们完全在自身的环境中训练智能体——或者说，在他们称之为环境的“梦境”中训练——然后再将结果迁移回真实比赛。[4]

因此，我们可以尝试先对世界模型给出一个近似定义——在第3节中，我会尝试给出一个更具操作性的定义——即：一种基于对环境如何呈现及如何变化的观测（例如视频、图像、声音、传感器读数）进行训练的系统；在可能的情况下，还会基于行动记录及这些行动所产生的后果进行训练。这样一来，世界模型就提供了一种工具，可以询问特定环境中将会发生什么，以及在大多数情况下，如果在那里采取某种行动，将会发生什么。最终，它们还可以向前运行，生成一些从未被观测到的轨迹，因为这些轨迹太罕见、不可行、昂贵或危险，难以进行探索。

在这一工作性描述中，我们可以区分出两个特征；在探讨那些使世界模型格外引人关注的治理、伦理及法律相关方面时，我们还会回到这两个特征：

> **1) 大多数世界模型都是动作条件化的。**“接下来通常会发生什么？”是一个简单的预测，而“如果采取了这个动作，接下来会发生什么？”则是一种复杂的预测，

[2]KJW Craik，《解释的本质》（剑桥大学出版社，1943年）第5章。[3]关于从 Craik 经由早期人工智能发展至当代系统的脉络，参见斯坦福以人为本人工智能研究院（HAI），《世界模型与空间智能时代：治理超越语言的人工智能》（议题简报，2026年），第2–3页。[4]D Ha 和 J Schmidhuber，《世界模型》（arXiv 预印本 1803.10122，2018年），后以《循环世界模型促进策略演化》为题发表于《神经信息处理系统进展》第31卷（2018年）。

这种预测需要对因果关系（以及反事实性）有更深入的表征，这正是世界模型开发者所承诺的。[5]

2）世界模型可以替代其所建模的环境。软件和硬件都可以在世界模型中接受训练和测试 （正如 Ha 和 Schmidhuber 的智能体在自己的梦境中学习一样），因为模拟环境充当了现实的代理，在其中可以反复演练决策。这是一种新形式的模拟，基于习得的规律（即从视觉数据中推断出的规律），而非显式规则。 世界模型所遵循的规则隐含在神经网络的权重中。正如我将在接下来的章节中论证的那样，这是世界模型所带来的最具独特性的治理和监管挑战之一。

2.2. 世界模型带来了语言模型所不具备的什么

现行监管框架是在考虑传统人工智能（或大语言模型）的背景下起草的，而对本文的第一个质疑是：世界模型不过是另一种模态中的生成模型。更有力的质疑则认为，语言模型已经拥有世界模型；例如，一个经过训练、用于预测关于世界的文本的系统可能会恢复世界的结构；研究人员还发现，仅使用奥赛罗棋局记录训练的模型内部存在棋盘的表征，以及大语言模型内部存在地理位置和历史时间的表征。[6] 如果确实如此，那么二者的区别只是程度上的差异，而一套能够涵盖语言模型的监管制度也会一并涵盖世界模型。

这一反对意见缺乏说服力，也有许多理由促使我们认为，世界模型与语言模型在性质上有所不同。首先，具体而言，针对这一反对意见较强的版本，认知科学的证据表明，人类世界模型并非源自语言。我们知道，婴儿在开口说话之前很久，就已经能将物体表征为坚实且在运动中持续存在的东西（正如第2.4节所述，持续性如今也是评估人工世界模型的标准）；而这种认知贯穿我们的一生，以至于成年人的一种常见体验是，将物理场景处理成近似的内部模拟，而非语言推理。这是因为大脑的语言网络在功能上有别于支持世界知识的系统；人们因失语症而丧失语言能力后，仍能对物体（或数字）进行推理，这一事实便证明了这一点。由此可见，掌握语言形式——语言模型显然具备这一能力——与理解运用语言所需的世界知识是可以分离的；后者通过感知和动作习得，必须由文本之外的信息提供。[7]

---

[5]斯坦福大学以人为本人工智能研究院（注3）3：“世界模型的一个核心目标是反事实推理：预测某个动作的后果，包括模型从未接受过明确训练的动作。” [6]K Li 等，《涌现的世界表征》（国际学习表征会议，2023年）；W Gurnee 和 M Tegmark，《语言模型表征空间与时间》（国际学习表征会议，2024年）。[7]ES Spelke 和 KD Kinzler，《核心知识》（2007年）10《发展科学》89；PW Battaglia、JB Hamrick 和 JB Tenenbaum，《模拟作为理解物理场景的引擎》（2013年）110《美国国家科学院院刊》18327；E Fedorenko、ST Piantadosi 和 EAF Gibson，《语言主要是交流工具，而非思维工具》（2024年）630《自然》575；K Mahowald 等，《在大语言模型中区分语言与思维》（2024年）28《认知科学趋势》517；S Harnad，《符号落地问题》（1990年）42《物理学D辑》335；LW Barsalou，《具身认知》（2008年）59《心理学年度评论》617。相反，M Abdou 等，《语言模型能否在没有具身基础的情况下编码感知结构？》（计算自然语言学习会议，2021）；E Pavlick，《大语言模型中的符号与落地》（2023年）381《英国皇家学会哲学汇刊A辑》20220041。该论证并非表明语言模型不具备任何世界结构——这是一个经验问题，而是说明它们所携带的世界结构为何会是不完整的；其基础是欠决定性，而非反表征主义，并且不影响第2.1节中“世界模型”这一含义。

这并不意味着语言模型对世界无话可说，也不意味着它在某种程度上不能成为对世界某些部分的模型，但我们需要谨慎看待这类模型能够描绘什么。这一观点在哲学上的先驱，在某种程度上是维特根斯坦晚年对其《逻辑哲学论》中图像理论的否定；该理论认为，命题的结构映照着事实的结构。维特根斯坦晚年主张，词语的内容来自共同实践中的使用，而非对世界的映照。[8] 因此，我们有充分理由得出结论：语言模型是对世界如何被描述的模型，而描述无法充分决定其所描述的事物。我们应当牢记，这种欠决定性同样存在于仅用视频训练的模型中，无论这些模型有何用途。将世界模型与渲染器区分开来的，是动作数据，而不是更多的视频素材。

对于这种“较弱”形式的反对意见，一个能体现语言模型与世界模型差异的相关方面在于评估标准：语言模型预测下一个词元，而世界模型预测环境的下一个状态，通常是在给定一个动作的情况下，[9] 评判它的标准不是输出看起来是否正确，而是球是否落在它所说的位置。

将世界模型视为另一类事物的另一个关键原因是，人们期望世界模型能够随着时间推移，维持一个连贯的环境状态，而不仅仅是针对给定提示词生成一串词元；并且，视野之外的事物仍然持续存在。响应式且持续存在的环境表征是通过观察习得的。它不像电子游戏那样预先构建：在电子游戏中，每个物体都由设计者事先放置。[10] 正如斯坦福大学以人为本人工智能研究院的文件所强调的，政策制定者必须思考“[……] 一个学习环境是否足够贴近现实，足以用来训练系统、测试系统，或指导现实世界决策”。[11] 这种差异也反映在语言模型的治理中，因为这里并没有真正与之对应的事物：我们依据现实评估语言模型，而世界模型可能充当现实的替身，大多数复杂问题也由此产生。

最后，也是最显而易见的一点，是训练数据。语言模型以文本进行训练，这些文本大多从互联网上抓取而来，随后再进行微调；而世界模型则以视频、图像、三维扫描数据、传感器流，以及至关重要的行动及其结果记录进行训练。后者根本无法通过抓取获得，只能通过在真实地点操作机器来收集。[12] 这也引发了法律层面的担忧，例如数据保护概况问题（例如，持续采集真实环境、人脸、车牌、室内空间和运动模式）。另一个担忧涉及世界模型所使用的最重要的数据，有时称为动作数据。由于这类数据稀缺，开发者往往通过先用完全不包含动作的录像对模型进行预训练来弥补，之后才对模型进行调整。例如，V-JEPA 2 先用超过一百万小时的互联网视频进行预训练，之后再用不到 62 小时的机器人交互数据进行调整（第2.4节）。有时，开发者也会让模型从视频片段本身推断动作（正如 Genie 从没有记录任何输入的游戏视频中学习如何响应玩家输入）。然而，正如我将在第6.1节中论证的，这会影响我们在多大程度上能够相信世界模型对动作后续结果的预测。

---

[8] L Wittgenstein，《逻辑哲学论》（1921）2.1–2.19、4.01；L Wittgenstein，《哲学研究》（1953）§§ 23、43；关于计算语言学中的形式—意义之分，参见 EM Bender 和 A Koller，“迈向自然语言理解：数据时代的意义、形式与理解”（计算语言学协会年会，2020年）。[9] 斯坦福大学以人为本人工智能研究院（注3）第3页；Pierson Ferdinand LLP，“超越大语言模型：为何人工智能世界模型需要专属法律手册”（客户提示，2026年10月29日）。[10] 斯坦福大学以人为本人工智能研究院（注3）第3页。[11] 同上，第2页。[12] Pierson Ferdinand（注10）“训练差异”；斯坦福大学以人为本人工智能研究院（注3）第12页，并指出，“稀缺的带动作标签的交互数据无法简单地从互联网上抓取”，这“构成了极高的准入门槛”。

2.3. 世界模型与其他模拟器

世界模型从数据中学习环境的预测性表征，并可利用这些表征模拟环境可能如何演变。从这个意义上说，它们可以充当学习型模拟器，也可以作为更大型模拟或决策系统的部件。[13] 因此，世界模型属于计算机模拟这一领域。半个多世纪以来，该领域一直用于设计飞机、预测天气、测试药物、规划城市和重建事故；其认识论、伦理和证据地位也早已得到深入探讨。[14] 在这一领域中，我们应将世界模型与相邻方法进行比较——也就是显式编程模拟器和物理引擎、数字模型、影子模型与数字孪生，以及合成人口和基于智能体的模型。

第一项比较涉及动力学的规定。在显式编程的模拟器中，建模者提供算法和/或行为规则，以确定某个模拟状态或一组状态如何变化。如前文已多次提及，世界模型则是从数据中学习预测映射。开发者仍会选择训练目标和一些架构约束，因此建模假设是世界模型的核心特征。

第二项比较涉及模型与已识别目标系统之间的关系。按照一种常用的分类法，数字模型与其物理对应体之间没有自动化数据交换；数字影子会接收来自物理对应体的自动化更新；而在连接程度最高的情况下，数字孪生则具备双向集成。[15] 数字孪生可以纳入学习型预测部件，世界模型也可以针对特定环境进行训练，或在数字孪生中使用。相比之下，世界模型不必与任何特定物理系统保持实时连接，通常是因为没有可供连接的物理系统。这两种安排都无法解决预测有效性的问题：接收运行数据本身并不能验证数字孪生，而缺少持续连接的对应体也不意味着世界模型无法测试。因此，例如，研究人员可以（而且通常建议）将世界模型的预测与未用于训练它的观测进行比较，以此测试该模型。在可行的情况下，建议测试在模拟中学到的策略能否在模型所代表的环境中奏效：例如，在学习型仿真中接受游戏训练的智能体，之后可以在原版游戏中接受测试。[16]

第三个也是最后一个比较，涉及个体及其互动的表征方式。已知的模拟技术各有用途，而世界模型则表征着不同的内容，同时又以某种方式将这些用途汇集起来。以合成人口为例：它通常会说明表征了哪些对象，例如人工个体或家庭，其属性能够再现特定目标所需的部分人口统计数据。但它本身并不描述这些对象会做什么；描述它们行为的任务则由

---

[13]Ha 和 Schmidhuber（注5）；D Hafner 等人，《梦境控制：通过潜在想象学习行为》（国际学习表征会议，2020年）。[14]E Winsberg，《计算机模拟时代的科学》（芝加哥大学出版社，2010年）；WL Oberkampf 和 CJ Roy，《科学计算中的验证与确认》（剑桥大学出版社，2010年）；关于建模与模拟的伦理，FL Shults、WJ Wildman 和 V Dignum，《计算机建模与模拟的伦理》，载于《2018年冬季模拟会议论文集》（2018年）第4069页；更早些时候，JP Sullins，《伦理与人工生命：从建模到道德主体》（2005年）7 Ethics and Information Technology 139，论及模拟是“一种实现复杂思想实验的强大媒介”。[15]W Kritzinger 等人，《制造业数字孪生：分类文献综述与分类》（2018年）51(11) IFAC-PapersOnLine 1016，数字孪生的分类即源于此；A Fuller 等人，《数字孪生：使能技术、挑战与开放研究》（2020年）8 IEEE Access 108952。[16]Ha 和 Schmidhuber（注5）；Hafner 等人（注14）。

模拟。[17] 基于智能体的模型则会分别表征各个智能体，并模拟它们如何随时间采取动作和相互作用，同时为它们设定行为。这些行为可以是预先编程的、通过学习获得的，或兼具两者。一项著名且颇有意思的实验由 Park 等人开展：生成式智能体将语言模型与明确设计的记忆和规划机制相结合，由此产生个体动作（及对话），并涌现出集体行为。[18] 世界模型并不会真正将每个个体表征为独立的智能体，因为它学习的是环境如何变化，通常是响应动作而变化，并以此为基础进行预测。

视频生成世界模型说明，环境可以在不为其中描绘的每个实体单独指定智能体模型的情况下进行表示。例如，Genie 可以生成可动作控制的视觉环境。[19] 不过，视频只是世界模型可能的输出之一。Dreamer 利用紧凑潜在状态空间中的想象轨迹来学习行为。[20] 无论如何，当生成场景描绘人物时，不应将其视觉上的可信度误认为是模型能够预测人们对某项政策或其他干预会如何响应的证据。此类主张需要通过观测以及适合所预测行为的衡量指标来验证。

2.4. 应用

世界模型最先进的应用领域是机器人学，主要有两种形式。第一种形式中，机器人在模型内规划：它会在移动之前预测几种候选动作的结果，并执行预测结果最符合目标的那一种。已发表的最典型案例是 Meta 的 V-JEPA 2（2025年）。它源自 Yann LeCun 在 2022年提出的设想：模型应预测场景的抽象表示，而不是逐像素进行预测。[21] 值得注意的是，V-JEPA 2 先在超过一百万小时的互联网视频上进行预训练，之后仅凭不到 62 小时的机器人交互数据，就能在两个从未见过的实验室中，让机械臂执行规划好的抓取与放置任务。整个过程无需特定任务训练或奖励。[22] 第二种形式中，世界模型生成训练数据，也就是说，机器人通过示例学习动作，例如如何抓取或行走；但由于在现实世界中收集示例既缓慢又昂贵，世界模型便会根据单个记录示例生成数千种合理的变化（例如光照和周围环境）。这大致就是 NVIDIA Cosmos 平台的用途。该平台于 2025年1月以开放模型形式发布，此后不断扩展。NVIDIA 表示，1X、Agility Robotics、Figure AI 和 Skild AI 都在使用它来训练人形机器人和通用机器人。[23] 这些描述来自 NVIDIA；这些公司都没有公布结果。在这种用途以及下文所述的测试用途中，世界模型生成的内容被称为

---

[17]C Dwork、K Greenewald 和 M Raghavan，《基于多维多重集求和的合成普查数据生成》（arXiv 预印本 2404.10095，2024年）。[18]JS Park 等人，《生成式智能体：人类行为的交互式拟像》，载于第 *36* 届 *ACM* 用户界面软件与技术年度研讨会论文集（2023年）。[19]J Bruce 等人，《Genie：生成式交互环境》，载于第 *41* 届国际机器学习会议论文集（2024年）。[20]Hafner 等人（注 14）。[21]Y LeCun，《通向自主机器智能之路》（版本 0.9.2，OpenReview，2022年6月27日）。[22]M Assran 等人，《V-JEPA 2：自监督视频模型实现理解、预测与规划》（arXiv 预印本 2506.09985，2025年）。这仍是实验室研究：LeCun 的公司 AMI Labs 于2026年3月融资约10亿美元以推进这一方法，但截至2026年11月尚未发布任何产品（公司于2026年3月10日发布公告，据媒体报道）。[23]NVIDIA 关于 Cosmos 世界基础模型平台的新闻稿，发布于2025年1月6日和2025年3月18日。

至少在法律学术领域，这类数据被称为合成数据，即“人工数据，由计算机模拟或算法生成，具有分析价值”。[24]

与机器人学相关的是自动驾驶。在这一领域，世界模型更像一个测试平台：驾驶软件的评判标准是它应对罕见事件的能力，而实际驾驶中这类事件太少，无法用于测试，因此模型会生成这些事件。2026年2月，Waymo宣布推出 Waymo World Model。该模型基于 Google DeepMind 的 Genie 3 构建，并针对驾驶场景进行了适配。[25] 给定文本提示或一段录制的行车视频后，该模型会生成 Waymo 车辆在相应场景中所能感知到的摄像头图像和激光雷达读数，从而让驾驶软件能够在其中运行，仿佛置身道路之上。Waymo表示，该模型旨在覆盖车队从未遇到过的罕见情形；这是世界模型最成功的用途之一，例如龙卷风、洪水、动物出现在车道上以及近碰撞等。其依据是，Genie 3 在海量通用视频数据上进行的预训练，使其能够推演公司自身记录中未曾出现的事件。根据公开资料，这是学习环境首次被用于以商业规模测试安全关键系统的应用之一。

第三种有文献记载的应用是交互式内容：世界模型生成一个可供人游玩的环境，或一个可供人编辑的三维世界，例如用于游戏、电影和设计。在这里讨论的几种应用中，这一类与普通的人工智能视频生成最为接近，OpenAI 的 Sora 是最广为人知的例子。Google DeepMind 的 Genie模型最初通过游戏画面进行训练，之后又利用涵盖更广泛主题的视频进行训练，从而生成能够响应玩家动作的环境。借助 Genie 3（2025年8月），玩家可以在模型实时生成的世界中移动；如果玩家想要改变场景中的某些内容，也可以在模型运行时通过输入文本来实现。场景大体上能够连贯维持几分钟。2026年1月，Google DeepMind 向订阅用户开放了该模型，并将其命名为 Project Genie；2026年10月，该公司发布了 Genie 4，它还能生成声音。[26] 这类编辑用途的一个例子是 Marble，由李飞飞创立的公司 World Labs 于2025年11月发布。向 Marble 输入文本、图像或视频后，它会生成一个三维场景，供用户进一步编辑。[27]

还有其他应用，但大多仍处于较早阶段。例如，有人尝试将语言模型智能体与世界模型或类似世界模型的规划模块结合起来，让智能体在采取某个动作之前，先演练其后果。[28] 在其他领域，例如工程学和物理科学，人们提出利用世界模型在设计建成之前对其进行测试，并评估基础设施在压力下的表现；我们已经在数字孪生或数字影子的应用中见过这种方法。此外，还有一类相关的世界模型——它们仅在第 3 节所定义的广义意义上才算世界模型，因为它们更像结构化科学模拟器，而非以动作为条件的模型[29] ——已能根据数据而非大气物理学方程预测天气。[30]

---

[24]Michal S Gal 和 Orla Lynskey，《合成数据：数据生成革命的法律影响》（2024年）109《爱荷华法律评论》1087；此处引用的定义出自 I Cofone、KJ Strandburg 和 N Tilmes，《法律中的合成数据分类》（2025年）110《爱荷华法律评论在线版》217。[25]Waymo，《Waymo世界模型》（公司博客，2026年2月6日）。[26]Google DeepMind，《Genie 3：世界模型的新前沿》（2025年8月5日）；Project Genie，2026年1月29日发布的公告；Genie 4，于2026年10月27日发布的研究预览版。[27]World Labs，Marble发布公告（2025年11月）。[28]M Parmar，《世界模型中的安全性、安全保障与认知风险》（arXiv预印本2604.01346，2026年）第3.3节。[29] Physical Intelligence Team、上海人工智能实验室，《世界模型的定义与路线图》（arXiv预印本2607.06401，2026年）第2.6节，其中将学习型天气模型描述为一种结构化科学模拟器，而非智能体世界模型。[30]R Lam 等，《学习高技能中期全球天气预报》（2023年）382《科学》1416。

另有一项提案涉及证据，例如道路事故的闭路电视录像。Kirfel 及其同事介绍了他们所称的“反事实世界模拟模型”，该模型首先利用这类多模态证据重建事发经过，然后展示如果驾驶员更早刹车，或行人抬头看路，事态会如何发展。作者设想调查人员和法院都可使用该模型，并为这两种用途制定了伦理框架。[31]

社会模拟领域也有相关应用，这大概是此处所述用途里发展最不成熟的一种；国防领域也有应用。斯坦福大学以人为本人工智能研究院的简报列举的例子包括：在合成环境中演练任务、在通信受损时管理后勤，以及为情报用途分析级联效应。[32]

综合以上，我们应如何判断这项技术的成熟度？世界模型已经作为产品出售，其中一些采用开放权重；正如机器人学和自动驾驶的例子所示，它们也已被用作训练和测试安全关键系统的环境。至于它们有多可靠，则更难判断。斯坦福大学以人为本人工智能研究院的简报，其作者之一经营着一家领先的公司；简报相当坦率地指出：“能够预测物理后果，并在多种真实环境中支持可靠规划与行动的通用世界模型，仍处于积极的研究前沿。”早期模型很快就会无法跟踪场景，例如物体会闪烁或消失；衡量模型究竟改进了多少，至今仍要看智能体能否移动一个物体、离开场景，再回来发现它仍在原处。[33]

从法律与治理视角来看，这些系统的名称与其实际能力之间的差距本身就值得分析，没有多少理由要等到这一差距消失。原因之一在于，技术成熟度是灵活且具有时效性的法规所采用的一项标准。还有一个理由是应当持续关注这一差距：系统已经被描述、销售并被依赖为能够呈现世界复杂性的世界模型，而有关它们已经学会世界运行规则的证据仍然是不完整的。

3. 世界模型的定义与分类法

综合第2节所确定的特征，世界模型可以作如下操作性定义：

> 1.“世界模型”是指满足以下两个条件的计算系统，或系统中可识别的组件：
>
> (a)它全部或主要通过数据学习了环境如何随时间变化，包括其中的实体如何移动、互动和作出反应；
>
> (b) 其设计或改造目的是利用所学内容，根据有关环境的一种或多种实际状态或假设状态的信息，预测随后可能出现的一种或多种状态。

[31]L Kirfel 等，《当人工智能遇上反事实：反事实世界模拟模型的伦理影响》（2025年）5 AI and Ethics 4593。[32]斯坦福大学以人为本人工智能研究院（注3）第10–12页。[33]同上，第3页。

许多世界模型的特点是，它们展示了环境如何响应刺激而发生变化，这就需要第二个定义：

> 2. “动作条件世界模型”是指能够将动作作为输入，并预测采取该动作后环境将如何变化的世界模型，因此不同的动作可能导致不同的预测。

在该定义中，“环境”是指物理或虚拟场景，包括其中的实体，其状态可能随时间发生变化。“状态”是指环境在特定时间的状况的任何表征，包括图像或视频帧等传感器观测，或系统已学习并用来替代传感器观测的内部表征。“动作”是指任何智能体对环境实施的实际或假设性干预，无论其以何种形式表示，包括控制信号、对动作的学习表征，或描述该干预的指令。最后，“预测”可以是单一结果、多个可能结果或概率分布，也可以仅在系统内部使用，包括用于比较或选择动作。

该定义的目标之一，是无论世界模型的架构或学习技术、其内部（潜在）表征的形式，或其输入或输出的媒介为何，均适用于该模型。无论其独立运行还是作为另一系统的一部分运行，该定义同样适用，后者是最常见的情况。关于其排除标准，如果系统用于预测环境变化的手段完全由人员指定的方程和参数值决定，则该系统不属于该定义的范围。[34]

3.1. 分类法

尽管这项技术相对较新，相关文献已经提出了几种分类法，其中一些似乎已经适合在监管机构和政策制定者之间传播。例如，李飞飞根据世界模型所产生的结果，将其分为渲染器、模拟器和规划器。

- 一个 渲染器生成观测结果，通常是图片或视频，供人眼观看，因此视觉上的逼真性是最重要的质量标准。为了说明渲染输出与“普通”人工智能生成内容相比的独特特征，李飞飞举的例子是无人机飞越城市的视频：也就是说，从上方看可能完美无瑕，但如果你试图驾驶穿过其中，建筑物就会分崩离析。[36]

- 一个模拟器会生成状态，例如场景几何结构和动力学的表示，以便程序和专业人员对其进行计算。在这种情况下，

---

[34] 如果人为指定的规则或方程完全规定了环境如何变化，而数据（如有贡献）唯一的作用只是估计那些由规则或方程预先指定、作为建模环境属性的量，则不满足第1(a)段的要求。学习系统的架构或数学运算由人类规定这一点本身，不足以使该系统落入这一排除范围。[35]X Li 等人，《具身人工智能世界模型综合综述》（arXiv 预印本 2510.16732，2025年），按照功能（决策耦合与通用）、时间建模和空间表征提出了一种分类法；J Ding 等人，《理解世界还是预测未来？世界模型综合综述》（arXiv 预印本 2411.14499，2024年），将该领域划分为构建内部表征以理解世界的模型，以及预测世界未来状态的模型。[36]F-F Li 与 World Labs 团队，《世界模型的功能分类法：渲染器、模拟器、规划器及连接它们的循环》（World Labs，2026年6月3日）。无人机示例出自李飞飞本人。

最相关的质量在于场景是否能够正确地“展开”，因此，例如，物体能够持续存在，事物也会按照应有的方式下落。为了说明模拟器与渲染器的不同用途，可以想象一下工程师使用的碰撞测试模型，而不是供观众观看的模型。•

规划器生成动作：根据它观察到的情况和目标，指示智能体接下来应该做什么；而衡量其质量的关键在于动作是否实现了目标。

这种分类法提供了一个初步的思路，但没有考虑这些系统内部是如何进行计算的。举例来说，使用这种分类法，就好比因为专家系统和神经网络都用于图像识别，便将它们归为同一类别。这也是促使人们提出其他分类法的原因之一，例如上海人工智能实验室 Physical Intelligence Team 提出的分类法，即双维分类法。在这种替代方法中，Li 提出的三种功能保留在一个轴上，同时增设第二个架构轴，以展示模型对世界的认知在模型内部是如何表示的。[37] 下图展示了这两个轴如何结合：

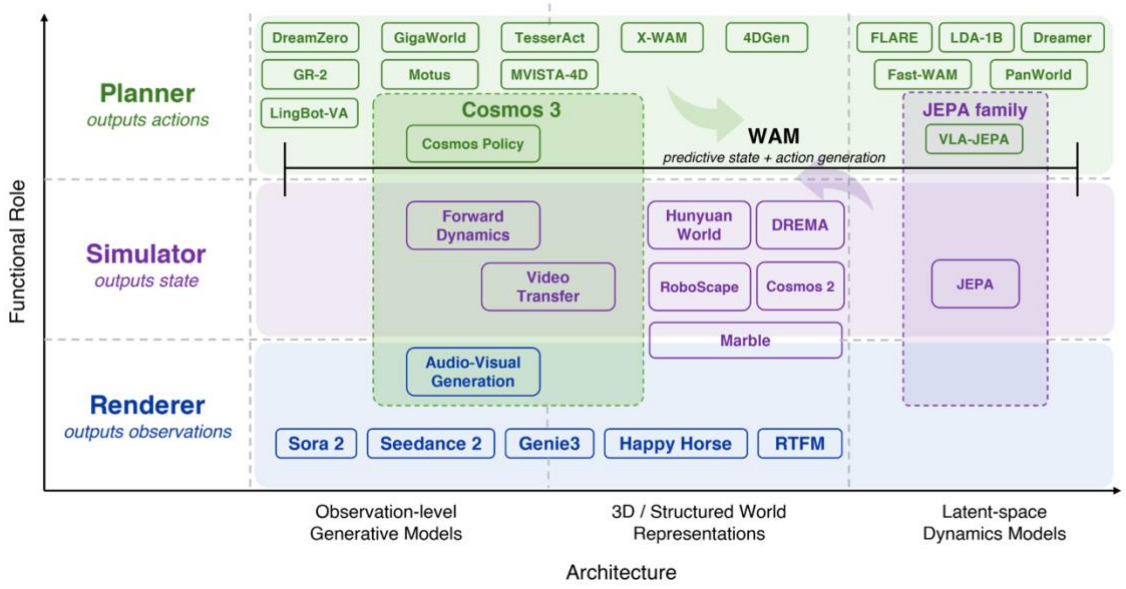


图1. 双维分类法。来源：上海人工智能实验室 Physical Intelligence Team，《世界模型的定义与路线图》（2026年），图5。

这两个轴构成了一个很好的起点，因为它们共同涵盖了一些与法律息息相关的技术特征：例如，功能决定了输出的使用者，而使用者反过来又决定或影响法律制度。简单来说，渲染器很可能会被视为一种内容技术，这意味着其输出会引发法律针对合成媒体提出的那些问题，例如其来源，以及对真实人物与地点的再现。相比之下，模拟器可以被视为现实（某些部分）的代理，因此其输出会引发许多与法律相关的问题，包括所生成证据的性质：例如，我们依据哪些标准得出结论，认定该代理足以忠实呈现据此作出判断的对象？甚至还可以问：有哪些保障措施能够避免将模拟输出直接带入现实，仿佛它是一次现实世界实验？规划器，

[37]上海人工智能实验室（注32）第2.4–2.5节：“因此，第一种分类法询问系统做什么，而第二种则询问相关世界知识是如何表示和计算的。二者是互补关系，而非相互竞争；通常应沿着这两个维度为模型分配一个或多个标签。”

最后，规划器可以被视为一种控制技术，使其进入由能够采取行动的机器所引发的繁杂法律问题领域，即产品安全以及对人身伤害的责任。[38] 分类法的重要性并不只在于涵盖各种功能，因为世界模型的架构也决定了哪些内容可以披露和检查（像素或 3D 场景可以查看，而潜在状态则无法查看）。换句话说，这些都是监管机构、审计员或认证机构必须审视的方面。

但我认为，从治理的角度来看，应当考虑一种不同的分类法。这种分类法受社会技术文献启发，提出少数几种类型，然后利用功能和架构来说明监管机构应当在每种类型中预期发现什么，以及应当进行哪些测试。从这个意义上说，分类问题是：谁或什么在对模型的输出采取行动？我们可以通过区分世界模型在应用中可能发挥的四种角色来构建分类法：控制、咨询、代理和内容。每种类型，即世界模型在应用中扮演的每种角色，都具有典型的功能和架构，但这些角色并非相互排斥。角色可能会重叠；当这种情况发生时，就应当将相关治理问题一并考虑。[39] 下面的图展示了我所提出的分类法：

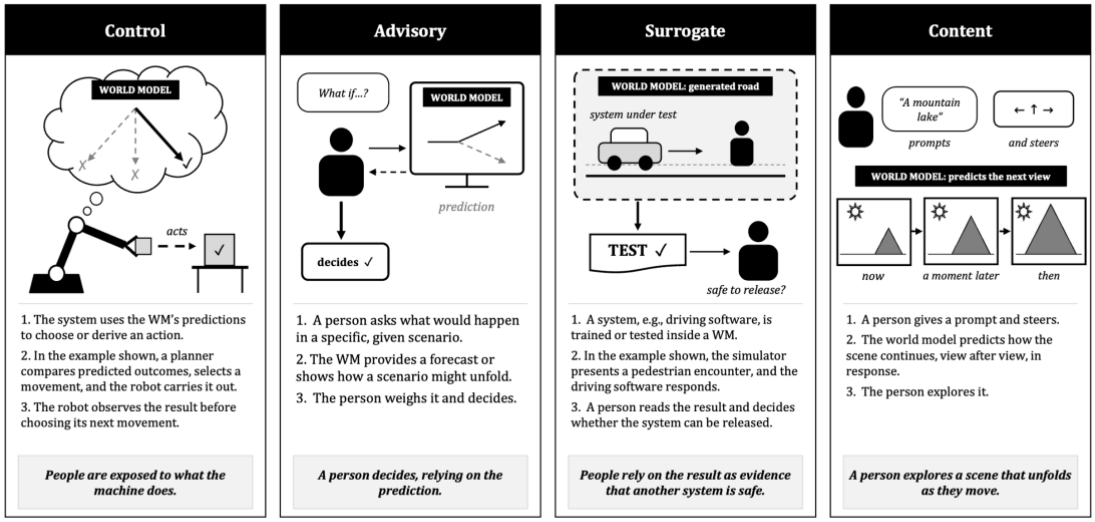


图*2*。基于角色的分类体系

1. **控制模型**。世界模型预测可能的（行动）方案所带来的后果，以支持在物理或虚拟环境中的自动化行动选择。例如，机器人的规划器利用预测轨迹来选择放置杯子时的移动方式，通常会选择成功率最高的动作，将杯子放到架子上。控制模型的架构可以是潜在空间架构，这意味着模型预测的是紧凑的内部状态，而不必生成图像；也可以是观测层级架构，即模型预测的内容就是机器人执行该动作时将会看到的内容。[40]

2. **咨询模型**。 这些模型 为决策者生成预测或演练，决策者可以是法官、操作员、政策制定者，甚至是软件代理的用户。根据前述分类，它们的

[38]Parmar（注29）第3.2节（范围说明）讨论了决策耦合模型与通用模型的不同风险特征，第8.4节讨论了自动化偏见。[39]一个密切相关的分类学先例，是 Bi等人《世界模型：人工智能中的预测性状态、生成式模拟与决策制定》（SSRN，2026年）中基于世界模型下游用途所作的分类。[40] Parmar（注29）第3.2节讨论了决策耦合模型与通用模型的不同风险特征，表3讨论了“铸造厂问题”，第8.4节讨论了自动化偏见；斯坦福大学以人为本人工智能研究院（注3）第12页讨论了有利于现有车队运营者的数据壁垒。

功能 将被视为模拟，有时也与规划结合。它们的架构十分灵活，因为可以属于三者中的任何一种。尽管大多仍处于实验阶段，但其中一些例子包括：为法院对道路事故进行反事实重建；在提出动作之前先进行演练的智能体；以及模拟人群将如何响应某项政策的模拟器。第2.4节提到的天气预报模型也属于此类，不过，尽管它们为人类决策提供信息，却不会预测某个动作的影响。[41]

3. **代理模型**。 它们 充当现实世界的替身，让另一个系统在其中进行观察、训练和测试。因此，其功能显然是模拟，而其架构属于观测层级或结构化（3D）。生成真实生活中的（或罕见的）情境，例如 Waymo世界模型所采用的驾驶场景，便属于这一类。

4. **内容模型**。 这些模型 生成供人观看、更重要的是供人主动探索的场景。因此，其功能是进行复杂的动态渲染，架构通常属于观测层级或结构化（3D）。面向使用者的一些内容模型包括 Project Genie 和 Marble；而像 Sora 这样的生成视频模型，只有在能够预测场景如何发展时，才属于这一类。

可以为这一分类法补充两点注意事项——或者说警示。一些世界模型可能采用开放权重，或授权给许多开发者使用，因此可充当基础（世界）模型。考虑到传播等风险，我们应对此更加谨慎。第二个问题是，模型是否锚定于真实地点或人物；若是，它们就更接近第2.3节讨论的数字影子和数字孪生，当然也会带来隐私和监控风险（相关文献包括 Dirk Helbing 和 Novelli 的研究）。还有一个相关问题，合成数据相关文献对此有所强调：一旦数据与其来源人物和地点脱钩，由于这些数据是人为生成而非采集所得，追溯同意和问责就会更加困难。[42]

无论如何，每种类型都有其风险特征，这会影响法律如何应对它；风险评估也需要仔细考量其特征（我们不能直接套用人工智能领域的方法），但探讨这一方面超出了本文的范围。下图总结了这些世界模型类型的一些最重要特征。

4. 世界模型的风险

在转向法律之前，先说明世界模型带来的若干最相关风险，并使用人工智能监管本身采用的术语，这很有帮助。《人工智能法案》将风险定义为“发生危害的可能性与该危害严重程度的组合”（第3(2)条）——这也是许多其他文件和风险科学文章中常见的定义——并要求高风险系统的提供者识别已知或可合理预见的风险（第9(2)(a)条）。[43] 就此而言，世界模型涉及的危害并不陌生，即人员伤害、财产损失以及错误决策。变化之处在于这些危害可能以何种方式发生，以及

[41]Kirfel 等人（注 33）关于“眼见为实”效应的讨论。[42]CD Whitney 和 J Norman，“虚假数据的真实风险：合成数据、多样性洗白与规避同意”，载于*2024*年*ACM*公平、问责与透明度会议论文集（2024年）。[43]Claudio Novelli 等人，“人工智能风险评估：针对《人工智能法案》的基于情景的比例方法”（2024年），3《数字社会》13。关于将风险规制作为人工智能治理方法的批评，包括它最适用于可量化问题、可能将政策决策呈现为技术决策等，参见 Margot E Kaminski，“规制人工智能风险”（2023年），103《波士顿大学法律评论》1347。

要预见这些危害有多困难。我们来试着列举三项相关风险：

1. 行动证据风险。我用这个标签来指称一种风险。据我所知，这一说法尚属新创：即使模型几乎没有或完全没有关于某个动作的可靠证据，人们仍依赖模型对该动作效果所作的预测。[44]2. 利用风险。[45] 这种风险是指机器或人依据模型高估了的预测采取行动。其根源在于预测的使用方式：每个预测都建立在前一个预测之上，因此错误会在一连串预测中不断累积，[46] 而被选中的动作是模型评估为最佳的动作；这往往也是它高估了的动作。[47] 3[48]. 虚假保证风险。这种风险是指：人们依据在某个世界模型内部进行的测试或训练，宣称某个系统安全，而该世界模型本身也存在同样的盲点，或根本没有学会相关情境。此处的危害虽属间接，却影响广泛，因为所有以这种方式获批的系统都会受到影响；如果同一个模型被提供给许多开发者，影响范围还会更广。[49]

5. 现行法律下的世界模型比较分析

So far, the only standard, appea 本文考察的世界模型文件，而只见于标准项目和技术出版物。一个案例是中国国家项目，名为“人工智能：世界模型评估标准”，于2026年5月28日在国家市场监督管理总局运营的国家标准平台上登记。[50] 该项目属于国家标准化指导性技术文件，旨在为仍在发展的领域提供技术指导，且不具强制性。另一个案例是IEEE P4216，这是一个旨在制定《人工智能世界模型评估标准》的项目。名称中的“P”表明这仍是一个项目，即IEEE标准协会已授权工作组起草该标准，但标准尚未正式出台。IEEE标准也是自愿采用的，由产业界、学术界及其他领域的专家通过协商一致制定。该项目涵盖多种输入类型的模型，例如图像、声音和传感器数据，并能生成或预测环境的变化。该项目旨在评估这些模型输出的一致性和真实性，以及这些输出对其他协同系统执行任务的支持程度。[51]

这两个项目都旨在建立评估世界模型的方法，这些方法或有助于评估模型的性能，但符合某项技术

---

[44] 数据表征或质量不佳当然并非数据科学中的新问题，但这里的问题不仅在于数据是否反映了使用场景，还包括数据是否呈现了动作所带来的后果。[45]Ha 和 Schmidhuber（注5）；T Kurutach 等人，《模型集成信赖域策略优化》（国际学习表征会议，2018年）；M Janner 等人，《何时信任模型：基于模型的策略优化》，载于神经信息处理系统进展第*32*卷（2019年）。[46]E Talvitie，《基于模型的强化学习自校正模型》，载于第三十一届*AAAI*人工智能会议论文集（2017年）。[47]A Kamat 等人，《学习世界：物理人工智能世界模型批判性综述》（预印本，2026年）。[48]F Liu 等人，《基于世界模型的具身人工智能安全保障：威胁、防御与评估的生命周期》（arXiv预印本 2607.28226，2026年）。另见 X Wang 等人，《迈向安全世界模型：风险、防御与评估》（arXiv预印本 2610.13372，2026年）。[49]Parmar（注29），表3。[50]国家标准信息公共服务平台（国家市场监督管理总局），项目 20263035-Z-469，《人工智能 世界模型评测规范》。[51]IEEE标准协会，P4216，《人工智能世界模型评估标准》，项目记录。另一项提及世界模型的IEEE项目P4198《工业人工智能中世界模型的应用架构指南》，见于IEEE SA标准委员会2026年9月25日的会议纪要草案。

标准并不自动意味着符合法律。其法律效力取决于适用的法律是否承认该标准，以及在何种条件下予以承认。[52]

例如，欧盟《人工智能法》[53]，就体现了此类认可：根据第40条第1款，符合协调标准的高风险人工智能系统或通用人工智能模型，若该标准的引用已在欧盟官方公报上公布，则推定其满足该标准所涵盖的法律要求或义务（但这并不意味着推定其符合整部《人工智能法》）。不过，协调标准由欧洲标准化组织应欧盟委员会的要求制定[54]。中国指导性文件的发布或IEEE标准的批准，本身都不会赋予其这一地位。只有在欧洲标准机构制定协调标准时借鉴了这些文件或标准，其内容才可能间接在欧盟产生这一效力。

本节其余部分将比较三个司法管辖区，并非因为它们的规则相似，而是因为它们将义务施加于不同对象：欧盟主要针对系统的使用和模型的通用性，加利福尼亚州针对最大规模模型的开发者，而中国针对所提供的服务；对于车辆，则针对产品。

5.1. 欧盟

5.1.1. 《人工智能法案》

《人工智能法案》（人工智能法案（AIA））（《条例（欧盟）2024/1689》，经《条例（欧盟）2026/1744》修订）对人工智能系统和通用人工智能模型进行监管。该法案对人工智能系统的定义有意保持宽泛，并借鉴了基本上技术中立的方法，涵盖在不同自主程度下运行的所有基于机器的系统，只要这些系统能够推断出输出（包括预测），并能够影响物理或虚拟环境（第3条第（1）款）。因此，基于世界模型构建的系统很可能符合该定义。出于同样原因，它也将属于经济合作与发展组织对人工智能系统的定义范围，该定义采用了几乎相同的措辞。[55] 至于模型本身属于受监管的系统、通用模型还是部件，则需要另行评估。要判断《人工智能法案》是否涵盖世界模型，第二个相关问题是其是否属于该法案的属地范围和事项范围；后者排除了某些研究与开发活动（第2条第（6）款和第（8）款）。第三个问题是相关义务何时适用，因为该法案的《规定》自不同日期起适用。[56]

然而，最重要的一点是，《人工智能法案》中的大多数义务取决于系统的用途。因此，预期目的有助于确定系统是否被归类为高风险系统（第3(12)条和第6条）；而其他规定则由系统的能力、用途和输出触发。这就是第3节所列的四种世界模型（如前所述，它们可能相互重叠）分别对应进入《人工智能法案》的四种可能路径的原因。

控制模型。这类世界模型的监管路径是产品安全。如果人工智能系统属于产品，或被设计为产品的安全部件，

---

52 Ronit Justo-Hanani，《欧盟〈人工智能法〉的风险导向方法：共同监管的益处与挑战》（2026年）9 Policy Design and Practice 171。[53]欧洲议会和欧盟理事会2024年6月13日关于制定人工智能协调规则的《条例（欧盟）2024/1689》（《人工智能法》）[2024] OJ L，2024/1689，经《条例（欧盟）2026/1744 [2026] OJ L，2026/1744》（《人工智能法案》）修订。除非另有说明，第5.1节引用的条款、序言条款和附件均为经修订的《人工智能法案》中的条款、序言条款和附件。[54] 欧洲议会和欧盟理事会2012年10月25日关于欧洲标准化的《欧盟条例（EU）第1025/2012号》[2012] OJ L316/12，第2(1)(c)条。[55]经济合作与发展组织，《关于人工智能的理事会建议》（OECD/LEGAL/0449，2019年5月22日通过，经修订）。[56]根据修订后的第113条，第三章第1至第3节通常分别自2027年12月2日起适用于附件III系统，并自2028年8月2日起适用于附件I系统。通用人工智能模型义务自2025年8月2日起适用；对于在该日期之前投放市场的模型，过渡期截至2027年8月2日（第111(3)条）。

且该产品受附件一所列法规管辖，并且需要进行第三方合格评定（第6条第1款），那么该系统就可能属于高风险系统。此类系统可能包含用于规划机器人运动的世界模型，尽管仅仅控制机械本身并不足以满足上述任一条件。接下来，一切取决于产品，因为附件一分为两份清单，两者的运作方式不同。第一份清单（A部分）列出了《人工智能法案》在其基础上增设自身要求的法律。如果人工智能系统是其中一部法律所涉产品的安全部件，《人工智能法案》针对高风险系统制定的规则便会直接约束该系统，同时还须遵守产品本身的一般安全规则。此时，提供者必须满足关于数据、人工监督、准确性和稳健性的要求（第10条、第14条至第16条）；部署者必须按照系统的指令使用系统，并指定具备相应能力的人员负责监督（第26条第1款至第2款）。欧盟委员会必须在2027年8月2日前通过授权法案，明确在行业法已提供同等或更高保护的情况下，可以限制适用哪些要求（第2条第13款）；截至本文撰写时，尚未通过此类法案。医疗器械相关法律列于此清单中（附件一A部分第11点），因此，这些义务适用于用于引导手术机器人的世界模型。第二份清单（B部分）则列出由各行业自行监管的法律，例如有关车辆和飞机的法律。对于受这些法律管辖的产品，《人工智能法案》的要求不直接适用（第2条第2款），只有在写入相关行业法后才会生效（第102条至第109条）。机械——涵盖大多数工业机器人和服务机器人的类别——于2026年从第一份清单移至第二份清单（附件一B部分第21点）。这意味着，利用世界模型规划运动的工厂机械臂或仓库机器人将遵循行业法路径，与下文讨论的车辆适用同一路径。

简言之，同一系统可能会受到两种不同的对待：也就是说，控制医疗器械的世界模型直接受《人工智能法案》要求的约束；而控制机器或车辆的世界模型——第2.4节所述的控制模型就属于这一类——只有通过其所属行业的法律才会受到约束，并且仅限于该法律已将《人工智能法案》的要求纳入其中的范围。因此，对于控制模型而言，首先要问的问题是其产品属于哪一清单。

咨询模型。这里的法律依据是《人工智能法案》所列的高风险用途（附件三）。对于这类世界模型，《人工智能法案》中最相关的是第8(a)点：旨在协助司法机关或代表司法机关开展事实和法律的调查与解释，并将法律适用于具体事实的系统，包括用于替代性争议解决的系统。例如，为法院重建事故场景而设计的系统可能属于这一范围，但须满足两个条件：这是其提供者所设定的用途，且第6(3)条规定的例外不适用。通用工具生成的图像，不会仅仅因为有人将其提交给法院，就因此具有这种用途。其他咨询用途可能属于其他条目，例如就业、信用度或获取公共福利。系统一旦被认定为高风险系统，就有两项保障措施可避免过度依赖。必须确保监督系统的人员能够意识到自动化偏见（第14条第4款(b)）。此外，受到基于该系统作出的决定影响的人，可以要求解释该系统所发挥的作用以及决定的主要要素，但此项权利有一定限制：它适用于除第2点以外的附件III系统，以及当事人认为对其健康、安全或基本权利造成损害、并产生法律效果或类似重大影响的决定（第86条第1至3款）。因此，《人工智能法案》规范了咨询模型的监督方式和解释要求。其模拟结果能否作为证据采纳，以及具有多大证明力，很可能仍由程序法和证据法决定。[57]

[57]关于从美国证据法的角度来看，在人工智能生成的证据获准采纳之前必须确立的有效性和可靠性，请参阅 Paul W Grimm、Maura R Grossman 和 Gordon V Cormack，《人工智能作为证据》（2021）19《西北大学科技与知识产权期刊》9。

代理模型。这类世界模型没有明确的专属监管路径。事实上，训练或测试另一系统本身并不构成《人工智能法案》第6条及附件一和附件三所规定的高风险用途。不过，《人工智能法案》仍可能通过三种间接方式涵盖代理模型。第一，同一模型也可能用于《人工智能法案》所监管的其他事项：它可能作为安全部件安装在产品中，可能旨在用于列明的高风险用途，也可能生成合成视频；按第50条第2款的字面规定，在后一种情况下须履行标记义务。代理模型也可能符合通用人工智能模型的定义。这取决于模型的通用程度，而非其是否用于训练或测试，因此，下文在讨论提供给众多开发者的世界模型时，将另行说明相关条件。第二，其输出可能受到针对使用这些输出进行训练或测试的高风险系统提供者所设义务的约束。[58] 这些义务可能要求审查合成数据和模拟测试。第三，《人工智能法案》还规定了高风险系统提供者依赖他人工具的情形。如果第三方提供人工智能系统、工具、服务或部件，且该系统、工具、服务或部件被“用于或集成到”某个高风险系统中，双方必须在书面协议中列明提供者为履行其在《人工智能法案》下的义务所需的信息、技术访问权限及其他协助（第25条第4款）。对于代理模型，这意味着，例如，一家公司从另一家公司获得世界模型许可，用于生成训练数据或测试场景时，可以要求对方披露该模型的构建和验证方式，以及已知局限性。但须注意若干限制：该规则仅涉及从第三方获取的工具，因此不涵盖公司自行构建并在内部使用的代理模型（不过《人工智能法案》仍可能适用，见上文）；此外，该规则不适用于依据自由和开源许可证向公众提供的工具，除非这些工具属于通用人工智能模型。简言之，《人工智能法案》没有专门针对代理模型本身的规则，但可通过上述三种方式对其进行规制：当模型还从事《人工智能法案》所监管的其他活动时，适用针对该模型提供者的义务；适用针对使用该模型进行训练或测试的系统提供者的义务；以及在模型由第三方提供时，适用双方必须签订的协议。不过，这三种方式都没有规定如何证明学习环境有效。

内容模型。这里的路径要经过《人工智能法案》的透明度规则，这些规则为两类不同主体规定了两项独立义务。第一项义务由提供者承担，也就是构建系统的公司。如果系统生成合成音频、图像、视频或文本，提供者必须确保输出带有机器可读标记，并且能够被检测为人工生成或操纵的内容（第50条第2款）。第二项义务由部署者承担，也就是使用该系统的个人或组织。如果内容属于深度伪造，部署者必须告知观众该内容是人工生成或操纵的（第50条第4款）。在人工智能法案中，深度伪造是指与现有人物、物体、地点、实体或事件相似，且看起来真实的图像、音频或视频（第3条第60款）。因此，这一定义也涵盖地点，这对世界模型而言十分重要：如果生成场景对真实街道或建筑的复现足够逼真，以至于看起来像是该地点的实拍画面，就符合这一界定。[59]

正如预期，广泛分发的世界模型也可能是通用人工智能模型，因此会产生一套适用于模型本身而非任何特定用途的独立义务。要达到这一点，该模型必须展现出显著的通用性，能够胜任广泛的不同任务，并融入各种下游系统

[58] 该提供者的训练、验证和测试数据必须与预期设置相关，具有充分代表性，并且在可能的情况下不存在错误且保持完整（第10条第（3）至（4）款）；其测试必须采用预先定义且适合预期目的的指标（第9条第（6）至（8）款）。[59] 执法领域存在例外；对于明显具有艺术性、创造性或虚构性的作品，则适用较宽松的规则，只需以不破坏作品的方式披露存在生成内容即可。

（第3条第63款）。相比之下，仅仅获得向许多开发者授权的许可，或以开放权重形式发布，本身并不能证明上述任何一点。

但根据《人工智能法案》，一个系统达到“通用”究竟意味着什么？答案在于规模和生成能力：参数至少达到十亿、通过自监督利用大量数据训练的模型，以及能够灵活生成文本、音频、图像或视频的大型生成模型（第98至99条序言）。欧盟委员会2025年7月发布的不具约束力指南将其转化为一项常规工作判定标准：训练计算量大于10^23 浮点运算次数，同时具备生成语言（以文本或音频形式），或根据文本生成图像或视频的能力。[60] 同一指南还以专门用于建模天气模式或物理系统的模型为例，说明其不属于该类别。对于世界模型而言，这会产生区分。能够根据文本提示生成视频的模型可能符合条件。然而，仅在潜在空间中进行预测，或仅根据动作和传感器数据进行预测的模型，则不太可能符合条件，无论其使用了多少计算量。

无论如何，如果一个世界模型符合条件，就会产生两层义务：提供者必须保留技术文档，并向下游提供者提供其理解该模型能力和局限性所需的信息（第53条第1款(a)至(b)项）；同时必须制定版权政策，并发布其训练内容摘要（第53条第1款(c)至(d)项）。根据自由和开源许可证发布的模型可免除第一组义务，除非其呈现系统性风险（第53条第2款）。第二层义务仅适用于所谓的“系统性风险”模型，即具有高影响能力的模型。如果训练计算量超过 10^25 浮点运算次数，则推定模型具备此类能力；提供者可以推翻这一推定，欧盟委员会也可以指定某一模型（第51至52条）。随后，提供者必须评估该模型，对其进行对抗性测试，评估并缓解系统性风险，报告严重事件，并确保模型安全（第55条第1款）。

总而言之，就这一点而言，《人工智能法案》具备涵盖世界模型的所有条件。其定义足够宽泛，可以将世界模型纳入其中；其义务则取决于系统的用途、输出内容以及模型的通用程度。控制模型通过产品安全规定进入适用范围；而对于机械和车辆（大多数控制模型都存在于这些领域），《人工智能法案》的要求只能通过行业法产生作用。咨询模型则通过高风险用途清单进入适用范围，该清单带来了有关监督和解释的规则，但将模拟的证据价值留给其他法律处理。代理模型没有专门的适用入口，只能根据模型的其他用途，以及训练或测试该模型的提供者而被纳入。内容模型主要通过透明度规则进入适用范围；这些规则要求输出具有可识别性。最后，通用人工智能制度以语言和由提示词生成的视频为着眼点，因此能够凭借世界模型的生成性特征将其纳入，但可能会遗漏仅预测物理结果的世界模型。

在上述所有情形中，世界模型最具特色的特征之一——一种用于取代真实环境的学习环境——只能通过有关代表性数据和充分测试的一般性要求得到涉及；而如何证明这种环境具有可信度，则是《人工智能法案》没有回答的问题。唯一处理这一问题的欧盟法律工具是针对单一行业制定的，下一节将对此进行考察。

5.1.2. 欧盟关于自动驾驶系统的《法规》

车辆类型在欧盟销售前，公共机构必须证明其符合适用的技术要求（这称为型式批准）。《条例（欧盟）2019/2144》规定了车辆的一般安全要求，并将技术细节交由欧盟委员会制定，而实施条例（欧盟）

[60]欧盟委员会，《关于〈条例（欧盟）2024/1689〉规定的通用人工智能模型义务范围的指南》C(2025) 5045 final（2025年7月18日），第17段及其后所列示例。

2022/1426[61] 是欧盟委员会针对能够自行驾驶的车辆（全自动驾驶车辆）所制定的规则手册。

根据《人工智能法案》，车辆法规（欧盟）2018/858号法规和（欧盟）2019/2144号法规列于该法案附件一的第二份清单中（B部分第18点和第19点）；该清单列出了保留自身规则控制权的行业。对于清单所列产品中的人工智能系统，仅适用《人工智能法案》的少数规定（即经修订的第2条第（2）款：第6条第（1）款、第60a条以及第102至112条；关于监管沙盒的规定，仅在高风险要求被纳入行业立法后适用）。这意味着，《人工智能法案》关于数据、监督和稳健性的要求并不直接约束自动驾驶系统。

实施法规（欧盟）2022/1426号针对的是某一车辆类型的自动驾驶系统（ADS），相关义务由制造商承担；制造商申请批准，并提交证据，随后由型式批准机构，或代表该机构行事的技术服务机构，根据附件二中的规范并依照附件三（2022/1426）所规定的程序进行评估。附件三第4部分在此按照2026年3月24日的合并文本阅读，是对我们最为重要的规定，因为它规定了“使用虚拟工具链进行自动驾驶系统验证的可信度评估原则”，其明确目标是证明用于验证驾驶系统的“任何仿真模型或虚拟工具链”具有可信度。制造商会部分通过模拟来测试驾驶系统，而世界模型可以用于这一目的；他们所使用的那套仿真工具，实际上正是《法规》所称的虚拟工具链。制造商还必须说明仿真涵盖哪些场景、模型在哪些方面已知存在局限、结果对其假设的敏感程度、通过哪些措施对模型进行了验证，以及模型具有多大不确定性，包括模型结构方面的不确定性（第3.3至3.4.5点）。在实践中，制造商必须证明模拟环境能够正常运行：一方面记录其用于验证仿真模型与现实相符的数据，另一方面证明通过哪些校准程序使模型参数适配这些数据。主管机关会审查这些文件，并可对制造商进行审计，既可以通过书面审查，也可以开展物理测试。评估结果会确定该工具的包络，即可以依赖其结果的条件范围。评估要求的严格程度取决于安全论证在多大程度上依赖该工具（第3.3.5点）。同时，还必须牢记一个明确的底线：模拟不能取代第3部分要求进行的物理测试（第2部分第4.2点）。

这是本文考察的欧盟框架中，最明确规定如何判定模拟环境是否足够可信、足以验证安全关键系统的一项框架。第4部分针对一般模拟而制定，适用于“任何仿真模型或虚拟工具链”，并未对那些从数据中学习其动力学、而非由工程师编写动力学的模型作出例外规定。因此，制造商若使用学习型世界模型验证其驾驶系统，就会受第4部分约束，但这是通过解释得出的结论：该条款并未提及此类模型，也未说明如何让某些要求适用于规则并非由任何人编写的模型，例如解释模型参数如何校准。

在此背景下，欧盟车辆法与《人工智能法案》不同，已经设有一套程序，用于询问模拟是否可信，并规定了文档记录义务、

[61]欧洲议会和欧盟理事会2019年11月27日关于机动车辆型式批准要求的《条例（欧盟）2019/2144》[2019]《欧盟官方公报》L325/1；欧盟委员会2022年8月5日发布的《欧盟委员会第2022/1426号实施条例》，规定了适用于全自动驾驶车辆自动驾驶系统（ADS）型式批准的统一程序和技术规范，以落实《条例（欧盟）2019/2144》[2022]《欧盟官方公报》L221/1，引自2026年3月24日合并版本。

审计权、明确的有效范围以及物理测试的最低要求。该程序仅适用于一个行业，而且将审查证据的权力留给了批准主管机关。

5.2. 加利福尼亚州

众所周知，在美国，尽管联邦政府曾多次试图另辟蹊径，但迄今为止，相关规则主要由各州制定；其中，加利福尼亚州的法规最为完善。该法规也与欧盟《人工智能法》颇为相似（虽需作适当调整），很可能正是受其影响。

《前沿人工智能透明度法案》（参议院第53号法案，2025年）[62]，于2026年1月1日生效，并不监管用途或产品，而是直接监管开发超大型模型的公司，并根据这些公司所开发的模型对其进行界定，主要采用三个标签。“人工智能模型”是指一种系统，能够根据输入推断如何生成“能够影响物理或虚拟环境”的输出（《商业与职业法典》第22757.11(b)条）；这一表述与欧盟对人工智能系统的定义十分接近。“基础模型”是指在广泛数据集上训练、旨在生成通用输出，并能够适应多种不同任务的人工智能模型（第22757.11(f)条）。“前沿模型”是指使用超过 10^26 整数次或浮点运算训练的基础模型——该数值是《人工智能法案》推定通用模型具有高影响能力的门槛的十倍——计算时包括初始训练，以及开发者之后进行的任何微调或其他重大修改（第22757.11(i)条）。

因此，世界模型属于加利福尼亚州《人工智能法案》所定义的“人工智能模型”；若其基于广泛的数据进行训练，且可适用于多种任务，则也属于“基础模型”，无论它是否生成语言。与此同时，符合这些定义本身影响不大，因为加利福尼亚州《人工智能法案》仅要求被其称为“前沿模型”的那部分基础模型的开发者履行主要治理义务（例如透明度和强制性事件报告要求）。大型前沿开发者——即年收入超过法定 5 亿美元门槛的开发者，包括其关联方——还必须制定、遵循并公布灾难性风险框架，每年对其进行审查，并提供额外的评估披露和内部使用报告（§§ 22757.11–22757.13）。

不过，训练计算量门槛并不能决定 SB 53 规定的所有义务。如果某个世界模型符合基础模型的定义，即使该模型未达到前沿模型门槛，只要其开发者凭借另一款模型符合前沿模型开发者的定义，相关举报人条款仍可能适用。开发者必须保护员工就涉及该世界模型的灾难性风险所作的符合条件的披露，并告知受涵盖员工其权利；大型前沿开发者还必须提供匿名内部举报渠道（《劳动法典》§§ 1107–1107.1）。这些义务本身并不要求公布该非前沿模型的训练计算量，也不要求验证其预测。

对于主要的模型治理职责，问题在于基础世界模型的累积训练计算量是否超过 10^26 整数或浮点运算。公众可获得的证据并不均衡，但也并非完全缺失。例如，最初的 Genie论文报告称，训练其动力学模型需要 6.6×10^{22} 浮点运算，低于门槛。[63] V-JEPA 2 报告的训练资源需求同样表明，就其所述的预训练而言，训练计算量估计远低于门槛。[64] 这些披露涉及特定部件或训练阶段，无法证明

[62]《前沿人工智能透明度法案》，《加州法规汇编》2025年，第138章（参议院第53号法案），新增《加州商业与职业法典》第22757.10–22757.16条及《加州劳动法典》第1107–1107.2条。第5.2节中未另行说明的条文编号均指《商业与职业法典》中的条文。[63]Bruce等人（注20）表12。[64] Assran等人（注23）第2.4节及图5。

与法定分类相关的完整累积计算量。扩展实验表明，有理由预期某些世界模型项目会采用更大规模的训练，但效率提升和架构差异使人们无法可靠预测特定模型是否会超过门槛，或何时会超过门槛。

需要特别指出的是，对于向加利福尼亚州居民公开提供的受规制生成式人工智能模型或服务，存在一项独立的披露义务，要求开发者声明在开发过程中是否使用了合成数据生成（《加利福尼亚州民法典》第3111(a)(12)条）。何时以及如何认定合成数据具有可靠性，则是另一个问题，披露义务对此并未作出明确说明。就此而言，值得注意的是，与欧盟自动驾驶车辆审批最接近的比较对象，是加利福尼亚州分阶段测试和部署许可程序：申请部署的制造商必须说明其已完成的安全论证，并提交所要求的测试证据（《加利福尼亚州法规汇编》第13编第3.7—3.8条，第228.08节）。因此，即使世界模型的开发者不属于SB 53的适用范围，将其用作内部模拟器仍可能影响许可决定。

5.3. 中国

中国法律可通过两条尤其相关的路径适用于世界模型：对生成式内容服务的监管，以及对自动驾驶车辆安全性的监管。

中国针对生成式内容服务的监管框架由三个部门发布的文件构成。每份文件监管的都是某一类服务，而非某一类模型，并且对服务的界定各不相同。

- 《互联网信息服务深度合成管理规定》（自2023年1月10日起施行）适用于在中国使用“深度合成技术”的互联网信息服务。[65]
- 《生成式人工智能服务管理暂行办法》（自2023年8月15日起施行）仅适用于在中国向公众提供的生成式服务（第2条）。[66] [67]
- 《人工智能生成合成内容标识办法》（自2025年9月1日起施行）则对已受前两项规定或算法推荐相关规则规制的提供者增设标识义务。[68]

由于这些规则针对的是服务，世界模型不会仅仅因为是世界模型就受到规制；只有在满足两个条件时才会受到规制：（i）有人利用它提供某项法律文书所述类型的服务；（ii）它生成的是该文书所指类型的内容。世界模型生成场景时，很容易满足第二个条件。中国法律将使用人工智能生成或修改内容称为“深度合成”，《规定》第23条将虚拟场景纳入其定义，列举三维重建和数字模拟等技术，并描述

[65]《互联网信息服务深度合成管理规定》（国家互联网信息办公室、工业和信息化部、公安部令第12号，自2023年1月10日起施行）（以下简称《深度合成规定》）。[66]Mimi Zou 和 Lu Zhang，《中国对生成式人工智能与大语言模型的监管路径》（2025年），《剑桥人工智能论坛：法律与治理》第1卷，e8，指出该办法不涵盖研发以及企业和行业应用；另见 Sara Migliorini，《中国生成式人工智能暂行办法：起源、内容与意义》（2024年），《计算机法与安全评论》第53卷，105985。[67]《生成式人工智能服务管理暂行办法》（国家互联网信息办公室等部门令第15号，自2023年8月15日起施行）（以下简称《暂行办法》）。[68]《人工智能生成合成内容标识办法》（国家互联网信息办公室、工业和信息化部、公安部、国家广播电视总局于2025年3月7日发布，自2025年9月1日起施行）（以下简称《标识办法》）。

参与者可以体验或与之互动的沉浸式场景。因此，由世界模型生成的可交互、可探索的环境符合这一表述。当然，这一表述也涵盖并非世界模型的系统，因为生成虚拟场景并不能证明系统已经学会环境如何变化。而且，它不太可能涵盖完全不生成内容的世界模型，例如仅仅选择机器人的下一步动作的模型。

第一项条件（i）即服务，决定了哪些用途属于规管范围。向公众提供可生成可探索世界的在线工具的公司，提供的服务属于这三项法律文件的适用范围。相比之下，仅在内部使用所构建模拟器的公司不受《暂行办法》约束；《暂行办法》明确将开发和应用生成式人工智能技术、但不向国内公众提供生成式人工智能服务的组织排除在外（第2条）。该公司是否也不受另外两项法律文件约束，则取决于它们各自的规定。

尽管此处的规则适用于服务，但其义务也通过三种方式延伸至服务背后的技术。

- 标识。模型生成的场景及其他内容必须带有技术标识；在公众可能感到困惑或受到误导的情况下，还必须带有醒目的标签（《规定》第16—17条）。特别是，《标识办法》规定了两种标签：在规定情形下使用的显式标签，例如可见或可听的提示，以及文件元数据中的隐式标签。该办法还对分发平台、应用商店和发布内容的用户规定了义务。若法律另有允许，且协议明确用户的责任并将日志保存至少六个月，提供者可应用户要求提供不带显式标签的内容；但仍须添加元数据标签（第9条）。• 技术治理。提供者必须定期审查和验证其生成式算法；若服务由世界模型驱动，该模型也包括在内，而某些工具还须接受安全评估（《规定》第15条）。一些义务也适用于提供底层技术的公司。• 准确性。提供者必须采取有效措施，提高生成内容的可靠性（《暂行办法》第4条第（五）项）和训练数据的质量（第7条第（四）项），例如评估数据标注的质量并抽查准确性（第8条）；并且必须应要求向监管部门提供有关其训练数据和算法的信息（第19条）。

这些义务涉及如何识别内容以及如何运营服务。其中包括提高准确性的一般义务，但没有规定任何程序来证明学习型仿真对于特定安全关键用途是有效的。如上所示，这些义务对内部用途的覆盖也并不均衡。因此，在公司内部用于训练或测试安全关键系统的世界模型，是这一途径最难妥善处理的情形。

第二种途径针对某一行业处理这一情形。GB 44721-2026 是一项关于自动驾驶系统安全性的强制性国家标准，于2026年夏季发布，并自2027年7月1日起适用。该标准规范的是一种产品，即车辆及其驾驶系统，并具有约束力：不符合规定的产品和服务不得生产、销售、进口或提供（《标准化法》第25条）。

该标准适用于车辆制造商，后者必须证明其驾驶系统具备安全性。据工业和信息化部称，他们必须通过安全保障流程、安全文件以及模拟、测试场和道路测试来做到这一点，并由第三方机构进行确认性测试。[69] 世界模型在该标准中

[69]工业和信息化部关于发布 GB 44721-2026 的声明（2026年8月4日），据新华社报道（2026年8月5日）。

有两种适用方式：一种是制造商将其构建到驾驶系统本身中，另一种是 制造商将其用于仿真测试，即作为道路的代理。

因此，从我们在第3.1节提出的分类法中的四种类型来看，中国法律对世界模型采取了不同的处理方式。内容模型受到直接监管，被视为一种面向公众生成内容的服务：其输出必须加以标记和标注，提供者必须审查其算法并采取措施提高准确性。[70]

控制模型作为其所控制产品的一部分受到监管。正如我们所见，对于车辆，强制性标准要求证明自动驾驶系统是安全的。[71] 对于工业机器人，相关安全标准早已实施，依据的是2006年和2011年针对机器人及其保障措施的国际文件。[72] 对于人形机器人，相关标准仍在规划中；计划草案并非强制性国家标准，但将世界模型列为标准化对象。[73] 这三类产品均受《产品质量法》约束。根据该法，如果产品不符合适用的个人安全或财产安全国家标准或行业标准之一，或造成不合理危险，即属于缺陷产品。[74]

咨询模型在决策环节受到监管。在法院，人工智能的结果仅可作为补充参考，[75] 证据规则会检验电子材料是否真实；必要时可进行专家鉴定，但不会要求判断模拟是否正确。[76]

最后，只有在产品规则要求进行模拟时，代理模型才会受到监管；自2027年7月起，这意味着车辆。[77] 内部开发并使用的模拟器不在《暂行办法》的适用范围内；它是否也不受其他服务规则约束，则取决于这些规则的条款。[78] 在所考察的各项法规中，都没有关于如何证明学习环境有效性的一般规则。

6. 尚未涵盖的内容，以及适应所需的条件

前文分析阐明，世界模型将纳入主要现行人工智能法规的适用范围，部分原因在于世界模型与其他机器学习系统一样，都是产生预测的系统；另一部分原因在于相关规则以一般性措辞撰写，并不取决于某项特定技术。不过，另一个问题是这些监管框架是否充分。

我将指出现行规则在哪些方面未能涵盖世界模型，但不会主张制定专门的《法规》。[79] 我先从世界模型的训练数据谈起（第6.1节），因为接下来的两个问题都取决于这些数据：世界模型作为预测器的准确性（第6.2节），以及作为其他系统训练或测试环境时的验证（第6.3节）。在每一节的末尾，我都会尝试

---

[70]《深度合成规定》（注67）第15—17条及第23条；《标识办法》（注70）第3—5条；《暂行办法》（注69）第4条第5项及第7条第4项。[71]GB 44721-2026（注71）。[72]GB 11291.1-2011，《工业环境用机器人：安全要求 第1部分：机器人》（与 ISO 10218-1:2006 相同）；GB 11291.2-2013，第2部分：机器人系统与集成（与 ISO 10218-2:2011 相同）。[73]工业和信息化部，《国家人形机器人产业标准体系建设指南》（2026年版），征求意见稿（2026年8月24日）。[74]《中华人民共和国产品质量法》第46条。[75]最高人民法院《关于规范和加强人工智能在司法领域应用的意见》（2022年12月8日）。[76]最高人民法院《关于民事诉讼证据的若干规定》（2019年修正，2020年5月1日起施行），第14条、第93条及第94条。[77]GB 44721-2026（注71）。[78]《暂行办法》（注69）第2条。[79] Lyria Bennett Moses，《如何思考法律、法规与技术：将“技术”作为监管对象的问题》（2013年）5《法律、创新与技术》1。

概述具体建议，并尽量采用足够普遍的表述，以免局限于某个特定司法管辖区。

以下分析中一个共同的关切是：大多数人工智能系统会做出一次性预测，可以根据观察到的案例进行核验；而世界模型预测某个动作会造成什么后果，以及其影响将如何逐步展开。这就是下面提出的一些补救措施具有时间性特点的原因。

6.1. 动作数据

至此应当已经清楚，世界模型通常缺乏开发者能够加以描述并披露的明确因果模型，因为没有人编写它的规则。开发者可以记录的一项内容，是模型如何学会将动作与结果联系起来的证据。但在法律推理中，这里有一个重要区别：动作是如何被选择的，以及动作是如何被识别的。据我所知，这一区别在因果推断和机器学习领域广为人知，但将其作为记录训练数据这一法律义务的内容来处理，尚属首创。

第一个问题区分了实验性数据与观察性数据。在实验中，研究人员可能会有意且具体地指示车辆在同一测试场地的不同位置刹车。如果车辆以相同速度起步，且道路状况保持相近，研究人员便可以探究延迟发出指令会如何影响停车位置。不过在这种情况下，证据的说服力来自比较设计，而不是仅仅因为有人选择了该动作：指示车辆刹车本身并不能证明刹车造成了什么后果。在观察性数据中，例如车队日常记录，驾驶员会出于各自的原因选择何时刹车；例如，驾驶员可能先看到道路旁有个孩子，而后车辆摄像头才记录到这个孩子。记录显示她减速后发生了什么，但若要估计她保持车速会发生什么，就需要对促使她减速的情形作出某种（额外的）说明。换言之，只有在相关情形得到充分记录且存在可比案例时，观察性数据才能支持这类推断。[80]

第二个问题涉及如何识别动作本身。车辆日志可能记录软件何时发出制动指令、刹车何时启动，也可能只记录车辆如何移动；这三者是不同的信息。如果日志中只有视频，开发者可能会根据车速下降来推断车辆正在制动，尽管汽车也可能是因为道路上坡而减速。Genie 展示了虚拟环境中的这一问题：它从游戏视频中学习，却无法获取玩家的指令，而是识别场景中的变化，并据此控制场景。这使生成场景具有交互性，但推断出的控制信号是否对应玩家实际采取的动作，是另一个需要审慎判断的问题，不应想当然地认定。针对这类推断动作的实验表明，如果视频中出现了与真实动作相关、但并非由真实动作导致的变化，推断结果的质量就会下降；而只要为一小部分数据提供记录的行动，质量便会显著提高。[81]

这说明了哪些内容值得记录。对于安全主张所依赖的动作，开发者应记录这些动作是如何选定的，以及它们是如何

[80]有研究记录了一个相关问题：从记录的示范中学习如何行动的系统，可能会把仅仅伴随专家动作出现的事物误认为动作的原因，驾驶场景中也会出现这种情况：Pim de Haan、Dinesh Jayaraman 和 Sergey Levine，《模仿学习中的因果混淆》，载于神经信息处理系统进展第*32*卷（2019年）。 [81]A Nikulin 等人，《存在干扰因素时，潜在动作学习需要监督》，载于第*42*届国际机器学习会议论文集（PMLR 267，2025年），46427。

被记录或推断出来的，以及它们是在什么条件下被观察到的。[82] 如果答案未知（互联网收集的视频往往如此），文档应如实说明。这样，主管机关便可以了解，例如，某个模型是在干燥路面上通过大量精确记录的制动指令进行训练的，但在冰面上的此类指令很少，进而判断在哪些方面有必要开展更多测试。凡是依赖模拟器测试驾驶系统的人，也应披露驾驶系统与模拟器所用数据或预训练模型之间任何已知或可合理查明的重叠之处，前提是这种重叠与所测试的行为有关。如果两者都从同一车队学习制动行为，那么模拟成功或许只是在重现两个系统共有的错误；即使使用的是不同车队，如果两者都不常在冰面上运行，这个问题仍可能没有答案。

这一数据背景是否受到法律规制？部分受到。以《人工智能法案》为例，其中已有若干《规定》为此类信息提供了依据。在适用高风险要求的情况下，第10条第2款(b)项涉及数据的收集方式及其来源；附件IV第2(d)点则要求在技术文档中说明训练数据的来源、选择情况，以及在相关情况下的标签。此外，附件IV第2(a)点要求提供者描述其对第三方提供的预训练系统或工具的使用情况及使用方式，这至少涵盖了前述部分重叠内容。根据第53条第1款(a)项和附件XI第1节第2(c)点，通用模型提供者也承担类似的文档义务，但须遵守第53条规定的限定条件。因此，缺口在于这些规定未作说明的部分；指导文件应予补充说明，即对于以动作作为条件的模型而言，有用的来源说明应当是什么样的。例如，“车队视频”能够指明来源，却无法让读者判断开发者是否能够获取制动指令，还是从视频素材中<styles id='1'>推断出这些指令。从这一意义上说，本提案在适用情况下，为现有文档义务提供了更精确的内容；但若要将其扩展至内部模拟器的每一位供应商，则需要单独的法律依据。

在中国和加利福尼亚州，这一差距更大，因为两地的数据披露规则并未涵盖此类情形。中国的《暂行办法》要求训练数据来源合法；如果数据经过标注，还要求制定标注规则并检查数据质量（第7至8条），但这些规定针对的是在中国大陆面向公众提供的生成式服务。加利福尼亚州要求向加州居民公开提供生成式人工智能系统或服务的开发者，发布其训练数据集的概要说明，包括数据来源或所有者（《民法典》第3111条）。仅使用模拟器测试自有车辆的制造商通常不属于这两套制度的适用范围。对这类制造商而言，真正相关的问题是：当其向车辆主管机关提交模拟结果时，必须提供哪些证明；第6.3节将再次讨论这一问题。

> **建议1（动作数据的来源）**。如果安全主张依赖于世界模型对动作的预测，开发者就应记录训练数据中的动作是如何选取、记录或推断的，以及这些动作是在什么条件下观测到的，并说明上述信息中哪些未知。
>
> **建议2（共享依赖披露）**。任何依赖世界模型来训练或测试系统的人，都应披露系统所使用的数据或预训练模型与世界模型所使用的数据或预训练模型之间已知或可合理查明的任何重叠，

[82]该提案将记录数据集构成和收集方式的既有做法扩展至动作数据：Timnit Gebru 等人，《数据集数据说明书》（2021）64(12)《ACM通讯》86。

> 只要这种重叠与安全性主张有关。

6.2. 控制、咨询和内容世界模型中的准确性

世界模型会预测环境的下一状态。由于它通常会向前运行多个步骤，因此每个预测状态都会自然地成为下一次预测的起点。在此基础上，控制模型和咨询模型会比较可能采取的动作，而内容模型则会生成一个随时间延续的场景。在这两种情况下，由于连续潜在状态之间存在这种依赖关系，一个错误预测的状态可能会产生累积误差。[83] 在包括大语言模型在内的、逐步生成预测的系统中，累积误差十分常见；这与《人工智能法案》中的相关规定有关，例如第9条第(8)款、第15条第(1)款和第(3)款要求高风险系统的提供者根据适合预期目的的“预先定义的指标和概率阈值”进行测试，以确保达到适当的准确性水平，并在使用说明中声明这些指标和阈值。在中国语境下，关于自动驾驶系统的标准要求进行测试，包括模拟测试；而《暂行办法》第4条第(5)款仅要求公共生成式服务的提供者采取有效措施，提高其生成内容的准确性。《加利福尼亚州法规》并未规定普遍的预测准确性声明义务，但世界模型的预测误差可能会被纳入其关于评估和降低灾难性风险的规定（第22757.12条），且仅适用于大型前沿开发者，并且只有在这些误差促成了《人工智能法案》所列的某一灾难性事件时才适用。

主要问题在于，这类规则没有说明预测需要向前检查多远才能与准确性指标进行核对；鉴于模型可能在向前预测一步时非常准确，而预测五步时却大错特错，这一点至关重要。如果使用说明中只出现第一个数字，就可能误导用户，使其误判所选动作的可靠性。造成这种情况的部分原因如前所述：每次预测都建立在前一次预测的基础上；另一部分原因则在于通常所说的模型利用。举例来说，规划器会根据模型对每个动作相对于规划器所接收的目标（例如，在不发生碰撞的情况下到达目的地）的预测得分，选择某些动作而非其他动作。估算总会包含一些误差，而由于规划器会挑选得分最高的动作，它往往会选择模型高估的动作（并忽略被低估的动作）。世界模型相关文献则从模型的角度描述了这一弱点：由于世界模型大多使用成功示范进行训练，因此即使某些动作会失败，它们也倾向于预测这些动作会成功。[84] 这两种弱点叠加在一起：规划器会选择模型高估的动作，而模型尤其会高估那些会导致失败的动作。模型对基于其优化的系统所作的性能预测，与该系统随后在现实中表现出的性能之间的差距，正是衡量这一问题的直接指标，却几乎从未被报告。[85] 这类错误通常会在系统接受真实世界测试时暴露出来；但如果世界模型取代了真实世界测试，错误就会难以发现（详见第6.3节）。尤其值得注意的是，这些错误不仅会使世界模型的预测能力变差，还会影响依赖这些预测的保障措施。在最近的一项实验中，一个使用世界模型进行规划的模拟机器人臂收到了一组需要遵守的规则（例如，不得将物体推入禁区），

[83]Talvitie（注48）。[84]Kamat 等（注49）第6.5和6.8节及表23。关于一手证据，参见 T Yang 等，《MiraBench：评估机器人世界模型中的动作条件可靠性》（arXiv 预印本 2605.29360，2026年），文中将这一失效称为“乐观偏差”；另见 J Xu 等，《世界模型梦想成功：诊断并修复机器人世界模型中的失败不敏感性》（arXiv 预印本 2610.09134，2026年），他们报告称，在四个已发布的模型中，经过验证的失败会得到类似成功的预测。[85]Y Yu 等，《具身决策中的世界模型应如何评估？以决策为中心的立场》（arXiv 预印本 2606.15032，2026年）第3.6和5节。

而且，每次动作执行前，都会根据模型的预测检查其是否符合这些规则。 在对场景有完美了解的情况下，机器人有93%的时间遵守了规则；而依赖世界模型时，遵守规则的比例仅为64.7%，并且对其施加的纠正中，超过一半针对的是从未发生过的违规行为。[86]

因此，弥合这一差距的方法，是调整此类规则的适用或解释方式，使准确性能够在规划器选择其动作所依据的同一时间跨度（即其规划时域）内，按相关时间间隔进行测量和声明，并随着该时间跨度的延长，跟踪任何恶化或改善。出于相关原因，还应针对规划器实际选择的动作检查准确性，因为测试集上所有动作的平均准确性，可能会对实践中的错误率造成误导。这一提议基于技术文献中近期提出的、指向同一方向的观点：应根据模型预测的提前量来报告错误率。应向模型展示已知会失败的动作，以检查它是否会对这些动作预测成功（这种技术有时被称为乐观偏差检测）[87], ，并且应比较实际针对该模型进行优化的系统的预测性能与真实性能（以衡量有时所称的可利用性差距）。[88]

这里需要作一项澄清：结果层面的衡量应优先，因为预测准确性即使是在整个规划时域内报告的，也不足以单独作为依据。研究发现，它并不总能反映使用该模型进行规划的系统的性能，而预测误差较低的模型仍可能错误地排列两种动作方案的优先顺序。[89] 因此，提供者首先应声明：模型对安全主张所依据的结果进行预测的准确程度，例如某一机动是否会导致碰撞；以及系统的预测性能与系统在现实中表现出的性能之间存在多大偏差。正如其倡导者所承认的那样，后者很难在真实硬件上大规模测量，[90] 因此，应根据安全主张在多大程度上依赖该模型，按相应比例提出要求，正如欧盟车辆规则已经针对模拟作出的规定（第5.1.2节）。

鉴于累积误差和模型利用在机器学习领域早已为人熟知，因此，这属于适用这些规则的标准中的细节问题；前文提及的各项提案，包括第5节开头提到的正在制定的标准，均表明此类标准可以涵盖哪些内容。事实上，同一文献还指出，任何提出世界模型的人都应事先说明其所针对的任务、（规划）时域以及提供该模型的用途。根据《人工智能法案》，这只意味着明确提供者本就必须声明的预期目的（第3(12)条）。

最后，同样的推理应当从批准前的测试延伸至系统获批后的使用阶段（有时称为上市后监测）。世界模型会明确预测接下来会发生什么，之后可以将这些预测与实践中实际发生的情况进行比较，因此，两者之间的差距可以持续衡量系统在实际使用中的准确性，也能表明模型是否偏离了测试时所处的条件。[91] 《人工智能法案》已经要求高风险系统保留日志，以支持

---

[86]D Waldner 等人：《基于世界模型的规划与法律推理立法》（arXiv 预印本 2609.15113，2026年）第6节（问题1、4和5）。 [87]Yang 等人（注86）。 [88]Kamat 等人（注49）；Yu 等人（注87）。 [89]N Lambert 等人，《基于模型的强化学习中的目标错配》，载于第二届动力学与控制学习会议论文集（PMLR 120，2020年），第761页；LM Bhamidipaty 等人，《不完美的世界模型可被利用》（arXiv 预印本 2605.15960，2026年）。 [90]Yu 等人（注87）。 [91]A Sattar，《评估世界模型：传统人工智能基准为何失效》（Medium，2026年）<
https://
medium.com/@adnansattar09/evaluating-world-models-why-traditional-ai-benchmarks-fail-65777077dc07>。

上市后监测（第 12 条第（1）–（2）款），并要求提供者在高风险系统的整个生命周期内收集并分析相关系统的性能数据，且方式须“与人工智能技术的性质相称”（第 72 条第（1）–（2）款）。对于世界模型而言，这意味着记录预测结果及实际观察到的结果，并将两者之间不断扩大的差距视为采取纠正措施的依据（第 20 条第（1）款）。[92]

> **建议3（预期用途的准确性）**。世界模型的开发者应声明其误差如何随预测时距而变化，并说明这些数据适用于哪些任务和用途。任何基于该模型构建系统的人，都应针对该系统实际选择的规划时域和动作，声明模型对安全主张所依据的结果预测得有多准确，以及预测表现与实际表现之间的差距。
>
> **建议4（监测使用中的预测）**。在能够观察到结果的情况下，系统提供者应对实际采取的动作，在预测期内按适当比例抽取样本，将预测与结果进行比较。提供者应预先明确如何衡量差异，以及在何种情况下需要启动审查。

6.3. 替代世界模型的验证

由其他系统训练或测试人工智能系统，本身并不是什么新鲜事：例如，大语言模型可以生成文本，进而用于训练更多语言模型。然而，在某些条件下，这种做法会重现并放大错误，这一点已有观察记录。[93] 正如我在第2.4节指出的，用于训练或测试另一系统的世界模型（又称替代模型，第3节）会生成合成数据；许多法律制度对此尚无准备，因为其数据治理的设计初衷是处理收集到的数据。与后者不同，合成数据的可靠性在很大程度上取决于生成这些数据时所采用的真实值假设是否有效。[94] 与此类似，世界模型带来的问题在于，如何验证用于测试或训练的模拟环境是否适合该用途；这就需要检查其真实值假设，因为这些假设并非由人撰写。

设想一下，驾驶软件在世界模型中接受测试：该模型生成一个驾驶场景，并在该环境中测试软件的行为（环境给出的响应也始终由模型生成）。现在再设想一下，该模型未能学会湿滑路面对汽车制动的具体影响，例如汽车需要多长时间才能停下。尽管这个生成场景在抽象层面上包含了人们可以检查和验证的要素，但其视觉上的可信度不足以证明汽车会停在真实汽车应当停下的位置；正如人工智能生成的流畅文本具有可信度，也不能证明文本是否正确。[95]

---

[92]在美国银行监管中，将预测与结果进行比较是一项既有实践，其中“结果分析”包括按照与模型预测期限相匹配的频率进行回溯测试，这属于模型验证的一部分；并且，模型投入使用后，验证工作仍会继续：美联储理事会和货币监理署，《模型风险管理监管指引》（SR Letter 11-7，2011年4月4日）第 V 节。该指引于2026年4月被 SR 26-2 取代。[93]I Shumailov 等，《在递归生成的数据上训练时，人工智能模型会崩溃》（2024年）631 《自然》755。[94]Gal 和 Lynskey（前引文献 25）；Cofone、Strandburg 和 Tilmes（前引文献 25）。[95]这一点也经过了测试：当多个世界模型被投入真实任务时，其视觉质量并不能保证成功，更重要的是它们能否对动作作出正确响应。参见 J Zhang 等，《世界中的世界：闭环世界中的世界模型》（国际学习表征会议，2026年）。

对于“传统”人工智能和语言模型而言，通常存在系统之外的某种参照物，可据此对输出进行比较和验证，例如真实案例、人类专家读者、原始数据或书面方程；而由于多种原因，验证世界模型要复杂得多。其中一个主要原因是，模拟的输出即使以图像或视频的形式呈现，也不能仅凭观察来判断，因为它描述的是一个复杂的物理情景，有时还只是一个假设情景，例如测试中的系统采取动作后，在后续某个潜在动作的对应情况下，物理上会发生什么。首先，正如前文所述，[96] 这要求建立新的基准测试，以认证世界模型所学到的内部环境是否足够接近现实，从而能够指导现实世界决策。在这一点上，技术文献已经开始说明此类基准测试应当衡量什么：不是场景看起来有多真实，而是例如预测是否会像现实那样发生变化，尤其是在相同的初始场景下，动作发生改变时，预测是否也会改变（即“动作跟随”或“可控性”）；以及那些具有物理真值的量，例如速度和距离，是否会得出正确结果。[97]

这些测试要求能够参考现实，不过所需现实世界证据的多少，取决于相关主张以及世界模型的用途。[98] 例如，制造商无需重现一次近碰撞，就能检验模型对停车距离的预测，因为在空旷赛道上、相似条件下采集的制动测量数据可能已经足够。[99] 有些关系即使不知道两次测试的确切结果，也能加以检验：例如，如果车辆以已知的恒定速度继续行驶一秒，然后在均匀路面上施加同样的制动，那么它在这一秒内额外行驶的距离可以独立计算。不过，最棘手的情况是第二类原因：这类情况使验证变得困难，也正是世界模型更有用之处——创造在真实数据中很罕见，或在实践中运行过于危险或昂贵的情境。[100] 此外，世界模型给出的并非单一答案，而是一整串未来时刻，其中有些时刻是动作条件化的，因此它最终可能进入的情境数量过于庞大，无法逐一测试。

当我们将这些困难与《人工智能法案》中的相关规则结合起来解读时，问题尤为突出。这些规则要求训练、验证和测试数据能够代表预期用途，并要求数据治理实践关注数据背后的假设，评估数据的可用性和适用性，识别相关缺口并说明如何弥补这些缺口（第10(2)(d)、(e)和(h)条以及第10(3)至(4)条）。《人工智能法案》还将测试数据定义为能够对系统进行“独立评估”的数据（第3(32)条）。在机器学习实践中，只有在使用与训练数据分开保存的数据进行评估时，

---

[96]斯坦福大学以人为本人工智能研究院（注3）。[97]Zhang 等人（注97）；Kamat 等人（注49）；Yu 等人（注87）。[98]美国食品药品监督管理局对医疗器械申报中的计算模型采取的做法，是让模拟所需的证据规模与对模拟的依赖程度相匹配。所需的可信度取决于“模型风险”，即模型对决策的影响与错误决策所造成后果的结合：美国食品药品监督管理局，《评估医疗器械申报中计算建模与模拟的可信度：致行业和美国食品药品监督管理局工作人员指南》（2023年11月）。该指南以 ASME V&V 40-2018 为基础，适用于基于物理学的模型或机理模型。[99]对于无法进行实地演练的事故，美国的核安全监管也采取类似做法：识别并排序影响结果的现象，并要求仿真模型体现每个重要现象：美国核管理委员会，《监管指南 1.203：瞬态与事故分析方法》（2005年12月）。[100]将模型输出与现实进行比较，并针对模型预定用途开展评估，这就是模拟文献所称的运行验证。相关文献还区分了两种情况：一种是能够收集真实系统的数据，另一种是无法收集数据、只能采用较弱验证形式的情况：RG Sargent，“仿真模型的验证与确认”（2013年），第7卷，《仿真学报》，第12页。前文提到的罕见情形属于第二类。

评估才是独立的。[101] 不过，这两项要求都是针对采集的数据制定的；当数据由生成方式产生时，便难以适用。通常通过将数据与真实案例进行比较来评估其代表性，但在少数使用世界模型生成数据的情形中，此类案例十分稀少，甚至不存在。通常通过将两组数据分开来确保独立性，但即使谨慎地将两组数据分开，只要它们由同一个世界模型生成，或由从相同记录中学习的模型生成，两组数据仍然相互依赖，因为它们共享相同的假设和缺口。换言之，当数据由生成方式产生时，其代表性和独立性取决于生成数据的环境，而这正是缺口所在。问题在于，《人工智能法案》并未涵盖这一环境；该法案仅在检测偏差和沙盒中提及合成数据，将其视为真实数据的替代品（第4a(1)(a)条和第59条），却从未说明何时可以将此类数据视为证据。在《人工智能法案》及其他文件中，提及合成数据的规则关注的是其来源，而非有效性：例如，欧盟委员会针对通用模型的模板会询问是否使用了合成数据，以及由哪些模型生成（见第5.1.1节）；加利福尼亚州要求开发者披露合成数据的使用情况（见第5.2节）；《人工智能法案》和《中国标识措施》均要求对生成内容加注标识（见第5.1.1节和第5.3节）。

那么，应如何解读立法中的这一沉默？上述分析表明，这种沉默可能成为限制性解释的来源，而限制性解释也可能是填补空白的一种方式：也就是说，生成证据的效力应与其可通过真实数据核验的程度成正比。出于同样的原因，如果测试是在系统接受训练的同一环境中运行，就不应被视为具有独立性；对于没有真实数据的情形，模拟场景本身不应证明任何事情。[102] 如果系统与环境仅共享数据或预训练组件，那么问题主要在于程度，因为大多数系统都会使用部分相同的公共数据集和预训练模型。在这些情况下，只有当共享内容与受测行为有关时，才应推定其缺乏独立性；而如果能够证明测试仍可捕捉相关错误，这一推定就应被推翻，建议7提供了一种实现方法。这种担忧并非空穴来风，也有实证依据，不过相关证据来自一个相邻领域：研究发现，作为评估器使用的大语言模型会偏袒相似模型。[103]

这里最大的局限在于，《人工智能法案》的规则仅直接适用于附件一第一份清单所列的产品以及附件三所列的用途，而其中并未记载将世界模型用作测试环境的情况。对于车辆和机械——这两类是世界模型测试最常见的领域——这些规则仅在行业法下适用。因此，问题在于，对车辆规则作同样的解读是否有依据。

在这一意义上，欧盟车辆法规包含了一些保障措施，尽管目前尚不清楚这些措施将如何适用于像世界模型这样的学习型模拟器。根据现行规则，制造商必须提供用于验证模拟的数据和场景信息，而这些数据必须充分反映车辆的运行域。因此，可信度评估的结果界定了模拟可用于支持批准的包络范围（附件三，第4部分）。超出该包络范围后，模拟结果无法得到验证，也不能用于支持批准。

---

[101] T Hastie、R Tibshirani 和 J Friedman，统计学习要素（第2版，Springer，2009年）第7章。[102] IEEE Std 1012-2016《系统、软件和硬件验证与确认标准》从技术、管理和财务层面界定了检查系统者与构建系统者之间的独立性；美国银行监管机构同样要求由不负责开发或使用模型的工作人员开展验证：SR Letter 11-7（注94）第V节。[103] S Goel 等人，〈伟大的模型想法相似，这削弱了人工智能监管〉，载于第*42*届国际机器学习会议论文集（PMLR 267，2025年），19621。

不过，在这里，法律措辞仍有一定的模糊空间，也可能存在一个空白。首先，该包络建立在用于将模拟与现实进行验证的数据之上，而《法规》并未说明模拟器接受训练所使用的数据；对于世界模型而言，这些数据决定了它究竟在哪些范围内能够被预期正常工作。其次，《法规》没有说明应如何绘制这一包络，这同样十分重要：世界模型可能已经在任何速度下的干燥路面以及低速下的湿滑道路上接受过检查，却从未在湿滑道路上的高速行驶条件下接受检查。如果将这些条件分别列出，就会掩盖这一空白，因为“高速”和“湿滑道路”都会显示为已检查，而从示例中学习的模型可能会严重错误地预测一个它从未见过的条件组合。因此，包络中应仅列出那些实际上已通过与实际驾驶进行对照而检查过、且模型在学习过程中接触过的条件组合。

此外，《法规》本身也存在一种张力：它允许使用模拟，“尤其适用于在测试场地或实际驾驶条件下难以实现的场景”（附件三第二部分）。对于传统模拟器而言，在常规条件下得到验证的方程，虽其依据本身仍有争议，但至少能为将其应用于不那么常规的情形提供一些基础；而对于学习型模拟器，这一依据则更具争议，因为其在训练和验证范围之外的行为，正是上述研究所质疑的。换言之，《法规》鼓励在学习型模拟器最不可靠的情况下使用模拟，却没有说明如何在这种情况下确定其包络。这就需要作出调整，遵循我在谈及《人工智能法案》时提到的同一原则：应按上述方式划定包络，而模拟困难场景的鼓励不应超出这一范围。

然而，这两种解读仍未明确：在缺少真实数据的地方，也就是包络的边界处，应当如何准确评判一项测试。一种可能有助于解决这一问题、但（据我所知）在法律文件中鲜少甚至从未被考虑过的技术，是变异测试。在软件工程中，人们会故意在程序中引入小缺陷，再检查测试能否发现这些缺陷，以此评估一组测试的质量；这种做法也已应用于机器学习系统。[105] 第6.2节提到的建议——向世界模型展示已知会失败的动作——遵循同样的逻辑，而且可以从模型扩展到整个测试：将存在已知缺陷的受测系统版本（例如，在湿滑路面上制动过晚的驾驶软件）放入世界模型中进行同一项测试；如果这些版本通过了测试，就说明该测试无法发现这类缺陷。在上述例子中，即使没有任何真实记录，也能发现测试遗漏了湿滑路面。不过，我们也应当意识到，这项技术有一定局限：它只能揭示测试对人为植入的缺陷视而不见，却无法发现意料之外的缺陷；而有缺陷的版本未能通过测试，也说明不了多少问题，因为它的失败本身就是由模型生成的。不过，这项技术在法律上的依据已经初步具备。根据《人工智能法案》，测试必须“确保”高风险系统能够按照其预期目的持续稳定地运行，并符合“预先定义的指标和概率阈值”（第9条第(6)款和第(8)款）；无法发现重要缺陷的测试，很难称得上能够确保这一点。测试能够发现的植入缺陷占比，可以作为其中一项指标，并列入技术文档中的“验证和测试程序”（附件四第2(g)点）。《人工智能法案》中唯一提及对抗性测试（即红队测试）的条款，同样着眼于寻找失败，而非确认成功，但其针对的是具有系统性风险的通用模型，而非测试本身是否充分（第55条第1款(a)项及附件XI第2节），并且

[104]Oberkampf 和 Roy（注 15）。关于传统模拟在此用途上的局限，参见 A Stocco、B Pulfer 和 P Tonella，《注意差距！自动驾驶系统虚拟测试与实体测试可迁移性研究》（2022年），《IEEE软件工程汇刊》<https://doi.org/10.1109/TSE.2022.3202311>。作者将一辆小型实体车辆与其数字孪生进行比较，发现许多测试结果可以在两者之间迁移，但并非全部。[105]Y Jia 和 M Harman，《变异测试发展分析与综述》（2011年），《IEEE软件工程汇刊》37：649；L Ma 等，《DeepMutation：深度学习系统的变异测试》（IEEE软件可靠性工程国际研讨会，2018年）。

模型本身，而非测试是否充分（第55条第1款(a)项及附件XI第2节）。与上文讨论的测试一样，变异测试因此可以通过实施《人工智能法案》的标准和指南，纳入现有监管框架，用于开展欧盟车辆规则已要求进行的可信度评估；也可以作为制造商必须在其加利福尼亚州安全论证中说明的一种验证方法。只有在要求世界模型供应商承担此类义务时，才需要制定新规则，因为所考察的监管框架均未对其规定此类义务。[106]

对于自动驾驶系统的中国标准，同样会出现这些问题；而加利福尼亚州法规的规定则更加不明确：也就是说，是否认为前沿模型可能带来灾难性风险，由开发者自行评估决定（§ 22757.12(a)(2)–(5)、(c)(2)），且评估方法也由开发者自行确定。

> **建议5（模拟结果的证据效力）**。世界模型生成的证据，其支持安全主张的力度，应与相关条件或其部件在多大程度上已通过真实数据进行核验，以及在多大程度上涵盖于模型的训练数据中相称。对于尚未经过上述核验或未被涵盖的情况，可以利用生成场景发现系统可能失效的情形，以便开展调查，并在确认后予以纠正；但生成场景本身不能证明系统不会失效。
>
> **建议6（验证证据的独立性）**。如果一项测试是在被测系统接受训练的环境中运行的，则不应将其视为独立测试。如果二者仅共享与被测行为相关的数据或预训练组件，则应推定其不具备独立性，除非能够证明该测试可以检测出相关错误。
>
> **建议7（检测安全相关故障）**。依赖在世界模型内部运行的测试的人员，应证明该测试能够使存在缺陷的系统暴露出故障，例如，将根据其危害分析或某项标准选取的、已知存在缺陷的系统版本，置于同一测试中运行。他们应报告任何未检测出的缺陷及其对相关主张的影响。检测出这些缺陷可以证明，在这些条件下测试能够捕捉到它们，但不能证明该测试足以表明系统是安全的（《建议5》所赋予权重的条件除外）。

7. 结论

本文的核心问题是，人工智能法规是否已准备好应对世界模型。为此，我考察了这一主题上发展最为成熟的三个法律环境——欧盟、加利福尼亚州和中国——并指出，它们实际上都能够涵盖世界模型，但仍存在需要填补的重要空白。我发现的主要不足涉及用于训练或测试其他系统的世界模型：在这类情况下，安全性的证据本身由模型生成，而现行关于数据的规则原本是为“传统方式”收集的数据设计的，因此可以扩展到涵盖生成数据的来源，却无法涵盖何时可以信任这些数据。更一般地说，这项分析表明，对世界模型的治理归根结底是对证据链的治理。

最后，我想强调一个超越规则本身的考量：世界模型可能会拓展治理所需的专业能力。如果人工智能法规设立的治理方案和机构

[106]在汽车功能安全领域，已有先例将要求附加于工具本身，而不只是产品：对软件工具所需的信心，取决于其故障是否可能引入错误或未能检测出错误，以及这些错误在其他环节被发现的可能性有多大：ISO 26262-8:2018《道路车辆—功能安全—第8部分》第11条。

是围绕某类专业知识构建的——例如，《人工智能法案》要求国家主管机关具备人工智能技术、数据、网络安全、基本权利以及健康与安全风险方面的专业能力（第70(3)条）——那么，这种能力或许已不足以应对当前需求。评估和治理世界模型，意味着要判断模拟是否可信、预测运动在物理层面是否正确，以及机器依据预测采取行动时会如何表现；这些问题更多涉及建模、物理学和机器人学，而不仅仅是数据和软件。

最后需要指出，我在本文中得出的这些结论仍是暂定的，因为这项技术尚不成熟，而且目前关于其应用的大部分信息都来自开发它的公司。