美国加州推进人工智能安全法案 引发科技领袖分歧

加利福尼亚州议会批准了一项备受争议的人工智能安全法案，要求企业确保其技术不会造成重大伤害。

这项名为SB 1047的法案得到了一系列美国知名人士的批评和赞扬。现在，这项法案将回到州参议院进行确认投票，可能进行修改，此后将提交给州长加文·纽森（Gavin Newsom）。

该法案已经在州众议院以41票赞成、9票反对获得通过。

SB 1047法案规定，开发强大人工智能模型的公司必须采取“合理的谨慎措施”，确保其技术不会造成“严重伤害”，比如大规模伤亡或超过5亿美元的财产损失。这项法案由加利福尼亚州参议员斯科特·维纳（Scott Wiener）提出，于今年5月在州参议院获得通过。该法案将要求企业采取预防措施，比如引入可以随时关闭其技术的“终止开关”。该法案还要求将模型提交给第三方测试，以确保将严重风险降至最低水平。

此外，该法案还将为希望分享安全隐患的人工智能公司员工提供举报人保护。不遵守该法案的公司可能会被加州总检察长起诉。

“创新和安全可以携手并进，而加州正在引领这一潮流，”维纳在声明中说，“通过这次投票，州众议院迈出了真正具有历史意义的一步，积极主动地开展工作，以确保一项令人兴奋的新技术在发展过程中保护公众利益。”他还称该法案是一项“简单易行的常识性措施”，将许多大公司已经做出的安全承诺编纂成法律。

在过去的几个月里，该法案遭到了许多主要科技领袖、初创公司和风险投资家的激烈反对，他们认为该法案代表了政府对一项仍处于起步阶段的技术的过度干预，可能会扼杀加利福尼亚州的科技创新。

上周，OpenAI公开表示反对该法案，认为此类政策应在联邦层面而非州层面实施。

包括美国众议员罗·卡纳（Ro Khanna）、旧金山市长伦敦·布里德（London Breed）等政界人士也对该法案表示反对，他们对科技行业的担忧表示认同，即该法案可能会阻碍加州在人工智能创新方面的领导地位。

但在这次投票前的最后阶段，该法案也赢得了人工智能领域一些重要人物的支持。8月27日，埃隆·马斯克（Elon Musk）出人意料地表示了支持，尽管他说这是一个“艰难的决定，会让一些人不高兴”。OpenAI的竞争对手Anthropic以注重安全而著称，也谨慎地支持了这项法案，并在上周表示，“收益可能大于成本”，而且可以切实地实施这些条款。在维纳实施了该公司建议的几项修正后，Anthropic对该法案更加支持。

面对批评者，维纳为该法案进行了辩护，强调其条款仅适用于花费超过1亿美元训练大型模型或花费1000万美元微调模型的公司，这将使大多数规模较小的初创公司免受影响。这位议员还表示，虽然他会支持联邦立法，但国会在科技监管方面历来行动迟缓，在缺乏全国性行动的情况下，他认为加州有责任发挥带头作用。

现在，随着SB 1047法案提交给州长纽森，有关该法案的争论还将继续。OpenAI、科技孵化器Y Combinator和风险投资公司Andreessen Horowitz（SB 1047的批评者）都注册了游说人员为该法案工作。

曝OpenAI“草莓”今秋发布，前身为神秘Q*模型

今天，外媒The Information援引知情人士称，OpenAI将在今年秋天推出代号为“草莓”（之前被称为Q*）的新模型。

“草莓”项目就是盛传已久的神秘Q*模型，据传是此前戏剧化的OpenAI政变关键原因之一。它展现出了解答未见过的数学问题、复杂编程的更强大能力，当时让AI安全的研究人员都为之震惊。

这一项目当下最重要的用处有两个：一是改进ChatGPT等现有产品，二是帮助OpenAI开发下一代旗舰模型Orion。这些都指向了OpenAI想要保持自己在大模型领导地位，并获得更多收入机会的决心。

不过，近日OpenAI还被曝出已经向美国国家安全官员展示了“草莓”项目，这可能是这家创企安全部门高管频繁离职背景下，想要让产品提高透明度的举措。

本月初，OpenAI CEO Sam Altman在推特暗戳戳秀自家花园5颗草莓的时候，就有网友根据草莓成熟时间跳了预言家：GPT-5可能在未来4-6周内确认发布。这正好和现在秋天可能发布的时间点对上了。

可能在不久的将来，我们就能看到“草莓”项目被直接集成到了ChatGPT里。

一、稳住大模型霸主地位，“草莓”能帮新旗舰模型减少幻觉

“草莓”项目是OpenAI新取得的重要技术里程碑，能解决以前从未见过的数学问题，经过训练可以解决涉及编程的问题，回答产品营销策略等主观问题、解决复杂字谜游戏都可以。

一直没有浮出水面的“草莓”模型，近段时间冲到大众视野的迹象变得更加明显。

除了Altman秀5颗草莓，还有OpenAI研究员Trevor Creech发文在OpenAI吃晚餐，盘子里是草莓。

新模型脚步可能越来越近的现实情况下，这可能是OpenAI想要保持自己在大模型领域霸主地位的举措，毕竟它的竞争对手已经拿出了不少与OpenAI最新旗舰模型GPT-4性能相当的模型。

尽管目前OpenAI的模型仍是企业和AI应用程序开发者的首选，但来自谷歌、xAI、Anthropic和Meta等其他企业的模型正在诸多排行榜上迅速赶上OpenAI。

因此，OpenAI的前景在一定程度上取决于它最终能否推出一款代号为Orion的新旗舰模型。该模型旨在改进其去年年初推出的旗舰模型GPT-4，后者于去年年初推出。

而“草莓”项目的发布就对Orion的训练至关重要——它可以为Orion生成高质量训练数据，减少幻觉。

“草莓”模型可以帮助OpenAI克服获取高质量数据的限制，从而利用从互联网上提取的文本或图像等现实世界数据来训练新模型。

智能体创企Minion AI首席执行官、GitHub Copilot前首席架构师Alex Graveley认为，使用“草莓”模型生成更高质量的训练数据可以帮助OpenAI减少其模型产生的错误数量，即所谓的幻觉。该模型之所以能够做到这一点，是因为“训练数据中的歧义较少，所以它猜测的次数较少”。

除了下一代旗舰模型，“草莓”项目的推出也能改进OpenAI的现有产品。

OpenAI内部正在通过“提炼”过程来简化和缩小“草莓”模型，以便在Orion发布之前将其用于提升现有产品的性能。这种“草莓”模型的较小、简化版本，能够在保持与较大模型相同性能水平的同时，更易于操作且成本更低。

一个显而易见的想法是将“草莓”模型改进的推理能力融入ChatGPT中。这可能意味着用户虽然获得了更准确的答案，但速度会变慢。因此，这可能不适用于SearchGPT搜索引擎等用户希望能获得立即响应的工具，但非常适合对时间不太敏感的用例，例如修复GitHub中的非关键编码错误。

那是不是在不久的将来，ChatGPT用户能够根据请求的时间敏感度来自主选择打开或关闭“草莓”模型。

二、OpenAI收入告急？今年夏天已向政府官员展示

面对资金实力雄厚的科技大公司以及疯狂吸金的创企劲敌，OpenAI需要开辟更多的收入机会。

尽管相比于一年前OpenAI业务增长飞速，目前其向企业销售的API和ChatGPT订阅收入增长了约两倍，达到每月2.83亿美元，但OpenAI每月的亏损可能更高。The Information基于此前未披露的内部财务数据和参与该业务的人士的分析，OpenAI今年可能亏损高达50亿美元。

知情人士透露，Altman希望为公司筹集更多资金，并寻找减少损失的方法。作为与微软达成商业合作的一部分，OpenAI自2019年以来已从微软筹集了约130亿美元，该合作将持续到2030年。但合作条款可能会发生变化，包括OpenAI如何向微软支付租用云服务器的费用以开发其模型，这也是目前是OpenAI最大的成本支出。

鉴于现有的ChatGPT等对话式AI在航空航天和结构工程等数学密集型领域并不擅长，解决棘手数学问题的模型可能是一个潜在的有利可图的应用。同时，数学推理的改进也可以帮助模型更好地推理对话查询，例如客户服务请求。

谷歌和一些初创公司也在开发推理技术。上个月，谷歌DeepMind的模型在国际数学奥林匹克竞赛中击败了大多数人类参赛者；Anthropic最新的模型可以编写更复杂的代码、回答有关图表和图形的能力；还有一些创企通过将问题分解为更小的步骤来提高推理能力，但这种方法速度慢且成本更高……

因此，“草莓”模型的发布可能会为推理技术带来新的思路。

Altman今年5月就曾透露“我们觉得我们拥有足够的数据来开发下一个模型。我们已经进行了各种实验，包括生成合成数据。”他当时可能指的就是Orion模型。

一位直接知情人士称，OpenAI在今年夏天向国家安全官员展示了“草莓”模型的能力。

此次演示可能是OpenAI努力向美国政策制定者提高透明度的一部分，近几个月来，OpenAI联合创始人之一John Schulman等几位安全部门的高管离职，这也使得业界对于OpenAI的技术安全更加关注。

结语：OpenAI寻求开辟更多收入机会

推出“草莓”模型是OpenAI是在大模型产业中永无止境的战斗的一部分，它要领先于其他资金雄厚的竞争对手，稳固自己在大模型领域的霸主地位。此外，这项技术还对未来的产品产生影响，如解决多步骤任务的智能体等。

OpenAI等大模型玩家同样希望借此能在这一领域开辟出更多的收入机会，以支持高昂的大模型训练费用。

OpenAI的竞争对手正在进行推理研究

公司	推理研究	投入资金
谷歌DeepMind	开发模型解决数学、几何问题	>每年24亿美元
Anthropic	新模型Claude 3.5 Sonnet；在编码、解释图表方面的推理能力	70亿美元
Imbue	开发AI代理帮助用户编写代码、完成任务	2.32亿美元
H	开发协作解决垂直领域问题的AI代理	2.2亿美元
Cognition Labs	开发AI代理编码	1.96亿美元
Magic	开发AI代理编码	1.45亿美元
Harmonic	开发模型解决复杂数学问题	0.5亿美元**

注：**根据证券申报文件估计