# **2023 出圈的大模型 未来已来？**

原创 Eagle 当下瞬间进入历史 *2023年12月30日 12:51* *新加坡*

2023年末，Sam Altman当选时代周刊年度CEO，Ilya Sutskever入选自然杂志10大年度科学人物，如果没有之前举世沸腾的公司混乱，也许2023年可以被轻松的称作OpenAI年。

OpenAI乱剧收场的时候也留下了两个悬念，Ilya未来何去何从？Q\*项目是什么鬼？

Q\*从名字上会自然让人联想到经典强化学习的Q-learning算法。如果真的有Q\*项目，并且让OpenAI董事会和首席科学家感到压力，技术上可能会有两个原因。

***强化学习（Reinforcement Learning，RL）***

强化学习的惊人威力为世人所知是AlphaGo在和李世石对弈的一盘棋中，AlphaGo的第37手棋（Move37）超出了所有在场人类围棋高手的想象。

DeepMind和OpenAI两强早期也把RL作为实现AGI的主要方法，但OpenAI在Ilya的坚持下转向大语言模型（LLM），帮助他们取得了现在领先的位置。

当时以Geoffrey Hinton为代表的上代深度学习领袖对RL也一直不看好，RL的反馈信号过于稀疏导致学习效率低可能是他们不看好的主要原因，但Richard Sutton和弟子David Silver保持着信念。

RL在ChatGPT中的作用再次令人瞩目，预训练大语言模型早在三年前就已经让学界认识到潜力，但模型的行为难以控制。通过人工提供适量偏好数据，RL可以将预训练模型调教成人们喜欢的对话机器人，偏好数据提供的成本比标注成本低很多，也让这个方法更加可以扩展，这就是人机协同的训练方法RLHF。

从OpenAI出走创业的Anthropic团队更向前走了一步，考虑到将来这个Loop中的人类可能还是瓶颈，他们用AI替代了人类，形成所谓RLAIF，效果非常好。但也开启了恐怖的未来方向，因为这个过程模型完全自动化迭代升级。Anthropic创新的安全措施主要是引入一组原则构成的宪章（Constitution）来管控模型的行为，增加了模型透明度和可解释性，但这个方法也可以被Bad Actor用于其他目的。

Anthropic给的一个安慰性的说法是RL的门槛比较高，年中Andrej Karparthy公开演讲也说过RL很难训练。

目前公开的信息中，RL主要还是用于对齐预训练模型。但主要玩家都提到RL也可以用于基础模型能力提升。Meta Llama 2和DeepMind的研究领导Oriol都说过，Oriol给了数据参考，AlphaGo中RL的算力约占95%，而现在大模型中RL部分只约占5%。

未来的超级模型（Superhuman Intelligence）似乎已经在地平线浮现。

***超级对齐（Super Alignment）***

如果超级模型出现，如何确保超级模型对人类友好，这就是所谓超级对齐问题。

前些天OpenAI发布了他们关于超级对齐研究的第一篇结果还非常初步的论文，能明显感觉到问题的急迫性。他们的超级对齐研究团队今年刚成立，目前要解决的问题比初步成果更多更难，这可能也是他们设立千万美元奖金呼吁更多研究的原因。

文章中他们的一个判断很吓人，他们认为超级模型会在10年内出现，但超级对齐还是一个开放的困难问题。

如何用一个智力低的模型（包括人类）去让智力远超于自己的模型和自己的价值系统及意图对齐，是一个理论上和现实世界中没有答案的科学问题。

支撑OpenAI的对齐研究领导Jan Leike的一个信念是，超级模型解决困难问题是NP难度的，但低智力模型控制对齐超级模型是一个容易一些的P问题。当然也有更乐观一点的科学家认为，超级模型会帮我们解决这个问题。考虑到这个问题的极其重要性（High Stakes），多数科学家呼吁更谨慎的态度，呼吁增加研究投入，甚至放缓模型能力提升速度。

***未来模型发展的一些猜测***

**多模态**

年底发布的Gemini模型最令人注意的是处理多模态的方法，所有模态信号都统一成无差别的Token输入到统一的模型，DeepMind说是他们第一家这么做。

GPT4V的多模态能力也很强，尤其是图像顺序的恢复的能力让人惊讶，但如何实现的信息很少。

单从视觉理解看，拿一张经典的幽默场景的图片测试GPT4V，发现还不能达到人类的水平。但考虑到海量的视觉数据和不断提升的算力，这可能只是一个很短的时间问题。

未来的多模态信息对基础模型的能力是否有提升？尤其是智能能力，Ilya认为语言已经足够，视觉进来更好，但不是必须。但前两年LeCun认为视觉是模型常识构建的核心，他用的是幼儿早期学习的例子解释，不知道他这个观念是否修正？

**模型架构**

专家模型（MoE，稀疏模型）是否会成为下代的标准架构？这个问题是从一篇对GPT4的推测文章开始变热的，那个推测的逻辑是如果用密集模型（Dense Model）推理的延时在现有硬件条件下是不现实的。

那么延伸的问题是专家模型除了用于解决系统问题，是否还是提升基础模型能力的方法？公开的MoE模型不多，从Llama团队出去创业的Mistral AI开发的模型是个典型。

Transformer架构是现在所有模型的基础，会不会有大的演进？Ilya认为它还有很大的利用空间，当然也有优化的空间。

目前图像生成模型Diffusion是一个独立的架构，会怎么演进？会合并到transformer中吗？

**新的应用场景**

聊天机器人，代码助手，文案写作，图像视频生成，信息检索和处理等是目前主要流行的场景。

大模型能力的提升会解锁更多的场景，核心的逻辑是新能力和场景需求的匹配，以及对应的产品型态。

在这个意义上，未来已来。