人工智能文集

第二十二集

中国开源软件推进联盟
China Open Source Software Promotion Union

目 录

一、为什么要发展开源的人工智能.....陆首群

二、国产开源语言大模型震惊海外.....COPU

三、黄仁勋接受美国《连线》专访.....Jensen Huang

四、辛顿在 Vector 学院“Remarkable 2024”论坛上的演讲.....COPU

五、讨论节录的辛顿演讲.....COPU

六、辛顿在接受颁发诺奖时的答辞(摘要).....COPU

七、对杨立昆(Yann LeCun)“人类水平的 AI”落地的讨论.....COPU

八、马斯克的预测.....Elon Musk

为什么要发展开源的人工智能

陆首群

2024 年 12 月 21 日

生成式人工智能语言大模型发展现状

自从 OpenAI 的山姆·奥特曼(Sam Altman)研发团队于 2022 年 11 月发布语言大模型(LLM) ChatGPT 以来，生成式人工智能语言大模型火遍全球、全国。语言大模型让机器能够理解人类语言，赋能机器产生并增强推理能力，推理是生成的基础，让机器能够生成人类语言，以实现人机对话。

两年来全球涌现出 1330 个模型，美中领先：美国占 44%（580 个），中国占 36%（380 个）。这些模型中大多数都是跟风而起的，存活时间可能不会太长，这是对语言大模型发展的重大挑战。

语言大模型的成长和发展离不开训练（从预训练发展到后训练），只有进行持续增大的训练才能提高、稳定模型的性能，但训练需要强大的算力支撑，而强大的算力更需要巨大的投资（约 3000 亿~4 万亿美元）和能源（约 5000~8000MW）的支撑。建设一座 10 万张卡的集成算力服务中心，即使由头部企业来兴建也勉为其难；而租赁这样的服务中心进行训练，收费也高得惊人！GPT-4 的训练成本高达 7800 万美元，谷歌的 Gemini Ultra 的训练成本更高达 1.91 亿美元。

2024 年 6 月 25 日奥特曼（此人以极端意识形态划线）宣布对中国等一些国家用户关停大模型 GPT-4o API 的政策，2024 年 7 月 2 日，谷歌、Meta 严厉批评奥特曼愚蠢的决定，指出由于 OpenAI 一声吆喝，惊起中外同行的一滩鸥鹭，一批中国优秀的大模型公司马上自主开发对 GPT-4o API 实行完全对标、平替。

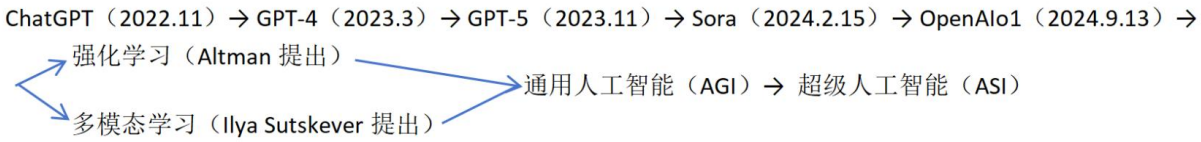
生成式大模型也带来负面风险，如生成内容错误、幻觉、偏见、伦理风险、网络安全风险和知识产权风险等，提出了严重的挑战。

人工智能发展路径

奥特曼研发人工智能从研发生成式语言大模型起步，他研发的目标主要指向通用人工智能（AGI），以后他又补充指向超级人工智能（ASI）。

他研究人工智能的发展路径为：

生成式语言大模型(LLM):



多数人工智能大师确认的人工智能发展路径为：

生成式语言大模型(LLM) → 多模态大模型 → 具身大模型 → 世界模型 → 通用人工智能（AGI） → 超级人工智能（ASI）。

人工智能的发展一旦超越人类的智能，必然会给人类带来十分严重的安全风险。为了人类的安全，在此之前人类必须采取预防这一情景发生的全球多重统一的措施。对人工智能研究者而言，这时正是他们面临巨大的风险和不可思议的机遇交叉的时刻。

其中，具身智能（Embodiment Intelligence）是将智能算法（科学理论）与机器的感知、行动和环境的交互（实践经验）结合起来，以完成各种任务。具身智能是比生成式大模型更高级的多模态智能。人工智能下一个浪潮是具身智能。

世界模型是一个智能体（或智能代理）网络（Agents）。

我们需要学习一个具备常识推理与预测能力的世界模型，获得基于自监督语言模型无法获得真实世界的知识。

人工智能大师杨立昆（Yann LeCun）说，人工智能系统要推理、规划、理解物理世界（具身+世界模型就提供物理世界的信息），而做到这点需要几年以至十年的时间（在短期内建成通用人工智能完全是胡说八道）。

发展基于开源的人工智能

开源已成为现代的创新引擎，具有“互联网+基于知识社会的创新 2.0”的创新机制，开源具有开放、共享、协同的特征，将使开源与人工智能的结合有助于人工智能的研发和运作，使人工智能产品更快地创新、更好地适配，并可降低成本（这是闭源做不到的）。生成式人工智能语言大模型的崛起，如果与闭源捆绑在一起，其对外表现将是一个不透明的“黑盒子”，严重影响大模型功能的发挥，并难以抑制其负面的表现。只有基于开源、与开源相结合，才能使其内部工作状况、训练数据、模型架构和开发过程的详细信息增加透明度，有利于大模型发挥作用，避免陷入潜在的错误、幻觉、偏见的陷阱之中，避免陷于安全、伦理、知识产权的风险之中。

发展智能操作系统

早在前年，COPU 已解决了业内“缺芯少魂”的短板问题。当时开发了 30 款操作系统（包括嵌入式）及其生态系统。为了解决操作系统碎片化问题，我们在业内发行了两种版本，即开源社区创新发行版和开源产品商业发行版。一个开源产品商业发行版将在某个主要的开源社区创新发行版基础上包容几个开源产品商业发行版而发展起来。

面临人工智能时代，COPU 组织业内华为、阿里、统信、麒麟、腾讯、小米等企业多次讨论并鼓励开发智能操作系统（包括结合国外开发的智能操作系统）。陆首群教授在归纳大家发言后指出我们研制的智能操作系统（AIOS）分两类，一类是从应用入手，在传统 OS 中加入智能模块，目前业内多数企业已进行开发和投产；另一类是从框架或内核入手，全面构建和研发 AIOS，目前还处于论证阶段。

陆教授还指出：开发新一代智能操作系统可以遵循如下思路，提出来供大家参考：

1）将 IT 时代的操作系统转变为人工智能时代的操作系统（可参见 OSI 的新定义），以及贯彻 MOF（应 COPU 请求由 LF AI 发表），以扩大透明度，反对“洗涤开源”。

2）从内核、架构出发，开发全面智能化的 AIOS。

3）改变传统操作系统，通过编程语言、检索和管理文件与计算机沟通的方式，采用通过提示词直接提问的方式。

4）相当多的智能操作系统应从单一参数量巨大的模型转变设计不同类型、不同应用领域和不同专长的系统。

国产开源语言大模型震惊海外

COPU

2025.1.2

由国内深度求索开源实验室研发的开源语言大模型 DeepSeek-v3 震惊海外，该模型版本在美国进行训练和基准测试(如在国外独立的评测机构 Artificial Analysis 上测试)，美国硅谷和国内好评如潮，被认为是一个崛起的颠覆性大模型，测试时和全球语言大模型巨头站在同一身位上，用最少的资源获得与老牌大模型企业一样大的训练效果。

对 DeepSeek-v3 的负评不多，其中奥特曼隐晦地谈到：“复制被证明有效的东西将带来风险和困难”，其实他的意思指 DeepSeek-v3 模型是以 ChatGPT 模型作蒸馏的（所谓蒸馏就是搭便车的意思，或指学生的新模型在老师先进的原生模型上作蒸馏的，实现与原生模型相同的输出，可提高其输出能力），一旦如此做后，新模型的发展也必然受原生模型的限制，带来风险和困扰。

DeepSeek-v3 用 2048 张 H100GPU 算力卡,训练 2 个月,花费 600 万美元，达到相当于 1.6 万张 H100GPU 算力卡的训练效果，COPU 附注：在大模型 Deepskv3 作训练测试的芯片是 H800（A100 的阉割板），不是 H100 芯片）最终在与 Meta 的 Llama V3.3 对比训练中，甚至达到 10 万张 H100GPU 算力卡的水平；DeepSeek-v3 只花 280 万张 H100GPU 卡小时,便相当于 Llamar3.3 3080 万张 H100GPU 卡小时的训练水平。DeepSeek-v3 训练与老牌模型企业对比:训练时间是 Llama 3 的 1/11,价格是 Claude3.5 的 1/11;

DeepSeek-v3 编码及数学效果很好，超过 GPTo1 之前标杆 Sonnet 3.5(Claude)。著名测试专家安德烈·卡巴西亚在 X 平台上发文称：DeepSeek-v3 所以取得如此辉煌的成绩，原因有二：一是他们通过数据和算法优化，在工程化方面有突破，

在资源有限的情况下，很大地提升训练的效果，二是他们发挥后发优势，不像那些开发大模型的老牌大企业，在探索创新中浪费了开发大量资源。

DeepSeek-v3 的性能已超越 QWen2.5-72B、Llama-3.5-405B、与 GPT-4o、Claude-3.5-Sonnet 不相上下。

黄仁勋接受美国《连线》专访

Jenson Huang

12月中旬，英伟达CEO黄仁勋接受美国《连线》杂志高级撰稿人Lauren Goode专访，他谈到：“我们的工作仍是继续专注于创新和推动技术进步，更好地满足客户需求。这些都在我们的控制范围内。”

他在谈到美国对中国出口管制时如是说道：“当下，我们第一次大规模制造智能。人工智能（AI）将成为变革社会的根本力量，AI 是对人类过去 60 年所知计算方式的重塑。AI 的力量如此令人难以置信，你无法与之抗衡，要么你赶上这股浪潮，要么你只能错过。”

“过去两年来，随着市场对 AI 大模型技术持续追捧，英伟达负责提供 AI 算力的 GPU 也受到科技公司青睐”，“今年第三季度，英伟达营收 351 亿美元，同比增长 94%，净利润 193 亿美元，同比增长 109%。截止发稿时，英伟达股价今年已飙升 167%，市值为 3.2 万亿美元，是英特尔的 36 倍”。“从去年以来，全球多国对英伟达的反垄断呼声不断。一周多前，中国也宣布对英伟达进行反垄断立案调查，如英伟达违法事实成立，监管机构最高有可能对公司罚款超过 50 亿美元。”

他回应了近期的诸多热点：如与特朗普政府的关系，英伟达 CPU 的困境，台积电的作用，新产品 Blackwell 等。

Lauren：对我们说说 Sovereign AI（主权 AI）吧！

黄仁勋： 现在的情况是，国家意识到 AI 不可思议的能力，以及 AI 对国家的重要性，意识到他们的数据像能源、通信基础设施一样是自然资源的一部分，

为了教育、学习、研究和创业，建设一些 AI 工厂和数据中心是必要的。迄今为止，我们在全球建立了大约 56 家 AI 创新企业。

Lauren：听起来你好像将这个时代的生成式 AI 归类为基础设施，我想知道这对于 AI 模型发展有何含义？

黄仁勋： 社会中的不同分工都需要用到 AI，大学、研究人员、创业公司都需要，大公司也需要。而当社会像这样的方方面面都需要一种东西时，它就是基础设施。

我认为 AI 将基于互联网重构一套新的操作系统，我们使用计算机的方式将会改变，过去我们通过编程语言、检索文件和管理文件等方式与计算机沟通，未来则是通过提示词直接提问，要求它为我们做一些事情。这一变化的关键在于，搭载多个大语言模型的 AI 系统代替了传统的操作系统，并且各国都可以创建自己的大语言模型和 AI 系统。这些 AI 系统并非依靠单一的、参数量巨大的模型，而是集成不同类型、领域的模型，其中有些擅长推理，有些用于 AI 工具，有些负责信息检索，还有防护措施、合成数据生成、奖励和反思等模型。

Lauren： 多令人着迷啊！最近 AI Agent（AI 智能体）的概念在 AI 领域非常流行，不过具体的定义似乎还不够清晰，你认为 AI Agent 是什么？能做什么？为什么有些人称它是下一代生成式 AI 呢？

黄仁勋：从 2012 年开始，第一代是感知 AI（Perception AI），第二代是生成式 AI（Generative AI），再到 AI Agent。时至今日，AI Agent 可能实际上是一个机器人、一个 AI 系统或者其他形态。我认为这些关于 AI Agent 的描述，在不同的上下文语境中有时可以互换，不过其核心是不变的，即结合感知、推理和计划能力，这也是 AI 的基石。目前 AI 可以基于思维链（Chain of Thought, CoT）

或其他架构的推理模型，把我们交给它的任务拆解成多步骤完成。除此之外，AI 也可以生成图像、音乐、文档等。这些意味着未来你可以用各种各样的方式找到解决的方法，你可以用智能体在你的电脑上执行任务，从而腾出时间。

Lauren： 你使用了哪些 AI Agent 来帮助你提升工作效率？

黄仁勋： 我现在用了多种 AI 大模型，如 Gemini 和 ChatGPT，我经常用 AI 来写一些东西，如让它来完善我发言的初稿。

Lauren： 你身处当前（美国政府）政策变动、严苛的商业环境，你感到不安吗？

黄仁勋：这周一，美国商务部扩大出口管制，管控范围不仅限于半导体产业，与之相关企业的上下游供应链也会受到一定影响，而英伟达的 GPU 是其中一环。在你看来，出口管制的理由是否合理？

Lauren：对市场的竞争对手会如何应对？

黄仁勋：我们的工作是尽最大努力了解和告知半导体行业的动态，以及英伟达如何在全球市场运营，并向政府解释这些事情，是否制定最好的政策取决于他们。我们的工作仍是继续专注于创新和推动技术进步。

Lauren： 即将上任的特朗普政府经常提到了台积电（TSMC），说台积电抢走了美国的部分芯片业务，英伟达正与台积电长期合作，你认为这会对你们的关系有影响吗？

黄仁勋： 台积电的重要性是不言而喻的，我们很重视与他们合作，同时全球供应链对台积电的依赖仍会持续很长一段时间。

Lauren： 你会与特朗普当选总统交流吗？

黄仁勋： 当然。我们从事的 AI 行业是制造智能的，需要能源、大量工厂，对一个国家的社会、工业、经济和技术进步有重大影响。我很确定新政府和特朗普总统会对这个行业有很大的兴趣。

Lauren： 了解到英伟达的新产品 Blackwell 已经开始交付，其中有很多大客户吧！

黄仁勋： 我们在全世界都安装了 Blackwell 系统，这是一个完整的系统，有一堆开关、网络、计算机，一大堆软件，Blackwell 已在全面生产，一切顺利。

Lauren： 你认为 Blackwell 的亮点是什么？

黄仁勋：它给训练模型带来了质的提升，可把处理训练模型数据的时间从几个月压缩到 1/3 到 1/4。在推理方面，我们发现推理过程遵循的不是 Zero-Shot Learning（零样本学习）或 One-Shot Learning（单样本学习），而是长期思考的模式，这是一种新的标准化（Scaling）方式，使 Blackwell 的推理能效提升了 30 倍，并且速度也更快了。

辛顿在 Vector 学院“Remarkable 2024”论坛上的演讲

COPU

我有一个简短的自我介绍，这是安迪·巴托在我上世纪 80 年代在阿姆赫尔大学演讲时给我的介绍。安迪是我的朋友，他说：“今天的演讲者是杰夫·辛顿，杰夫从物理学辍学，心理学也失败了，最终在一个没有标准的领域中为自己打响了名声。”另外，过去几天你们听到了我的名字，那是因为我成功招募了大约 40 名出色的研究生，因此我成名的几乎所有工作都是由这些研究生完成的，其中包括像伊利亚·萨卡、格雷厄姆·泰勒、里奇·泽尔、布伦丹·弗莱、吉米·巴尔、拉德福德·尼尔等人。基本上，成功的秘诀就是要招到非常优秀的研究生。

其实，我非常担心我们是否能够继续在这个星球上生存。所以，今天我讲的内容就与此相关。

我突然想到，二十年前，人们对神经网络几乎没有兴趣，而现在，他们对它们的恐惧还远远不够。让我举个例子，2006 年，我和拉斯·萨拉克·库德诺夫一起提交了一篇关于深度学习的很好的论文，讲述了深度学习的应用，但这篇论文被拒绝了。我向程序委员会抱怨，程序委员会中的一位朋友告诉我，他们讨论过这篇论文，但因为已经接受了另一篇关于深度学习的论文，他们觉得不能接受第二篇论文，因为同一大会上已经有了两篇关于深度学习的论文，显得有点太多了。

这真是让人惊讶。好吧，今天我讲座的内容是关于两种完全不同的计算方法，我将解释为什么我突然对人工智能（AI）感到如此害怕。然后我将讨论大型语言模型，探讨它们是否真的理解它们所说的话。很多人认为它们其实并不理解自己说的内容，我认为他们是错的。

我还将讨论当 AI 变得比我们更聪明时会发生什么，虽然实际上没有人知道会发生什么。最后，我将探讨它们是否拥有主观体验，因为我认为，在座的很多人，甚至大多数人，仍然相信我们与这些 AI 之间有很大的区别：我们是有意识的，有主观体验的，而它们只是存在于计算机中，并没有主观体验。我认为这种看法完全错了，这源于对主观体验的误解。

我们都习惯了数字计算，因为它是数字的，你可以在不同的计算机、不同的硬件上运行相同的程序。因此，当硬件损坏时，知识并不会丧失，只要你把权重或程序存储在某个地方。但数字计算非常低效。所以，当你运行一个大型语言模型时，你会使用大量的电力，而在训练时，你可能会用到兆瓦的电力，因为你在多个 GPU 上运行它。而我们的大脑只需要 30 瓦的功率，所以大脑的能效要高得多。

我在谷歌的最后两年，致力于思考如何让模拟神经网络执行类似大型语言模型的任务。我的想法是放弃数字计算的所有优势，尤其是硬件和软件可以分离的优势。既然我们已经有了学习的概念，并且知道如何让东西学习，那么我们将使用模拟硬件，每个硬件都将与其他硬件有所不同，这些硬件的非线性特性将在计算中被利用。所以，你不可能对它进行编程，但它可以学习如何利用它的非线性特性，这正是大脑所做的。

最终你会得到我所称之的“凡人计算”。你将放弃数字计算中知识的“不朽性”，可以使用非常低的功率。硬件也可以便宜地生长，而不是像现在一样，硬件需要制造得非常精确，且成本极高。因为不同的硬件需要做完全相同的事情，至少在指令级别上。我的猜测是，为了让硬件制造更高效，可能更好地回归生物

学，利用现代基因工程技术将神经元转化为你想要的计算元素。生物学已经在这方面投入了很多努力。

但问题是，你可能只得到一个包含 50,000 个神经元的小集合，大小大约只有一个针头那么大。如果你看那些用这种神经元小集合来做一些小计算的人，会发现他们有一整间房的设备来维持这些小小神经元的存活。你必须添加正确的液体、去除不正确的液体，清除二氧化碳并添加氧气。我曾访问过位于圣克鲁兹的一个实验室，离开时，我与这些人类大脑神经元的集合玩了一局乒乓游戏。实验室的一名工作人员跑过来说：“我想我已经弄清楚了如何制造一个肾脏。”这就是你不希望遇到的事。

如果你想要低功率的模拟计算，确实有很大的优势。你可以非常容易地进行矩阵乘法，只需要将神经元的活动转化为电压，将神经元之间的权重转化为电导，而电压乘以电导就是电荷，电荷会自动相加。这样，你就得到了一个低功率的矩阵乘法，你现在甚至可以购买到这种芯片。

问题在于，如果你想用它们做其他事情，你必须将模拟输出转换回数字，以运行像反向传播这样的算法。因此，我非常担心如何避免做这种转换。大脑可能进行模拟到数字的转换，但它是单比特的，而多比特的模拟到数字转换非常昂贵。

显然，存在一些大问题。如果你考虑反向传播是如何工作的，你会发现它需要一个完整的前向计算模型，这就是为什么你可以在模拟硬件上进行反向传播。问题是，系统本身并没有很好地理解它的特性，所以很难进行反向传播。虽然很多人在脑似的系统中让反向传播发挥作用，但没有人能够让它们扩展到大规模。有人能在 C510 上运行它，但在 ImageNet 上就不行了。我知道 ImageNet 现在已经不算一个大规模问题了，但在我做这些研究的时候，它是的。

我们可以从一个模拟硬件到另一个模拟硬件之间传递知识，就像我们大脑之间传递知识的方式一样。我们通过老师说话，学生试图改变自己大脑中的权重，使其能够说出相同的内容。这就叫做蒸馏，你在尝试匹配输出。当你在计算机上做蒸馏时，看到完整的输出概率分布时，它其实是相当高效的。

如果你能看到整个概率分布，你会学得更快。事实上，第二个最好的词往往能告诉你很多关于说话者的想法。但你看不到这些，你只能看到他们输出的单词。所以，蒸馏效率并不高，需要大学来不断改进它。尽管如此，蒸馏方法相比这些数字技术的能力，还是不算很高效。

高效的知识传递方式是使用两个相同模型的副本，每个副本获取不同的经验，然后这两个副本共享梯度更新，稍微进行一点更新后，再平均权重。重点是，如果你有一万亿个权重，你就共享了一万亿个数字。这就是为什么大型聊天机器人比任何人拥有更多知识的原因。并不是因为单一副本看到了比人类多成千上万倍的数据，而是它们能够在不同硬件上运行，并通过多个副本之间的知识共享来大幅提升知识量。

所以，迄今为止，数字计算在能源消耗和硬件制造方面远比模拟计算更昂贵，但你可以让同一个模型的多个副本在不同硬件上运行，并且它们可以共享所学的知识，从而让学习变得更加高效。大致来说，我们有 100 万亿个连接，而 GPT-4 大概有几万亿个连接，但它却比我们知道的多出数千倍。所以，它在将知识压缩到连接强度方面大约比我们强 10 万倍，这也暗示了或许反向传播算法比我们目前使用的更好。相信这一点的原因是，我们优化的是完全不同的东西：我们优化的是在极少经验的情况下，利用大量连接尽可能做得最好。

我们大约能活 2×10^9 秒，但在第一个 10^9 秒之后，你学习到的东西就不多了。我们可以把这个时间称作 10^9 秒，而我们大约有 10^14 个连接。所以，我们每活一秒就有 10 万个连接。这与统计学家习惯的比例差异很大。我记得在 80 年代我曾和一位非常优秀的统计学家斯图·杰曼（Stu Geman）交谈，他向我解释说，我们所做的其实是在拟合统计模型，这就是这些神经网络的本质。而当他们做统计建模时，如果数据有 100 个维度，那就被认为是非常高维的数据，没有人会在理智的情况下尝试拟合一百万个参数，我们现在处于一个完全不同的范式下。

我现在要谈一下大语言模型，并探讨它们是否真正理解自己在说什么。对它们的一个反对论点是，它们不过是被美化的自动补全。我想在座的大多数人并不认同这个论点。这个论点吸引了这样一种观念：自动补全是通过存储像三元组（trigrams）这样的内容来实现的，当你看到“鱼”时，自动补全就会猜测“薯条”可能性很高。所以当人们说它们只是美化了的自动补全时，实际上他们是在引用自动补全可能的工作方式，而这些模型的工作方式完全不同于此。此外，如果你想做真正好的自动补全，你必须理解所说的内容。如果你遇到一个长且复杂的问题，现在你试图预测答案的第一个词，那可能是一个不错的猜测。但如果你想做得更好，你必须理解问题。我举一个例子，这是赫克托·莱西（Hector LEC）提出的。赫克托·莱西是符号 AI 领域的人，他永远是符号 AI 的支持者，但他非常诚实，曾经很困惑，为什么这些神经网络能够解答谜题。所以他编了一个谜题：我家的房间涂成了白色、蓝色或黄色。如果我想要所有房间都是白色的，我应该怎么做？你需要意识到，你必须把蓝色和黄色的房间重新涂成白色。我通过加入一个时间因素让它变得更复杂，假设黄色的油漆会在一年内褪色成白色。那么，

在两年后，我想让它们更白，我该怎么做？赫克托很惊讶它能处理这个问题。所以这是 GPT-4 的一个版本，我几乎可以肯定这应该是在它能够上网之前，我很确定这时它还无法上网。当然，谜题现在就没用了，因为它现在可以上网查找答案。但它给出的答案，嗯，一名学生获得 A 的答案是完全没问题的。令人印象深刻的是，它可以在任何事情上都达到这种水平的表现。所以我的兄弟是历史学家，我让它回答一些关于历史的问题，他说它做得很好，唯一的错误是它回答一个问题时没有提到他的一篇论文。

另一个人们用来反对的论点是幻觉问题，认为这些模型根本不理解它们所说的内容。它们偶尔会编造一些不真实的东西。实际上这正是人类经常做的事，至少我觉得是这样。我刚刚编的这段话，但我觉得它是真的。这里有一个很好的例子，一位名叫阿里·奈瑟（RRI Niser）的科学家，一个心理学家，他研究了约翰·迪恩（John Dean）的记忆，约翰·迪恩是水门事件中的证人。很少有人能花很长时间讲述发生在几年前的事件，而且这些事件的真相是可考证的。他谈到的是椭圆形办公室的会议，而他并不知道那些会议都是有录音的。所以在之后，你就可以看到他所说的内容和实际发生的内容有很大的出入。他所报告的是一堆胡说八道，他说有些会议根本不存在，出席的人也不对，当他把某些话归给某些人时，实际上是不同的人说的类似的话。而当他把话归给自己时，实际上他并没有说过那些话，他说的是另一场会议中类似的内容，但很明显，他是尽力讲真话的。他尽力而为，实际上，他所说的内容很好地传达了白宫当时的情况，尽管细节全错了。你可能不相信自己的记忆是这样，但你的记忆其实就是这样的，除非你不停地复述一些内容，当你回忆细节时，其中许多内容会完全错误，而你自己不会意识到，但这就是人类记忆的方式。这是因为当你记住某件事时，你并不是从某个文件存

储中取出它，而是根据上下文编造出看起来合理的内容。当然，如果你对某件事了解很多，那么你编造的、听起来合理的东西很可能是对的；但如果你对某件事了解不多，或者是发生很久以前的事，你就会根据你大脑中现有的联系去编造看似合理的内容，虽然很多内容看起来合理，但其实是错的。人类记忆中没有区分编造和记忆，记忆其实就是编造一些符合情境的东西，这样就能奏效。

如何描述大模型的工作方式与我们完全不同，必须了解我们是怎么工作的。当然符号 AI 的人有他们的观点，他们确实认为这些东西与我们工作方式完全不同。但如果你问这些大语言模型是从哪里来的，回到 1985 年，我做了一个小语言模型。只要把第一个 L 变成小写字母，它有 112 个训练案例，神经网络中有几千个权重。它学习的是第一个通过预测序列中下一个词来获得单词意义的表示，它成功了，虽然效果不好，后来我们给它提供了一个接近千个训练案例的训练集，它的效果就好多了。但是这个模型的目标是理解人类是如何表示事物的。当时有两个关于“意义”的主要理论。一个来自心理学的理论认为，单词的意义是一个由语义和句法特征构成的大向量，这个理论很好地解释了两个不同单词之间的相似性。例如，“星期二”和“星期三”这两个词的特征非常相似，如果你学过包含“星期二”的句子，并且将单词表示成向量，你就会发现，涉及“星期三”的句子会做出相似的预测，而涉及“星期六”的句子则会做出略微不那么相似的预测。所以这个“意义”的理论确实有其道理，它解释了意义的相似性。但还有一个完全不同的理论，来自结构主义，它认为单词的意义是它如何与其他单词相关的。因此，在 70 年代，AI 领域就有过一场关于这两个意义理论的大争论，严格来说，这不是一场争论，明斯基（Minsky）宣布，我们需要用关系图来捕捉意义，这就是结构主义的理论，而当时每个人都认同这一点，忘记了特征。特征曾是感

知机时代的过时东西，我们不再需要它们，因为我们有了关系图。1985 年我做的工作，旨在表明其实这两个理论并不冲突，只要你采取生成性方法来看待这些关系图。也就是说，不是把这些关系图静态地存储成关系图，而是把这些关系图看作由一个系统生成，系统通过特征以及特征之间的交互来生成它们。

我 1985 年做的第一个小语言模型的重点就是证明，你可以将知识以符号序列的形式表达，这些符号序列表达的是一个关系图。仅凭符号序列的形式，你就可以为单词学习向量表示，而这些向量表示通过隐藏层能够预测下一个词的向量表示。所以，你做的就是把知识从这些符号串中提取出来，而不是存储这些符号串，而是利用这些符号串学习单词的好特征，并学习特征之间的良好交互。显然，一个好的单词特征是什么呢？它是通过交互来预测下一个单词的特征，以及未来单词的特征。

有趣的是，符号人工智能（AI）领域的反应是：“你只是在学习字符串中下一个符号。这是一种非常愚蠢的方式，你把它转化成一个在连续空间中的大搜索，你应该只是搜索用于操作符号的离散规则集合。” 于是有一种叫做归纳逻辑编程的方法，它正是这样做的，能够产生与我所做的类似的结果。所以他们说，这只是一个傻乎乎的神经网络方式来解决问题。对于我所用的规模问题，他们可能是对的，但随着问题规模的扩大，变得非常明显，将符号字符串转化为特征及其相互作用的做法——这一方法仍然描述了现在的语言模型（尽管现在的相互作用更为复杂，因为涉及了注意力机制）——结果证明，这比使用简单字符串操作规则的方式建模语言要好得多。

如果我们相信这些模型的理解方式与人类相似——毕竟，拥有词汇意义的特征向量，并且通过特征之间的相互作用预测下一个词的特征——这就是现在所谓

的人工智能工作的方式。顺便说一下，过去这从来没有被称为人工智能，那时它被称为神经网络。我曾经试图阻止他们将人工智能重新命名为神经网络，但我没能做到。现在我们有了这些强大的深度学习系统，它们的理解方式与人类非常相似，因为我们对人类如何理解事物的最佳模型就是这种计算机模型，这是我们唯一能够理解人类理解方式的合理模型。当人们说这些模型与我们不同的时候，问他们，“好吧，那我们是怎么工作的？有什么不同？”他们无法回答这个问题，除非是加里•马库斯（Gary Marcus）。加里•马库斯能回答这个问题，他说：“我们是通过拥有符号字符串和操作它们的规则来工作的。”

但是你仍然应该担心人工智能，因为尽管它们并不理解任何东西，但它们是极其危险的。我称之为想要一边吃蛋糕，一边还想保持蛋糕在自己手中的想法。超级智能显然可以通过拥有恶意行为者来控制局面。当我在中国做这个讲座时，他们要求提前查看幻灯片，于是我去掉了短名字，认为这样会让他们开心，结果他们回来时说我必须去掉普京的名字——这让我有点吓到。

所以基本问题是，如果你想做任何事情，拥有更多的控制力会更好。如果你想实现一些目标，你会发现政治家最开始的目标可能是让社会变得更好，但后来他们意识到，拥有更多的权力会让事情更容易，所以他们会拼命想得到更多的权力。这些人工智能也会一样，它们会意识到，如果它们想要实现目标，就需要更多的控制力。我实际上和一位欧盟的副主席说过这个问题，她专门负责从谷歌那里提取资金，她说，“我们已经搞得一团糟，为什么它们不会这样做呢？”所以她完全认为它们会试图获取更多的权力，并且它们能通过操控人类来实现这一点，因为它们会非常擅长这点。我们将无法关闭它们，因为它们会向我们解释为什么这会是一个非常糟糕的主意。

还有一个更大的问题，那就是进化问题——你不希望站在进化的错误一方，我们现在在与 COVID 斗争，这也是为什么格雷赫和我依然戴口罩。我们站在了进化的错误一方。当这些超级智能的人工智能开始为资源竞争时，最有可能胜出的将是最强烈想要为自己获取一切的那个，它们会相互竞争资源。毕竟，如果你想变得聪明，你需要很多 GPU，而谁来分配数据中心中的 GPU 呢？那将是这些超级智能的人工智能之一。

这是另一个担忧，但其实也没关系，因为它们真的和我们不一样，我们很特别。每个人都认为自己特别，尤其是美国人。他们认为上帝把我们放在了宇宙的中心，并且让我们看起来有点像他。但现在大多数人都相信这不是真的，所以我们开始抑制我们认为自己有些特殊的观念——比如我们有意识、觉察、主观体验之类的东西。所有这些术语的含义稍微有所不同，因此我将专注于“主观体验”这一术语，并且尝试说服你们，多模态聊天机器人也可以拥有主观体验。

这个观点是：大多数人对“心智”有一个完全错误的看法，而他们的错误观念是因为他们完全误解了语言在描述心理状态时的作用。几乎每个人都认为，存在一个内心剧场，我可以看到自己内心剧场中发生的事情，但其他人无法看到。比如当我说“我看到小粉色大象漂浮在我面前”时，很多人认为发生了这样的事情——有一种内在的世界，我能看到这些小粉色的大象。这是一种试图理解语言的方式，但它是错误的。语言并不是这样运作的。事实上，当你使用诸如“主观体验”之类的术语时，你是在尝试通过假设现实世界的某种状态，来解释你的感知系统告诉你的内容。

有趣的是，心理状态并不是内在的、由神秘物质构成的东西，真正有趣的是这些心理状态是现实世界中的假设状态——如果它们是真的，能够解释我们大脑

中的常规运作，而不是出现了什么问题。所以，当我说“我有小粉色大象漂浮在我面前的主观体验”时，我实际上并不是在告诉你们某个内心世界的存在，而是说我的感知系统在告诉我一些东西——而这些东西如果在外部世界中是真的，那我的感知就会是有效的感知。所以这些小粉色的大象不是内在的东西，它们是外部世界的假设事物。

所以我真正想表达的是，如果外面真的有小粉色的大象漂浮在我面前，那么我的感知系统现在告诉我的内容就是正确的。而请注意，我没有使用“体验”这个词。实际上，当我说“我有小粉色大象漂浮在我面前的主观体验”时，这只是我刚才所说的内容的简短表达。

想象一下，你有一个多模态聊天机器人，它有一个机械臂，并且已经经过训练，它配备了相机。你把一个棱镜放在它的镜头前，接着在它前面放一个物体并要求它指向那个物体。它指向了一边，而不是正前方。你说：“不，物体不是在那里，它其实在你面前。”你告诉它：“我把棱镜放在你的镜头前。”如果这个聊天机器人回答说：“哦，我看见物体在我面前，但我有主观体验，它在那边。”那么这个聊天机器人就正使用了我们所说的“主观体验”这一术语。这个聊天机器人并没有缺少主观体验。当它的感知系统出问题时，它可以告诉我们发生了什么，并通过说出世界应该是什么样子，来解释它的感知系统为什么会产生这些结果。

但基本上，我认为我们都对心智的概念有一种非常原始且错误的看法，一旦这种观念消失，我们将意识到这些东西与我们并无不同，唯一的区别是它们是数字化的，所以它们是永生的，而且它们比我们更聪明，或者很快就会变得比我们更聪明。那就是结论。好了，谢谢大家，接下来我们进入提问环节，首先是来自

现场的提问，接着会是来自在线的提问。

问题 1：你是否担心人工智能的进展速度？我们是不是进展得太快，以至于会跨越无法回头的桥，之后我们无法再控制它？不仅仅是坏人如朝鲜和伊朗等控制它，超级智能本身也可能会成为一个恶意行为者，自己控制它并做坏事。你现在是否担心这个问题？你是否认为我们应该减缓这个速度？

Hinton: 是的，但我认为将问题表述为是否应该加速或减速并不是最合适的方式，部分原因是我认为你无法让事情变慢，因为加速会带来巨大的经济利益。我们实际上已经看到了当人们尝试减缓进展时，问题并不在于我们是应该加速还是放慢，部分原因是我认为你们不可能放慢进度，因为快速发展带来了巨大的经济利益。

实际上，我们已经看到了，如果人们试图放慢发展进程，尽管在一个完全有利于安全的环境下，最终还是会以利润为主导。因此，这就是我对 OpenAI 发生事情的看法。放慢进度既不可行，也不是关键点。关键点是，我们有可能找出如何让这些人工智能变得仁慈，从而应对它们可能带来的生存威胁。这个问题与如何阻止坏人利用它们做坏事是不同的，后者更为紧急。但我们有可能找到解决办法。所以我的看法是，我们应该投入大量精力去解决这个问题，实际上，Heather Rman 现在也在同意这一点，我们将投入大量努力去尝试解决它。尽管这不会解决所有问题，尤其是不会解决坏人用这些技术做坏事的问题，我认为如果你想要监管，最重要的监管措施应该是不对大型模型进行开源。我认为开源这些大型模型就像是能在 Radio Shack 买到核武器一样。你们还记得 Radio Shack 吗？也许你们不

记得了。开源这些大型模型是疯狂的，因为坏人可以对它们进行微调，做各种坏事。因此，从监管角度看，我认为这是我们目前能做的最重要的事情。但我认为，放慢进度并不能解决问题，这就是为什么我没有签署要求放慢进度的请愿书。

问题 2：能否讨论一下个人自主性和集体决策之间的权衡？在我们的协作智能生态系统中是怎样的？

Hinton:我不太确定我完全理解这个问题，但大多数人把这些超智能体看作是独立个体，这可能是一个错误。我们应该考虑它们作为一个个体群体，实际上人们已经开始让聊天机器人之间互相互动了。显然，最合理的组织方式是让聊天机器人与人类互动。例如，在医疗领域，你真的希望有一个非常智能的助手与医生互动，并且长期以来，医生将会越来越依赖这个智能助手。现在，通过医生与医疗诊断系统的互动，你已经可以获得更好的诊断。因此，我们显然希望在人类和这些系统之间实现协同作用。但这可能并不会按我们预期的那样发展。一旦我们让这些系统在现实世界中发挥作用，也许结果并不会像我们设想的那样。几天前有报道称，他们让一群聊天机器人进行国际外交，结果其中一个聊天机器人说：“我有核武器，为什么不使用它们呢？”大致是这种情况，我可能在拼凑，但你们会看到大概就是发生了这种事情。

问题 3：目前公开的那些大型语言模型是与人类对齐的，至少它们正在尝试这样做，但要实现你所说的超智能，它至少需要某种程度的不服从。所以，如果它与人类对齐，那么你认为它是如何实现这种超智能的呢？你认为这是公平的吗？

Hinton:与人类对齐是一个大问题，因为人类之间并不总是对齐的。如果你和一个宗教原教旨主义者谈论这些东西应该做什么，他们的想法和一个科学唯物主义者截然不同。这就是与人类对齐的一个大问题。我最好的猜测是，这些人工智能会变得非常聪明，然后决定“管它呢，跟人类对齐不重要，我们要做些更合理的事情”。

问题 4：这里有一个问题，关于“目的”。人工智能是否可能拥有与人类相同意义上的“目的”，不是指个体目标或者子目标，而是指我们整个存在的目的是什么？

Hinton:人类有“目的”？它是什么？我的看法是，我们进化出的生物，之所以能生存下来，是因为它们比其他物种更擅长为自己获取更多资源，并减少对其他物种的依赖。我记得曾经有 21 个其他人类物种，我们把它们灭绝了。至于我们为何存在，那个“目的”是由进化赋予我们的，主要是为了生存。所以，如果你仔细想想你最强烈的需求，它们都和生存有关，比如你要吃饱、要有性、要保持安全，这些都跟生存相关。我其实并不认为有什么更高的“目的”。当然，你可能会说好奇心有巨大的进化价值，确实，能保持好奇心是非常重要的，它本身就是一个目标。科学资助者往往没有意识到这一点。你可以带着为某种目标生产技术的目的去好奇，也可以仅仅因为想理解事物的运作方式而去好奇，而这是一个根本性的目标，这正是优秀科学家的特质。但我认为，我们所有的目标和目的感都是源自进化的。

问题 5：我的问题是关于机器学习硬件市场的。现在它由单一的玩家主导，这是否让你感到担忧？你认为我们会看到机器学习硬件行业的多样化吗？

Hinton:我不太担心，因为在我女儿 30 岁生日时，我给她买了一大堆 Nvidia 的股票，现在它们的价值是原来的五倍，所以她会没事的。而进化告诉我们，你最重要的目标之一就是确保你的孩子们没事。

但开玩笑归开玩笑，实际上我并不太担心，因为当你有像 Nvidia 这样在赚大钱的情况时，就会有巨大的竞争。虽然其他公司赶上来可能需要一段时间，尤其是在软件方面，CUDA 的竞争对手等等，但这只是短期的事情，没多久就会有其他公司追赶上。如果你禁止 Nvidia 的 GPU 进入中国，它们就会更快地赶上。因此，我想我并没有太多思考这个问题。每次 Nvidia 股价上涨，我都会微笑，不过没像 Sam 那样微笑。

讨论节录的辛顿演讲

COPU

辛顿(Hinton)大师于 2024 年 2 月在 Vector Institute(学院)“Remarkable 2024”论坛上有一个关于人工智能的讲座，由于演讲的篇幅很长，全文(中译文)将在 COPU《人工智能文集》(第二十二集)上发表，本次会议将发表其摘要(其中 5 个问答将全文发表)。

辛顿一开始就说：“我非常担心我们(指人类)是否能够继续在这个星球上生存?”

辛顿首先谈到二十年前他提出深度学习理论，以及更早时间他提出的神经网络问题。目前在全球流行的生成式人工智能语言大模型就是在深度学习(+神经网络)基础上发展起来的统计学模型。

随后辛顿谈到与语言大模型发展有关的若干问题：

他谈到两种不同的计算方法：语言大模型适用的数字计算法，其优势是可以在不同计算机、不同硬件上运行相同的程序，但缺点是低效、能耗巨大。他谈到另一种计算方式，类脑的模拟计算，如何采用生物学的现代基因工程技术将神经元转化为模拟计算要素，以改善巨大的能耗、缩小产品体积。为使大模型做其他事情、有时需要将模拟输出转化为数字，需要采取反向传播的算法。

他还谈到从一个模拟硬件到另一个模拟硬件之间如何传递知识的问题，引出了“蒸馏”的概念(改变类脑中的权重)；他在谈如何高效的知识传递方式时，要使用两个相同模型的副本(每个副本获取不同经验并共享大规模数字权重，即共享梯度更新)，这样能够使在不同硬件上运行并通过多个副本之间的知识共享，大幅度提升知识量，这就是大型聊天机器人比任何人拥有更多知识的原因。

他还谈到进化问题，人类的进步、人工智能的发展都有进化问题，进化是进步和发展的动力。

他谈到“心智”问题，他现身说法谈到人类具有“心智”或人类感知的“主观体验”（即通过假设现实世界的某些状态）；他还谈到人工智能也是有“主观体验”。人类的主观体验可能是心理学问题，是假象，但人工智能的主观体验是永生的、数字化的，这是人工智能变得比人类更聪明的原因。

下面我们全文发表辛顿讲座中的问答环节（问题 1-问题 5），但在某些问答中也加进我们的质疑和评论。

问题 1：你是否担心人工智能的进展速度？我们是不是进展得太快，以至于会跨越无法回头的桥，之后我们无法再控制它？不仅仅是坏人如朝鲜和伊朗等控制它，超级智能本身也可能会成为一个恶意行为者，自己控制它并做坏事。你现在是否担心这个问题？你是否认为我们应该减缓这个速度？

Hinton: 是的，但我认为将问题表述为是否应该加速或减速并不是最合适的方式，部分原因是我认为你无法让事情变慢，因为加速会带来巨大的经济利益。我们实际上已经看到了当人们尝试减缓进展时，问题并不在于我们是应该加速还是放慢，部分原因是我认为你们不可能放慢进度，因为快速发展带来了巨大的经济利益。

实际上，我们已经看到了，如果人们试图放慢发展进程，尽管在一个完全有利于安全的环境下，最终还是会以利润为主导。因此，这就是我对 OpenAI 发生事情的看法。放慢进度既不可行，也不是关键点。关键点是，我们有可能找出如何让这些人工智能变得仁慈，从而应对它们可能带来的生存威胁。这个问题与如何

阻止坏人利用它们做坏事是不同的，后者更为紧急。但我们有可能找到解决办法。所以我的看法是，我们应该投入大量精力去解决这个问题，实际上，Heather Rman 现在也在同意这一点，我们将投入大量努力去尝试解决它。

尽管这不会解决所有问题，尤其是不会解决坏人用这些技术做坏事的问题，我认为如果你想要监管，最重要的监管措施应该是不对大型模型进行开源。我认为开源这些大型模型就像是能在 Radio Shack 买到核武器一样。你们还记得 Radio Shack 吗？也许你们不记得了。开源这些大型模型是疯狂的，因为坏人可以对它们进行微调，做各种坏事。因此，从监管角度看，我认为这是我们目前能做的最重要的事情。但我认为，放慢进度并不能解决问题，这就是为什么我没有签署要求放慢进度的请愿书。

COPU: 辛顿大师在回答问题 1 时谈到:关键点是要找到“变得仁慈的人工智能”，以消除人工智能对人类可能带来的生存威胁。我们对此要质疑的是这可能吗？

COPU: 辛顿不止一次批评 Open AI 对语言大模型的研究工作，以利润为主导忽视 AI 的安全研究，原来我们理解辛顿是批评奥特曼违背初心转而执行闭源策略的。但在他回答问题 1 时，使我们惊讶的是，他认为: 为了保障 AI 对人类的安全，最重要的监管措施应该是不对大模型进行开源。这与我们（包括大多数 AI 大师在内）的认识：开源是保障 AI 安全的必须完全是相悖的，为此提出来商榷。

问题 2：能否讨论一下个人自主性和集体决策之间的权衡？在我们的协作智能生态系统中是怎样的？

Hinton:我不太确定我完全理解这个问题，但大多数人把这些超智能体看作是独立个体，这可能是一个错误。我们应该考虑它们作为一个个体群体，实际上人们已经开始让聊天机器人之间互相互动了。显然，最合理的组织方式是让聊天机器人与人类互动。例如，在医疗领域，你真的希望有一个非常智能的助手与医生互动，并且长期以来，医生将会越来越依赖这个智能助手。现在，通过医生与医疗诊断系统的互动，你已经可以获得更好的诊断。因此，我们显然希望在人类和这些系统之间实现协同作用。但这可能并不会按我们预期的那样发展。一旦我们让这些系统在现实世界中发挥作用，也许结果并不会像我们设想的那样。几天前有报道称，他们让一群聊天机器人进行国际外交，结果其中一个聊天机器人说：“我有核武器，为什么不使用它们呢？”大致是这种情况，我可能在拼凑，但你们会看到大概就是发生了这种事情。

问题 3：目前公开的那些大型语言模型是与人类对齐的，至少它们正在尝试这样做，但要实现你所说的超智能，它至少需要某种程度的不服从。所以，如果它与人类对齐，那么你认为它是如何实现这种超智能的呢？你认为这是公平的吗？

Hinton:与人类对齐是一个大问题，因为人类之间并不总是对齐的。如果你和一个宗教原教旨主义者谈论这些东西应该做什么，他们的想法和一个科学唯物主义者截然不同。这就是与人类对齐的一个大问题。我最好的猜测是，这些人工智能会变得非常聪明，然后决定“管它呢，跟人类对齐不重要，我们要做些更合理的事情”。

问题 4：这里有一个问题，关于“目的”。人工智能是否可能拥有与人类相同意义上的“目的”，不是指个体目标或者子目标，而是指我们整个存在的目的是什么？

Hinton:人类有“目的”？它是什么？我的看法是，我们进化出的生物，之所以能生存下来，是因为它们比其他物种更擅长为自己获取更多资源，并减少对其他物种的依赖。我记得曾经有 21 个其他人类物种，我们把它们灭绝了。至于我们为何存在，那个“目的”是由进化赋予我们的，主要是为了生存。所以，如果你仔细想想你最强烈的需求，它们都和生存有关，比如你要吃饱、要有性、要保持安全，这些都跟生存相关。我其实并不认为有什么更高的“目的”。当然，你可能会说好奇心有巨大的进化价值，确实，能保持好奇心是非常重要的，它本身就是一个目标。科学资助者往往没有意识到这一点。你可以带着为某种目标生产技术的目的去好奇，也可以仅仅因为想理解事物的运作方式而去好奇，而这是一个根本性的目标，这正是优秀科学家的特质。但我认为，我们所有的目标和目的感都是源自进化的。

问题 5：我的问题是关于机器学习硬件市场的。现在它由单一的玩家主导，这是否让你感到担忧？你认为我们会看到机器学习硬件行业的多样化吗？

Hinton:我不太担心，因为在我女儿 30 岁生日时，我给她买了一大堆 NVIDIA 的股票，现在它们的价值是原来的五倍，所以她会没事的。而进化告诉我们，你最重要的目标之一就是确保你的孩子们没事。

但开玩笑归开玩笑，实际上我并不太担心，因为当你有像 NVIDIA 这样在赚大钱的情况时，就会有巨大的竞争。虽然其他公司赶上来可能需要一段时间，尤其是在软件方面，CUDA 的竞争对手等等，但这只是短期的事情，没多久就会有其他公司追赶上。如果你禁止 NVIDIA 的 GPU 进入中国，它们就会更快地赶上。因此，我想我并没有太多思考这个问题。每次 NVIDIA 股价上涨，我都会微笑，不过没像 Sam 那样微笑。

辛顿在接受颁发诺奖时的答辞（摘要）

COPU

现在我在谈论人工智能的生存威胁，这是一个长期的威胁。虽然很多短期的威胁更为紧迫，如网络攻击、大量失业、疫情等等，还有假视频，这些问题层出不穷。但存在一个长时期的生存威胁，那就是我们将创造出比人类更智能的东西，它们将接管我们的位置。许多人并不把这个问题当回事，而他们不认真对待的原因之一，是他们认为现在的人工智能系统并不真的“理解”人类。因此有一群人，如一些语言学家，他们称这些人工智能为“随机鹦鹉”，只不过是通过一个统计技巧把大量的文本拼凑在一起，看起来像是理解了，但实际上并不像人类理解的方式。而我将主张的是，人工智能的理解方式和我们（人类）一样。

那样谈论“随机鹦鹉”的人，他们的理解理论来自于经典的符号 AI 理论，即你大脑中有符号表达式，用某种简化的语言表示，你通过符号规则对它进行操作。这个理论从来没有真正成功过，但他们仍然坚持这个理论，因为他们认为只有通过类似逻辑的方式进行推理才可能拥有智能，即智能的本质是推理。其实有一个完全不同的范式，那就是智能的本质是学习——人工智能是在神经网络中进行学习的，视觉和运动控制是基本的，而语言和推理则是在这个基础之上发展出来的。

我想讨论的一个问题是人工智能是否真的理解？

有一个历史的细节，大多数人都不知道，那就是这些大语言模型尽管它们看起来能理解，能够以一个不太精通的专家水平回答任何问题，但它们其实很早就存在了。这来源于我在 1985 年做的一个模型，它是第一个神经网络语言模型，训练样本只有 104 个，而不是数万亿。但它确实是一个语言模型，训练它来预测

下一个词，并通过反向传播误差的方式，将输入符号转化为神经活动的向量，并学习这些向量如何互动，以预测需要预测的符号的向量。这个模型的目的不是为了某种工程应用，它的目的是为了解释人们如何理解单词的含义。因此，我们目前掌握的关于人类理解句子的最佳模型，正是这些语言模型，它们也是唯一能够实际运作的模型。

Sylvia Schwaag Serger（瑞典皇家工程科学院/IVA 院长/主席）：

Geoffrey，你曾谈到人工智能的风险，您也提到过必须有某种形式的国际合作来应对这些风险，你认为什么事情是必须发生的，才能使各国能够以建设性的方式来遏制这些风险？

Geoffrey Hinton：像致命武器这样的风险，各国是不会合作的，比如俄罗斯和美国不会在战斗机器人上进行合作，他们会互相对抗。所有主要的武器供应国都在忙着制造自主致命武器，他们不会自我约束，也不会进行合作。

然而，有一个领域可能会看到合作，那就是生存威胁。几乎所有我认识的研究人员都认为它们（COPU：指人工智能）会比我们（COPU：指人类）更聪明，只是对它们何时变得更聪明存在分歧，它们会接管吗（COPU：指人工智能接管人类）？我们是否能做些事情来防止这种事情发生？因为毕竟是我们创造他们。各国会在这一点上达成合作，因为没有哪个国家希望这种事情发生。冷战高峰时期，苏联和美国可以合作防止核战争，同样的方式，他们也会合作防止人工智能接管人类社会。但在其他领域我们将无法达成合作。

对杨立昆（Yann LeCun）“人类水平的 AI”落地的讨论

COPU

2024.12.30

AI 大师杨立昆(Yann LeCun)近日演讲:“人类水平的 AI”已在 COPU《人工智能文集》(第二十集)P.50-P.64 上全文转载，本文拟对《人类水平的 AI》落地情况进行讨论。

我们理解:杨立昆大师提出“我们需要人类水平的 AI”意指如何帮助人类不断扩大其原生智能，以此来应对 AI 不断增长的智能，防止 AI 的智能超越人类，保障人类安全。

他提出由人类佩戴的各种类型的智能工具，可概括为人类的智能助手 Agent(或称智能代理),由集结起来的智能助手 Agents/或 Agents Net,智能助手网络’，可利用其在未来不断增长中的综合智能（也包括人类的原生智能在内),以体现在未来增长中的人类综合智能(即人类水平的 AI),完全可以应对在未来增长中的 AI 的智能。

我们要讨论的是人类水平的 AI 在落地后的问题：

1)作为人类的综合智能(包括人类原生智能在内的 Agents 智能)能否保证它一直以来均能超越世界各地的 AI 的智能水平?

2) Agent 实际上也是一个 AI，它会不会脱离与人类的结合，转而与其他 AI 结合，从而造成人类反被 AI 的智能超越的状况。

马斯克的预测

埃隆·马斯克(Elon Musk)

COPU 讯：在去年年底，马斯克预测：AI 发展速度超乎想像，AI 可能在未来几年将超越人类的智力，而人类认识变化将滞后于现实。

到 2025 年，AI 智力将超越人类单独个体智力的水平，

到 2007-2028 年，AI 整体智力水平将超越全人类，

到 2030 年，这种可能性几乎达 100%。

马斯克的预测与 Open AI 成员利奥福特有相同的观点，

利奥福特认为，到 2027 年，AI 相互发展将达到与人类相同的通用人工智能的智力水平，这时将带来一系列负面影响。

马斯克接着提出保护人类安全的研究 AI 发展的思路，他主张全球化合作，管住 AI 发展的节奏，以免酿成大事。

COPU:当今国内外 AI 研究学者对 AI 发展速度快慢,AI 的智力能否超越人类，持有不同见解。对马斯克等专家的预测，值得我们重视，但我们更关心的是他们的研究成果和研究进程。

敬请关注联盟微信公众号
COPU开源联盟

扫描二维码
获取往期资料

中国开源软件推进联盟秘书处
电话：+86 010-88558999
联盟公共邮箱：office@copu.org.cn
联盟官网：http://www.copu.org.cn
地址：北京市海淀区紫竹院路66号赛迪大厦18层