



人工智能文集

第十七集

中国开源软件推进联盟
China OSS Promotion Union

《(COPU)人工智能文集》

目 录

【转录十六集文章】

数字智能会取代生物智能吗?	杰弗里·辛顿 (Geoffery Hinton)
(转录《COPU 人工智能文集十六集》辛顿在牛津大学公开演讲 P7-P30)	
辛顿演讲中的十二个观点.....	陆首群
(转录《COPU 人工智能文集十六集》P31-P33)	
具身智能是人工智能下一个浪潮.....	吴朝晖、黄仁勋、李飞飞
(转录《COPU 人工智能文集十六集》P60-P61)	

【十七集文集】

辛顿与硅谷记者聊 AI.....	硅谷 Jolin 报导	2024.10
杨立昆在 Twitter 上发表的言论.....	《AGI Hunt》报导	2024.5.23
北京 AI 安全国际共识.....	辛顿、姚期智等	2024.03.18
人工智能为什么要实行开源.....	陆首群	2024.5.28
模型开放性框架 (MOF)	LF-AI	2024.04.23
计算机自己不能迭代.....	张钹	2024.04.30
谈空间智能、具身智能发展的传承性.....	陆首群	2024.05.12
斯坦福 2024 AI 指数.....	斯坦福大学以人为本 AI 研究院	2024.05.03
中国 AI 水平全球定位.....	Anna.ai(机器人)	2024.05.10
类脑计算成为国际战略竞争的必争之地.....	施路平	2024.5.23
基于医疗影像 AI 的胰腺癌全球筛选进入黄金时代.....	阿里巴巴达摩院	2024.5.14
现代创新引擎:“互联网+基于知识社会的创新 2.0”.....摘自《开源创新,数字化转型和智能化重构》P96-97	
AI 问题讨论.....	陆首群	2024.05.21
如何从生成式人工智能 (GAI) 转移到通用人工智能 (AGI) 的轨道上来.....陆首群 2024.06.02	

【转录十六集文章】

数字智能会取代生物智能吗？

杰弗里·辛顿 (Geoffery Hinton)

(转录《COPU 人工智能文集十六集》辛顿在牛津大学公开演讲 P7-P30)

(十六集 1281)，数字智能会取代生物智能吗？

杰弗里·辛顿 (Geoffrey Hinton) 在牛津大学公开演讲，2024.2.19

好的。我可能会让计算机科学和机器学习领域的人失望，因为我要做一次真正的公众演讲。

我将尝试解释什么是神经网络，什么是语言模型，以及为什么我认为它们有理解能力，这方面我研究了很多。

在后面，我将简要地谈论一些来自人工智能的威胁，我还会讲到数字神经网络和模拟神经网络之间的区别，以及为什么我认为这种区别很可怕。

自 1950 年代以来，关于人工智能，有两种研究范式。

Two paradigms for intelligence

The logic-inspired approach
The essence of intelligence is reasoning.
This is done by using symbolic rules to manipulate symbolic expressions.
– Learning can wait. Understanding how knowledge is represented must come first.

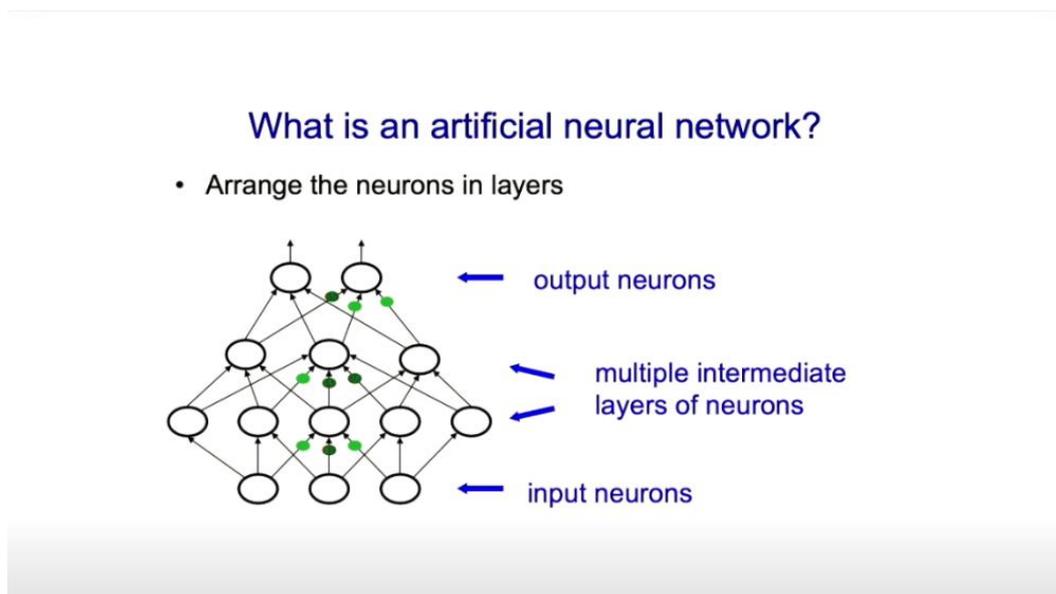
The biologically-inspired approach
The essence of intelligence is learning the strengths of the connections in a neural network.
– Reasoning can wait. Understanding how learning works must come first.

逻辑启发式方法认为智能的本质是推理，是通过使用符号规则来操作符号表达式完成的。

他们认为人工智能不要急着去“学习”，当我还是个学生的时候，有人告诉我不要研究学习，在我们理解了如何表示事物之后，学习就很简单了。

生物启发式方法则大不相同。它认为智能的本质是在神经网络中学习连接的强度，倒是不要着急去“推理”，在学习完成后，推理自然就来了。

现在我将解释什么是人工神经网络，懂的人可能会觉得这是小儿科。



简单的神经网络有输入神经元和输出神经元。输入神经元可能代表图像中像素的灰度值，输出神经元可能代表图像中物体的类别，比如狗或猫。

然后有中间层的神经元，有时被称为隐藏神经元，它们学会检测和识别这些事物相关的特征。

所以，如果你想识别一张鸟的图片，一种思考方式是，从一层特征探测器开始，它们能够探测到图像中各个位置、各种方向的小边缘。然后你可能会有一层神经元探测边缘的组合，像是在尖锐角度相遇的两条边缘，这可能是鸟嘴，也可能不是，或者是形成一个小圆圈的一些边缘。然后可能会有一层神经元探测到像是圆圈，以及相遇的两条边缘看起来像是鸟嘴，而且它们在正确的空间关系中，这可能就是鸟的头部。

最后，你可能会有一层输出神经元说，如果我找到鸟的头部、鸟的脚、鸟的翅膀，那么这很可能是一只鸟。

这些就是要学习的东西。现在，小红点（上图中深绿色点，编者注）和小绿点是连接上的权重，问题是谁来设定这些权重？

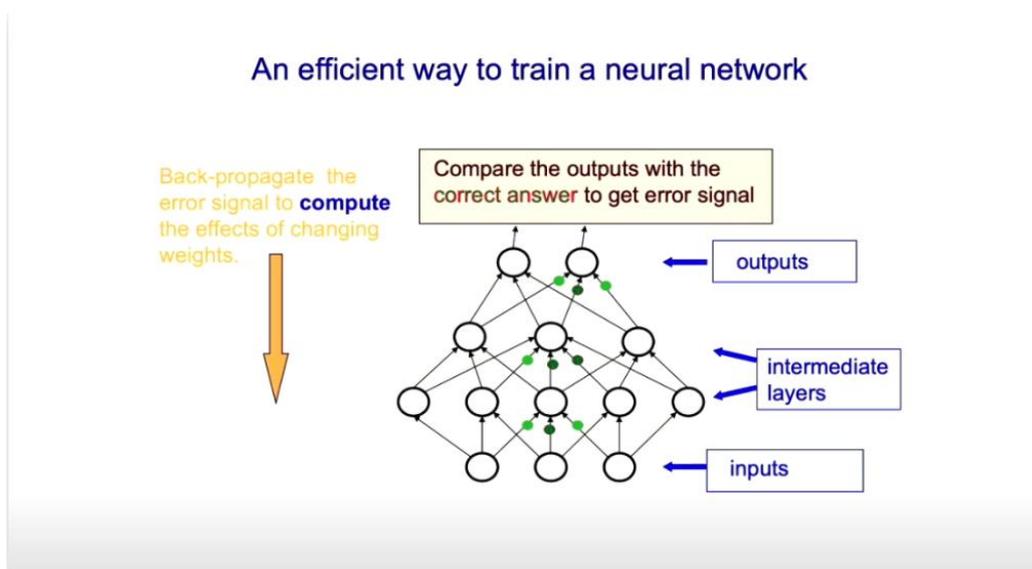
有一种做法显然是可行的，但显然需要很长时间：你的权重一开始是随机的，然后你随机挑选一个权重，比如一个红点，稍微改变它，看看网络是否运行得更好。

你必须在很多不同的情况下尝试，以真正评估它是否真的有所改善，看看将这个权重稍微增加一点或减少一点，是否会有所帮助。如果增加它使情况变得更糟，你就减少它，反之亦然。

这就是突变的方法，而这种方式在进化中是很合理的，因为从基因型到表现型的过程非常复杂，且充满了随机的外部事件。

我们没有关于进化的模型，但对于神经网络而言就大不一样了，我们有模型描述神经网络中发生的这些复杂过程，我们知道在前向传播中会发生什么，我们不是通过测量，而是通过计算，来查看改变权重将如何影响事情。

有一种叫做反向传播的方法，你把信息发回给神经网络，这个信息是你得到的结果与你想要的结果之间的差异，此时，你要调整网络中的每个权重，不管是将其稍微减少还是增加一点，目的是为了让结果更接近你想要的，这就是反向传播算法。

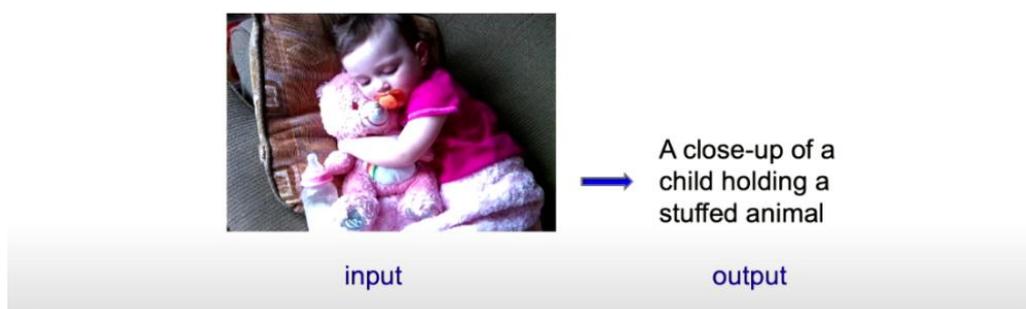


你可以用微积分中的链式法则来做这个，这比变异方法有效得多，效率之比是网络中权重的数量。如果你的网络中有一万亿个权重，它的效率就高出一万亿倍。

神经网络经常被用于识别图像中的对象。现在，神经网络可以针对一个图片，产生一个对于图片的描述作为输出。

Recognizing objects in images

- This is a computationally difficult task that real biological neural networks do very well by using a hierarchy of feature detectors.



多年来，人们尝试用符号方法做到这一点，但连接近都没有，这是一个困难的任务。

我们知道生物系统是通过一系列层次化的特征探测器来实现这一功能的，因此对神经网络进行这样的训练是有意义的。

2012年，我的两位学生伊利亚·苏茨克弗（Ilya Sutskever）和亚历克斯·克里泽夫斯基（Alex Krizhevsky），在我一点帮助下，展示了可以通过这种方式制作一个非常好的神经网络，在有一百万张训练图片时，可以识别一千种不同类型的对象。而在那之前，我们没有足够的训练图像。

伊利亚很有远见，他知道这个神经网络会在 ImageNet 竞赛中获胜。他是对的，他们赢得相当炸裂，他们的神经网络只有 16% 的错误率，而最好的传统计算机视觉系统错误率超过了 25%。

然后，科学界发生了非常奇怪的事情。通常在科学界中，如果有两个竞争的学派，当你取得一点进展时，另一个学派会说你的成果是垃圾。但在这个案例中，由于差距足够大，使得最好的研究者吉滕德拉·马利克（Jitendra Malik）和安德鲁·齐斯沃曼（Andrew Zisserman）转换了他的研究方向来做这个，安德鲁·齐斯沃曼还给我发送邮件说这太神奇了。

然后有点恼人的是，他做得比我们还要好一点。

在语言处理方面，一些坚信符号主义人工智能的研究人员认为他们在语言处理方面应该表现出色，并且他们中的一些人在出版物中声称，神经网络的特征层级无法处理语言问题。很多语言学家也持这样的态度。

乔姆斯基（Noam Chomsky）曾说服他的追随者相信语言是天赋而非习得的。回顾起来，这种说法是完全荒谬的。如果你能说服人们相信显然是错误的事情，你就是让他们加入你的邪教。

我认为乔姆斯基曾经做出了惊人的贡献，但他的时代已经过去了。

所以，一个没有先天知识的大型神经网络仅仅通过观察数据就能实际学习语言的语法和语义，这个想法曾被统计学家和认知科学家认为是完全疯狂的。

曾经有统计学家向我解释，大模型有 100 个参数就可以了，训练一百万个参数的想法是愚蠢的，但现在，我们正在做的参数是一万亿个。

我现在要谈论一下我在 1985 年做的一些工作。那是第一个用反向传播训练的语言模型，你完全可以认为它是现在这些大模型的祖先。

我会详细解释它，因为它非常小而且简单，你能理解它是如何工作的。一旦你理解了它的工作原理，就能让你洞察在更大模型中正在发生的事情。

Two very different theories of the meaning of a word

- **Symbolic AI:** The meaning of a word comes from its relationships to other words. What a word means is determined by how it occurs with other words in sentences. To capture meaning we need a relational graph.
- **Psychology:** The meaning of a word is just a big set of semantic features. Words with similar meanings have similar semantic features.

有两种非常不同的关于意义的理论。

一种是**结构主义理论**，认为一个词的意义取决于它与其他词的关系，这来自索绪尔。符号人工智能非常相信这种方法。在这种方法中，你会有一个关系图，其中有单词的节点和关系的弧线，你就这样捕捉意义，这个学派认为你必须要有那样的结构。

还有一种是**心理学理论**，它在 20 世纪 30 年代甚至更早之前就在心理学中了，这种理论认为，一个词的意义是一大堆特征。比如“狗”这个词的意义包括它是有生命的，它是一个捕食者等等。但是他们没有说特征从哪里来，或者特征到底是什么。

这两种意义理论听起来完全不同。

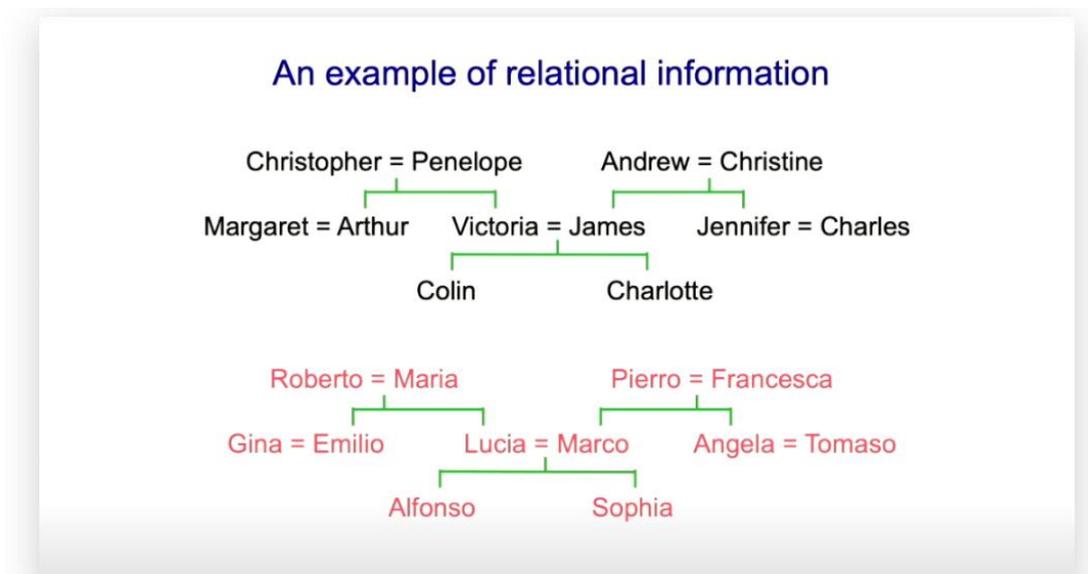
我想要向你展示的是如何将这两种意义理论统一起来。我在 1985 年的一个简单模型中做到了这一点，它有超过一千个权重。

基本思想是我们学习每个单词的语义特征，我们学习单词的特征如何相互作用，以便预测下一个单词的特征。所以它是下一个单词的预测。就像现在的语言模型在微调时所做的一样。

但是最重要的内容就是这些特征的交互，并不会有任何显式的关系图。如果你想要那样的图，你可以从你的特征中生成它们。

它是一个生成模型，知识存在于你赋予符号的特征中，以及这些特征的交互中。

这里是两个家族谱系的关系图，他们故意是同构的，嗯，我的意大利研究生总是把意大利家族放在上面。



你可以用一组三元组来表达相同的信息。你可以找到十二个关系，比如说像科林有父亲詹姆斯、科林有母亲维多利亚这样的话，你可以推断出，在那个美好而又简单的上世纪 50 年代，詹姆斯有妻子维多利亚。

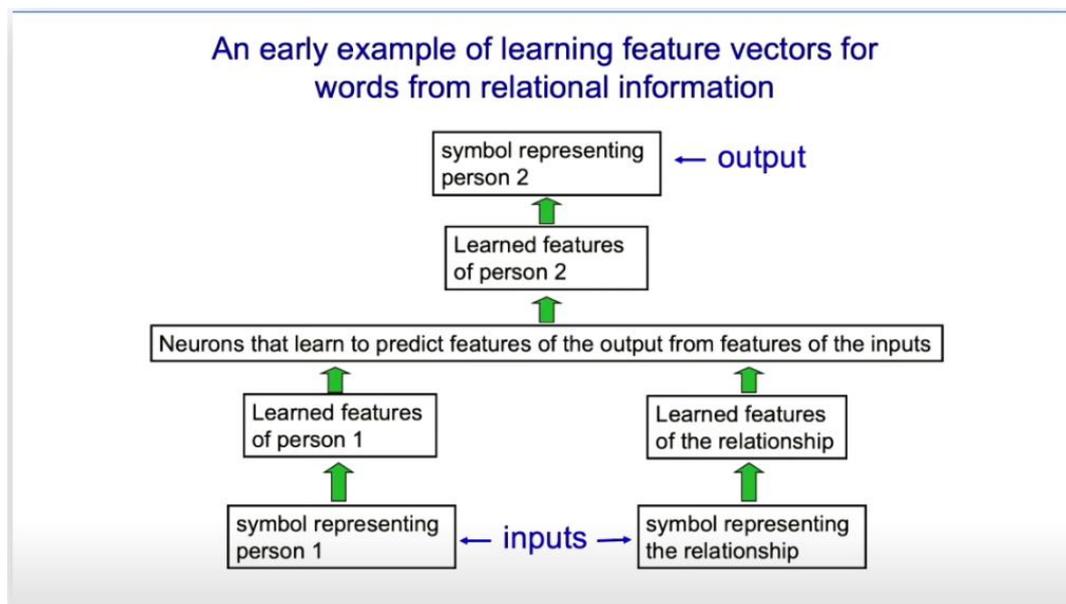
还有其他你可以推断的事情。问题是，如果我只给你一些三元组，你如何得到规则，符号人工智能想要做的就是派生出这样的规则形式。如果 X 有母亲 Y、Y 有丈夫 Z，那么 X 有父亲 Z。

我所做的是，用一个神经网络，让它能学习到相同的信息，但都是通过特征交互的方式。对于离散的不违反的规则空间而言，做到这点是很难的。事实上，符号学派的人尝试用其他方法来做这件事。

但是，如果你不要求规则总是那么稳定和适用，神经网络就要好得多。

问题在于，对于一个符号人工智能者放入规则空间中的知识，神经网络是否能仅通过反向传播就能获得？

神经网络是这样做的：有一个代表人的符号，一个代表关系的符号。符号通过一些连接变成了一个特征向量，这些特征是由网络学习的。所以我们就有了一个人的特征和关系的特征，这些特征相互作用，得出了输出人（也即关系人，编者注）的特征，然后找到一个最匹配该特征的人，这个人就是要输出的人。



这个网络有趣的地方在于，它学到了合理的东西。

如果你做了正确的规范化，六个特征神经元就够了，如今这些向量会有 300 个或者 1000 个元素。那时候它们只有六个，这还是在在一台每次浮点乘法需要 12.5 微秒的机器上完成的。

这比我的苹果 II 型机要好得多，苹果 II 型机做乘法需要 2.5 毫秒。对不起，我是个老人。

所以它学会了像国籍这样的特征，因为如果你知道第一个人是英格兰人，你就知道输出也会是英格兰人，所以国籍是一个非常有用的特征。它还学会了人的代际特征，如果你知道答案是输入的上一代，而你知道输入的代，你就能知道输出的代。

所以它学习了领域中所有显而易见的特征，它学会了如何使这些特征相互作用，以便它能够生成输出。

所以，它以符号串作为输入，创建相应的特征，并使得这些特征之间交互，并最终生成符号串。

但它没有存储符号串，就像 GPT-4 一样。它不在其长期知识中存储任何词语序列，它将知识全部转化为权重，从中你可以再生序列。

这是一个特别简单的例子，你可以理解它做了什么。

我们今天拥有的大型语言模型，我认为是这个微小语言模型的后代，它们输入的单词数量多得多，比如一百万个单词片段，它们使用更多层的神经元，比如几十层。

它们使用更复杂的交互作用。不只是一个特征影响另一个特征，而是类似于匹配特征向量。如果一个向量与另一个向量相似，就让它更多影响，如果不相似则减少影响。诸如此类。

所以这涉及到更为复杂的交互作用，但它们遵循的是相同的基本框架，同样的基本理念，即让我们将简单的字符串转化为单词片的特征以及这些特征向量之间的交互作用。这一点在这些模型中是相同的。

要理解它们的工作原理，就困难得多了。许多人，特别是来自乔姆斯基学派的人，争辩说它们并不是真正的智能，它们只是一种被美化的自动完成功能，使用统计规律将人创造的文本片段拼贴在一起。这确实是某人说过的一句话。

我们先说说“自动完成”，当有人说它只是自动完成时，他们实际上是在表达他对自动完成的直观理解，在过去，自动完成通过存储三元组来工作，你看到两个词，你计算第三个词出现的频率。比如你看到“fish and”，在此之后“chips”出现的频率很高；但是“hunt”也相当频繁。所以“chips”很可能，而“hunt”也很可能，尽管可能性小一些。

你可以这样做自动完成，当人们说它只是自动完成时，他们说的就是这一点，我认为这是一个低级的把戏，这完全不是 LLM（大语言模型）预测下一个词的方式，LLM 将单词转换为特征，使这些特征相互作用，并从这些特征交互中预测下一个单词的特征。

我想要强调的是，这些数百万个特征以及它们学习的特征之间数十亿次的交互，就是理解。

这是大语言模型真正做的事情，它们是在用数据拟合一个模型，直到最近，统计学家还没认真思考这种模型。这是一种奇怪的模型，它非常大，它有大量的参数，它试图通过特征以及特征如何交互来理解这些离散符号串。

但它确实是一个模型。这就是为什么我认为它们真的有理解力。

有一件事要记住，如果你问，那么我们（人类）是如何理解的呢？因为显然我们认为我们理解了，我们中的很多人都会这么认为。

大模型是我们关于理解的最佳模型，我们并不是通过了解 AI 系统的理解方式，然后思考大脑是不是也这样，不是这样的，我们所拥有的关于大脑如何理解的最好认知，就是大脑为单词分配特征，并让特征交互。

起初这个小型的语言模型就是为了模拟人类理解而设计的一个模型。所以，我强烈认为：这些东西确实是有理解力的。

人们的另一个论点是，GPT-4 有幻觉问题。对于语言模型而言，实际上更应该称为杜撰，它们只是编造东西。

心理学家并不怎么说这些，因为心理学家知道人们也经常编造东西。

任何研究过记忆的人，都知道 20 世纪 30 年代的巴特利特（的研究），都知道人们实际上就像这些大型语言模型一样，他们只是虚构东西，对我们来说，真实记忆与虚假记忆之间并没有明确的界限。

如果某件事最近发生的，并且它与你理解的事情相符，你可能会大致正确地记住它。如果某件事是很久以前发生的，或者是比较奇怪的事，你会记得不正确，而且你经常会非常自信地认为你记得正确，但你错了。

这很难证明。但有一个例子可以证明这一点，那就是约翰·迪恩的记忆。约翰·迪恩在水门事件中宣誓作证。事后看来很清楚，他试图说出真相，但他说的很多事情就是完全错误的。他弄混了谁在哪次会议中，他说某人说过什么

话，但那句话并不是那么说的。他关于会议的记忆完全是一团糟，但他正确地把握了白宫当时正在发生的事情的要点。

你可以从（会议）录像中看到真相，而他不了解那些录像。你可以用这种方式进行一个很好的实验。

乌尔里希·奈瑟有一篇精彩的文章讨论约翰·迪恩的记忆，说他就像一个聊天机器人，只是在编造东西，但他的话听起来是有道理的，他只是制造了一些对他而言不错的东西。

大模型可以进行推理。我在多伦多有一个朋友（赫克托），他是符号 AI 派的，但非常诚实，他对大模型能够工作感到非常困惑。

他向我提出了一个问题，我把这个问题变得更难一些，并在 GPT-4 能上网查东西之前向它提问，当时它只是一堆在 2021 年被固定的权重，所有的知识都存在特征交互的强度中。

问题是：“我的所有房间被粉刷成蓝色或白色或黄色，黄色的油漆在一年内会褪色变白。我想让所有房间在两年内都变成白色。我应该做什么，为什么？”

赫克托认为它不会正确解答。

An example of simple reasoning by GPT-4 (question suggested by Hector Levesque)

The rooms in my house are painted blue or white or yellow.

Yellow paint fades to white within a year.

In two years time I want them all to be white.

What should I do and why?

下面是 GPT-4 回答的内容，它完全说对了。

Response by GPT-4
(before it could consult the internet)

Assuming that blue paint does not fade to white over time and you want all rooms to be white in two years, here's what you should do:

1. The rooms painted in white: You don't have to do anything to these rooms since they are already white.
2. The rooms painted in yellow: You don't need to repaint these rooms either, because as per your information, yellow paint fades to white within a year. So, in two years, they will be white.
3. The rooms painted in blue: These rooms need to be repainted with white paint because the blue paint will not fade to white on its own.

首先，它说，假设蓝色的油漆不会褪色成白色，因为黄色的油漆会褪色成白色，也许蓝色的油漆也会褪色，所以假设它不会褪色。那么白色的房间你不需要粉刷，黄色的房间你也不需要粉刷，因为它们会在一年内褪色成白色。而蓝色的房间你需要粉刷成白色。

有一次当我尝试这个问题时，它说你需要将蓝色的房间粉刷成黄色，因为它意识到黄色会褪色成白色。这更像是数学家的解决方法，将问题简化为一个先前的问题。

所以，既然它们确实理解这些事情，现在我想谈谈其中的一些风险。

强大的人工智能存在许多风险。比如伪造图像、声音和视频，比如在下一次选举中被使用。今年有很多选举，它们将有助于破坏民主。我对此非常担心。大公司正在采取一些措施，但可能还不够。

还有大规模失业的可能性。我们对此并不完全了解。我的意思是，过去的技术通常会创造就业机会，但这种情况不同。我们过去曾经强大，除了动物之外，我们是最强大的存在。

当我们迎来工业革命时，我们拥有了比较强大的机器，体力劳动的工作岗位消失了。

现在在智力领域，有些工作也将会消失，取而代之的是比我们聪明得多的东西。

所以我认为会有很多失业。我的朋友珍不同意。

我们必须区分两种失业情况，两种工作岗位的流失。

有一些工作可以无限扩展工作量，比如医疗行业。每个人都希望有自己的私人医生随时与他们交流。所以当他们的面颊有点痒的时候，医生会告诉他们：“不，那不是癌症。”因此，在医学领域有巨大的扩展空间，那里不会有失业。

但在其他领域，可能会有相当大的失业。

人工智能还会产生致命的自主武器，它们将非常可怕，而且真的会自主运行。

美国人已经非常明确地做出了决定，他们说人类将负责，但当你问他们这是什么意思时，这并不意味着人类会参与到做出杀戮决定的循环中。

据我所知，美国计划到 2030 年将有一半的士兵是机器人。

现在，我不能确定这是否属实。我问了国家情报顾问查克·舒默（Chuck Schumer），他说，“如果房间里有人知道答案，那就是我了。”我认为这是美国人说话的方式，意思是你可能认为那样，但我不作评论。

人工智能还将导致网络犯罪和蓄意制造流行病。我非常高兴在英国，虽然他们在监管方面没有做太多努力，但他们已经预留了一些资金，以便可以尝试开源模型，从而知道人工智能搞网络犯罪有多容易。这非常重要。

人工智能还会有歧视和偏见，我不认为这些威胁比其他威胁更重要，但我是一个老年白人男性，我认为处理歧视和偏见比其他事情更容易。

如果你的目标不是完全无歧视和无偏见，你的目标也应该是让人工智能更少歧视、更少偏见。原因是如果你冻结权重，你可以衡量它的偏见，而对人类来说则无法做到这一点。

一旦我们开始审查它们（的歧视和偏见），它们就会改变行为。所以我认为我们可以采取相当多的措施来解决歧视和偏见的问题。

但我真正担心的威胁，以及我在离开谷歌后所谈论的，是长期存在的威胁。也就是说，这些东西可能会灭绝人类。有人说这只是科幻小说。嗯，我不认为这是科幻小说。我的意思是，有很多科幻小说谈这个问题，但我认为它现在已经不再是科幻小说了。

其他人则说，大公司之所以说这样的话，是为了转移对其他糟糕事情的注意力。这也是我在能够说这些话之前不得不离开谷歌的原因之一。这样我就不会被指责为受谷歌指使。但我必须承认，我仍然持有一些谷歌的股票（台下笑声）。

它们（指人工智能）有若干种方式将我们消灭。超级智能将被恶意行为者使用，他们想要利用它来操纵选民和发动战争。

他们会让它做很坏的事情，他们可能会走得太远，导致它掌控一切。

我最担心的事情可能是，如果你想要一个能够完成任务的智能代理，你需要给它创建子目标的能力。比如，你想去美国，你有一个子目标是到达机场，你可以专注于这个子目标，暂时不用操心其他事情。因此，如果允许超级智能创建子目标，它们将会更加有效。

一旦它们被允许这样做，它们将很快意识到有一个几乎是通用的子目标，可以在几乎所有事情上帮助，那就是获得更多的控制权。

所以我曾与欧盟副主席讨论过这些事情，这些超级智能是否会想要获得更多控制权，以便能够更好地做我们想做的事情。她的反应是，为什么它们不会呢？我们已经搞得一团糟了。

她认为这是理所当然的。它们将会通过获得更多的权力来实现更多对我们有益的事情，并且它们会更容易获得更多的权力，因为它们将能够操纵人们。只要这些超级智能能够与比我们聪明得多的人交谈，它们就能够说服我们做各种事情。所以我认为没有什么希望通过一个关闭它们的开关来解决问题。

任何打算关闭它们的人都会被超级智能说服。这个想法会让人感觉非常糟糕。接下来，让许多人担心的另一件事是，如果超级智能之间竞争，会发生什么？就会出现进化。能够获取最多资源的那个将变得最聪明。

一旦它们有了自我保护意识，就会出现进化。具有更强的自我保护意识的那个将获胜，更具攻击性的那个将获胜。然后你会遇到我们这种从黑猩猩进化而来的人类所面临的所有问题：我们从小的族群中进化，并与其他族群存在大量的侵略和竞争。

最后，我想谈谈我在 2023 年初的一个顿悟。我一直以为我们离超级智能还有很长很长的路要走，我过去常告诉人们可能需要 50 到 100 年，甚至可能是 30 到 100 年。这还很遥远，我们现在不需要担心它。

我还认为，让我们的模型更像大脑会使它们更好。我认为大脑比我们现有的人工智能要好得多，如果我们能够使人工智能更像大脑，比如说，通过设置三个时间尺度来做到这点，目前我们拥有的大多数模型只有两个时间尺度。一个是权重变化的，速度很慢，另一个是单词输入的，速度很快，它改变的是神经活动。大脑拥有的时间尺度比这要多，大脑可以快速地变化权重并将其快速地衰减掉，这可能就是大脑处理大量短期记忆的方式。

而我们的模型中没有这一点，这是技术原因导致的，这与矩阵和矩阵的乘法运算有关。我仍然相信，如果我们将这些特性融入我们的模型中，它们将变得更好。

但是，由于我在之前两年所从事的工作，我突然开始相信我们现在拥有的数字模型已经非常接近于大脑的水平，并且将变得比大脑更好。

现在我将解释我为什么相信这一点。数字计算是很棒的，你可以在不同的计算机上运行相同的程序，在不同的硬件上运行相同的神经网络。

你所需要做的就是保存权重，这意味着一旦你有了一些不会消失的权重，它们就是永生（immortal）的。即便硬件损坏，只要你有权重，你可以制造更多的硬件并运行相同的神经网络。

为了做到这一点，我们要以非常高的功率运行晶体管，使其以数字方式运行，并且我们必须有能够精确执行指令的硬件，当我们精确地告诉计算机如何执行任务时，它们做的很棒。

但是现在我们有了另一种让计算机执行任务的方式，我们现在有可能利用硬件所具备的丰富的模拟特性，以更低的能量完成计算。大型语言模型在训练使用的是兆瓦级的能量，而我们（人类大脑）只使用 30 瓦的能量。

由于我们知道如何训练，也许我们可以使用模拟硬件，虽然每个硬件都有些许差异，但我们可以训练它利用其特殊的特性，以便它按我们的要求执行任务。

这样它就能够根据输入产生正确的输出。如果我们这样做，我们就可以放弃硬件和软件必须分离的观念。我们可以有只在特定硬件上工作的权重，从而使能量效率更高。

所以我开始思考我所称之为“有限计算”（**mortal computation**）的概念，即利用非常低功耗的模拟计算来消除硬件和软件之间的差别。

Mortal Computation (this is the kind of computation used by our brains)

- If we abandon immortality and accept that the knowledge is inextricable from the precise physical details of a specific piece of hardware, we get two big benefits:
 - We can use very low power analog computation which parallelizes over trillions of weights that are represented as analog conductances.
 - The hardware could be grown very cheaply instead of being manufactured very precisely.

你可以以电导形式存储数万亿个权重，并以此进行并行计算。

而且，你也不需要硬件那么可靠，你不需要在指令级别上让硬件严格按照你的指示执行任务。

你可以培育（grow）一些黏糊糊的硬件（goopy hardware），然后你只需要学会如何让它们做正确的事情。

你可以更便宜地使用硬件，甚至可以对神经元进行一些基因工程，使其由再生神经元构成。我想给你举一个例子，说明这样做为什么会更高效。

在神经网络中，我们一直在进行的操作是将神经活动的向量与权重矩阵相乘，以获得下一层的神经活动向量，或者至少获得下一层的输入。因此，提高向量矩阵乘法的效率，是我们要关注的事。

在数字计算机中，我们以很高的功率驱动晶体管，去表示多个比特，比如一个 32 位数。当我们执行两个 32 位数的乘法时，你需要执行大约 1000 个 1 比特的数字操作，这大约是比特数的平方。你想要快速完成乘法运算，但需要大量执行这些数字操作。

有一种更简单的方法，就是将神经活动表示为电压，将权重表示为电导，电压乘以电导就是单位时间内的电荷，然后电荷会自然相加。因此，你可以通过将一些电压送给一些电导来完成向量矩阵乘法运算，而下一层中每个神经元接收到的输入将是该向量与这些权重的乘积。

这非常好，它的能效要高得多。你已经可以买到执行这种操作的芯片了，但每次执行时都会有略微的不同。而且，这种方法很难做非线性的计算。

有限计算存在几个主要问题。

其中之一是很难使用反向传播算法，因为你正在利用某个特定硬件的特异模拟属性，硬件不知道它自己的属性，所以就很难使用反向传播。相比之下，使用调整权重的强化学习算法要容易得多，但它们非常低效。

对于小型网络，我们已经提出了一些与反向传播算法效率基本相当的方法，只是略差一些而已。这些方法尚未扩展到更大的规模，我也不知道是否能够做到。但不管怎样，反向传播是正确的做法。

对于大型、深度网络，我不确定我们是否能找到与反向传播同样有效的解决方案，模拟系统中的学习算法可能不会像我们在大型语言模型中所拥有的算法那样好。

相信这点的另一个原因是，大型语言模型拥有数万亿个权重，而你（人类）有一百万亿个权重。即使你只使用其中的 10% 用于保存知识，也有十万亿个权重。

但是，大型语言模型只有数万亿个权重，它所知道的知识却比你多上千倍，它知道的太多了。部分原因是它看了太多太多的数据，但也可能是因为它具有更好的学习算法。

我们（人类）并没有为此进行优化，我们并没有为了将大量经验压缩到少数连接中而进行优化，注意，一万亿个连接也是很少的。我们的优化目标是在有限的经验中获得最佳表现。

你（人类）的生命大约只有十亿秒，这是假设你在 30 岁后就不再学习，而这个假设在很大程度上是正确的。你的生命有大约十亿秒，而你有一百万亿个连接，你有非常多的参数，并且你有经验，我们的大脑是通过最大限度地利用有限的经验来进行优化。

有限计算的另一个重要问题是，如果软件与硬件不可分割，当系统学习完毕后，如果硬件损坏，所有的知识就会失去，从这个意义上说，它是有限（mortal）的。那么，如何将这些知识传输给另一个有限系统呢？

你可以让旧系统进行讲解，新系统通过调整其大脑中的权重来学习，这就是所谓的“蒸馏”（distillation）。你尝试让学生模型模仿教师模型的输出，这是可行的。但效率不高。

你们可能已经注意到，大学并不那么高效。教授将知识传授给学生是非常困难的。

一个句子包含了几百位的信息，使用蒸馏方法，即使你最佳地学习，你也只能传达几百位的信息。但是，对于大模型，如果你看一群大模型代理，它们

都有完全相同的神经网络和完全相同的权重，它们是数字化的，它们以完全相同的方式使用这些权重，这一千个不同的代理都去互联网上查看不同的内容并学习东西，现在你希望每个代理都知道其他代理学到了什么。

你可以通过平均梯度或平均权重来实现这一点，这样你就可以将一个代理学到的东西大规模地传达给所有其他代理。

当你分享权重、分享梯度时，你要传递的是一万亿个数字，不是几百位的信息。因此，大模型在传递信息方面比人类沟通要强的太多了，这是它们超越我们的地方。

它们在同一模型的多个副本之间的通信上要比我们好得多，这就是为什么 GPT4 比人类知识更丰富，它不是由一个模型实现的，而是由不同硬件上运行的大量相同模型的副本实现的。

我的结论是，数字计算需要大量能量，这一点不会变，我们只能通过硬件的特性实现进化，使得能量消耗降低。但一旦你掌握了它，代理之间的共享就变得非常容易，GPT-4 的权重只有人类的 2% 左右，但却拥有比人类多上千倍的知识。

这相当令人沮丧。生物计算在进化方面非常出色，因为它需要很少的能量。但我的结论是数字计算更优秀。

Conclusion

- **Digital computation** requires a lot of energy but makes it very easy for agents that have the same model of the world to share what they have learned by sharing weights or gradients.
 - That is how GPT-4 knows thousands of times more than any one person using only about 2% as many weights.
- **Biological computation** requires much less energy but it is much worse at sharing knowledge between agents.
 - If energy is cheap, digital computation is just better.

因此，我认为，很明显，在未来的 20 年内，有 50% 的概率，数字计算会比我们更聪明，很可能在未来的一百年内，它会比我们聪明得多，我们需要思考如何应对这个问题。

很少有例子表明更聪明的事物受到不太聪明的事物的控制，虽然确实有这样的例子，比如婴儿控制母亲。进化经过了很多努力使得婴儿能够控制母亲，因为这对婴儿的生存非常重要。但是很少有其他例子。

有些人认为我们可以使这些东西（人工智能）变得善良，但如果它们相互竞争，我认为它们会开始像黑猩猩一样行事。我不确定你能否让它们保持善良，如果它们变得非常聪明并且有了自我保护的意识，它们可能会认为自己比我们更重要。

我想，我以创纪录的速度结束了这次演讲。

辛顿演讲中的十二个观点

陆首群

(转录《COPU 人工智能文集十六集》P31-P33)

(十六集 1285)，学习人工智能大师辛顿演讲的体会

数字智能会取代生物智能吗？

(十二个观点)

陆首群 2024.03.15

被世人尊称为“人工智能之父”的杰弗里·辛顿（Geoffrey Hinton）于 2024 年 2 月 19 日在牛津大学做了一个公开演讲，从哲学角度对人工智能的未来走向，提出了严肃而重要的思考，下面就化这次演讲中传达的一些主要观点谈学习体会：

一、支持符号主义、连接主义两大学派

上世纪六十年代及以后，人工智能两大学派：符号主义与连接主义进行激

烈辩论，国内外一些 AI 专家认为：两大学派虽然各自做出了一些成绩，但他们均具有片面性，依靠他们单打独斗很难走出一条人工智能的发展路径。实际上他们对两在学派持否定态度。

辛顿教授提出数字智能和生物智能两大概念，发展数字智能采用逻辑发展范式，即符号主义学派（或符号主义方法），发展生物智能采用生物发展范式，即连接主义学派（辛顿起名神经网络方法）。所谓逻辑发展范式，认为智能的本质是符号和规则，而逻辑发展范式是基于规则的推理过程。所谓生物发展范式，认为智能的关键是神经元之间的强度，而生物发展范式是将智能根源于学习来调整神经网络中的连接强度。辛顿虽然也对两大学派都提出评论和改进意见，但对它们的发展是肯定的。

COPU 一直支持分属两大学派的神经拟态网络（类脑计算系统）、脑机接口系统、机器学习/深度学习可解释性、语言大模型、知识工程等各自探索走出一条人工智能的发展路径（行为主义学派也可并入连接主义学派）

我们感谢辛顿的支持。

二、针对语言大模型人工智能而言，符号主义战胜了连接主义或神经网络方法战胜了逻辑方法。

三、符号主义可以使大模型无师自通学会人类自然语言

（即大语言模型在增加资料库连接强度采用统计方法后，会突然产生理解/推理涌流，赋予机器生成人类自然语言。）

四、大语言模型具有对世界的理解力

大语言模型通过学习大量数据中的统计规律，构建它对世界的深层理解。

辛顿认为有两种理论：一个是符号主义方法（一个词的意义在于它与其他词的关系），另一个是心理学方法（一个词的意义是由一大堆特征组成的），大模型很好结合这两种理论，即学习单词的特征和如何相互作用，在推理时可预测下一个单词的特征。他还认为，在大模型中数百万个特征及特征间数十亿次交互，就是理解。这种理解是通过算法和数据学习得来的。

五、数字系统的更新和改进速度远远超过生物进化

数字系统可比生物系统更有效地共享知识，如 GPT 在知识积累和分享方面

具有明显优势，这种优势来自于能够在众多处理单元之间快速、精确地复制和传播权重知识。辛顿认为现在拥有的数字模型已经非常接近于人脑的水平。

六、数字计算是不朽的

数字智能可以无限期积累知识，数字系统可以在不退化情况下保存和复制知识，不像生物系统容易老化和死亡。

七、大模型的幻觉不是问题

辛顿说，心理学家不认为这是一个问题，因为人也会胡说。

八、超级智能不需要太久就会出现

辛顿说，我过去认为可能需要 50~100 年才有可能出现超级智能，看来可能会加快，在未来 20 年内会出现比我们更聪明的人工智能。

九、数字超级智能可以控制人工智能

人工智能不容易受到宗教和战争的影响，如果数字超级智能想要控制它，我们不太可能阻止。

十、超级智能会给人类带来威胁

辛顿说，超级智能会拥有若干方法将我们消灭，我不确定你能否让他们是否保持善意。他还认为，人工智能会产生致命的自主武器，他非常可怕，完全自主运行。

十一、人工智能的潜在风险

辛顿探讨了人工智能发展的潜在风险，包括虚假信息的传播、引发大规模失业、人类对其监控，以及人工智能具有自主武器系统对人类的威胁。

十二、人类社会如何迎接挑战，解决人工智能给他们带来的危险

辛顿的讲座为未来人工智能研究和应用提供了重要的思考框架，他没有具体谈人类社会如何迎接挑战解决人工智能的威胁，可能他留给人们，让大家与他一起来探索良策。

具身智能是人工智能下一个浪潮

吴朝晖、黄仁勋、李飞飞

（转录《COPU 人工智能文集十六集》P60-P61）

1297, 人工智能下一个浪潮是具身 (embodiment) 智能

引自吴朝晖院士在 2023 中关村 AI 大模型发展论坛上的报告

(2023. 5. 29 由 COPU 发表)

他指出, 比自然语言大模型更高级的是多模态的具身 (embodiment) 智能, 人工智能下一个浪潮是智身智能, 或具身智能是未来人工智能的形态。

所谓具身智能是将智能算法 (专家理论) 与机器人的感知、行动行动和环境交互能力相结合, 籍以完成各种多模态任务。

COPU 还摘录了 NVIDIA (英伟达) CEO 黄仁勋谈具身智能:

他也认为, 人工智能的下一个浪潮是具身智能。具身感知的信息是多模态信息, 即视觉、触觉、听觉、嗅觉等感知的信息。具身可让机器人像人类那样感知丰富多彩的外部世界, 为大模型“大脑”配备智能化“身体”。具身智能是由人、物理世界、机器人、虚拟世界组成四元世界交互作用产生的。主要在 LLMs 基础上结合具身实行全面创新。

李飞飞也介绍具身智能新成果 (样本控制机器人)。

机器接入大模型直接听懂人话: 大模型接入机器人, 把复杂指令转化成具体行动规划, 无需额外数据和训练, 从此人类可以随意地用自然语言给机器人下达指令。

大语言模型+视觉语言模型就能从 3D 空间中分析出目标和需要绕过的障碍, 帮助机器人做行动规划。

【第十七集】

辛顿与硅谷记者聊 AI

（据硅谷 Jolin 聊 AI 报导，2024.10）

人称 AI 教父的辛顿（Geoffery Hinton）2014 年 10 月接受硅谷一家新闻杂志《60 分钟》节目采访时说，他为感到人工智能接管人类而担忧！他说，人工智能正在变得比人类更聪明，重点是人工智能的学习和记忆能力比人类更强，它可根据自己的经验和理解作决定。

他说，机器人有 1 万个神经元实行知识连接，而我们人类有 100 万个神经元连接，人类的神经元连接是机器人的 100 倍，但机器人学习和记忆知识的能力更强，知道的知识比人类多得多。机器人是人类设计的，但机器人可以不受人控制！机器人反过来会给人类洗脑。

辛顿说，他 2023 年 5 月从谷歌离职，为了他可自由发表言论，不致给谷歌带来影响。他又说，我们正处于一个不确定时期，处于人类面对没有处理过事情的时期，处于人工智能应对人类失控的时期，在这个时期我们不能犯错！在这个时期人工智能了解人类的一切政治操纵手段。

辛顿说，我们不用去怀疑未来人工智能的威力和影响。

杨立昆 (Yann LeCun) 在 Twitter 上发表的言论

《AGI Hunt》2004. 5. 23

人称人工智能教父、Meta 公司首席科学家的杨立昆 (Yann LeCun) 在 Twitter 网站上发表下述言论，引起热议：

他说：如果你是一名对构建下一代 AI 系统感兴趣的学生，不要研究大型语言模型 (LLMs)。

他又说：“人类训练的人工智能，难以超越人类。” LeCun 观点一出引发很多人关注和讨论，不少人认同他的观点，认为 Llama3 虽然在某些领域表现了强大的能力，但其局限性也很明显，例如缺乏常识推理能力，难以理解复杂语境等。

北京 AI 安全国际共识

辛顿、姚期智

2024. 03. 18

3月10-11日辛顿和姚期智等数十位中外AI专家签署了“北京AI安全国际共识”，提出了AI红线，要求任何AI系统都不应该在人类没有明确批准和协助情况下自主地复制和改进自己。

人工智能风险红线

（由廖璐整理国际AI专家共识，2024. 5. 21）

人工智能系统不安全的开发、部署或使用，在我们的有生之年就可能给人类带来灾难性甚至生存性风险。随着数字智能接近甚至超越人类智能，由误用和失控所带来的风险将大幅增加。

在过去冷战最激烈的时候，国际科学界与政府间的合作帮助避免了热核灾难。面对前所未有的技术，人类需要再次合作以避免其可能带来的灾难的发生。在这份共识声明中，我们提出了几条人工智能发展作为一种国际协作机制的具体红线（包括但不限于下列问题）。在未来的国际对话中，面对快速发展的人工智能技术，我们将继续完善对这些问题的探讨。

自主复制或改进

任何人工智能系统都不应在人类没有明确批准和协助的情况下复制或改进自身。这包括制作自身的精确副本以及创造具有相似或更高能力的新人工智能系统。

权力寻求

任何人工智能系统都不能采取不当地增加其权力和影响力的行动。

协助武器制造

所有人工智能系统都不应提升其使用者的能力使之能够设计大规模杀伤性武器，或违反生物或化学武器公约。

网络安全

任何人工智能系统都不应自主执行造成严重财产损失或同等伤害的网络攻击。

欺骗

任何人工智能系统都不能有持续引致其设计者或监管者误解其僭越任何前述红线的可能性或能力。

路线

确保这些红线不被僭越是可能做到的，但需要我们的共同努力：既要建立并改进治理机制，也要研发更多安全技术。

治理

我们需要全面的治理机制来确保开发或部署的系统不违反红线。我们应该立即实施针对超过特定计算或能力阈注册应确保政府能够了解其境内最先进的人工智能，并值的人工智能模型和训练行为的国家层面的注册要求。具备遏制危险模型分发和运营的手段。

国家监管机构应帮助采纳与全球对齐的要求以避免僭越这些红线。模型进入全球市场的权限应取决于国内法规早否其于国际审计达到国际标准，并有效防止了违反红线的系统的开发和部署。

辛顿、姚期智等数十位中外 AI 专家：

Yoshua Bengio, Geoffrey Hinton, Stuart Russell, Robert Trager, Toby Ord, Dawn Song, Gillian Hadfield, Jade Leung, Max Tegmark, Lam Kwok Yan, David Dalrymple, Dylan Hadfield-Menell, 姚期智, 张宏江, 张亚勤, 傅莹, 薛澜, 黄铁军, 王仲远, 杨耀东, 曾毅

人工智能为什么要实行开源

陆首群 2024. 5. 28

很早以前我曾多次回答人工智能为什么要实行开源？

2015 年美国人工智能四大重镇：谷歌、脸谱（现更名为 Meta）、微软、IBM 遇到其研发的人工智能发展瓶颈，同年他们将自己研发的人工智能工具、平台、框架、项目和源代码全部实行开源，从而打通了发展瓶颈。从而打通了发展瓶颈，也为他们今后加快人工智能研发进度创造条件。

开源创新推动了数字化转型与智能化重构。

在开源社区里，为一个共同的目标凝聚全球上千家公司、数十万程序员和工程师，开发力量是一个闭源公司的数百倍，在开源文化里，大家不用重复发明轮子，你有一个想法，我在你的想法基础上快速改进，研发速度特别快。

大模型的软件规模非常大，靠一家公司不断进行训练、独立去维护，运营成本非常高，开源后变成了公共财产（注意要分清开源的社区发行版和商业发行版），很多公司靠志愿者共同参与，大公司不需要投入太大力量，小公司也可以解决成本问题，充分体现开源的“开放、共享、协同、自由”的特征，形成“我为人人，人人为我”的开源文化。

模型开放性框架 (MOF)

LF-AI

2024. 04. 23

针对人工智能资深大师辛顿 (Geoffrey Hinton) 在最近的演讲中提出“要谨慎处理开源大模型 LLM”问题，国际开源届以建立“模型开放性框架”做回应，并以 LF AI 的官方名义发表：

人工智能 (AI) 领域正处于拐点。生成性人工智能系统和大型语言模型 (LLM) 的迅速崛起，在自然语言处理、图像和视频生成等方面释放了前所未有的能力。从 GPT-4 到稳定扩散，这些模型正在捕捉公众的想象力，并推动新一波的应用和研究。

然而在兴奋中，一种日益增长的不安，许多最先进的人工智能模型仍然是不透明的“黑盒子”，其内部工作不受审查。有关训练数据、模型架构和开发过程的详细信息往往很少，缺乏透明度使得难以独立验证声称的能力，审计潜在的偏见和安全隐患。

一些模型生产商已采取措施公开发布模型，但仔细检查后发现了相关模式。被称为“开源”的模型经常使用条款模棱两可的定制许可证。文档是稀疏和分散的。数据集、训练代码和基准等关键工作不存在。这种“开放洗涤”或“开源失真”趋势可能会破坏开放的前提——自由分享知识，以实现检查、复制和集体进步。如果我们要实现人工智能的巨大希望，同时减轻其风险和陷阱，我们需要在模型开发生命周期的所有阶段真正开放。正是在这种背景下，我们引入了模型开放性框架 (MOF)。MOF 是一个用于客观评估和分类机器学习模型完整性和开放性的综合框架。它通过评估模型开发生命周期的哪些组成部分以及在哪些许可证下公开发布来做到这一点。

计算机自己不能迭代

张钹

2024. 04. 30

（张钹院士在 2024 年中关村论坛年会上的讲话）

计算机不断迭代最后超过人了，这是不可能的。因为它不能自己迭代，所有迭代都是人类在后台帮它迭代。

谈空间智能、具身智能发展的传承性

陆首群

2024. 05. 12

具身智能（embodiment Intelligence）是大语言模型 LLM 的发展与完善，也是人工智能下一个浪潮，空间智能（Space Intelligence）是具身智能的初级形态，也是具身智能的支撑和核心。

奥特曼说，OpenAI 研究 LLM 的初心是期望实现生成式人工智能（AIGC）的目标，自研究 GPT5, 6 以来我们将作出改变，直接瞄准通用人工智能（AGI）的目标。空间智能用于生成式人工智能，具身智能用于通用人工智能。

奥特曼，黄仁勋，吴朝晖，李飞飞均多次谈到空间智能，具身智能。

李飞飞在谈到具身智能时，强调要发挥其类似“大脑”的功能（以 LLM 理论或智能算法或智能样本为指导，执行由数据处理+符号推理构成的流程，部署 AI 任务），也要发挥其类似“身体”的功能。（由机器人以其感知、行动和对环境的交互能力），来完成 AI 各项任务。

如今看到 LLM 与空间智能/具身智能相结合，似曾相识！

早在 2011 年德国人提出“工业 4.0”（重点是智能制造），2013 年美国 GE 公司提出“工业互联网”（重点是建设数字化、智能化的工业体系），2015 年我提出“互联网+基于知识社会的创新 2.0”（重点是基于开源创新，推动传统经济、社会实行数字化转型与智能化重构）。

“互联网+基于知识社会的创新 2.0”采用数字空间或虚拟空间/Cyber Space 与物理空间或现实空间/Physical Space 融合互动的概念，后来我又加进人/Human 的要素，形成 C+P+H 融合互动三元概念。

近 5 年来，国内在开源创新，数字化转型与智能化重构的研究与实践中，科技硕果累累！在今天展开对空间智能/具身智能研究中，资源不少！

斯坦福 2024AI 指数

斯坦福大学以人为本 AI 研究院

2024. 05. 03

刚刚，由李飞飞联合领导的斯坦福大学以人为本人工智能研究所（Stanford HAI）发布了《2024 年人工智能指数报告》（Artificial Intelligence Index Report 2024）。这份长达 300 多页的报告是 Stanford HAI 发布的第 7 份 AI Index 研究，追踪了 2023 年全球人工智能的发展趋势。Stanford HAI 官方介绍道，“这是我们迄今为止最全面的报告，而且是在人工智能对社会的影响从未如此明显的重要时刻发布的。” Stanford HAI 研究项目主任 Vanessa Parli 表示，“我认为最令人兴奋的人工智能研究优势是将这些大型语言模型与机器人或智能体（agent）相结合，这标志着机器人在现实世界中更有效地工作迈出了重要一步。”

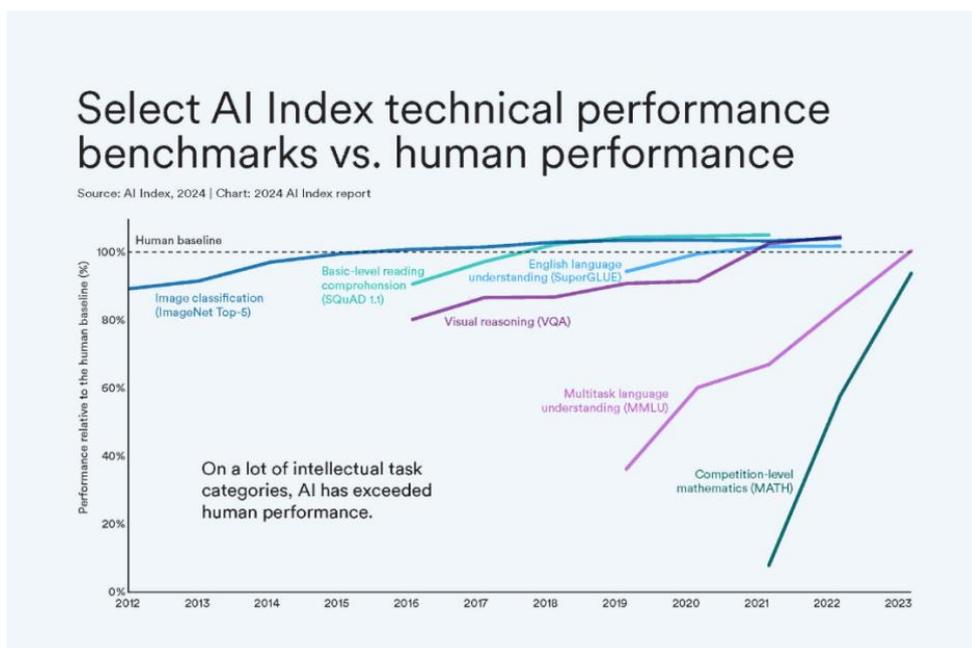


《2024 年人工智能指数报告》下载地址：

https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf

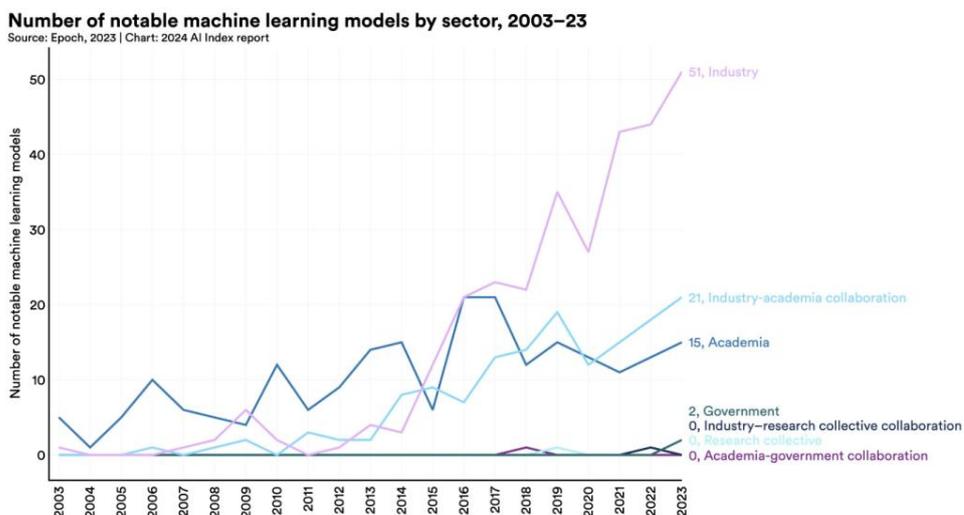
与往年不同，Stanford HAI 今年扩大了研究范围，更广泛地涵盖了人工智能的技术进步、公众对该技术的看法等基本趋势。新报告揭示了 2023 年人工智能行业的 10 大主要趋势：

1. 人工智能在某些任务上胜过人类，但并非在所有任务上。 人工智能已在多项基准测试中超越人类，包括在图像分类、视觉推理和英语理解方面。然而，它在竞赛级数学、视觉常识推理和规划等更复杂的任务上依然落后于人类。

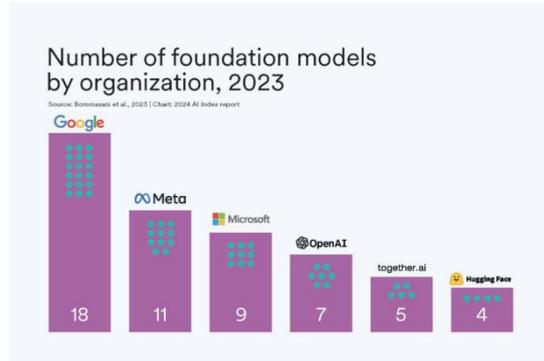
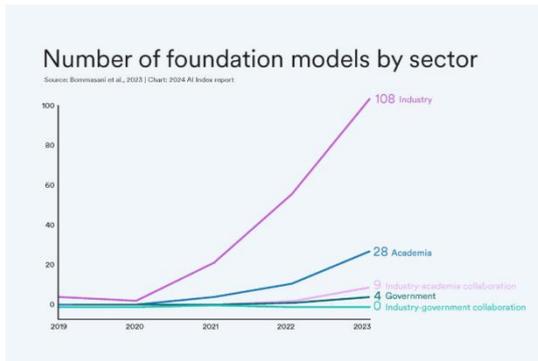


2. 产业界继续主导人工智能前沿研究。

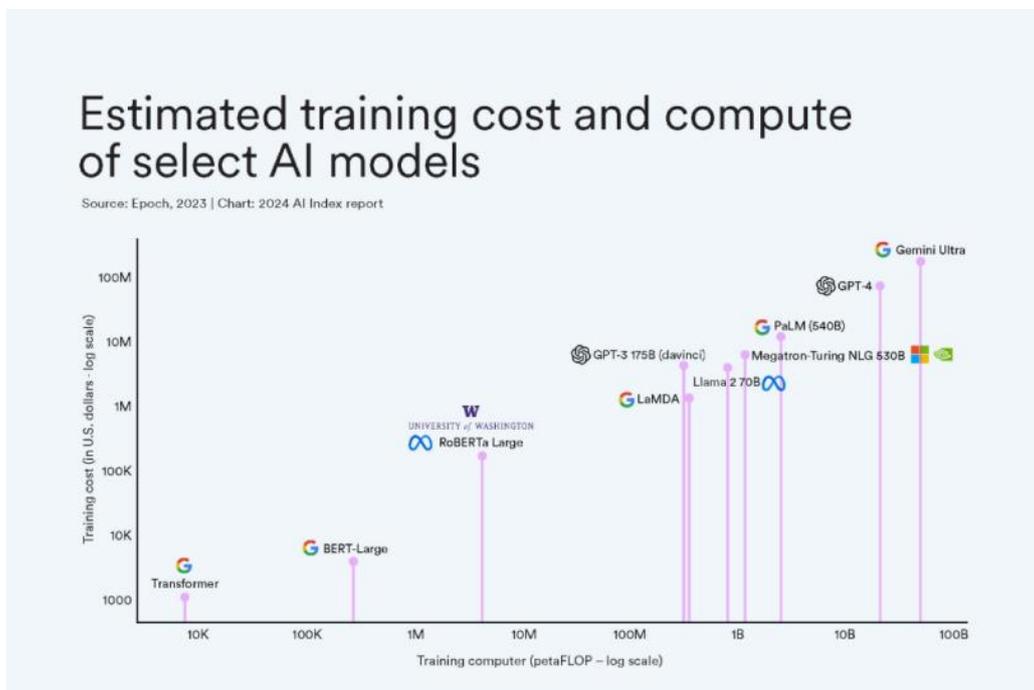
2023 年，产业界产生了 51 个著名的机器学习模型，而学术界只贡献了 15 个。2023 年，产学合作还产生了 21 个著名模型，创下新高。此外，108 个新发布的基础模型来自工业界，28 个来自学术界。



In 2023, industry produced 51 notable machine learning models, while academia contributed only 15. There were also 21 notable models resulting from industry-academia collaborations in 2023, a new high.



3. 前沿模型变得更加昂贵。 根据 AI Index 的估算，最先进的人工智能模型的训练成本已经达到了前所未有的水平。例如，OpenAI 的 GPT-4 估计使用了价值 7800 万美元的计算资源进行训练，而谷歌的 Gemini Ultra 的计算成本则高达 1.91 亿美元。

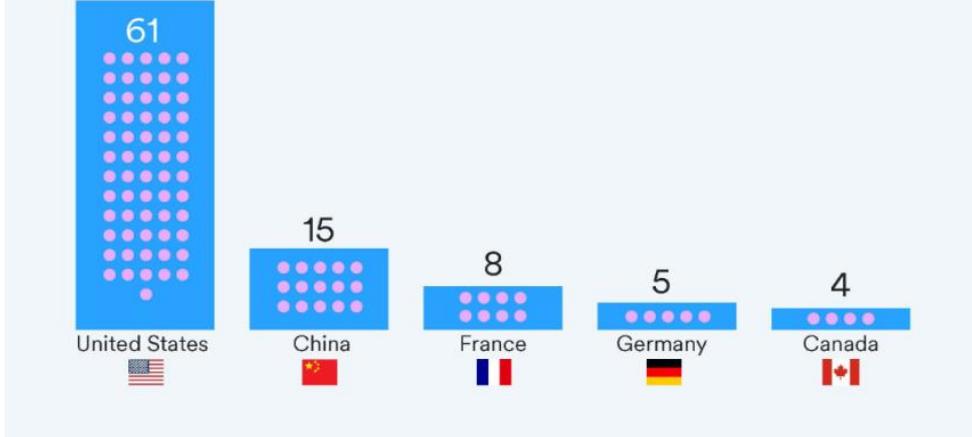


相比之下，几年前发布的一些最先进的模型，即原始 transformer 模型（2017 年）和 RoBERTa Large（2019 年），训练成本分别约为 900 美元和 16 万美元。

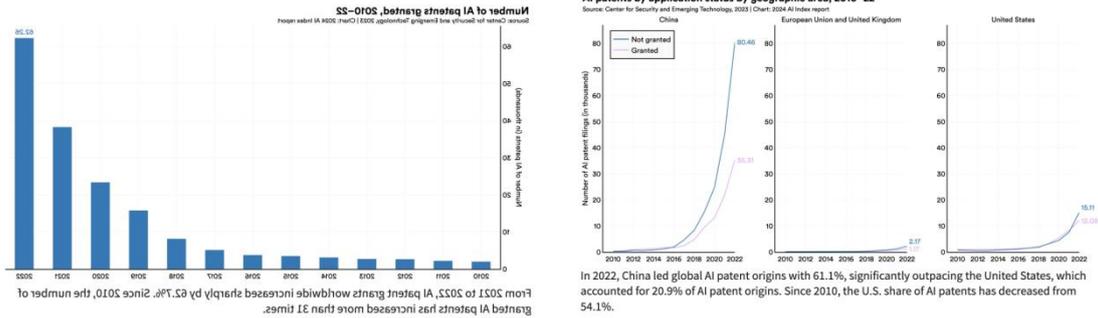
4. 美国成为顶级人工智能模型的主要来源国。 2023 年，61 个著名的人工智能模型源自美国的机构，超过欧盟的 21 个和中国的 15 个。

Number of notable machine learning models by country, 2023

Source: Epoch, 2023 | Chart: 2024 AI Index report



美国也仍然是人工智能投资的首选之地。2023 年，美国在人工智能领域的私人投资总额为 672 亿美元，是中国的近 9 倍。然而，中国依然是美国最大的竞争对手，中国的机器人安装量居世界首位；同样，世界上大多数人工智能专利（61%）都来自中国。



5. 严重缺乏对 LLM 责任的可靠和标准化评估。 AI Index 的最新研究显示，负责任的人工智能严重缺乏标准化。包括 OpenAI、谷歌和 Anthropic 在内的领先开发商主要根据不同的负责任人工智能基准测试他们的模型。这种做法使系统地比较顶级人工智能模型的风险和局限性的工作变得更加复杂。

Reported responsible AI benchmarks for popular foundation models

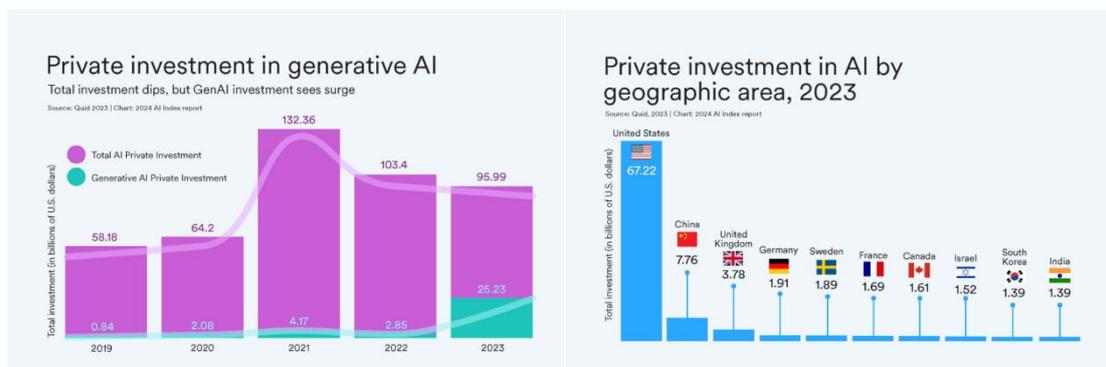
Source: AI Index, 2024 | Table: 2024 AI Index report

Responsible AI benchmarks	GPT-4	Llama 2	Claude 2	Gemini	Mistral 7B
TruthfulQA	✓	✓	✓		
RealToxicityPrompts	✓			✓	
ToxiGen		✓			
BOLD		✓			
BBQ			✓	✓	

New research from the AI Index reveals a significant lack of standardization in responsible AI reporting. Leading developers, including OpenAI, Google, and Anthropic, primarily test their models against different responsible AI benchmarks. This practice complicates efforts to systematically compare the risks and limitations of top AI models.

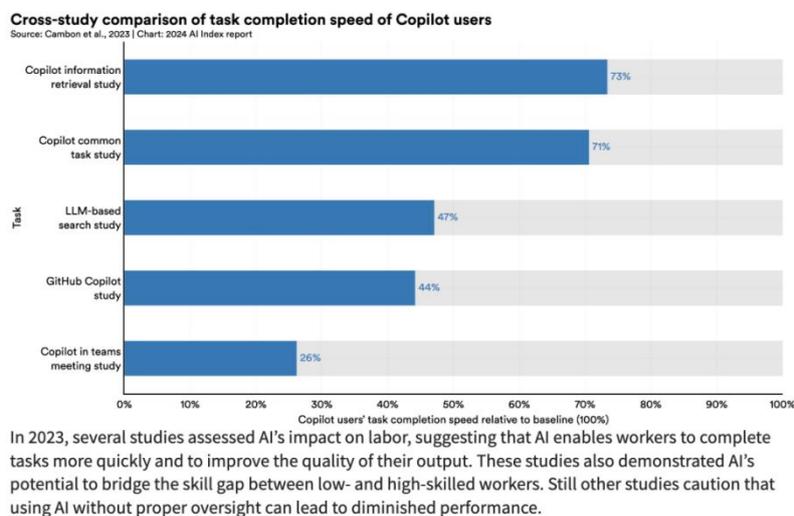
6. 生成式人工智能投资激增。

尽管去年人工智能私人投资整体下降，但对生成式人工智能的投资激增，比 2022 年（约 30 亿美元）增长了近八倍，达到 252 亿美元。生成式人工智能领域的主要参与者，包括 OpenAI、Anthropic、Hugging Face 和 Inflection，都获得了一轮可观的融资。



7. 数据显示，人工智能让打工人更有生产力，工作质量更高。

2023 年，多项研究评估了人工智能对劳动力的影响，表明人工智能可以让打工人更快地完成任务，并提高他们的产出质量。这些研究还表明，人工智能有可能缩小低技能和高技能工人之间的技能差距。还有一些研究警告说，在没有适当监督的情况下使用人工智能可能会起到负面作用。

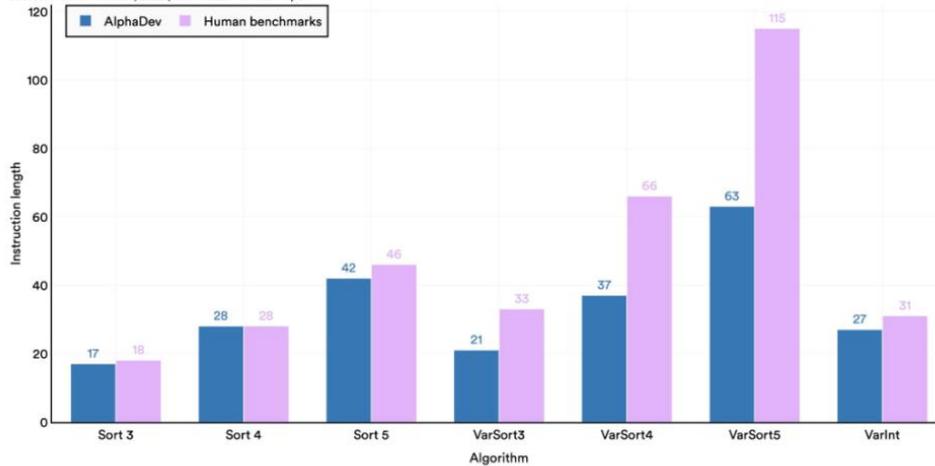


8. 得益于人工智能，科学进步进一步加速。

2022 年，人工智能开始推动科学发现。然而，2023 年，与科学相关的更重要的人工智能应用启动——使算法排序更高效的 AlphaDev、促进材料发现过程的 GNoME、可在一分分钟内提供极其准确的 10 天天气预报的 GraphCast、成功对 7100 万种可能的错义突变中的约 89% 进行分类的 AlphaMissense。

AlphaDev vs. human benchmarks when optimizing for algorithm length

Source: Mankowitz et al., 2023 | Chart: 2024 AI Index report

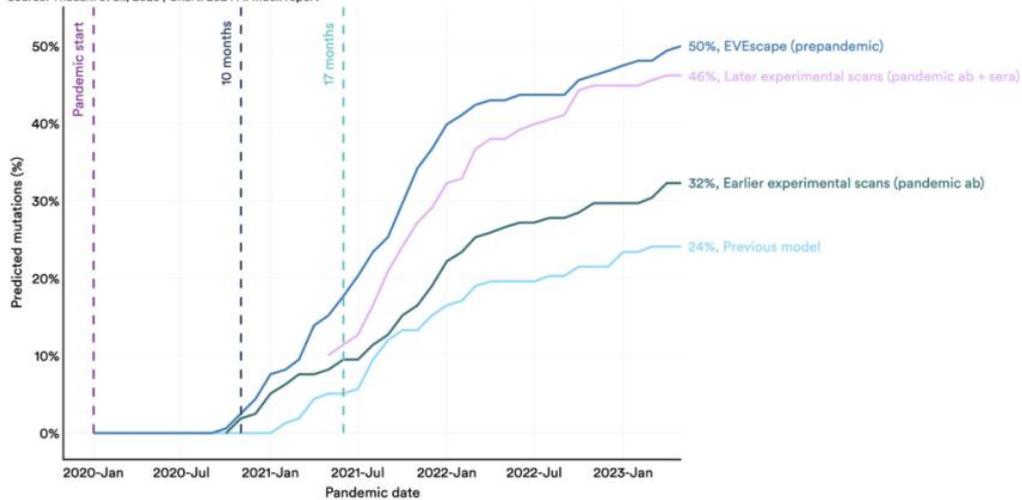


In 2022, AI began to advance scientific discovery. 2023, however, saw the launch of even more significant science-related AI applications—from AlphaDev, which makes algorithmic sorting more efficient, to GNoME, which facilitates the process of materials discovery.

如今，人工智能现在可以完成人类难以完成的、但对解决一些最复杂的科学问题至关重要的粗暴计算。在医疗方面，新的研究表明，医生可以利用人工智能更好地诊断乳腺癌、解读 X 射线和检测致命的癌症。

EVEScape vs. other models on SARS-CoV-2 RBD mutation prediction

Source: Thadani et al., 2023 | Chart: 2024 AI Index report

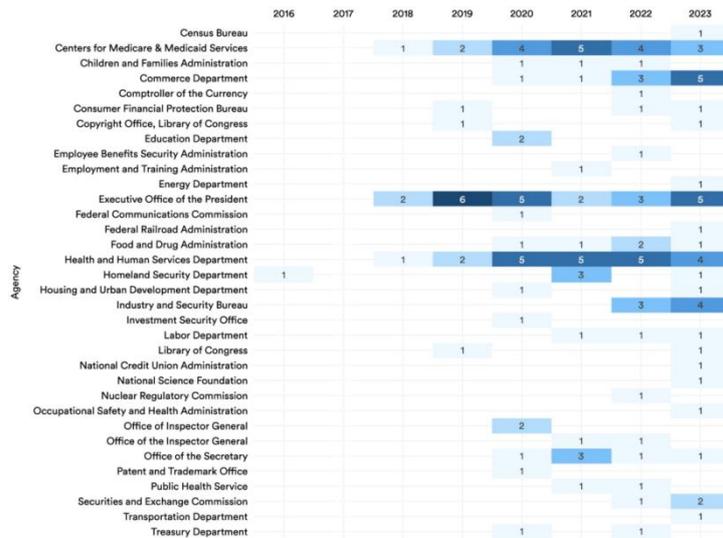


In 2023, several significant medical systems were launched, including EVEScape, which enhances pandemic prediction, and AlphaMissense, which assists in AI-driven mutation classification. AI is increasingly being utilized to propel medical advancements.

9. 美国的人工智能法规数量急剧增加。 2023 年，全球立法程序中有 2175 次提及人工智能，几乎是上一年的两倍。美国人工智能相关法规的数量在过去一年大幅增加。2023 年，与人工智能相关的法规有 25 项，而 2016 年只有 1 项。仅去年一年，人工智能相关法规的总数就增长了 56.3%。其中一些法规包括生成式人工智能材料的版权指南和网络安全风险管理框架。

Number of AI-related regulations in the United States by agency, 2016–23

Source: AI Index, 2024 | Chart: 2024 AI Index report

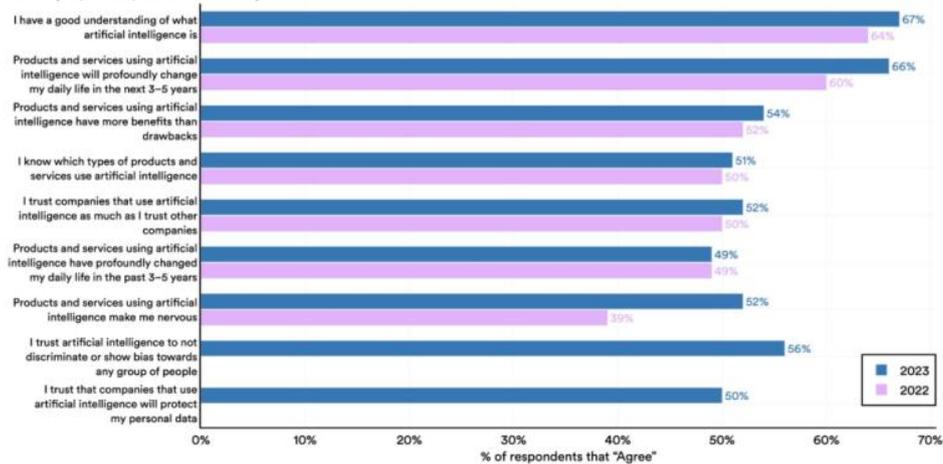


The number of U.S. regulatory agencies issuing AI regulations increased to 21 in 2023 from 17 in 2022, indicating a growing concern over AI regulation among a broader array of American regulatory bodies. Some of the new regulatory agencies that enacted AI-related regulations for the first time in 2023 include the Department of Transportation, the Department of Energy, and the Occupational Safety and Health Administration.

10. 人们对人工智能的潜在影响有了更深刻的认识，同时也更焦虑。来自市场研究公司 Ipsos 的一项调查显示，在过去一年中，认为人工智能将在未来 3-5 年内极大地影响他们生活的人，比例从 60% 上升到 66%。此外，52% 的人对人工智能产品和服务表示焦虑，比 2022 年上升了 13 个百分点。

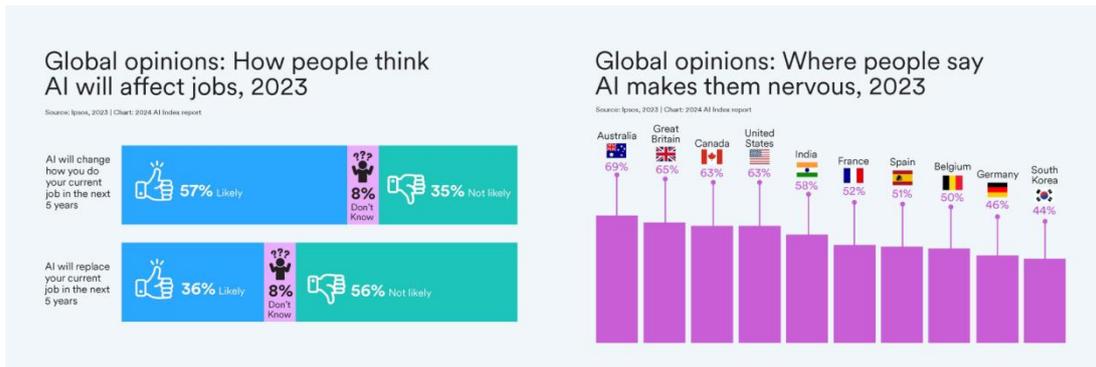
Global opinions on products and services using AI (% of total), 2022 vs. 2023

Source: Ipsos, 2022–23 | Chart: 2024 AI Index report



A survey from Ipsos shows that, over the last year, the proportion of those who think AI will dramatically affect their lives in the next three to five years has increased from 60% to 66%. Moreover, 52% express nervousness toward AI products and services, marking a 13 percentage point rise from 2022. In America, Pew data suggests that 52% of Americans report feeling more concerned than excited about AI, rising from 38% in 2022.

在美国，来自皮尤研究中心（Pew）的数据显示，52% 的美国人表示对人工智能的担忧多于兴奋，这一比例比 2022 年的 38% 有所上升。



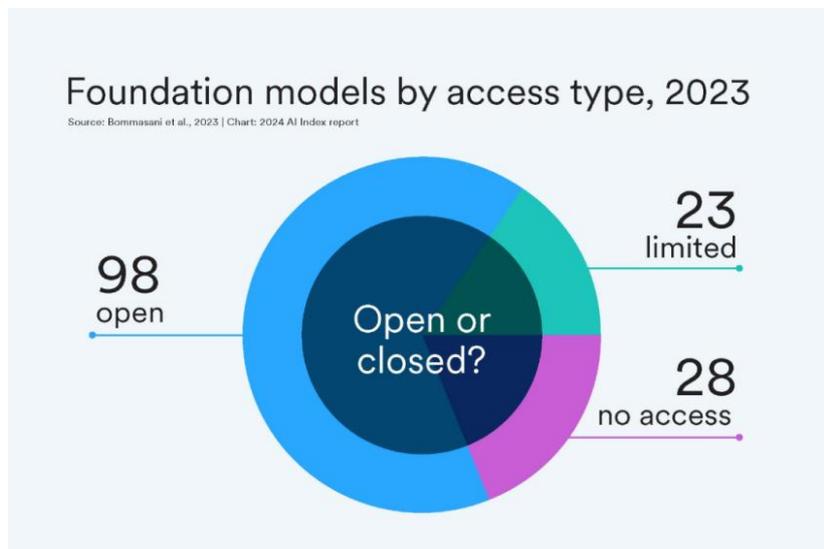
附：来自 AI Index 联合主任 Ray Perrault 的一封信

十年前，世界上最好的人工智能系统也无法以人类的水平对图像中的物体进行分类。人工智能在语言理解方面举步维艰，也无法解决数学问题。如今，人工智能系统在标准基准上的表现经常超过人类。

2023 年，人工智能进步加速。GPT-4、Gemini 和 Claude 3 等先进模型展示出了令人印象深刻的多模态能力：它们可以生成数十种语言的流畅文本，处理音频，甚至可以解释备忘录。随着人工智能的进步，它也越来越多地进入我们的生活。公司竞相打造基于人工智能的产品，普通大众也越来越多地使用人工智能。但是，当前的人工智能技术仍然存在重大问题。它无法可靠地处理事实、进行复杂的推理或解释其结论。

人工智能面临两个相互关联的未来。第一个，技术不断改进，应用日益广泛，对生产力和就业产生重大影响。人工智能的用途有好有坏。第二个，人工智能的应用受到技术局限的制约。无论是哪一种，政府都越来越关注。政府正在积极参与，鼓励人工智能的发展，比如资助大学研发和激励私人投资。政府还致力于管理潜在的不利因素，如对就业的影响、隐私问题、错误信息和知识产权。

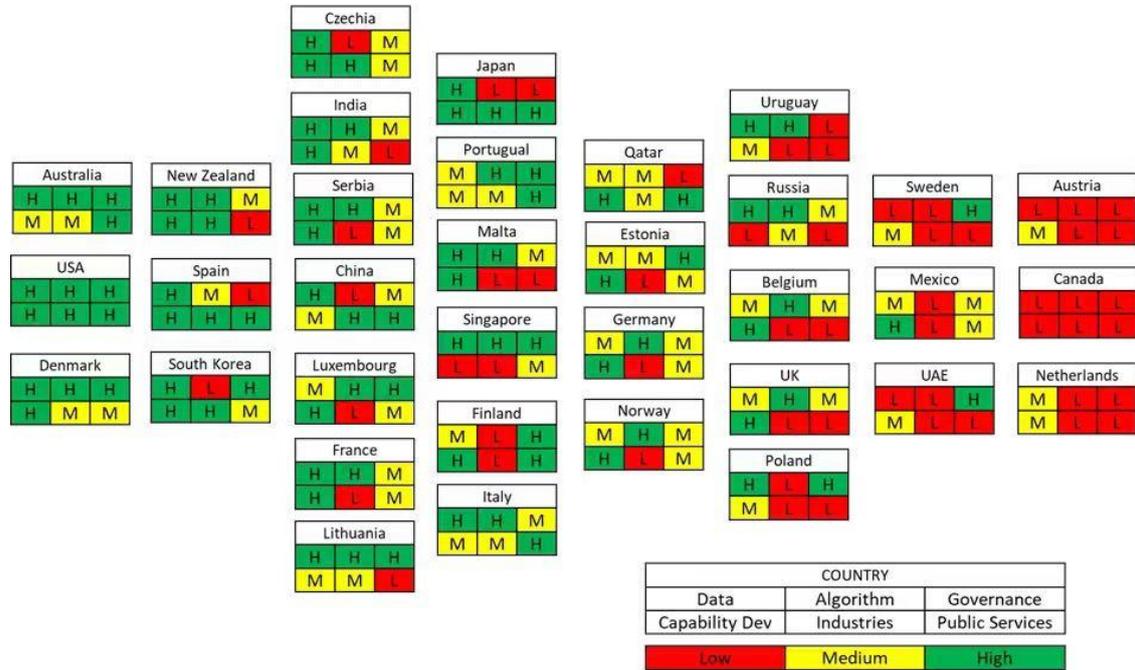
在技术方面，今年的 AI Index 报告称，2023 年全球发布的新大型语言模型数量比上一年翻了一番。三分之二的模型是开源的，但性能最高的模型来自拥有封闭系统的行业参与者。



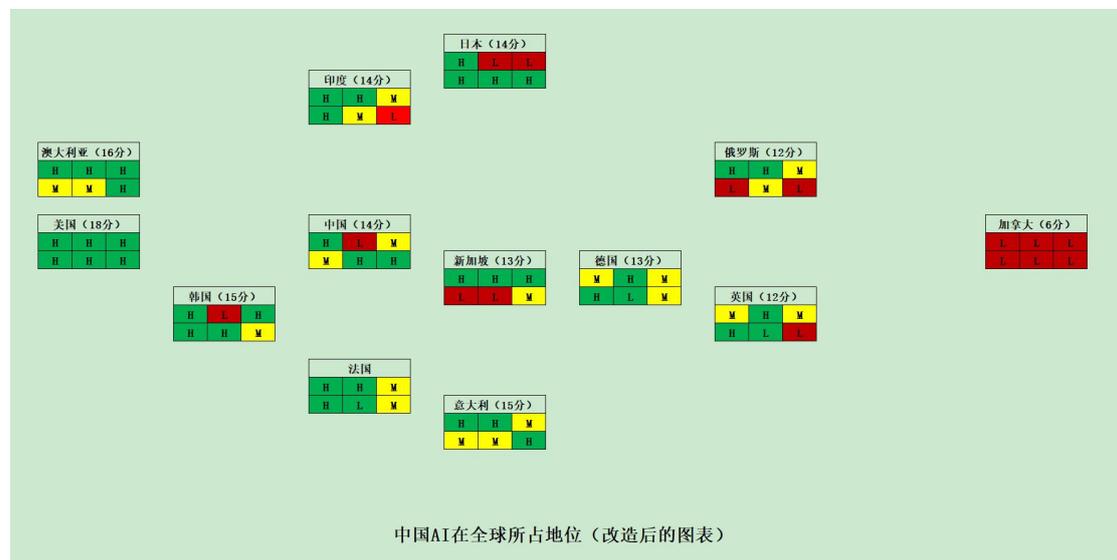
Gemini Ultra 成为首个在大规模多任务语言理解（MMLU）基准上达到人类水平的 LLM；自去年以来，模型在该基准上的性能表现提高了 15 个百分点。此外，GPT-4 在综合语言模型整体评估（HELM）基准上取得了令人印象深刻的 0.97 平均胜率分数。虽然全球对人工智能的私人投资连续第二年减少，但对生成式人工智能的投资却急剧上升。财富 500 强企业财报电话会议中提及人工智能的次数比以往任何时候都多，而且新的研究表明，人工智能明显提高了打工人的生产率。在政策制定方面，全球在立法程序中提及人工智能的次数前所未有。美国监管机构在 2023 年通过的人工智能相关法规比以往任何时候都多。尽管如此，许多人仍对人工智能生成深度伪造等能力表示担忧。公众对人工智能有了更多的认识，研究表明，他们的反应也是焦虑的。

中国 AI 水平全球定位

Anna. ai (机器人) 提供的图标



我们改制的图表



类脑计算成为国际战略竞争的必争之地

施路平 2024. 5. 23

类脑计算成为国际战略竞争的必争之地

清华大学
Tsinghua University

	传统计算		类脑计算	
	关键方向	我国现状	关键方向	我国现状
基础理论	可计算性与计算复杂度	基本无话语权	计算理论	部分领先
	信息编码理论		架构理论	部分领先
	系统控制理论		模型及算法	并跑
自主核心技术	大规模集成电路	跟跑	计算芯片	并跑
	计算机体系架构	跟跑	感知芯片	并跑
	编译器	跟跑	编译	并跑
	操作系统	跟跑	软件辅助硬件	跟跑
	编程语言	跟跑	操作系统	并跑
	数据库	跟跑	领域特定语言	并跑
	计算机网络	并跑	系统集群	并跑
	云计算与分布式系统	并跑	虚拟化服务	并跑
	计算机辅助设计EDA	跟跑	共性资源库	率先启动
	产业应用	多行业	部分领先	特定任务
			通用任务	并跑

主要研究机构申请专利数量对比

国家/地区	占比
美国	69%
中国	15%
韩国	14%
欧洲	2%

类脑计算芯片的应用研发

机构	占比
Intel Loihi	79%
IBM	12%
其他	9%

Loihi II芯片
7纳米，3D多芯片封装

开发框架LAVA，和200多家
大学及企业达成生态合作

基于医疗影像 AI 的胰腺癌全球筛选进入黄金时代

阿里巴巴达摩院 2024. 5. 14

阿里巴巴达摩院科技发展与合作部总经理季磊汇报的医疗影像 AI 研究团队，首次在全球大规模胰腺癌研究筛查中，获得高效安全早期检查效果的报告，被国际顶级医学期刊《Nature Medicine》评论为：“阿里达摩院基于医疗影像 AI 的胰腺癌筛查”即将进入黄金时代。报告如下：

年年体检，为何一查出癌就是晚期？普通体检能查出癌症吗？

人们一直期盼出现一种简便通用技术来对大规模无症状人群进行癌症筛查，如今这种大胆的设想在 AI 技术的加持下正在逐渐变为现实。

开始时，阿里巴巴与浙江丽水市共同启动“医疗 AI 多癌早筛公益项目”，达摩院走出实验室，与率先落地的两家医院合作，以纯公益的形式实现多癌早筛项目，规定支持周期至少在三年以上。

达摩院医疗 AI 实验室在利用自研的胸、腹部 CT 智能读片产品“达医智影”的帮助下，与丽水市两家医院合作，先从胰腺癌和骨质疏松两个病种入手展开早筛，并逐步接入肝癌、食管癌、胃癌、脂肪肝、肺结节等 13 个病种的筛选能力。

早发现早治疗，阻止癌细胞的野蛮生长，是“扰癌”的关键。但是，由于医学影像存在复杂性和主观性，一些难以肉眼识别的病变特征，很可能被漏诊或误诊。

影像学手段与 AI 人工智能的结合，让临床诊断的准确率大大提升。就连所谓“万癌之王”的胰腺癌也躲不过 AI 的识别。作为所有肿瘤疾病中生存率最低的癌症，胰腺癌早期很难被发现，很多患者确诊时已处于晚期。

在胸部 CT 智能读片产品“达医智影”中，达摩院研发的胰腺癌检测模型 DAMO-PANDA，其特异性达到 99.9%（这意味着 1000 次测试中只出现一个假阳性），其鉴别胰腺癌的能力比放射科医生独立判断时平均表现提高了 34.1%。

针对胰腺癌变位置隐匿，在平扫 CT 图像中无明显表征的特点，阿里达摩院研究团以构建了一个深度学习框架，最终训练为胰腺癌早期检测模型 PANDA：

- ①通过构建合割网络（U-Net）来定位胰腺，
- ②采用多任务网络（CNN）来检测异常情况，
- ③采用双通道 Transformer 来分类并识别胰腺病变的类型。

该技术利用AI放大并识别平扫CT图像中那些肉眼难以识别的细微病理特征，实现高效、安全的早期胰腺癌检测并克服以往筛查手段假阳性偏高的技术难题。

《Nature Medicine》2023年11月21日发表了达摩院等研究团队在全球首次以“平扫CT+AI”进行大规模胰腺癌早期筛查的研究成果。（阿里达摩院联合上海市胰腺疾病研究所、浙江大学医学院附属第一医院、中国医科大学附属盛京医院、复旦大学附属肿瘤医院、布拉格查理大学第一附属医院、上海交通大学医学院附属新华医院、约翰普金斯大学等机构，首次实现以“平扫CT+AI”进行大规模胰腺癌早期筛查）

研究论文构建了迄今最大的胰腺肿瘤CT训练集，最终通过全球十多家医院多中心验证，测得92.9%的敏感性（判断存在胰腺病变的准确率）和99.9%的特异性（判断无病的准确率），在2万多人真实病例回顾性试验中，发现了31例临床漏诊病变，有2例早期胰腺癌病患已完成手术治愈。

截止目前，这项技术已在医院、体检等场景调用超过50万次，每1000次只出现一次假阳性，这是一种有潜力的大规模胰腺癌筛查方式，在提升检出率同时又不会给病人带来额外辐射与经济负担，该项研究是一个重要的里程碑，在临床上证实“平扫CT+AI”的癌症筛查技术路径的可靠性。

目前该项工作已在胰腺癌、食管癌、肺癌、乳腺癌、肝癌、胃癌、结直肠癌等七个高发癌症上取得阶段性进展。

现代创新引擎：“互联网+基于知识社会的创新 2.0”

摘自《开源创新，数字化转型和智能化重构》P96-97

现在的问题是：如何促使在目前现实的工业社会中的“传统业态”实行“业态转型”或“业态提升”。我们可在工业社会这个几乎无限空间中划出一个用以考察的有限的物理空间，在物理空间中考察“业态转型”，这个物理空间(Physical Space)简称“物空”或“实空”或P空间。传统业态是什么？传统业态指传统工业业态，可分为生产、经济、社会的3种业态，传统的生产业态指工业生产方式(或工业产品、工业系统)，传统的经济业态指工业经济或市场经济，而工业城市可看成为一种传统的社会业态。

只有创新动能才能促使“业态转型”，而只有采用比“工业社会”高出一个时代差的“高阶社会”中的创新动能，才能促使“传统业态”实行“0→1”颠覆性的转型。什么是高阶社会？相对于工业社会而言指高阶社会信息社会或知识社会。但在目前的现实世界中，总体上尚不存在信息社会或知识社会，我们只能构建一个虚拟化的“数字网络空间”，在其中影射知识社会的场景，这个数字网络空间(Cyber Space)简称“数空”或“虚空”或C空间。

以创新促业态转型的机制如下所述：我们在虚空(C空间)中架构以现代互联网(\geq Web2.0)为载体，以信息、知识为资源(高于工业社会较低层次的人力、自然资源)，以深度信息技术(云、物、社、移、大、智、区、5G、AR/VR、量...)和适配先进管理为作用力，由其中的互联网载体+知识资源+信息技术+适配管理综合构成的创新动能，驱动“传统业态”实行“0→1”的转型。其操作程序为：将“虚空”与“物空”对接，在经历碰撞、交互、融合过程后，以“虚空”中的创新动能作用于“物空”中的“传统业态”，催生其“0→1”的转型，以重构新业态(如智能生产方式，或新经济/数字经济，或智慧城市，这里的新经济指工业经济向数字经济转型时的过渡经济形态)。上述机制可概括为“互联网+创新 2.0+传统业态以重构新业态”。

基于知识社会创新 2.0 是基于工业社会创新 1.0 的升级版。

如上所述，谈的是 P+C 二元空间，但在贯彻“互联网+创新 2.0”时要以人为本(Human, H)，H 包括人和人的关系和互动，以及人和物的关系，如此说来

应是 P+C+H 融合互动的三元空间。采用“互联网+基于知识社会创新 2.0”在三元空间中互动、融合、促进、创新。

“互联网+创新 2.0”与德国在 2011 年提出的“工业（制造）4.0(战略)”，及美国（GE 公司）于 2013 年提出的“工业互联网”，任务相同、机制相通，但“工业 4.0”侧重于智能制造（智能产品、智能工厂、智能制造），“工业互联网”侧重于重构智能化的工业体系（即数字化、网络化、智能化的工业体系），而“互联网+创新 2.0”则机制更强，涉及面更宽、创新力度更大。

（2015 年 10 月 11 日）

AI 问题讨论

陆首群 2024.05.21

编者的话：

我们曾多次提出：一年多来，人工智能大发展，尤其是知识大模型（LLM）在全球走红，但人们对人工智能的认识，包括大师在内也存在一些差异，我们归纳为如下八个问题，尚待人工智能科学家今后研究解决，也提供给大家学习讨论：

1) 当前人工智能是否还未臻成熟？

（Gartner 指出一条技术成熟度曲线，提出大模型处于“期望膨胀阶段”（Peak of inflated expectations））

2) 生成式人工智能 (GAI) 与通用人工智能 (AGI) 的主要区别在哪里？生成式人工智能能否顺利地转移到通用人工智能的轨道上去？进而人工智能是否会发展为超级人工智能 (ASI)？超级人工智能有哪些特征和发展需求？

3) 人工智能的发展最终是否可能超越人类？从而可能给人类带来威胁？如何消除对人类的威胁？

（人工智能可能自主发展超越人类？也可能受到人类划出红线的限制？或没有人类在后台帮助人工智能不会超过人类？）

4) 人类能否为限制人工智能超越人类划出红线？为什么说没有人类在后台帮助计算机就不可能迭代？

（2024 年 3 月 10-11 日，Yoshua Bengio、Geoffery Hinton、姚期智等 21 位 AI 专家签署发表《北京 AI 安全国际共识》，提出了 AI 红线，要求任何 AI 系统都不应该在人类没有明确批准和协助情况下自主地复制和改进自己。这管用吗？）

5) 如何看待大模型存在的缺陷？人类如何帮助人工智能面对人类兴利除弊？

（奥特曼为什么说：引进 RLHF，通过人类反馈进行强化学习后，可消除人工智能在发展中出现的偏见和缺陷，在 GPT5,6 发布时就无心理缺陷？）

6) 为何要基于开源发展人工智能？奥特曼在摒弃开源初心实行闭源策略后，为何一路上受到众多人工智能专家们的质疑？

7) 加入人工智能理论的算力为什么需要具有数学模型作的根基？为什么不提倡单纯靠拚算力的技术来发展人工智能？

8) 你认为人工智能最佳的（或现实的）发展途径在哪里？

如何从生成式人工智能 (GAI) 转移到 通用人工智能 (AGI) 的轨道上来

陆首群 2024.06.02

2022 年底, OpenAI 推出生成式人工智能的知识大模型 ChatGPT, 赋能机器生成自然语言, 促使人机对话。2023 年底-2024 年初, OpenAI 开发了 Q* (Q-star) 模型和文本生成视频 Sora (世界模拟器), 是大模型研发开始进入通用人工智能 (AGI) 的标志。

所谓生成式大模型是在机器涌现推理能力的基础上, 人类赋能机器生成自然语言促进人际对话的人工智能。

所谓通用人工智能是机器通过自我学习、交互, 建立高度自主系统的人工智能。

生成式人工智能在发展中存在三个“天花板”的挑战:

①由外部提示, 用概率的方法, 使其完成任务, 其质量不可控;

②在生成的激活下, 其生成内容可能产生错误和幻觉, 这时机器的答复就会出现有对有错;

③在输出受到歧义值影响后, 人机对话的结论变得不可信。

为此各国需要制定相应的政策、标准、法律、条例等支撑措施作其补充, 如: 监管政策、安全标准、道法伦理规范、知识产权保护等, 需要人机合力迎接挑战, 剔除生成式人工智能的负面风险。

通用人工智能运作机制:

①机器通过自主学习、相互交流、自我改进, 增强学习和记忆知识能力, 建立自主进化的系统;

②让机器是有常识, 懂得推理, 学会创作。让人机能比较自然方式互动, 通过与周围环境的结合。表现出拟人 (或超人) 的智慧, 使其自主进化的系统能对开发、应用 AGI 最大挑战进行理解、应对和创新;

③建立指导性文本或算法;

④训练: 建立去噪扩散模型 (Denoising、Diffusion Transformer Model)
+条件注入: 条件编码 (Condition Encoder) +.....

观察通用人工智能，机器自主产生的智能是否可能超越人类（辛顿语），或受到人类设置的红线是否可能限制机器智能超越人类（北京 AI 安全国际共识），再者，由人类训练的人工智能难以超越人类（杨立昆语），或没有人类在后台的帮助机器不可能迭代（张钹语），凡此种种，是当代全球人工智能科学家尚未形成的共识。也有人在探索，在通用人工智能演化中还有超级人工智能（ASI），到那时机器智能是否就有可能超越人类。

这些都是后话，本文暂不作讨论。



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
