从感知到认知，“知识方程”能否通向强人工智能

人工智能（AI）大模型的诞生，让2023年成为人类历史上一个重要年份——通用人工智能元年。这意味着以智能革命为代表的第四次工业革命已然到来，人类站到了智能时代的门槛上。

人类所经历的前三次工业革命，都令人类文明实现了新的飞跃，也对世界格局产生了深远影响。长远来看，人工智能在很多方面的智慧能力将可能超过人类，但目前通用人工智能还需迈过多道门槛，才能真正实现强人工智能。

日前，中国科学技术大学知识计算实验室提出了新的知识模型“知识方程”，并以此为基础建立新型专家系统，通过与深度学习的结合，尝试突破现有通用人工智能的技术瓶颈。

图/视觉中国

■周熠

人工智能（AI）诞生至今只有短短60多年，但其发展经历了几度跌宕起伏。近年来，深度学习技术为AI带来了新的革命，其中包括我们熟悉的AlphaGo、ChatGPT等。

当前AI技术在很多任务上取得了超越人类的成效，包括人脸识别、语音识别、字符识别等，在机器翻译、问答和医疗诊断等领域也交出了令人相对满意的答卷。可以说，AI已经迈入能够大规模落地应用的阶段。

不过，当我们试图将人工智能再向前推进，就会发现要克服其现有缺陷，还有待技术模式的创新与突破。

当下大模型面临4个关键缺陷

OpenAI公司推出的ChatGPT是一个AI聊天机器人程序，更是一个人工智能自动生成内容（AIGC）的工具。作为一个对话系统，ChatGPT具备出色的多功能性，无论是畅聊多种话题、解答数学题，还是提供礼品选择建议、制定行程规划，都可从容应对。因此，从某种意义上来说，ChatGPT具备了广泛的应用潜力和灵活性，可以说是一个通用人工智能（AGI）程序。

尽管在某些方面（例如逻辑和语义理解等）的评测表明，ChatGPT并非在所有领域都比现有的最佳模型更强大，但现有的最佳模型可能只是针对特定任务而设计，而ChatGPT则是一个通用模型。

其实，多年前人们就认识到大模型的巨大潜力，但其实际发展速度仍比预想中快了许多。ChatGPT大模型甫一问世，很快就在应用层面受到极大关注。半年后，中国就涌现出了百余个大模型。

目前，大模型的应用主要分为生成式人工智能（AIGC）、大模型辅助工具、个人智能交互3类。其中，个人智能交互尤其值得关注。任何真正能促进交互的技术和产品都能产生巨大价值。这种交互不仅包括“人—人”（通过机器），也包括“人—机”，甚至包括“机—机”。而人工智能，包括大模型，有望在这方面取得颠覆性突破。

然而，目前大模型的应用之路并不如想象的那么顺利。根本原因在于，虽然大模型很强大，但仍然在技术上存在一些关键缺陷。

首先，大模型有时会犯事实性错误，也就是可靠性问题（俗称“幻觉”）。例如，它会弄错诗词作者。从原理看，大模型的答案选择基于概率，所以很难保证百分百正确。这一问题在目前许多领域的大模型中都存在，也是大模型面临的最重要的挑战之一。

第二，大模型的数学和逻辑推理能力仍然需要加强。虽然GPT-4在某些考试中表现优异，但在面对一些精心设计的逻辑推理问题时，大模型的回答与随机答案相差无几。因为在进行深度推理时，即便大模型每一步的预测准确率都高达95%，但是当推理到20步时，最终的准确率将会是0.95的20次方，即不到36%，这是一个无法令人满意的结果。

第三，大模型的形式语义理解能力有待提升。虽然大模型在一定程度上能够实现语义理解，但要想真正从意义和形式上完全理解语言背后的意义，还有很大的改进空间。

最后，大模型作为一个黑盒模型，存在一些通用弱点。比如，其可解释性、可调试的能力较弱等。

通向强人工智能或需另辟蹊径

大模型开启了通用人工智能落地应用的窗口。但正如前文所说，技术上的一些关键缺陷意味着它与通用强人工智能之间尚有很大距离。要缩短这个距离，至少有两条不同路径值得探索。

第一条路径就是继续沿着大模型现有的发展路线向前走。AI诞生不过60多年，GPT真正开始训练至今也只有5年。如果让大模型再发展5年、50年、500年，它会取得怎样的进步？这是一个值得思考的问题。

沿着现有技术路线，大模型的发展在两个关键点上会遇到一定的挑战。

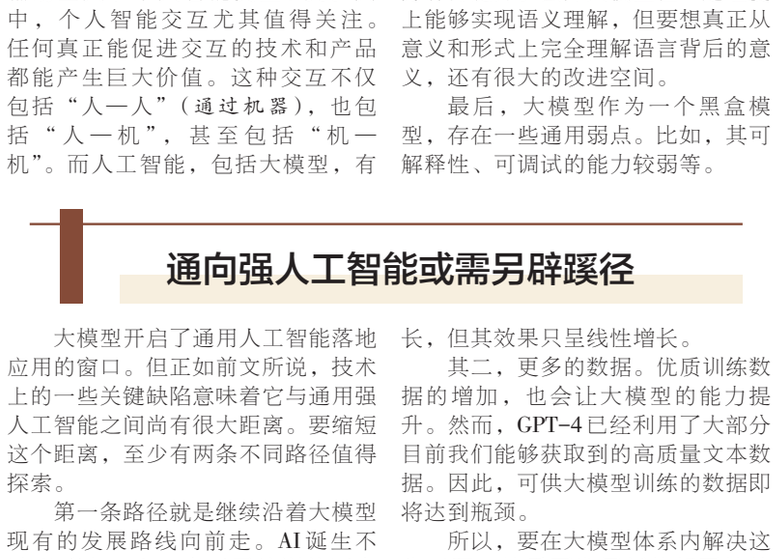
其一，更多的参数。参数量的增加，会让大模型的能力提升。摩尔定律表明，计算能力每18个月到24个月翻一番，而目前大模型的参数量正以三四个月翻番的速度增长。因此，计算能力很快会跟不上模型的发展需求。而且，尽管参数量呈指数级增长，但其效果只呈线性增长。

其二，更多的数据。优质训练数据的增加，也会让大模型的能力提升。然而，GPT-4已经利用了大部分目前我们能够获取到的高质量文本数据。因此，可供大模型训练的数据即将达到瓶颈。

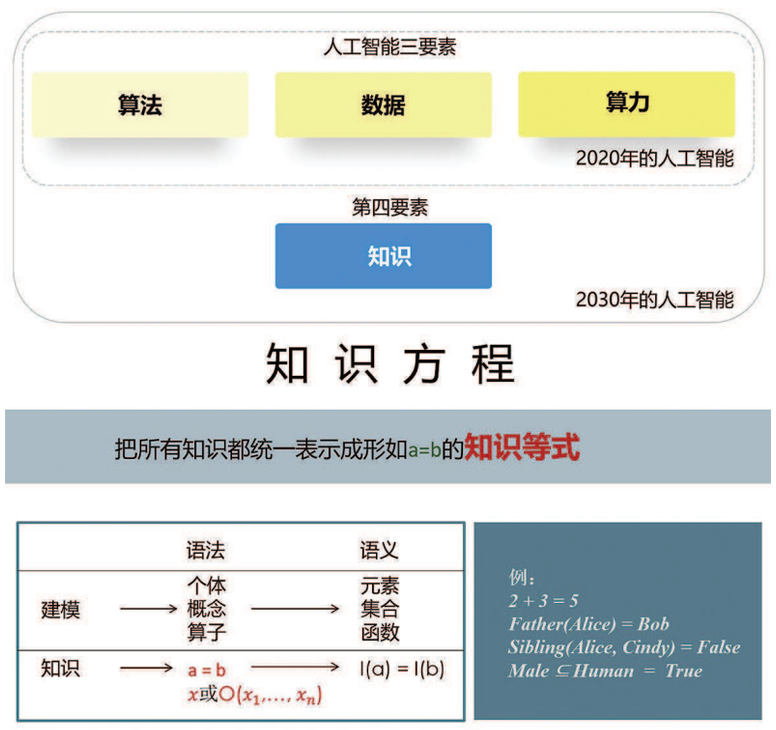
所以，要在大模型体系内解决这些问题，就需要发展新的颠覆性技术，来突破结构化信息、陈述性事实、长链条推理、深度语义理解等方面遇到的瓶颈。

另一条通往通用强人工智能的路径则有很大不同。

当前AI正在经历从感知智能向认知智能的范式转变。众所周知，人类拥有两套推理系统，即直觉思考的快系统和理性思考的慢系统。快系统是一种底层、快速、下意识、不加思索便可即刻得到结论的推理方式，就像人们在家里闭着眼睛也能找到洗手间的位置；而当我们面对陌生环境，想要找洗手间时，则需要依赖慢系统进行慎思，这种推理相对较慢、能耗较高，但更精确。



图/视觉中国



（图片来源：中国科学技术大学知识计算实验室）

目前的大模型更多涉及到的是快系统层面的推理，慢系统推理能力表现还不佳。所以，人们自然而然就会想到，能否将这两个系统结合起来。

事实上，上一波AI浪潮就是由专家系统驱动的。专家系统是一种类似于人类慢系统的推理方式，它以符号的方式把专家的知识输入机器，再通过自动推理，使得机器能够像专家一样自动回答问题。

专家系统与大模型各有所长。前者在精确度、可解释性、逻辑推理能力、语义理解能力等方面表现更佳，而后者在通用性、泛化性、不确定性知识、学习能力等方面更具优势。因此，专家系统与大模型有机结合，正好可以取长补短，这或是通往通用强人工智能的一条更好路径。

融合两大推理系统探索未来智能

中国科学家在专家系统与大模型结合的道路上，已经开始了探索。中国科学技术大学知识计算实验室提出了新的知识模型“知识方程”，在此基础上建立起新型专家系统，并将其与深度学习相融合。

简言之，知识方程分为建模和知识2个层面。在建模层面，知识方程将领域对象统一抽象成为个体、概念、算子3类语法元素，它们之间可以相互转换、相互融合。在知识层面，知识方程将所有知识统一表示成为形如“a=b”的知识等式。基于知识方程，我们提出了基于新的数据与知识双轮驱动的、结合大模型与推理引擎的智能信息系统范式。

随着ChatGPT等大语言模型的兴起，在原有以数据库为核心的信息系统之上，大模型可从暗数据库（文本、图像、视频等）中挖掘有效信息，并在一定程度上进行推理与（辅助）决策。

事实上，这是信息系统的一次范式革命。在所有数据中，暗数据占到绝大部分。传统的信息系统必须通过一些手段（包括人工、自然语言处理、计算机视觉技术等），将“暗”数据转换成数据库中的“明”数据才能使用。这种转换往往由于工程和成本等问题，只能处理暗数据中的极小部分。而大模型可以直接基于暗数据得以输出，具有很强的暗数据处理能力。

基于大模型的信息系统类似于人类用于直觉思考的快系统，能在一定程度上直接基于大数据进行推理与决策。但由于大模型目前存在的技术缺陷，在很多应用场景并不能直接满足应用要求。中科大知识计算实验室所提出的大模型增强技术，可构建领域本体与知识库，在此基础上融合大模型与知识推理引擎，研发知识数据双驱动的智能信息系统框架，将思考的快系统与慢系统结合起来。与单纯的大模型相比，该框架具有正确性、可靠性、可解释性、可调试性等优点，能够显著提高大模型在各行各业的应用价值。

除了数据库和暗数据库，该系统还可有效利用知识库的信息。因此，该框架有望引领大模型之后的又一次信息系统范式革命，也将成为智能信息系统的新形态。

从应用角度看，通用强人工智能无论在广度，还是在深度方面，都是现有的大模型技术无法比拟的。从长远来看，人工智能在很多方面的智慧能力可能会超过人类，不仅是计算、记忆和存储等基础智能，还可能包括决策、预测、创新等高阶智能。随着基于计算的大模型和知识推理引擎的不断发展，AI也将越来越接近甚至超越人类，这将在极大程度上推动生产力。

（作者为中国科学技术大学教授、知识计算实验室主任）

人工智能会『吃掉』自己吗

■杨馥溪/编译

随着人工智能（AI）生成的内容充斥互联网，它正在破坏未来模型训练的数据。当AI“吃掉”自己时，会发生什么？

得益于生成式人工智能的蓬勃发展，普通人也可随时使用计算机程序来生成文本、计算机代码、图像和音乐。与此同时，新的AI模型开发需要更多数据进行训练，这些由AI生成的内容可能很快会进入训练新模型的数据集。一些专家表示，这将在无意中引入错误，并随着每一代模型的诞生而不断积累。

越来越多证据显示，人工智能生成的文本，即使被引入训练数据集的量很少，最终也会对训练中的模型产生“毒害”。而目前，几乎还没有有效的“解毒剂”。英国爱丁堡大学信息学院计算机科学家里克·萨卡尔说：“虽然现在或几个月后，这可能还不是问题，但我相信，几年后这将成为一个必须要面对的问题。”

AI生成数据“毒害”已真实存在

AI模型以自身产生的数据“毒害”自身的状况，可能有点类似于核试验带给人类的困境。

自1945年人类引爆第一颗原子弹后，数十年的核试验使得大量放射性尘埃进入地球大气层。而当这些空气进入新制造的钢材时，就会增高这些钢材的辐射性。

对辐射特别敏感的钢材应用而言，例如盖革计数器（一种用于测量放射性辐射的探测器），就必需使用低辐射金属。因此，人们只能抢购日益减少的低辐射金属，比如在旧船残骸中寻找1945年前生产的钢铁废料。

一些业内人士认为，类似的循环将在AIGC中重演——研究人员不得不寻找没有被“污染”的训练数据。

AI模型是如何“中毒”的？研究人员将一些由AI生成的语料作为训练数据，“喂”给一个正在训练的语言模型，然后使用它所输出的结果再来训练新模型，并重复这一循环。他们发现，模型每迭代一次，错误就会叠加一次。当人们要求第10次被训练出的模型写出有关英国历史建筑的内容时，它“吐出”的却是有关豺兔的一堆胡言乱语。

英国牛津大学机器学习研究员伊利亚·舒迈洛夫及其同事称这种现象为“模型崩溃”。他们在语言模型、生成手写数字和区分概率分布等模型中，都观察到了这种现象。“即使在最简单的模型中，这种情况也已经发生。”舒迈洛夫说，“我向你保证，在更复杂的模型中，也肯定已经发生了”。

在最近的一项预印本研究中，萨卡尔及其在西班牙马德里和英国爱丁堡的同事，用一种名为扩散模型的AI图像生成器进行了类似的实验：第一个模型可以生成可识别的花朵或鸟类，但到了第三个模型，所生成的图片就变得模糊不清了。

萨卡尔说，其他测试也表明，即使是部分由AI生成的训练数据集也是“有毒”的。他解释说：“只要有一部分训练数据源自人工智能所生成的内容，就会产生问题。”但更多具体细节还有待研究。

目前研究表明，模型在其数据的“尾部”（模型训练集中出现频率较低的数据元素）所受到的影响最大。由于这些尾部包含的数据与“标准”相去甚远，模型崩溃可能导致AI输出的结果失去“人类数据”特有的多样性。

令舒迈洛夫特别担心的是，这会加剧模型对边缘群体的既有偏见，“我们需要加紧努力，来遏制这种情况的发生”。

阻止“模型崩溃”尚需求解

无可辩驳的事实是，AI生成的内容已经开始进入机器学习工程师们所习惯于获取训练数据的领域。以语言模型为例：即使是主流新闻媒体也已经开始发布人工智能生成的文章，一些百科网站的编辑也希望使用语言模型为网站生成内容。

瑞士洛桑联邦理工学院（EPFL）学者维尼亚明·韦谢洛夫斯基认为，人类正处于这样一个拐点，“许多我们用来训练模型的现有工具，很快就会被AI生成的文本‘喂饱’”。

有迹象表明，AI生成的数据也可能通过其他途径进入模型训练。韦谢洛夫斯基及其同事通过统计分析发现，已有约1/3的医学研究摘要有ChatGPT生成文本的痕迹。

EPFL小组的研究成果于上个月发布在预印本服务器arXiv.org上。不过，机器学习工程师们也提出反驳。EPFL的研究生马诺埃尔·奥尔塔·里贝罗认为，使用ChatGPT对文本数据进行注释更加便捷且效果更佳。

面对模型崩溃的威胁，机器学习工程师该怎么办？答案可能相当于盖革计数器中的战前钢铁：已知不受（或尽可能不受）AIGC影响的数据。

例如，萨卡尔提出了采用“标准化”图像数据集的想法。这些数据集将由人类进行策划，因为人类知道这些数据集的内容仅由人类创作组成，并且可供开发人员免费使用。

一些工程师可能想打开互联网档案馆，查找AI热潮之前的内容，但舒迈洛夫并不认为使用历史数据是一种解决方案。首先，可能没有足够的历史信息来满足不断增长的模型需求。另外，这些历史数据不一定能反映不断变化的世界。

“如果你想收集过去100年的新闻，并试图预测今天的新闻，这显然是行不通的，因为技术和时代都已经发生了变化。”舒迈洛夫说。

因此，我们面临的挑战可能更为直接：从合成内容中分辨出人工生成的数据，并过滤掉后者。不过，即使有了这方面的技术，这也远不是一项简单的任务。正如萨卡尔指出的那样，如果Adobe Photoshop允许用户使用人工智能生成技术编辑图像，那么这样编辑出来的图像到底是不是人工智能生成的呢？