人机大战赛前思考

> 上个月，谷歌计算机系统AlphaGo战胜欧洲围棋冠军樊麾的新闻又一次让人工智能上了头条。



上个月，谷歌计算机系统AlphaGo战胜欧洲围棋冠军樊麾的新闻又一次让人工智能上了头条。今天开始，AlphaGo将升级和人类的围棋对决，与更胜一筹的韩国著名棋手李世石对战5局。据报道，李世石对赢得比赛很有信心，认为AlphaGo只有三段棋手的水平，但他同时觉得，如果人工智能技术继续发展的话，再过一至两年，比赛的结果将很难预料。

很少有人料到，人工智能能和人类围棋高手一决高下的日子到得如此之早。现在的问题是：接下来会怎样？对关注科技多于关注围棋的人来说，计算机是否能在围棋上打败所有人已经不太重要，而重要的是:人工智能全面达到甚至超越人类智能的时候是否快到了？

两派对立观点

一种常见观点认为，既然人工智能通常是指“用计算机完成那些需要人类智能的任务”，那么随着计算机科技的进展，自然不断有任务会从“只有人能完成”的单子中被移到“计算机也能完成”的单子中。这种情况在历史上已经发生多次了，而“在围棋上达到职业棋手水平”只不过是最新的一次而已。这当然不会是最后一次，计算机也一定会越来越多地接管人类的脑力劳动，就像其它技术接管人类的体力劳动一样。但每个任务都有自己的特点，计算机在围棋上的进展不会直接对应于在其它领域中的进展。由于人能干但计算机不能干的事还数不胜数，人工智能全面达到甚至超越人类智能仍只是一种遥远的可能性，甚至可能根本就是天方夜谭。

这个观点在历史上能找到很多证据。大众常常误判人工智能的发展水平，其思路是“计算机连这么难的事都能干，别的事自然是不在话下了”，或者“计算机连这么容易的事都干不了，别的事自然是无从谈起了”。殊不知由于计算机和人脑在结构和机制上的各种重大差别，任务的难易程度在这两种系统中有非常不同的次序。 “人工智能已经达到几岁孩子的水平了？”是一个错误的问题，因为计算机在一些任务上早已超过所有人类，而在另一些任务上还远不如任何人，这怎么比？最容易误导大众的要算某些机器人或计算机系统的演示了。这些系统在某些精心设计的任务上的出色表现使得人们想象它们在其它任务上也已经达到了同样水平，实际上完全不然。“内行看门道，外行看热闹”在这里是再恰当不过了。

与此相反的观点认为这次的情况和以往不同。上面的分析的确适用于“深蓝”战胜国际象棋世界冠军的例子，但AlphaGo是基于“深度学习”技术的，其围棋知识不是设计者编进程序里的，而是系统自己从大

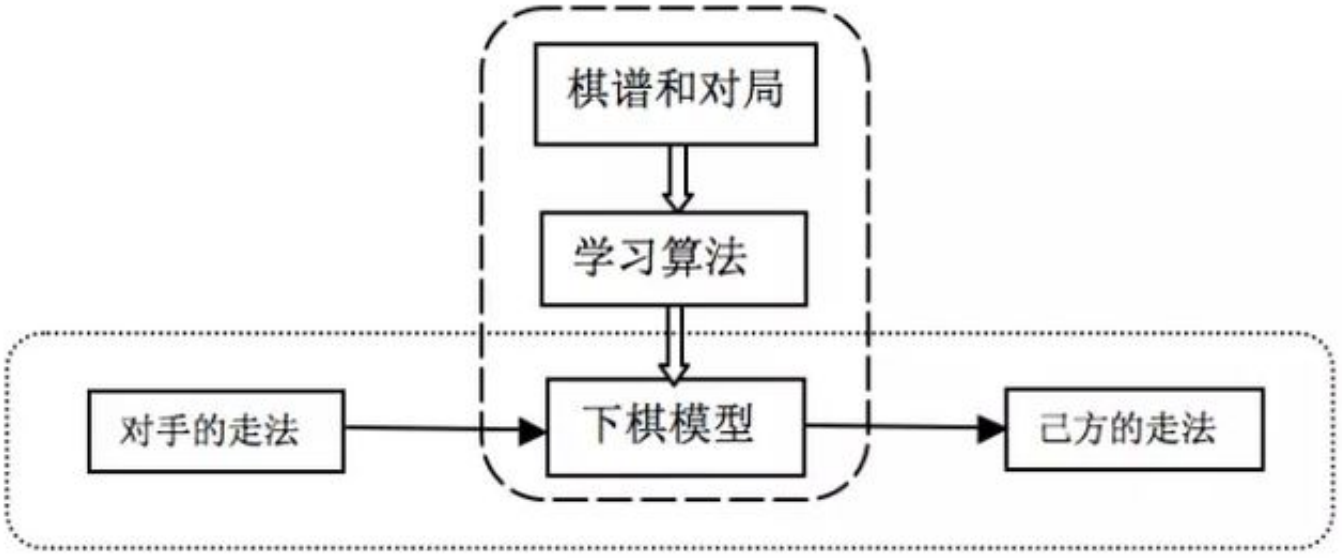
量的棋谱和对局学到的。这就是说系统的围棋水平会在没有人类介入的条件下不断提高。由于计算机的学习速度远高于人类，其水平超过所有人似乎只是个早晚的事。如果计算机能学围棋，它为什么不能通过学习在任何任务上赶上以至超过人类呢？

这种观点的代表作是库兹韦尔的《奇点临近》和博斯特罗姆的《超级智能》。据他们的看法，技术水平的进步是不断加速的。尤其是对人工智能而言，由于聪明的机器会设计更聪明的机器，这种自我强化的智能最终会远远超过人类的智能水平，甚至超出我们的理解能力。我们和它们的智能差别会像低等动物和我们的智能差别那样。这种预测让不少人不寒而栗，以至于把人工智能看作人类最大威胁的观点得到了不少科技界名人的响应。

“计算”与“学习”

我认为上述两种观点都是错的。关于人工智能的讨论常常陷入混乱，是因为其中的一些核心概念在该领域内的工作定义和其日常含义不同。在《当你谈论人工智能时，到底在谈论什么？》中，我列出了几种对“人工智能”的不同理解，而在《人工智能迷途：计算机的高技能等于高智能吗？》之中，我分析了把“智能”看成“解决问题的能力”和“学习的能力”所造成的不同后果。在这篇短文中，让我们看看“学习”在今天的人工智能领域中到底是什么意思。

在计算机科学中，“计算”这个词是有明确定义的。如果一个问题的解决过程在一个系统中表现为“计算”，那就是说对这个问题的每个实例（输入），系统的解答过程必须遵循某个预定的方法（算法），而且答案（输出）一定符合某个确定的标准。而与此相反，“学习”恰恰意味着对同一个输入，输出的质量随着经验的积累而逐渐提高，因而不是确定的。对同一个过程而言，这两个标签是不能同时适用的。在目前的“机器学习”研究中，“学习”及其效果通常是通过两个“计算”过程实现的，各有其输入和输出（详见参考文献[1]）。如下图所示，AlphaGo在“训练”阶段靠“学习算法”从训练数据（输入）中总结出一个“下棋模型”（输出）；在学成之后转入“实战”阶段，这时的输入就是对手的走法，而输出是由下棋模型（作为算法）选择的走法。尽管在实战阶段对系统的性能进行进一步的改善仍是可能的，但这种调整不会造成模型的根本性改变。



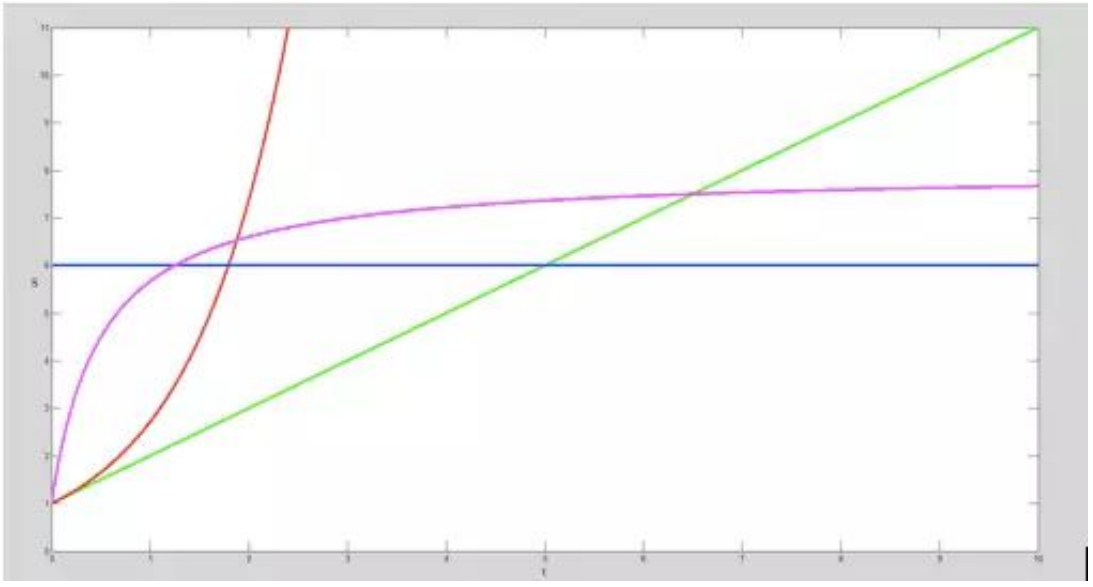
因此，这种学习预设了一个“计算”（在AlphaGo里就是理想的围棋模型）作为终点，也就是说每个局面都有一个最好的下法，而学习就是找到或接近这一下法的过程。一旦对每个局面都找到了最好的下法，学习的使命也就结束了。由于围棋的变化实在太多，学习的任务就很艰巨，但这并未改变上述对学习的理解。尽管“深度学习”的算法可以用在不同领域中，其每个具体应用仍是针对一个给定的输入输出关

系的。比如说AlphaGo还能在围棋上继续提高，但现在再教它学象棋已是不可能的，更不要说教它学兵法了。这显然和人类学习所体现的“触类旁通”或“功夫在棋外”完全不同。

现有的主流机器学习技术可以有效地学习很多问题的解决方法，或称“技能”，但不适用于问题和学习方法都无法预先确定，并且系统需要随时对新情况做出反应的情景。像我在《人工智能迷途：计算机的高技能等于高智能吗？》所介绍的，这种更具一般性的学习能力目前是“通用人工智能”（AGI）的研究课题。我设计的AGI系统“纳思”在参考文献[2]中有详细介绍，其基本想法是把系统的目标、知识、操作等表示成一个概念网络，其中所有成分都可以在新经验的影响下被修改。这样，“学习”就是系统内部的自组织活动，而不是以获得某个特定技能为目标的。这种学习机制由一组规则来实现，其中每条规则会被特定的情境所触发，而整个学习过程并非遵循一个有预定途径和终点的学习算法。纳思可以学习多种不同的技能，但往往没有一个明确的“学成”时刻，而且已经“学到的”知识仍可能被修改。在这种系统中，学习是终生的，是无时无处不在的，而且“学习”和“工作”通常是同一个过程的不同侧面，因此不属于上图所示的“机器学习”。

四类“智能系统”

为了描述的直观性，让我们假定一个系统的总技能水平可以度量，比如说是它能解决的问题的数量。这样一来，系统的学习能力就体现为这个量和时间的关系，而根据前面的讨论，我们可以区分四类不同的系统，如下图所示：



（1） 蓝线对应于一个完全没有学习能力的系统。这样的系统仍然可能有很高的技能（如“深蓝”），但其技能完全来自于初始设计，而和系统的经验无关。传统的计算系统都属于这一类，其优点是可靠、高效，但没有灵活性、适应性和创造性等。

（2） 紫线对应于一个学习能力有限的系统。这里“有限”是指系统的总技能会趋近于一个常量，即这个学习系统最终会转化成一个计算系统。现有的“机器学习”系统（包括AlphaGo）大体是属于这一类。这种系统在学习阶段有适应性，但模型形成后就基本不变了。

（3） 绿线对应于一个学习能力基本固定的系统。这种系统的“智能”（元能力）是先天设计的，且基本保持稳定，但在习得的“技能”（具体能力）的增长上没有限制，最后在输入输出关系上也不会转变成一个计算系统。纳思和其它一些AGI系统属于这一类。

（4） 红线对应于一个学习能力本身在增长的系统。这大致上就是库兹韦尔和博斯特罗姆所说

的“超人智能”了。

上图只是一个形象化的说明，而非对这些系统的严格刻画。以纳思为例，因为其中所有的学习都是以过去经验为依据的，所以在经验不足、思虑不周或时过境迁时也会犯错，这时候其总技能水平很可能不升反降了。这四类系统的实质性差别是：系统整体的输入输出关系是通过设计确定的还是通过学习得到的，以及系统的学习能力是否随时间而有根本性改变。

按我的观点（见《人工智能迷途：计算机的高技能等于高智能吗？》），“智能”应指学习能力，而非解决具体问题的能力（那是“技能”），因此上面四种系统的次序也就是智能由无到有，由低到高的次序。

在一个特定时刻，智能更高的系统未必技能也更高。而且从具体应用的角度看，并不是智能越高越好。实际上往往相反:在能满足使用要求的前提下，系统的智能越低，可靠性和效率通常越高。高智能只有在低智能不胜任的时候才是有效益的。当然，从理论研究的角度看，高智能系统更有意思，也更具挑战性。

有“超人智能”这回事吗？
按上述分类标准，人类大致属于“绿线族”。这是因为一个人在正常情况下总是可以学习新技能或修改已有技能的，但不能对自己的思维规律做根本性修改。不同的人的智力有高低及特征上的不同，但在这一点上和其它三类系统的差别还是明显的。

因为主流人工智能的目标是拥有或超过人类技能，其中大多成果属于“蓝线族”；机器学习以技能习得为己任，造出的系统基本是“紫线族”；“绿线族”目前是通用人工智能（AGI）的研究范围，应用技术尚未成熟。目前并没有人在研发“红线族”的系统，而其它三类系统的进展也不会“跨界”，因为它们对应于不同的技术，而并非同一个技术的不同发展水平。

以纳思为例，其设计中确定了目标、知识、操作等成分的表示方式，从已有成分中生成新成分的规则，在系统运行时分配计算时间和存储空间的策略等等，但系统中不需要有任何“先天技能”（尽管有为具体应用定制的可能）。这个设计包含了某种对智能的认识，而对系统来说这些认识就是“元知识”。在元知识指引和控制下，系统会从经验中总结出它自己的知识和技能，但这种学习不会导致对元知识（即系统设计）的根本修改，尽管局部的调整和补充是可以的。

人类智能当然未必是智能的最高形式，所以 “超人智能”是个有意义的概念，尽管我们至今尚未得到任何关于其确实存在的证据。但即使是作为一个纯粹理论上的可能性来说，目前相关的议论仍充斥着大量的概念错误，如“智能”与“技能”的混淆以及不同类的“学习系统”之间的混淆。在“奇点”这个提法背后的基本假设就是智能系统的智力会加速增长，以至于会造成质的差别。这个假设至少在人类历史的尺度上是没有证据的。和先秦诸子或古希腊先贤相比，今天的我们可以说知识和技能水平要高很多，但未必智力水平更高吧？就凭你会编程序或知道量子力学，就敢说自己比诸葛亮聪明？

在物种进化的尺度上说，人类智能的确是从低智能的动物中发展出来的，但这不意味着这种能力会无限制地继续发展。实际上，在任何领域中假定某个量以往的增长趋势会自然推广到未来的预测都是不可靠的。一个流行的说法是人工智能会通过“给自己编程序”来进化，但这是缺乏计算机知识所导致的误解。在设计一个计算机系统时，“程序”和“数据”的区分常常是相对的，而这二者的功能也或多或少可以互相替代。允许程序自我修改的程序设计语言早已存在几十年了（如Lisp和Prolog），但只要这种修改是基于系统过去经验的，那么这种“自我编程”和通过修改知识来改变系统行为并无本质区别(仅仅

是改变了“知识”和“元知识”的区分方式而已），而且未必是个更好的设计。这样的“自我编程系统”仍属于“绿线族”。

随着计算机的发展，人工智能系统的运行速度、存储容量、数据拥有量都会继续大幅提升，这会使得计算机在越来越多的技能上赶上以至超过人类，但这都不意味着计算机因此拥有了超人的智能。这就好比把同一个程序系统从一个普通计算机上移植到一台超级计算机上：这以后它自然在解决问题的能力上有极大提高，但这个系统并没有因此而变得“更聪明”。真能达到“红线族”的计算机系统在我看已经不是“人工智能”，而应该被称为“人工神灵”了，因为它们的工作原理不再是我们所谓的“智能”，而是像“如来掐指一算”那样完全超出我们的理解能力。和其它根据定义就超出我们理解力的对象一样，我们当然没办法证明这种计算机系统不可能存在，但忧心于如何防范它们也同样毫无意义。

我的观点是：通用人工智能将会在“元知识”层面上（即基本工作原理）和人类大致相当，既不更高也不更低。如果这种系统真能造出来，那说明我们已经基本搞清了智能是怎么回事。在那以后即使计算机在大量具体问题的解决能力上超过人类，并且可以通过学习进行自我提高，人类仍然可以根据其工作原理（包括其学习过程的原理）寻找用其利避其害的办法。至于这些工作原理到底是什么，得容我以后逐步解释，有干劲的读者也可以去读参考文献[2]和我主页上的几十篇文章。

参考文献

[1]Peter Flach, Machine Learning: The Artand Science of Algorithms That Make Sense of Data, Cambridge University Press, 2012

[2]Pei Wang, Rigid Flexibility: The Logic of Intelligence, Springer, 2006

版权声明

本文仅代表作者观点，不代表百度立场。

本文系作者授权百度百家发表，未经许可，不得转载。

- 面包财经 2