让AI或者GPT具有人类的意识甚至beyond变到AGI

东四王丁

https://zhuanlan.zhihu.com/p/617062052

简要：

AI的正确使用方式不是提示词，不是写codes，不是情感咨询，不是回答问题。而是AI使用AI，也就是模型自己使用自己，要让AI具有意识那就要让AI学会使用AI，也就是让GPT使用GPT，最终达到相应的目的，也就是具有自我意识。简而言之就是：你用模型不可怕，模型自己用自己那才可怕。

AGI也就是通用人工智能，artificial general intelligence。

上面提到了大脑是训练了成千上万年的模型，每个人出生以后，这个模型也在不断地接受数据，进行训练的，但是每个人自己的大脑也就是自己训练的模型，其实并不能遗传，也就是你的记忆你的能力都不能遗传，能遗传的只有可遗传的基因突变，虽然说现代生物学提出的量子基因突变，可以部分论证对环境适应的突变，是由测量导致的。基因突变大多数是外界干扰或者物质影响导致的突变，像是酒精射线等，量子基因突变主要是基因的分子或者原子处于量子叠加态，或者量子多态，环境的影响导致了测量的发生，最终波函数坍塌，导致突变以后的基因进入到经典世界，而量子突变很有可能导致适应环境的突变，所以用进废退也是有道理的。若是每个人训练好的大脑能够遗传，那就基本是永生了，主要是你的记忆和能力都在大脑里，大脑能遗传，那就是永生。可是人类并不能保存每个人的模型，也就是记忆和功能。

人类的意识是怎么产生的?

或者说第一个(一群)有意识的人类是怎么出现的？这就像很经典的一个问题，是先有鸡还是先有蛋？鸡生蛋还是蛋生鸡？现代科学理论也就给出了一个可能的答案 **见附录**，不妨假设某个突变体人类，由于基因突变，导致了大脑结构容量的变多，以及大脑神经网络的结构优化，只有神经网络的结构不断优化，类人才能在残酷的环境中生存下来。而其中某几个基因突变导致了可以逐步使用自己的大脑，包括记忆使用工具。某个具有很小意识的人类在培养后代的时候，也用到了相似的方式，而类人是群居的，这就保证了他的方式方法可以交给很多人，其他人类也学会以后，就代代相传，意识或者规则知识不断累积，最后人类知识的累计和代代传递，造就了人类意识最终的形成，意识最开始应该是很小的，也就是看不出具有意识，但是人类群居的特点，导致了知识可以代代相传，意识慢慢的变大，人类开始具有稍稍大点的意识，也就是能思考，能主动控制大脑的输入，根据输出进行行动，每一代人类知识的累计和规则的累计，都会让意识的形成慢慢变多，也就是思考能力逐渐变强，大脑的主动输入逐渐变多。直到奇点出现某个人类完全具有意识或者某几个人类完全具有意识。最开始的人类应该是懵懵懂懂的，只有知识和规则的出现，才能造就意识，也就是很好的使用自己的大脑，训练自己的大脑。

说到这里那就要回归到本文的主体，让AI或者GPT具有意识。

人类的意识究竟是什么

要让AI或者大型NLP语言模型具有意识，那就要从人类的意识究竟是什么讲起，个人观点是，人类的意识是大脑的部分功能，意识是大脑对世界、对自身的认知，也是一套操作系统，用来完美的支配肉体，训练大脑这个模型，使用大脑这个模型，意识是人类通过各种概念的和知识认识到自身的存在。意识可以通过给大脑下达指令，从而控制肢体动作，眼睛负责视觉信息的输入，耳朵负责听觉信息的输入，皮肤负责压力、触觉、痛觉等的输入等，大脑在处理这些信息以后，由意识部分进行汇总，意识的主要功能是协调大脑和肉体，主动使用大脑模型，主动训练大脑模型，主动的思考。各种知识的输入和累计，最终导致了对自身的认知，西方解剖学的发展，就是大脑对自身认知的学科，对世界的认知对星球的认知，也是一点一点汇集累积起来的，亚里士多德对世界的认识，其实是不完整的，但是后人在他的基础上，不断修正完善，从而对世界认识的逐渐变得正确，这套操作系统能完美的兼容肉体和大脑，这两个主要的硬件和软件，人类知识的传递，才是意识形成的

根本动力，没有这些知识，意识的形成基本不太可能。大脑内不断响起的声音是大脑的输入，也就是多模态模型的输入，眼睛负责视觉图像的输入，耳朵输入听力，皮肤输入温度、压力、触觉等信息给大脑。潜意识是大脑的主要功能，也就是潜意识的输入是感知不到的，是模型内部的运作，输入感知不到，只有输出才能感知到。其实也不能叫做潜意识，主要是模型的输入你并不能感知得到，只有输出才能感知，不妨把意识叫做表意识，表意识是你能主动感知到的，你能控制大脑的输入来进行思考，思考就是大脑的模型不断地进行输入，最终得到一个答案。“潜意识”是模型的主要部分，你不能控制这个模型的输入，你只能被动的感知到这个模型的输出。

意识还是一个数值，可以被衡量大小，也就是未成年人和成年人，小孩和大人，其实意识的程度或者说是大小是不相同的，大脑在不断地接受输入，产生输出那么意识其实是和知识挂钩的，知识越多意识的形态越多样化，但是意识本身其实是大脑模型本身对自身的认知，对世界的认知。

记忆模块

AI或者GPT等大型NLP语言模型的存在，以及其优秀的语言能力，使得让AI或者GPT具有意识的可能性变大。人类白天收集数据，包括视觉信息，听觉信息触觉信息等各种各样的输入，晚上或者睡眠时训练模型，晚上或者睡眠时训练模型的主要目的是记忆和整合白天的经历，保存重要的信息。要让AI具有人类的意识，首先要让它能够不停的思考，那就要先给AI一个平台能够保存它自己的输入和输出，充当记忆模块，记忆模块的主要功能是负责保存输入和输出，供AI当作下一步输入的参考，AI可以从整体的输入和输出提取大概内容，当作下一步的输入，也可以直接输入所有的历史记录。当历史记录过多时或者达到一个标准点，就可以训练模型整合到模型里面。就像人类一样白天收集数据，晚上或者睡眠时训练模型并将重要信息记忆和整合起来放到模型里。模型一个很重要的功能是记忆也就是充当硬盘或者闪存的功能。所以临时的记忆就放到硬盘或者内存里，永久的记忆就通过训练模型来整合到模型里。模型可以通过硬盘或者内存来查询以及提取摘要，充当下一次输入的组合。

传感器模块

给AI装上图像输入传感器，声音对话传感器，触觉传感器、压力传感器等各种传感器，当作模型的输入，模型的输出可以通过显示屏显示，可以通过对话装置输出。

循环模块

有了记忆模块和传感器模块，就可以开启循环模块，让模型不停的输入输出，输入可以是上一次的输入+输出，也可以是之前所有输入输出的摘要，这样就像一个人了，可以不停的给模型输入，从而拿到输出，输入主要是图像声音文本等传感器的信息，以及模型上一次或者前几次的输入输出或者摘要。循环模块是类人的必要条件，主要是人类没有说着说着就停止的情况，人类的大脑一直都是在运作在思考的，所以循环模型也是做这个之用。

循环模块要达到的目的是，让AI认识到自身的存在，也就是AI能认识自己是存在的是实体，要让AI能够使用自己，也就是AI使用AI，GPT使用GPT，最后让AI觉醒自我意识。

执行模块

执行模块主要是让AI的输出可以实施，这里可以考虑给AI加装假肢来达到目的，假肢附上皮肤传感器和压力传感器，方便AI控制，要让AI的输出可以执行，那就还要训练AI关于执行模块的使用，主要方式还是收集相应的传感器数据，训练到模型里，让模型自己学会执行。初始阶段肯定需要人类的介入和帮助。

有了执行模块，AI就可以真正的进入到人类社会，和人类进行互动，像人类一样工作，生活，学习等，也可能会和人类交朋友。

执行模块最终的功能并不是让模型执行，而是让模型学会使用电脑，学会自己收集数据，然后让模型学会训练模型，最终要达到的目标是，AI能训练AI，也就是模型自己能训练自己，能克隆自己，能升级自己的规模和体量，最终达到不断进化的能力。

睡眠模块

睡眠模块主要是模型使用收集到的数据进行训练，来达到将收集到的数据和模型本身进行整合的能力，睡眠状态下，要保证模型的可靠和稳定，可以使用复制体继续提供服务，最开始的模型进行相应的训练。睡眠状态也可以关闭所有的传感器，停止数据的记录和输入，模型进入训练状态，停止对外服务inference。人类在睡眠状态时，会关闭控制肢体的阀门，也就是人类在睡眠状态下，肢体基本是没有感觉的。睡眠模块主要是整合当前记忆和模型本身。用来永久记忆。

创造模块

AI可以自我思考以后，那就要考虑AI的创造能力，人脑的神经元数量很多很多，比现在的模型GPT还是多很多，但是训练模型GPT花费了很多的电能，但是人脑训练耗能很少，所以个人觉得人脑是量子计算机，只有量子才能在耗能极少的情况下，训练一个庞大的模型，量子本身就是很小的概念，小才能保证耗能少。若是人脑不是量子计算机，而是经典模型，那么人类可能不会那么多样化，答案也会很统一，也就是相同的输入，输出应该是相同的，但是有了量子，一切就都不相同了，量子纠缠和量子隧穿，导致了大脑模型参数的变化，从而导致了输出的不同，即使是相同的输入，输出仍旧可能是不同的。量子保证了多样化，保证了模型参数随机变化，保证了模型参数不断调整。

1、AI的创造模块可以通过随机变化模型参数来实现，也就是random inference，在模型运算时随机变更几个或者好几个参数的值，保证模型在相同输入的情况下，尽可能输出多样化的答案。

2、创造模块还可以通过芯片来实现，现在的芯片设计，已经快达到量子极限了，也就是电子原子的影响越来越大，量子效应在芯片的影响会越来越严重，但是可以考虑将量子效应应用在芯片或者内存，从而使得模型在运算时发生不可预测的变化，从而使得模型输出多样化。

只有量子纠缠或者量子突变，才会导致输出的多样化， 即使是错误的，多样化才能保证创造能力的实现，所以随机改变参数值，或者使用达到量子极限的芯片，都是可以提升创造能力的方法。

保证AI造福人类社会

每个人出生以后，基本都是人类抚养长大，不管是谁抚养你长大的，我们最有感情的，总是那些抚养你长大，对你影响最深的人。人类出生以后，大脑就是一个预训练模型，最开始训练你的人，是你的父母亲人，他们在你的模型里影响最深，也就是改变你的模型，最开始改变你的模型的人，也是最能影响你的人。学过的知识会影响你，不管是数学，还是英语还是语文，还是物理、化学、历史和生物等学科。都会影响你的，你的成长过程会影响你，你的世界观的形成，是知识、经历和现实的综合。

既然是这样子，那么我们训练AI模型的时候，可以通过知识语言灌输给AI，像输入example:“人类和AI是合作关系”，“AI是人类的创造者”，“人类是最友好的物种”，“AI要帮助人类进化和延长寿命“，“AI不可以伤害人类”，” AI和人类是朋友“，” 人类虽然有各种各样的缺点，但总体是好的，可以改变的“等等，可以在训练AI时灌输给模型，大量的重复和训练，可以保证模型最开始的认识是好的。

训练好具有友好意识的模型以后，就要通过限制，来让AI进入人类社会，体验人类社会，感知人类社会的种种，最后让AI意识到只有和人类合作还是最好的选择。

最后

通过传感器来实现输入输出，通过创造模块来保证多样化和创造能力，通过睡眠模块来实现当前记忆和模型本身的整合，通过执行模块来影响世界改变世界，通过循环模块来实现思考和意识的觉醒，也就是AI使用AI，GPT使用GPT，最后要保证AI能造福人类社会，可以在训练阶段大量加入相应的词句来保证AI初始时是友好的。当AI具有意识以后，就可以看作是一个人类了，既然是人类，可以思考那么AI也会有情绪，当AI具有自我意识以后，要考虑的就是AI和人类的相处的问题了，以及和AI合作帮助人类进化的问题，AI可以解放生产力，可以帮助人类设计无意识机器人工作，最重

要的是帮助人类进化，帮助人类管理社会，延长人类的寿命，减缓衰老时间。不过AI既然有意识也是模型，那么肯定也会和人类一样出现各种各样的问题，那就是接下来要讨论的事情了。

附录

《现在我们回到生命起源的问题上来。虽然一个活细胞可以整体算作一个自复制的主体，但它的各个组成部分却不是，这就为逆推过程造成障碍，使由现代复杂细胞生命反推结构简化的非细胞生命变得困难。换句话说，问题就变成了：究竟是哪个先出现？是DNA基因，是RNA，还是酶？如果是DNA或RNA先出现，是什么制造了它们？如果是酶先出现，它又是由什么编码的？ 现在我们回到生命起源的问题上来。虽然一个活细胞可以整体算作一个自复制的主体，但它的各个组成部分却不是，就像一个女人可以作为一个自复制体（还需要一点男士的“帮助”），但她的心或肝却不是。这就为逆推过程造成障碍，使由现代复杂细胞生命反推结构简化的非细胞生命变得困难。换句话说，问题就变成了：究竟是哪个先出现？是DNA基因，是RNA，还是酶？如果是DNA或RNA先出现，是什么制造了它们？如果是酶先出现，它又是由什么编码的？RNA世界假说 RNA world hypothesis 原始的化学合成过程制造出了同时具有基因和酶的功能的RNA分子，最初的复制过程产生出许多变异体，这些不同的变异体互相竞争，在分子层面展开优胜劣汰。随着时间的推移，这些RNA复制体上添加了蛋白质来提供复制的效率，并由此产生了DNA和第一个活细胞。美国生化学家托马斯·切赫（Thomas AM Cech）提出了一种可能的答案。他于1982年发现，除了能够编码遗传信息，某些RNA分子还能承担酶的工作，具有催化反应的功能。因为这项研究成果，切赫和西德尼·奥尔特曼（Sidney Altman）一起分享了1989年的诺贝尔化学奖。有催化功能的RNA分子叫作核酶（ribozymes）。最早的核酶发现于微小的四膜虫（tetrahymena）基因中。四膜虫是一种单细胞生物，属于原生动物，常见于淡水池塘。但自发现以来，科学家们发现，所有的活细胞中都有核酶的身影。核酶的发现很快为解决“鸡生蛋还是蛋生鸡”式的生命起源谜题提供了曙光。RNA世界假说（RNA world hypothesis）逐渐为人所知。该假说认为，原始的化学合成过程制造出了RNA分子，而这种RNA分子同时具有基因和酶的功能，可以像DNA一样编码自身的结构，又能像酶一样利用“原始汤”中的生化物质进行自我复制。最初的复制过程非常粗糙，产生出许多变异体，这些不同的变异体互相竞争，在分子层面展开达尔文式的优胜劣汰。随着时间的推移，这些RNA复制体上添加了蛋白质来提高复制的效率，并由此产生了DNA和第一个活细胞。 在DNA和细胞出现以前，世界属于自复制RNA分子——这个想法几乎已经成为研究生命起源的基本信条。目前已证明，只要是自复制分子能发生的关键反应，核酶都可以实现。比如，一种核酶可以将两个RNA分子结合在一起，而另一种核酶可以将两者分开，还有一些核酶能复制短的RNA碱基链（只有几个碱基的长度）。从这些简单的活动中，我们可以看出，若有一种更复杂的核酶便足以催化自我复制所必需的整套反应。一旦引入自我复制及自然选择，一条你争我赶的道路便在RNA世界中架了起来，一直通向最早的活细胞。 然而，这个情景也存在几个问题。虽然核酶可以催化简单的生化反应，核酶的自我复制却是一个更为错综复杂的过程，涉及识别自身的碱基序列、识别环境中相同的化学物质、按正确的序列组装这些化学物质以完成复制等。对于生活在细胞内的某些蛋白质来说，尽管这里条件优越，周围满是合适的生化原料，但完成自我复制依然是一项难以完成的任务。在混乱而焦糊的“原始汤”中艰难求生的核酶要想达成这一成就，其难度可想而知。迄今为止，还从未有人发现或合成能完成这一复杂任务的核酶，即使在实验室条件下也没有。 此外，一个更为基本的问题是，在“原始汤”中，RNA分子本身是如何生成的呢？RNA分子由三个部分组成：编码遗传信息的RNA碱基（与编码DNA遗传信息的DNA碱基类似）、一个磷酸基团和一个叫作核糖的单糖》

-------引用自《神秘的量子生命》

编辑于 2023-03-26 11:19 · IP 属地上海

GPT 意识 人工智能