Does AI actually understand us?

人工智能是否可以真正理解人类?

2023-04-03

People are funny. We're constantly trying to understand and interpret the world around us. I live in a house with two black cats. And let me tell you, every time I see a black, bunched up sweater out of the corner of my eye, I think it's a cat.

人是非常有趣的。我们一直在试图理解和解释我们周围的世界。我家里有两只黑猫。我可以告诉你，每当我的余光瞥见一件打结的黑色毛衣，我都会认为那是一只猫。

It's not just the things we see. Sometimes we attribute more intelligence than might actually be there. Maybe you've seen the dogs on TikTok. They have these little buttons that say things like walk or treat. They can push them to communicate some things with their owners. And their owners think they use them to communicate some pretty impressive things. But do the dogs know what they're saying?

不只是我们看到的东西。我们有时以为一些东西有超常的智慧，但实际上未必有。比如，你也许在TikTok上看到过狗狗的视频。上面有一些小按钮，写着“遛遛”或是“要吃的”。这些狗狗能用这些按钮和它们的主人交流。它们的主人也以为用这些按钮就能让狗狗做一些令人惊奇的事情。但狗狗知道它们在说什么吗?

Or perhaps you've heard the story of Clever Hans the horse. And he could do math. And not just like, simple math problems, really complicated ones, like, if the eighth day of the month falls on a Tuesday, what's the date of the following Friday? It's like, pretty impressive for a horse. Unfortunately, Hans wasn't doing math. But what he was doing was equally impressive. Hans had learned to watch the people in the room to tell when he should tap his hoof. So he communicated his answers by tapping his hoof. It turns out that if you know the answer to if the eighth day of the month falls on a Tuesday, what's the date of the following Friday, you will subconsciously change your posture once the horse has given the correct 18 taps. So Hans couldn't do math. But he had learned to watch the people in the room who could do math, which, I mean, still pretty impressive for a horse. But this is an old picture. And we would not fall for Clever Hans today. Or would we?

或许你听过《聪明的汉斯》的故事。这匹马居然能做数学题。不仅仅是简单的数学计算，而是非常复杂的问题，比如，如果一个月的第八天是星期二，那么下一个星期五的日期是什么？对于一匹马来说，这真是令人惊叹。不幸的是，汉斯并不是在做数学题。但它学会的事情也很了不起。汉斯学会了观察房间里的人，来判断它该在什么时候敲蹄子。它通过敲蹄子来“说出”它的答案。真实的情况是，如果你知道答案，就是“如果每个月的第八天是星期二，那么下一个星期五是什么日子”的答案，你会在汉斯正确地敲打18下的时候下意识地改变你的姿势。所以汉斯不会做数学。但它学会了观察房间里会做数学的人，这对一匹马来说还是很了不起的。但这只是一个古老的故事了。今天的我们已经不会再被聪明的汉斯骗到了。会吗？

Well, I work in AI. And let me tell you, things are wild. There have been multiple examples of people being completely convinced that AI understands them. In 2022, a Google engineer thought that Google's AI was sentient. And you may have had a really human-like conversation with something like ChatGPT. But models we're training today are so much better than the models we had even five years ago. It really is remarkable.

我在人工智能领域工作。我可以告诉你，事情很疯狂。有许多例子表明，人们完全相信人工智能能理解他们。2022年，一位谷歌工程师认为谷歌的人工智能有自我意识。你也可能试过与ChatGPT进行类似人类的对话。我们今天训练的模型比我们仅五年前的模型都要好得多。这真的很了不起。

So at this super crazy moment in time, let's ask the super crazy question: Does AI understand us? Or are we having our own Clever Hans moment?

所以在这个疯狂的时刻，让我们问一个疯狂的问题：人工智能是真的理解我们？还是我们又遇到了一匹聪明的汉斯？

Some philosophers think that computers will never understand language. To illustrate this, they developed something they call the Chinese room argument. In the Chinese room, there is a person, hypothetical person, who does not understand Chinese. But he has along with him a set of instructions that tell him how to respond in Chinese to any Chinese sentence. Here's how the Chinese room works. A piece of paper comes in through a slot in the door, has something written in Chinese on it. The person uses their instructions to figure out how to respond. They write the response down on a piece of paper and then send it back out through the door. To somebody who speaks Chinese, standing outside this room, it might seem like the person inside the room speaks Chinese. But we know they do not. Because no knowledge of Chinese is required to follow the instructions. Performance on this task does not show that you know Chinese.

一些哲学家认为，计算机永远不会理解语言。为了说明这一点，他们设计了一段被称为“中文房间”的论证。在中文房间里，有一个人，虚构的人，他/她不懂中文。但他/她眼前有一套指令，告诉他/她如何用中文回应任何中文语句。中文房间中会进行以下流程。门上的缝隙里递来了一张纸条，上面写着一些中文。这个人要用面前的指示想出回应的内容。他/她把回答写在一张纸上，再通过门上的缝隙传出去。对于说中文的人来说，站在这个房间外面，可能会觉得房间里的人会说中文。但我们知道他/她不会。因为遵循指示不需要学会中文。这项任务的表现并不能说明你懂中文。

So what does that tell us about AI? Well, when you and I stand outside of the room, when we speak to one of these AIs like ChatGPT, we are the person standing outside the room. We're feeding in English sentences. We're getting English sentences back. It really looks like the models understand us. It really looks like they know English. But under the hood, these models are just following a set of instructions, albeit complex. How do we know if AI understands us?

这个论证和AI有什么关系？当你我站在房间外面，与像ChatGPT这样的AI说话时，我们就是站在房间外面的人。我们输入的是英语句子，得到的是英语句子的反馈。看起来这些模型真的能理解我们。看起来它们真的像懂英语。但在系统的底层，这些模型只是遵循一套指令，尽管是很复杂的指令。我们如何知道AI能否理解我们?

To answer that question, let's go back to the Chinese room again. Let's say we have two Chinese rooms. In one Chinese room is somebody who actually speaks Chinese. And in the other room is our impostor. When the person who actually speaks Chinese gets a piece of paper that says something in Chinese in it, they can read it, no problem. But when our imposter gets it again, he has to use his set of instructions to figure out how to respond. From the outside, it might be impossible to distinguish these two rooms. But we know inside something really different is happening. To illustrate that, let's say inside the minds of our two people, inside of our two rooms, is a little scratch pad. And everything they have to remember in order to do this task has to be written on that little scratch pad. If we could see what was written on that scratch pad, we would be able to tell how different their approach to the task is. So though the input and the output of these two rooms might be exactly the same, the process of getting from input to output, completely different.

为了回答这个问题，让我们再回到“中文房间”的例子。假设我们有两个中文房间。其中一个房间里是真正会说中文的人，而另一个房间里的是个冒牌货。当真正讲中文的人拿到一张写有中文的纸时，他/她当然可以读懂。但是，当这位冒牌货拿到纸条时，他/她必须使用那套指令来作出回应。从外面看，可能无法区分这两个房间。但我们知道房间里面发生着一些根本不同的事情。为了说明这一点，假设在两个房间里的两个人，两个人的头脑中各有一个小草稿本。完成这项任务需要的所有记忆都必须记在小草稿本上。如果我们能看到写在草稿本上的东西，我们就能知道他们完成任务的方法有什么不同。因此，尽管这两个房间的输入和输出可能完全相同，但从输入转化为输出的过程完全不同。

So again, what does that tell us about AI? Again, if AI, even if it generates completely plausible dialogue, answers questions just like we would expect, it may still be an imposter of sorts. If we want to know if AI understands language like we do, we need to know what it's doing. We need to get inside to see what it's doing. Is it an imposter or not? We need to see its scratch pad. And we need to be able to compare it to the scratch pad of somebody who actually understands language. But like scratch pads in brains, that's not something we can actually see, right?

这跟人工智能又有什么关系呢？即使人工智能产生了完全合理的对话，像我们期望的那样回答问题，它也仍然可能是某种程度上的冒牌货。如果我们想知道人工智能能否像我们一样理解语言，我们需要知道它在做什么。我们需要深入内部，看看它在做什么。它到底是不是一个冒牌货？我们需要看到它的草稿本，并将其与真正理解语言的人类的草稿本进行比较。但是，大脑中的“草稿本”不是我们能随便看到的东西，对吧？

Well, it turns out that we can kind of see scratch pads in brains. Using something like fMRI or EEG, we can take what are like little snapshots of the brain while it's reading. So have people read words or stories and then take pictures of their brain. And those brain images are like fuzzy, out-of-focus pictures of the scratch pad of the brain. They tell us a little bit about how the brain is processing and representing information while you read.

实际上，我们可以在一定程度上“看到”大脑中的草稿。用fMRI或EEG这样的技术，我们可以在人阅读时拍下大脑的快照。在人们在读单词或读故事的时候，拍摄他们大脑的状态。这些脑成像的图片就像是模糊、失焦的草稿本照片，它们能告诉我们一些信息，阅读时的大脑是如何处理、表现信息的。

So here are three brain images taken while a person read the word apartment, house and celery. You can see just with your naked eye that the brain image for apartment and house are more similar to each other than they are to the brain image for celery. And you know, of course that apartments and houses are more similar than they are to celery, just the words. So said another way, the brain uses its scratchpad when reading the words apartment and house in a way that's more similar than when you read the word celery. The scratch pad tells us a little bit about how the brain represents the language. It's not a perfect picture of what the brain's doing. But it's good enough.

这里有三张大脑图像，对应一个人读到三个词时的情况：“公寓”、“房子”和“芹菜”。你一眼就能看出，“公寓”和“房子”的脑图像比“芹菜”的脑图像更为相似。当然，公寓和房子就词意来说本来就更相似，相比于芹菜来说。换一种说法，大脑在读到“公寓”和“房子”两个词时在草稿本上记录下的内容比读到“芹菜”时的草稿本更相似。这些草稿本向我们透露了一些大脑表示语言的方法。这种方法并不能完美展现大脑中发生的情况。但是足够用了。

OK, so we have scratch pads for the brain. Now we need a scratch pad for AI. So inside a lot of AIs is a neural network. And inside of a neural network is a bunch of these little neurons. So here the neurons are like these little gray circles. And we would like to know what is the scratch pad of a neural network? Well, when we feed in a word into a neural network, each of the little neurons computes a number. Those little numbers I'm representing here with colors. So every neuron computes this little number. And those numbers tell us something about how the neural network is processing language. Taken together, all of those little circles paint us a picture of how the neural network is representing language. And they give us the scratch pad of the neural network.

现在我们有了大脑的草稿纸。我们需要拿到AI的草稿纸。许多AI的内部是一个神经网络。而神经网络是由一个个小的神经元组成的。神经元就像这些灰色的小圆圈。我们想要知道神经网络的草稿纸长什么样？当我们给神经网络输入一个词时，每个小神经元都会计算一个数字。我用颜色来表示这些数字。每个神经元会计算一个数字。这些数字给我们展现了神经网络是如何处理语言的。总结起来，所有这些小圆圈为我们描述了神经网络表示语言的方法，展现了神经网络的草稿本。

OK, great. Now we have two scratch pads, one from the brain and one from AI. And we want to know: Is AI doing something like what the brain is doing? How can we test that?

太好了。现在我们有了两张草稿纸，一张来自大脑，一张来自人工智能。我们想知道的是：AI做的事情是否和大脑类似？我们要怎么判断呢？

Here's what researchers have come up with. We're going to train a new model. That new model is going to look at neural network scratch pad for a particular word and try to predict the brain scratch pad for the same word. We can do it, by the way, around two. So let's train a new model. It's going to look at the neural network scratch pad for a particular word and try to predict the brain scratchpad. If the brain and AI are doing nothing alike, have nothing in common, we won't be able to do this prediction task. It won't be possible to predict one from the other.

这是研究人员想出的办法。我们要训练一个新的模型。新的模型将检查神经网络对某个单词的“草稿”，并试图预测同一个单词的大脑草稿。顺便一提，这个过程也可以反过来。我们来训练一个新的模型。它会检查神经网络对特定单词的草稿，并预测大脑的草稿。如果大脑和AI所做的事情没有任何相似之处，没有任何共同之处，这项预测任务将无法完成。两者中的任何一个都无法预测另一个。

So we've reached a fork in the road. And you can probably tell I'm about to tell you one of two things. I'm going to tell you AI is amazing. Or I'm going to tell you AI is an imposter. Researchers like me love to remind you that AI is nothing like the brain. And that is true. But could it also be the AI and the brain share something in common?

现在我们到了一个岔路口。答案只会是以下两者之一：要么AI是非常惊人的；要么AI只是一个冒牌货。像我这样的研究人员特别喜欢说，人工智能与大脑完全不同。这是事实。但AI和大脑有没有相似点呢?

So we've done this scratch pad prediction task. And it turns out, 75 percent of the time the predicted neural network scratchpad for a particular word is more similar to the true neural network scratchpad for that word, than it is to the neural network scratch pad for some other randomly chosen word. 75 percent is much better than chance. What about for more complicated things, not just words, but sentences, even stories? Again, this scratch pad prediction task works. We're able to predict the neural network scratch pad from the brain and vice versa. Amazing. So does that mean that neural networks and AI understand language just like we do? Well, truthfully, no. Though these scratch pad prediction tasks show above-chance accuracy, the underlying correlations are still pretty weak. And though neural networks are inspired by the brain, they don't have the same kind of structure and complexity that we see in the brain. Neural networks also don't exist in the world. A neural network has never opened a door or seen a sunset, heard a baby cry. Can a neural network that doesn't actually exist in the world, hasn't really experienced the world, really understand language about the world?

我们进行了这项预测草稿的任务。结果发现，有75%的概率针对某一特定词语的神经网络草稿的预测结果会更类似针对这一词语的真实大脑神经网络草稿，而不是更接近于针对其他随机词语的大脑神经网络草稿。75%要远高于随机水平。那么对于更复杂的事物，不只是单词，还有句子，甚至故事呢？这个草稿预测任务得到了同样的结果。我们可以从大脑图像预测神经网络，反过来也可以。太有意思了。那么，这是否意味着神经网络和人工智能可以像我们人类一样理解语言呢？说实话，并不是。尽管这些草稿预测任务表现出高于随机的准确率，两者底层的相关性仍然非常弱。尽管神经网络的灵感来自于大脑，它们并不具备大脑呈现的结构和复杂性。神经网络也不存在于真实世界中。从来没有一个神经网络打开过门，看到过日落，听到过婴儿的哭声。一个并不真实存在于世界上、没有真正体验过世界的神经网络，能真正理解描述世界的语言吗?

Still, these scratch pad prediction experiments have held up, multiple brain imaging experiments, multiple neural networks. We've also found that as the neural networks get more accurate, they also start to use their scratch pad in a way that becomes more brain-like. And it's not just language. We've seen similar results in navigation and vision.

尽管如此，这些草稿预测实验仍然站得住脚——多个大脑成像结果，多个神经网络模型。我们还发现，随着神经网络变得更加准确，它们也以一种更像大脑的方式使用着草稿纸。这不仅仅是语言方面。我们在导航任务和视觉任务上也看到了相似的结果。

So AI is not doing exactly what the brain is doing. But it's not completely random either. So from where I sit, if we want to know if AI really understands language like we do, we need to get inside of the Chinese room. We need to know what the AI is doing, and we need to be able to compare that to what people are doing when they understand language.

人工智能所做的并不完全和大脑相同。但也不是完全随机。从我的角度看来，如果我们想真正知道AI能否像我们这样理解语言，我们需要进入那个“中文房间”，需要知道AI到底在做什么，需要能将AI的行为与人类理解语言的行为比较。

AI is moving so fast. Today, I'm asking you, does AI understand language that might seem like a silly question in ten years. Or ten months.

人工智能发展得太快了。今天我还在问大家，人工智能能否理解语言，可能十年以后，这个问题就会看起来很“傻”。也可能十个月。

But one thing will remain true. We are meaning-making humans. And we are going to continue to look for meaning and interpret the world around us. And we will need to remember that if we only look at the input and output of AI, it's very easy to be fooled. We need to get inside of the metaphorical room of AI in order to see what's happening. It's what's inside the counts.

但有一件事不会变化。我们是创造意义的人类。我们将继续寻找意义，解释我们周围的世界。我们需要记住，如果我们只看AI的输入和输出，我们就容易被骗到。我们需要真正深入人工智能里的那个“房间”，看到真正在发生的事情。房间里发生了什么才是最重要的。

Thank you.

谢谢。