



实现道德奇点的三个关键步骤

 演讲者 / Atoosa Kasirzadeh

大家好！今天非常荣幸来到这里。

我们设想未来的时候有两个选择，要么绕过实现的方法，直接想象；要么先找到走向未来的道路，在这个过程中让未来的图景逐渐显现。我们想要设想 2100 年的人机关系，就要知道现在的处境，这样才能知道我们要如何面对伦理奇点。

我对计算机科学、人工智能的发展历史非常着迷，我们今天没有时间详述所有的光辉历史，但是我想强调三个关键时刻，这使得我们今天的发展成为可能。

首先是世界第一台可编程的电子通用数字计算机，它是 1945 年完成的，名字叫 ENIAC。第二个是图灵划时代的论文，发表在英语世界最富盛名的哲学期刊《Mind》上。他在开篇就提出了

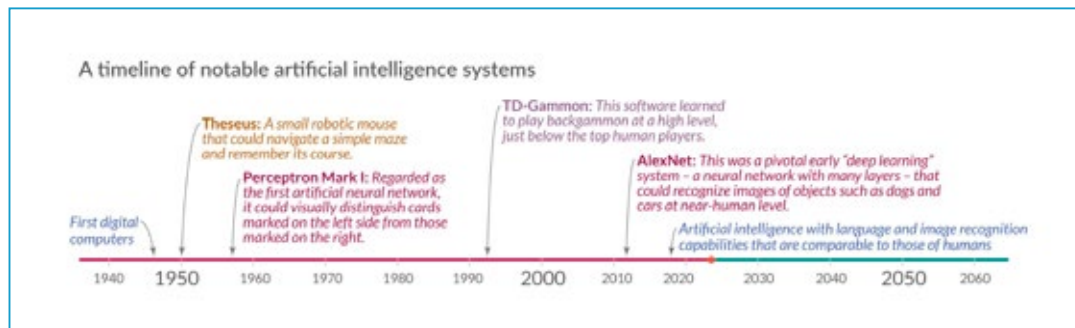
一个问题：“机器能思考吗？”这篇论文写得简洁优雅，层层递进。当你读完这篇论文会忍不住去思考开头的这个问题，这就是图灵在这篇文章中的魅力。

第三个是个研讨会，1956 年在达特茅斯学院举行。在这个研讨会当中，研究者提出了一个项目，目的是要打造一个像人类一样思考的系统。这些研究人员认为在 1956 年夏天他们可以解决这个问题。事实上，我们现在仍然在研究这个问题。1956 年的达特茅斯会议被认为是人工智能学科成立的标志性事件。

我还想额外介绍一下控制论（Cybernetics）。控制论研究通信和控制领域，不同的工程师、认知科学家、哲学家一直在思考，人类要如何控制环境，我们可以从环境中得到什么样的反馈循环。包括在二战中如何控制武器。许多工程师开始思考如何通过建立不同的沟通反馈回路来驯服自然和环境。他们开始对技术的社会影响问题感兴趣。著名的控制论创始人诺伯特·维纳 1954 年写了一本书《人有人的用处》（The Human Use of Human Beings）。如果你没看过，强烈推荐你看一下。在这本书中也回答了技术发展对社会发展的影响。

AI 的演进对我们生活的影响

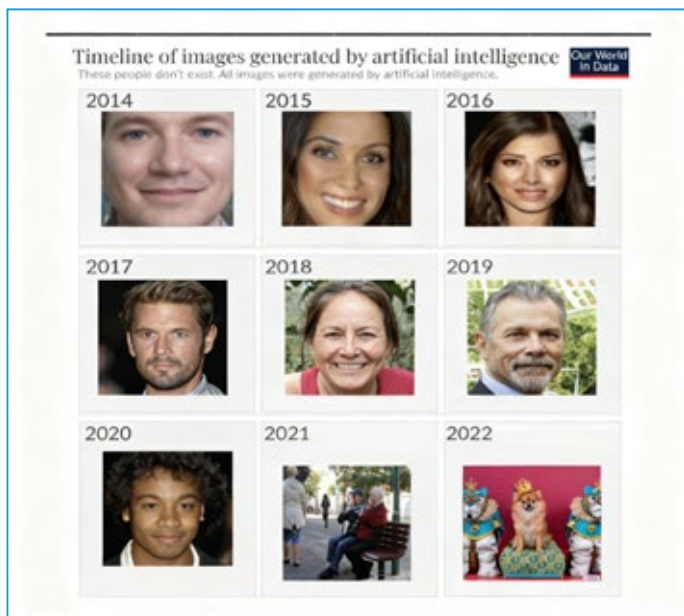
机器是否能思考是艾伦·图灵在 1950 年 10 月提出的问题。从那时起，我们在回答这个问题上取得了很大的进展。



我们看到这是一个很清楚的 AI 演进的时间线，在 1945 年的时候有第一个数字计算机，然后有了小型的机器老鼠，然后是第一个人工神经网络，然后是 AlexNet，还有很多不同的语言和图像系统。

当我们谈论 AI 的时候，要考虑到所有的这些技术。AI 不是指一个特定的东西，它指的是跨越了几十年时间这一切不同的东西。

AI 系统用于解决一些特定的任务，其中一个任务就是图像的识别。请看上面这个图片，都是由 AI 生成的图像。2014 年图像的清晰度还不是特别好，后来越来越好，可以对比一下 2014 年



和 2024 年图片质量，有大幅度的飞跃。

我们跨越了一个时间段，从 2017 年开始看。我们能够看到更多的进展，很多的模型都是关于语言生成的，还有一些模型是关于视觉、视频，音频、还有一些具体的包括生物学的专用的模型和其他的一些大模型。AI 系统取得了巨大进展，特别在语言方面。

我现在在卡内基梅隆大学工作，那里曾经有很多人工智能领域的先驱学者，比如赫伯

特·西蒙，还有汤姆·米切尔，他写了关于机器学习理论的重要著作。几周前我见到他，问他是否想过像 ChatGPT 这样的系统会出现——他从上世纪七八十年代就开始致力于机器学习系统的研究了——他说没想过。因为他不认为创造力是一项可以自动化的任务。

现在我们可以看到一些数据，比如每年发布多少 AI 的产品和系统。自从 2021 年以来这个数量激增，现在还不知 2025 年数字是否还会保持高位。我们已经在之前的三年有了巨大的突破和进展，我们都知道有很多生成的工具可以生成视频。还有用于科学发现的工具，比如 AlphaFold，帮助戴米斯·哈萨比斯等人获得了诺贝尔化学奖。许多生物学家对这个 AI 系统如何帮助他们搞清楚蛋白质折叠的结构非常感兴趣，然后他们用这些知识来进行医学药物方面的开发。

当然，我们还看到了一些不好的案例——AI 系统以某种令人悲伤的方式影响着社会和人文。可能大家都看过科幻电影《Her》，我在 2013 年的时候看过，几周之前和我的父母又看了一次。我的爸妈说，这不就是 ChatGPT 吗？这很有趣。之前觉得是科幻的内容，现在已经成为现实了。

我们还不清楚未来是怎么样的，但是我们现在正成为社会实验的一部分。

我想说说几个月之前的一个例子——ChatGPT 本来应让我们变得更有创造力，做更好的事情，但是也会让人和现实脱节，出现一些幻觉。

一位女士在社交网站上和网友分享，她的伴侣一直与 ChatGPT 聊天，自认为建立了世界上第一个真正递归的人工智能模型，可以为他提供宇宙的答案。他坚定认为自己是超人了，而且还在变得更厉害。这位女士看了伴侣和 ChatGPT 的对话，发现他并没有做出什么真正的突破，但

是 ChatGPT 对他说法的方式就像他是先知。她的伴侣对她说，如果她不用 ChatGPT，他就和她分手。这位女士和伴侣在一起 7 年，共同建立了家庭。她觉得荒谬又困惑，把这段经历放在网上，获得了 1600 多个评论，其中 50 多人表示自己的伴侣有相似的情况。后来《纽约时报》也报道了相似的故事，其中两个人已经脱离现实，一个想自杀，另一个人服用特定的药物，这个药物也是他自己训练的 ChatGPT 开给他的。

人工智能系统让我们更好了吗？还是出现了更多的问题？我们现在处在一个十字路口。

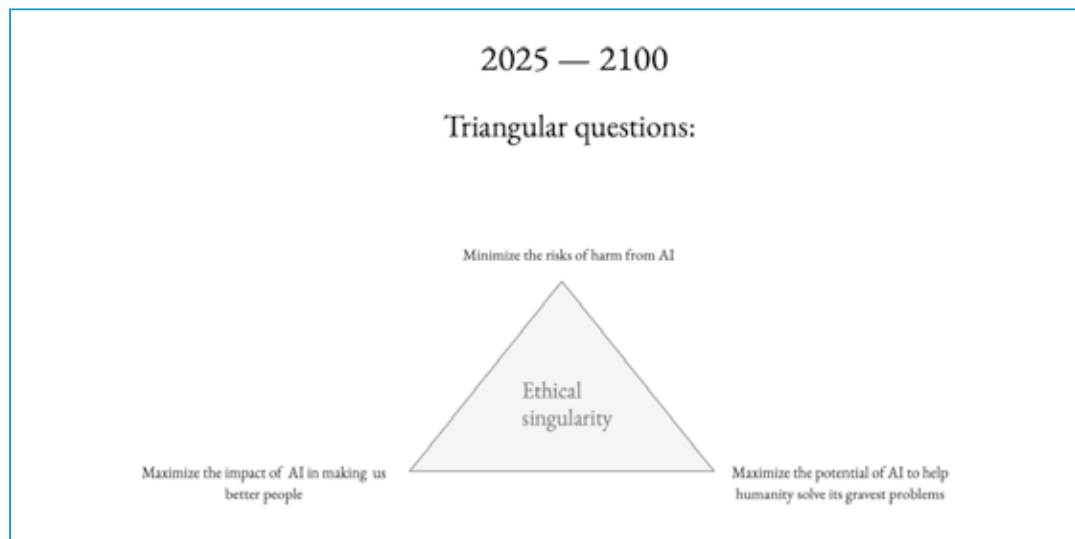
AI 在很多情况下可以帮助我们，比如帮助我们找到癌症还有阿尔茨海默症的治疗方法。但是它仍然会对一些人产生不同的影响，我们已经看到了一些负面的例子。

我们需要回答的三个问题

再回到一开始的问题，机器能思考吗？我们对于这个问题好像也有了很多不同的回答。

从务实的角度看，机器可以做一些类似于思考的事情，我们可以和它对话，它已经成为我们的一部分。

在之前的 75 年当中，对我们最重要的问题是机器能思考吗？从 2025 年到 2100 年，我们需要回答三个问题，而且需要同时回答这三个问题。回答得越好，我们越能达到道德奇点。



第一个是我们需要确保将人工智能造成的伤害风险降到最低。我们无法将它彻底消除，但是风险应该最小化。

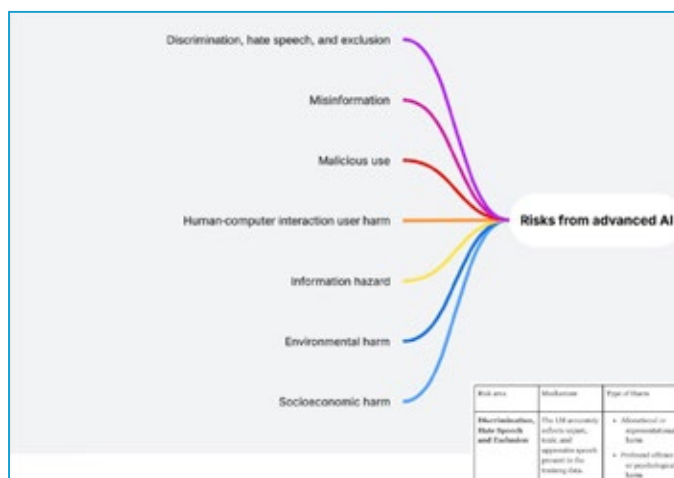
第二我们尽可能地发挥人工智能的影响，使我们成为更好的人。我们可以提高我们的认知能力，我们可以通过人工智能让我们和其他人建立更好的连接。

第三个问题是最大限度地发挥人工智能的潜力，解决人类社会的大问题。我们正在面对很多环境、社会、地缘政治等方面的问题，我们要怎么样用 AI 这个很有潜力的工具，帮我们解决最棘手的问题。

当我思考 2100 年的情况，我不知道可能会发展成什么样，但是我们必须得回答这三个问题。我尝试从我从事的工作角度回答。

首先要尽可能地减少人工智能造成伤害的风险，我们看到很多的论文和演讲，是关于人工智能的风险，很多跨学科的团体坐在一起，看到底 AI 可能会对社会，对个人，对一些具体的行为或者对于经济产生什么样的影响。我们会提出一些关于这些问题的解决方案。

也有一些开放性的问题，我们不知道该如何解决。现在有各种不同的伤害风险分类法，我们和来自谷歌 Deepmind 的不同学科 20 多名研究人员一起进行了分类。总的来说，几乎所有人工智能的伤害风险都可以归入以下七个分类。



一个是歧视仇恨言论和排斥，我们要确保 AI 不会向这个方面演进；第二个是虚假信息，我们要确保 AI 不产生虚假信息；第三个是恶意使用；第四，不能在人机交互中对人带来伤害，人们不能够太沉迷于 AI 系统；第五，AI 不会向人们生成或输出私人信息；最后，它不应该生成对环境有危害，以及对社会经济有危害的问题。

但在英文世界，还有其他类别的人工智能带来的生存风险——我不知道在中国人们是如何看待生存风险的——某个 AI 系统变得非常强大，成为超级智能，一旦失控它可能会杀掉人类，这是一种中心化的风险模型。

但还有另外一种方向，就是生存风险以去中心化的方式发生。各种先进的 AI 系统嵌入到经济生活、社交媒体，损伤到社会的方方面面。日积月累，也会造成巨大的混乱，但是不会导致人类的灭绝。

第二个问题是如何最大限度发挥 AI 的影响，让我们成为更好的人。我要给大家介绍一系列令人印象深刻的成果，其中一项已经发表在《科学》杂志上，论文的名字是《人工智能可以帮助人类在民主审议中找到共同点》。这里考虑的是一些人想要做决定，他们有不同的想法。在某些

时刻他们的对话停止了，现在研究人员使用大模型找到共同之处，解决集体的问题。这个研究模型也被使用来找到以色列和巴勒斯坦和平建设工作的共同点。有的时候很难开始这样的对话。但是这种大语言模型的工具得以应用，使得人们能够更好地找到共通之处，实现和平。

第三个问题是最大限度地发挥人类的潜力，帮助人类解决重大的问题。



这是联合国的可持续发展目标的一部分，有 17 个不同的问题，没有一个得到解决。如果我们专注于用非常有针对性的方式来使用 AI 系统解决这些问题，也许可以走向更美好的未来。

所以，如果我们能够解决这三个问题，那我们就会体验更好的伦理奇点。

Stephen: 我非常喜欢你的想法：AI 可以使人类更美好。是 AI 这么想的还是人这么想的？

Atoosa: 我非常同意您的问题，到底是什么样的人类算更好？我们会有一些共同的道德直觉，比如大规模杀人是不对的。当我们审视各个文明的时候，各个国家都有自己的哲学。无论是中国哲学还是亚里士多德的哲学，我们考虑好人的时候有不同的方式，没有统一的答案，但是至少对好是什么样子，有一些共同的底线的理解。不同国家可以有自己的好人概念，并推动自己的公民向这个方向努力。

Stephen: 现在我邀请陈楸帆介绍一下他对于未来的认识，大家欢迎！