郭毅可，香港浸会大学副校长，帝国理工学院数据科学研究所所长，兼任上海大学计算机学院院长，英国皇家工程院院士，欧洲科学院院士，中国人工智能学会名誉副理事长，英国计算机学会院士 (FBCS)，主要研究方向为机器学习、数据科学、大规模科学数据管理，为许多重大研究项目作出了重要贡献，是英国和欧洲的几个主要数据科学项目的领军人物，发表学术论文 250 余篇。

卷首语 Foreword

从做得多到做得对: 也谈人工智能的发展方向

郭毅可

人工智能发展到今天，人们在追求让机器尽可能多做事的路上走得很猛，也走得很远： 让机器写新闻，让机器作音乐，让机器改照片。正是因为这些发展，使得人们不禁担忧，生怕哪天机器做得太多了，人类都无事可做，要饱受失业之苦。但是，慢慢地我们发现我们对机器做得多的要求也许不难实现，但要求机器做得对，则不那么简单了。从“对抗攻击”到“深度作假”，一次又一次地证明，今天的技术不仅不牢靠，而且很容易用来做坏事。从埃塞俄比亚航空 302 号航班的坠毁，到自动驾驶车的事故，人们开始认识到机器是有行为的，而机器的行为也不一定是有益的，很可能会带来危害。于是，人们开始提出人工智能伦理性、有益性的问题，开始关注机器行为的合理性和正确性，开始认真地询问一些人工智能的基本问题：

机器行为的目的是什么：机器是不是能按照人们意图去改变世界？

机器行为的原则是什么：机器如何不违背人类的伦理和规范？

机器行为的结果如何验证：如何来衡量结果和目的的一致性？

机器行为如何来解释：如何来理解机器获得结果的过程？

这些问题是相互关联的，构成了人工智能的一个重要的分支——机器行为学。 机器行为的研究和机器学习的基础理论研究密切相关，譬如，我们如何向机器表达学习的目的，我们如何来验证机器行为结果的正确性，以及对机器行为过程的可解释性，这些都是今天机器学习研究中最困难也是最基本的问题。在人工智能领域的名著《人工智能一种现代的方法》中，对机器行为的阐述是全书的核心。两年前，在意大利我和该书的作者 Peter Norvig 聊天，谈到他的这本书，他说他正在写新的一版，新版的重点将是论述“效用函数”。效用函数是我们向机器阐述学习目的的表达形式，今天我们对机器表达对其行为目的的要求通常很简单，往往就是一个损失函数，反映的是对机器学习结果的一些简单的质量要求（如精确度、紧致度和稳定性等）。但是，如果我们要对机器行为有很多的要求，要向机器描述做得对的标准，那么这样的效用函数就会非常难定义，如果更深究一下这个问题的话，我们不禁要问“对学习目的或行为目的的表达，一定是要通过效用函数吗？”

今天的深度学习，说到底就是对一个由神经元网络所构成的非线性函数在大数据上作拟合，这种学习行为使得它在应用的普适性上有很大的优势，满足了我们对“做得多”的要求，但是，它的结果正确性无法得以验证，因为我们无法理解该结果生成的逻辑，无法解释学习的认知行为。所以它的稳定性和可靠性都是一个难题。我们可以在许多应用中用到深度学习的技术， 但无法对这样的智能加以对和错的评价。从这个意义上来讲，我们离图灵时代还很远，因为当我们向人和机器共处的黑屋提问，并无法区分得到的回答来自人还是机器，这时候我们可以加上一句： “请告诉我你是如何得到这个问题的答案？”，人是能回答这个问题的，而机器则一定很茫然。

当然，对于这些问题的研究，不仅仅是技术和科学，许多还是哲学性的思考，什么是“对”，这就是典型的哲学问题，从阿西莫夫的机器人三定律，到今天 Stuart Russell 的可证明有益人工智能的三大原则，我们无不在什么是“对”这个问题上进行探讨。更进一步，我们也会对于一些具体的应用上什么是“对”这个问题作出各种理解。但这是一个永远开放性的研究课题，因为 “什么是对” 是一个价值观的问题，对应的回答不可能是唯一的，也不可能是完全的，因为我们多少年来时时刻刻都在讨论这个问题，而人类文明史就是在这样的讨论中前进的。

所以，人类在赋予机器智能的时候，实际上也在对自己的智能行为作出审视，我们要求机器和我们人类一样：不是要做得多，而是要做得对。