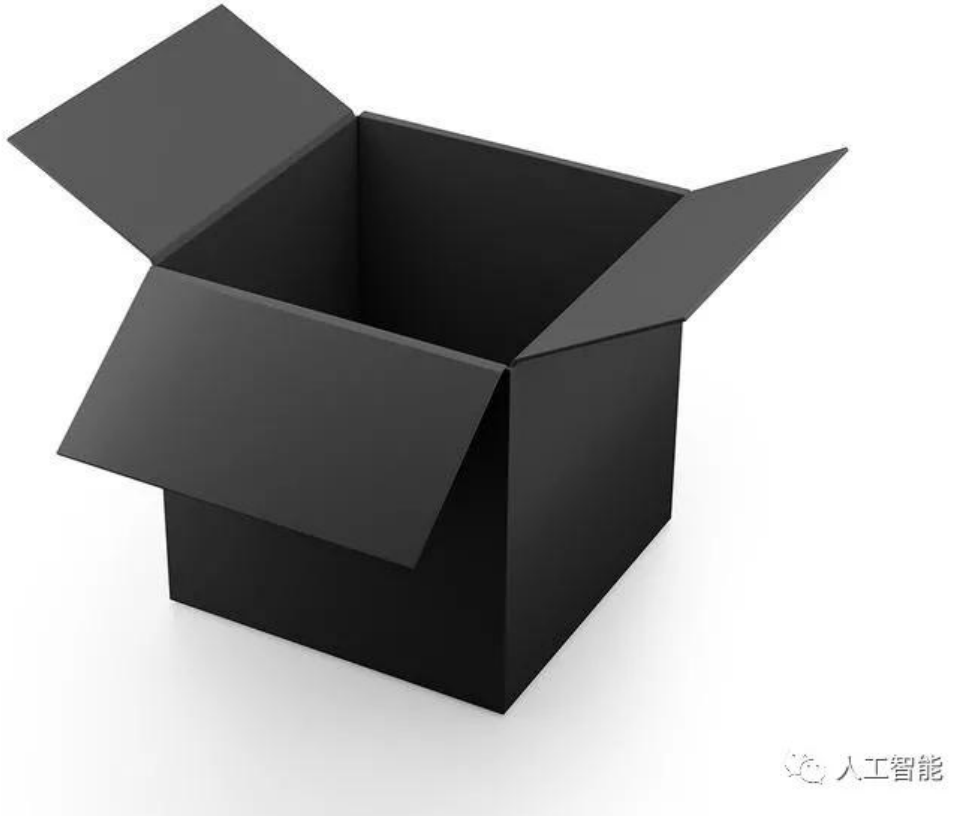
# 不要一棍子打翻所有黑盒模型，其实可以让它们发挥作用

人工智能2019-04-15

一直以来大家都对深度学习这样的黑盒系统多有诟病，即便深度学习的可解释性方面已经有所建树，怀疑和抵制之声仍然很多。但 CMU 材料科学与工程系教授 Elizabeth A. Holm 近期在《科学》杂志发表了一篇短评文章，少见地为黑盒系统给予一些肯定。这篇文章的视角也提醒我们重新考虑一下，一听说是黑盒系统就敬而远之是否是最好的做法。



曾经，科幻小说作家 Douglas Adams 假想人类建造出了有史以来最厉害的计算机，它的名字叫 Deep Thought ,它上面运行的程序可以解答人类能够提出的最深刻的问题「生命的意义是什么」、「宇宙为什么存在」,以及其它所有问题。在计算了 750 万年以后，Deep Thought 给出了一个回答：数字「42」。随着人工智能系统已经开始进入所有人类努力探索的领域，包括科学、工程以及医疗保健，如今人类也必须面对 Douglas Adams 在这个故事里巧妙地隐含的问题 :当我们不理解为什么会出现这个答案的时候，我们还有没有必要知道这个答案？一个黑盒系统到底好还是不好？

在我们学校大多数的物理科学和工学的教授同事们眼中，不使用深度学习这样的 AI 方法的最大原因就是他们不知道如何解释 AI 给出的答案是如何产生的。这个反对意见非常有力，其中隐含的顾虑可以包括实践、可以包括道德、甚至还可以包括法律。科学家们的使命、以及工程师们的职责都要求不仅仅能够预测会发生什么，还要理解它为什么会发生。一个工程师能够学会预测一座桥是否会倒塌，AI 系统其实也可以学会做同样的事情,但只有工程师才能通过物理模型解释清楚他的决定是如何做出的，然后和别人沟通交流、让他们评价他的思路。假设有两座桥，人类工程师认为一座桥不会塌，AI 认为另一座桥不会塌，那你会对哪一座桥更放心呢？

黑箱系统给出的答案无法完全令人信服的事情并不只发生在科学家和工程师身上。2018 年提出的「欧盟一般数据保护条例」GDPR 中就要求基于个人数据的自动决策系统能够为决策对象提供「对于涉及的决策逻辑的有意义的解释」。目前人们仍然在讨论这条要求如何在司法实践中落实,但是我们已经可以看到司法系统对于无法解释的系统的不信任。

在这种整个社会的怀疑氛围下，AI 研究人员们的行动很好理解，他们不再公开宣扬黑盒决策系统，但他们展开更多研究，尝试更好地理解黑盒系统是如何做出决策的——这也就是我们常说的「可解释性」问题。实际上，这也是计算机科学领域当今最大的挑战之一。

不过，一刀切地拒绝所有的黑盒系统也许鲁莽了一点。在现实中，科学家和工程师们，作为人类、也和所有其他人一样地，根据自己已有的判断和经验做出决策，就好像是来自他们自己大脑中的「深度学习系统」。所以，脑神经科学也遇到了和计算机科学一样的可解释性挑战。然而，对于人类做出的决策、给出的结论，我们常常不加防备地直接接受，也不去试着完全了解它们的来源过程。这样说来，AI 系统给出的答案也许值得考虑一下，它们也许也能发挥类似的益处；如果能确认的话，那我们还应该使用它们。

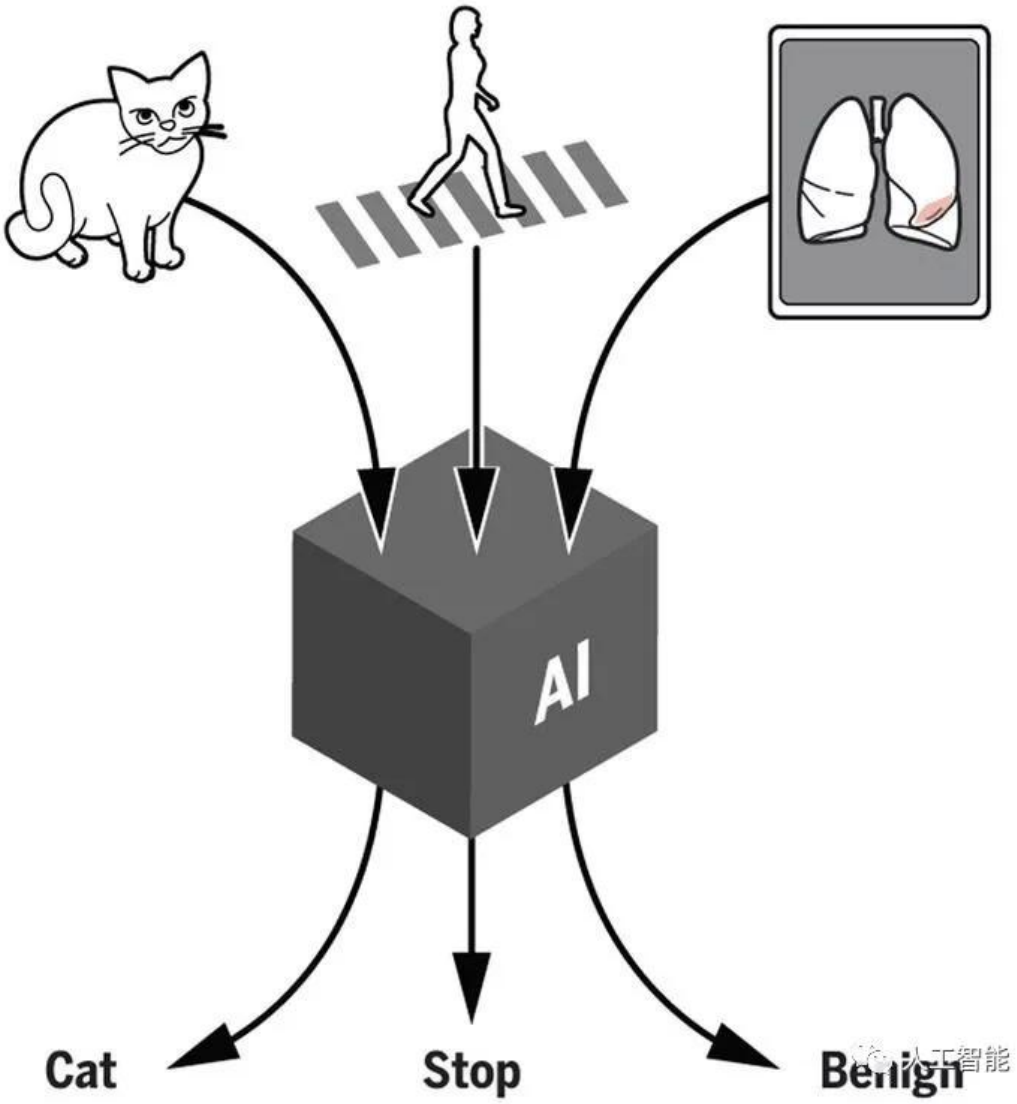
首当其中可以说的，也是最明显的，就是当错误答案带来的代价比正确答案带来的价值低很多的时候。定向广告投放就是一个典型的例子。从广告商的角度讲，投放了广告但是目标群体不想看的成本是很小的，但是成功的广告却能带来可观的收益。以我自己的研究领域，材料科学来说，图像分割任务通常都需要人类把材料图像中感兴趣的部分的复杂内部结构的边界手动勾画出来。这个过程成本很高，以至于不论是博士论文还是工业级的质量控制系统中一旦有需要图像分割的部分，都要让这部分所需的图像尽可能地少。如果换成 AI 系统，它就能很快完成大批量的图像分割任务，同时还具有很高的保真度（虽然并不完美）。在这里，完美的图像分割结果对于这些系统

并不是必需的，因为出现那么一些误分类的像素的代价要比没有 AI 系统时研究生们付出的时间精力低太多了。

第二个可以使用黑盒系统的例子也很明显，不过要更有活力一些。如果一个黑盒系统能产生最佳的结果，那我们就应当使用它。比如，在评估标准的平面医学影像时，经过训练的 AI 系统可以帮助人类影像科医生得到更准确的癌症评估结果。虽然这种情况下出现一个错误答案（不论是假正例还是假负例）的代价并不低，但在黑盒系统的帮助下我们可以达到其它任何方案都无法达到的高准确率，这就成为了当前最佳的解决方案。当然了，有人会说让 AI 看 X 光片本来就可以接受，部分原因是因为总会有人类医生检查 AI 给出的结果；让 AI 开车就会让人有更多顾虑，因为这个黑盒系统做出的决策能影响人的生死，但同时它却没有给人类留出干预的机会。即便这样，自动驾驶汽车也总有一天会比人类驾驶的汽车更安全，它们将会在事故率和死亡率上都做得比人类司机更好。如果取一些合理的指标来衡量，那么那一天一旦到来我们马上就会知道，但是是否让人类司机让位给 AI 司机会是整个社会的决定，需要考虑到人类道德观念、公平性、非人类实体的追责等许多方面。

但是需要说明的是，我们能列出这些情况并不代表黑盒模型在这些场景中就直接得到许可了。在上面两种情况中我们都假设了一个理想的黑盒子，有人对它的运行负责，而且能够它的代价，或者能够明确无误地定义什么是最好的结果。这两个假设都有落入误区的可能。AI 系统可能会有一系列的缺点，包括偏倚、在训练的领域外不适用、脆弱性（很容易被欺骗）。更重要的是，评估代价和最佳结果是一个复杂的决策问题，需要在经济性、个体需求、社会文化、道德考量等许多方面之中做出权衡。更糟糕的

是，这些因素可能是一环套一环的：一个有偏倚的模型可能会隐含着一些代价，可以表现为模型自己做出错的预测，也可以表现为外人对模型的公平性的评估不准确。一个脆弱的模型可能会包含一些盲点，在某些时候会产生错的离谱的糟糕决定。就像面对任何决策系统一样，使用黑盒系统的时候仍然要配合知识、判断力和责任心。



**根据定义，人类无法解释一个黑盒算法是如何给出某个具体的答案的。但是，当黑盒系统能够带来最佳的产出，或者给出错误答案的代价很小，或者能够启发新的思维的时候，它们仍然可以为我们带来价值。**

虽然 AI 的思考过程是带有限制的、可能包含偏倚甚至可能直接就是错误的，但它们毕竟和人类的思考方式有很大的区别，有可能可以揭示新的联系和新的方法。这样一来，黑盒系统就有了第三种可以使用的场景：作为引导人类思考和质疑的工具。比如在某项突破性的医学影像研究中，科学家们训练了一个深度学习系统来根据眼部照片诊断糖尿病性视网膜病变，得到的结果能够近似或者超过一组眼科专家的表现。更令人惊奇的是，这个系统还可以一并分析得出眼科诊断中不会涉及的别的信息，包括心脏病风险高低、年龄、性别等等。在此之前从来没有人注意过不同性别的视网膜之间会有什么区别，所以这个黑盒子系统的发现就给科研人员们提供了新的线索，可以进一步探究不同性别的视网膜之间到底有何区别。对于这些引发的问题的研究也就不再继续属于可解释的 AI 系统以及人类智慧的黑盒系统领域。

说了一圈，我们可以再来看看一开始提到的 Deep Thought 给出的答案「42」。我们没法用黑盒 AI 系统寻找因果关系、构建知识和逻辑系统以及达成理解，一个黑盒系统也没办法告诉我们桥为什么会塌、生命和宇宙的种种问题的答案是什么、以及解释世间的一切。至少目前，这些问题都属于人类智慧和逐渐发展的可解释 AI 的领域。但同时，我们也仍然可以用适当的方式接受黑盒系统。黑盒系统可以对科学、技术、工程、数学等等领域产生潜在且正面的影响，可以产生价值、优化结果以及启发创新。

via

science.sciencemag.org/content/364/6435/26 ,

Science 05 Apr 2019: Vol. 364, Issue 6435, pp. 26-27。

**- END -**