# 对人工智能提出正确的问题

在过去的几年里，关于人工智能（AI）将如何拯救或毁灭世界的讨论层出不穷。自动驾驶汽车会让我们活下来；社交媒体泡沫会摧毁民主；机器人烤面包机会剥夺我们加热面包的能力。

你可能很清楚，有些是胡说八道，有些是真的。但如果你没有深入到这个领域，你可能很难猜出是哪一个。尽管互联网上有无数的初级读物，供那些想学习人工智能编程的人使用，但对于那些不想成为软件工程师或统计学家的人来说，对这些想法以及它们所暗示的社会和道德挑战的解释并不多。

如果我们想作为一个社会对此进行真正的讨论，我们需要解决这个问题。所以今天，我们将讨论人工智能的现实：它能做什么，不能做什么，它在未来可能会做什么，以及它带来的一些社会、文化和道德挑战是什么。我不会涵盖每一个可能的挑战；其中一些，如过滤泡沫和虚假信息，是如此之大，他们需要自己的整篇文章。但我想给你足够多的例子，说明我们面临的真正问题，你将开始自己提出困难的问题。

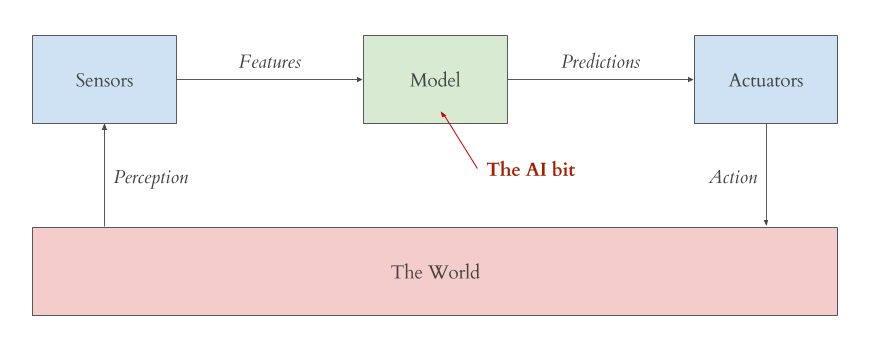
我给你一个破坏者开始：大多数最困难的挑战根本不是技术。人工智能最大的挑战往往是从写它开始的，它让我们必须非常明确地表达自己的目标，在某种程度上，几乎没有其他东西能做到——有时，我们不想对自己那么诚实。

#### 人工智能与机器学习

在我写这篇文章时，我将或多或少地交替使用“人工智能”（AI）和“机器学习”（ML）这两个术语。有一个愚蠢的原因，这些术语的意思几乎是一样的：它是“人工智能”在历史上被定义为“任何计算机还不能做的事情”。多年来，人们争论说，真正的人工智能需要下棋、模拟对话或识别图像；每一次事实上，球门柱被移动了。“人工智能”这个词实在太吓人了：它可能与我们定义自己的方式太接近了，也与我们作为人类的不同之处太接近了。因此，在某个时候，专业人士开始使用“机器学习”这个术语来避免整个对话，而这个词却被卡住了。但它从来没有真正卡住，如果我只说“机器学习”，我会听起来奇怪的机械 - ，因为即使是专业人士也一直在谈论人工智能。

所以让我们先来谈谈什么是机器学习，或者人工智能。从最严格的意义上讲，机器学习是“预测统计学”领域的一部分：“它是关于建立一个系统，这个系统可以获取关于过去发生的事情的信息，并利用它们周围世界的某种模型来预测在其他情况下可能发生的事情。这可以简单到“当我向左转动方向盘时，汽车也倾向于向左转”，也可以复杂到试图了解一个人的整个生活和品味。

你可以用这张图片来了解每一个人工智能的功能：



有一个系统有一些传感器可以感知世界，这些传感器可以是任何东西，从摄像机和激光雷达到一个网络爬虫看文件。还有一些其他的系统可以对世界起作用，从开车到展示广告到分类鱼。有时，这个系统是一台机器，有时它是一个人，他必须根据一些令人绝望的复杂或太大的事情做出决定，而不能像互联网的全部内容那样一下子就想起来。

要将两者联系起来，你需要一个盒子，它可以记录你对世界的看法，并给出建议，说明如果你采取各种行动，可能会发生什么。中间的那个盒子叫做“模型”，就像“世界如何运作的模型”，那个盒子是人工智能的一部分。

上面的图表中有一些额外的词，当专业人士讨论人工智能时，你可能会听到这些词。“特征”仅仅是原始感知的某种升华，这些感知的部分是模型设计者认为有用的在一些人工智能系统中，特征仅仅是原始的感知，例如，相机的每个像素所看到的颜色。如此大量的功能对人工智能是有益的，因为它不会对什么是重要的和什么不是重要的强加任何先入之见，但会使人工智能本身的构建变得更加困难；只有在过去的十年左右，才有可能构建足够大的计算机来处理这些问题。

“预测”是另一个结果：当你向模型展示一些特性时，它通常会给你一堆可能的结果，并且它对每种结果的可能性有最好的理解。如果你想让一个人工智能做出决定，那么你需要对它应用一些规则，例如，“选择最有可能成功的一个”，或者“选择最不可能导致灾难性失败的一个”。衡量可能的成本和收益的最后一条规则，对系统的重要性不亚于模型本身。

现在，你可以想象一个非常简单的“模型”，它给出的规则对许多用途来说都是很好的：例如，旧蒸汽机上的机械调节阀是一种简单的“模型”，它在一端读取压力，如果压力将杠杆推过某个设定点，它就会打开一个阀门。这是一个简单的规则：如果压力高于设定值，打开阀门；否则，关闭阀门。

这个阀门之所以如此简单，是因为它只需要考虑一个输入，然后做出一个决定。如果必须决定一件更复杂的事情，而这取决于数千或数百万的输入，比如如何控制汽车（这取决于你所有的视觉、听觉等等），或者哪一个网页能给你关于袋熊养殖的问题最好的答案（这取决于你是偶然感兴趣的还是专业的有袋牧马人，并且取决于网站是否是由一个狂热者写的，或者仅仅是想卖给你便宜的普通袋熊伟哥） - 你不会发现一个简单的比较，但是数百万，甚至数千万，需要做出决定。

人工智能不会感到无聊或心烦意乱：一个模型可以连续数百万或数十亿次对不同的数据片段做出决策，而且不会变得更糟（或更好）。

使人工智能模型与众不同的是，它们是为此而设计的。在任何人工智能模型中，都有一系列的规则来组合特征，每一个规则都依赖于成百上千甚至上百万个单独的旋钮中的一个，告诉它在不同的情况下如何权衡每个特征的重要性。例如，在一种称为“决策树”的人工智能模型中，该模型看起来像是一棵包含是/否问题的巨树。如果人工智能的工作是区分金枪鱼和鲑鱼，第一个问题可能是“图片的左半部分比右半部分暗吗？”？最后，根据对过去374个问题的回答，这个正方形中像素的平均颜色是橙色还是红色？这里的“旋钮”是提问的顺序，以及每个问题的“是”和“否”之间的界限是什么。

神奇之处在于：不可能找到正确的设置组合来可靠地区分金枪鱼和鲑鱼。他们太多了。首先，人工智能是在“训练模式”下运行的。人工智能是一个接一个的例子，每次调整它的旋钮，这样它就能更好地猜测下一步会发生什么，并在每次错误后修正自己。它看到的例子越多，看到的不同的例子越多，它就越能从偶然事件中分辨出关键。一旦经过训练，旋钮的值就固定下来，模型就可以投入使用，连接到真正的执行器上。

ML模型比人类执行相同任务的优势不在于速度；ML模型通常需要几毫秒的时间来做出决定，而这也是人类所需要的时间。（你在开车的时候总是这样做）他们真正的优势是他们不会感到无聊或心烦意乱：一个ML模型可以连续数百万或数十亿次地对不同的数据片段做出决策，而且不会变得更糟（或更好）。这意味着你可以将它们应用于人类在 - 方面非常糟糕的问题，比如为一次搜索对数十亿个网页进行排名，或者驾驶汽车。

（人类驾驶汽车很糟糕：这就是为什么2015年仅在美国就有人驾驶汽车。）。这些撞车事故中，绝大多数是由于分心或司机失误造成的，这些事情人们通常都做得很好，但在关键时刻却没能做到一次。驾驶需要极大的意识和在一小部分时间内做出反应的能力，如果你考虑到这一点，我们完全可以做到。但更糟糕的是，它要求我们能够连续几个小时这样做，而事实证明我们做不到这一点。）

当有人谈论在项目中使用人工智能时，他们的意思是将项目分解成上面绘制的组件，然后构建正确的模型。这个过程首先收集训练实例，这通常是任务中最困难的部分；然后选择模型的基本形状（例如“神经网络”、“决策树”等；这些是适合不同问题的基本模型）并运行训练；然后，最重要的是，找出损坏的地方并调整它。

例如，看下面六张图片，找出前三张和后三张的关键区别：

如果你猜到“前三个有地毯”，你说得对！

当然，如果你猜到前三张是灰色猫咪的照片，后三张是白色猫咪的照片，那你也是对的。但如果你用这些图像训练你的灰猫探测器，当模型试图给你的训练图片打分时，你可能会得到极好的表现，而在现实世界中，你的表现也很糟糕，因为模型实际上学到的是“灰猫是坐在地毯上的猫形物体”

这被称为“过度拟合”：当你的模型学习了训练数据的特性，而不是你真正关心的东西时。避免这一点是构建ML系统的人花大部分时间担心的。

#### 人工智能的优缺点

现在我们已经讨论了什么是人工智能（或ML），让我们讨论一下它到底在哪里有用或无用。

如果目标和实现这些目标的方法都被很好地理解，那么问题甚至不需要人工智能。例如，如果你的目标是“把这个车轮上的所有螺母拧紧到100英尺磅”，你所需要的只是一个能够拧紧和测量扭矩的机构，当扭矩达到100时停止拧紧。这就是所谓的“扭矩扳手”，如果有人提供给你一个人工智能扭矩扳手，正确的第一个问题是问他们为什么我会想要。这些是人工智能的蒸汽减压阀，你只需要一个简单的机械装置。

人工智能在目标明确但手段不明确的问题上大放异彩。这在以下情况下最容易做到：

* 可能的外部刺激的数量是有限的，因此模型有机会了解它们，并且
* 你必须控制的事情是有限的，所以你不需要考虑太多的选择，而且
* 刺激或决定的数量仍然很大，你不能仅仅写下规则；另外，这
* 很容易将你的一个行为与外界可观察到的结果联系起来，这样你就可以很容易地找出哪些起了作用，哪些没有起作用。

这些事情比看上去更难。例如，现在拿起一个坐在你旁边的物体 - 我用一个空的汽水罐来做。现在再慢慢地做一遍，看看你的手臂做了什么。

我的手臂在肘部快速旋转，将我的手从键盘上的水平移动到垂直，距离罐子几英寸，然后迅速停止。然后它向前移动，而我的手指比罐子稍大一点，比第一个动作慢一些，但还是有点快，直到我看到我的拇指在罐子的另一边，而不是我的其他手指 - 尽管我的其他手指被罐子挡住了视线。然后我的手指闭上，直到遇到阻力，几乎立即停止。当我的手臂开始抬起时，这一次是从肩膀开始（保持肘部固定），他们的握力无限地收紧，直到它牢牢地抓住了罐子，但没有变形。

事实上，我们走路时不会在困惑中摔倒在地，这比看上去要神奇得多。下次你穿过房间时，注意你走的确切路线，每次你弯曲或移动你的身体或把你的脚放在任何地方，除了直接在你面前。机器人学中称之为“运动规划”的问题确实很难解决。

这是一项非常困难的任务，以至于我们的大脑有两位数的质量百分比专门用于其他事情。这让我们觉得他们比实际要容易得多。这一类的其他任务是人脸识别（我们大脑的很多部分不是专门用于普通视觉，而是专门用于识别人脸）、理解单词、识别三维物体以及在不碰到物体的情况下移动。我们并不认为这些很难，因为它们对我们来说是如此直观，但它们是直观的，因为我们进化出了专门的器官，除了擅长这些之外什么也不做。

对于这一系列狭窄的任务，计算机做得很差，并不是因为它们做得比做类似的任务差，而是因为我们直觉上非常擅长它们，以至于我们对于什么是“可接受的性能”的基线非常高。如果我们的大脑中没有很大一部分除了识别面孔之外什么也不做，人们会像犰狳那样看待我们，这正是计算机所发生的事情。

（相反，人类被连接起来的方式使得其他任务人为地容易让计算机“做得足够正确”。例如，人类大脑被连接起来，在怀疑的情况下，假设某种或多或少有生命活动的东西实际上是有生命的。）。这意味着，与人类进行令人信服的对话并不需要理解一般的语言；只要你能将对话保持在一个或多或少集中的话题上，人类就会自动纠正任何不清楚的地方。这就是为什么语音助理是可能的。最著名的例子是，一个1964年的“人工智能”，它模仿了一个罗杰尔心理治疗师。它只会理解你的句子，让你多讲一些关于各种事情的事情，如果它被搞糊涂了，它会回到安全的问题上，比如“告诉我关于你母亲的事情”。虽然它的一半意思是开玩笑，但人们确实报告说，与它交谈后感觉好多了。如果你有权使用谷歌助手驱动的设备，你可以告诉它“好的谷歌；和伊丽莎谈谈”然后自己看看。）

要理解上面描述的最后一个问题 - 一个很难将你的直接行动与结果联系起来的情况 - 考虑学习玩电子游戏。有些动作的后果是显而易见的：当你本该走弯路的时候，你却走弯路，撞到了墙，游戏结束了。但当你在一场比赛中变得越来越好，你就会开始意识到“糟糕，我错过了那一次机会，五分钟后我会完蛋的”，并把这个决定归因于一个更晚的结果。你必须花很多时间去理解游戏的机制，然后你才能理解这种联系；人工智能也有同样的问题。

我们已经讨论了目标和方法被理解的情况，以及目标而不是方法被理解的情况。还有第三类，人工智能根本帮不上忙：目标本身不太清楚的问题。毕竟，如果你不能给人工智能一堆好的和不好的解决方案的例子，它会从中学到什么？

我们将在一瞬间更多地讨论这些问题，因为实际上像这样但我们认为不是最棘手的道德问题经常出现的问题。很多时候真正发生的事情是，要么我们不知道“成功”的真正含义（在这种情况下，你如何知道你是否成功？）或者更糟的是，我们确实知道，但并不真的想承认。计算机编程的第一条规则是他们不擅长自欺欺人：如果你想让他们做某事，你必须向他们解释你想要什么。

在我们进入伦理学之前，这里有另一种方法来划分人工智能的优缺点。

最简单的问题是在一个可预测的环境中有明确的目标，从一个非常简单的环境（一个凸耳螺母，你甚至不需要人工智能）到一个更复杂但可预测的环境（一个摄像头看着一条装配线，它知道一辆车很快就会出现，它必须找到车轮）。几年来我们一直擅长于自动化。

更困难的问题是在不可预测的环境中有明确的目标。驾驶汽车就是一个很好的例子：目标（以合理的速度从a点安全地到达B点）很容易描述，但是环境中可以包含任意多的惊喜。人工智能在过去几年才发展到这些问题真的可以被攻击的程度，这就是为什么我们现在攻击的是像自动驾驶汽车或自动驾驶飞机这样的问题。

另一种困难的问题是在可预测的环境中的间接目标，这些都是环境有意义的问题，但是你的行为和这些目标之间的关系非常遥远，就像玩游戏一样。这是另一个领域，我们在最近的过去取得了巨大的进展，与人工智能能够做以前难以想象的事情，如在围棋中获胜。

在棋盘游戏中获胜本身并不是很有用，但它为在不可预测的环境中实现间接目标开辟了道路，比如规划你的金融投资组合。这是一个更困难的问题，而且我们还没有在这方面取得重大进展，但我希望我们在未来十年里能够在这方面取得进展。

最后你遇到了最困难的情况，目标不明。人工智能根本无法解决这些问题；如果你不能告诉系统你想要它做什么，你就无法训练它。写小说可能就是一个例子，因为“好小说”是由什么构成的还没有一个明确的答案。另一方面，在这个问题的某些特定部分，目标可以被定义，例如，“写一本小说，如果以恐怖的形式推销，它会很畅销。”

这到底是对人工智能的一种好的还是坏的使用，取决于读者的智慧。

#### 伦理与现实世界

所以现在我们可以开始审视问题的实质了：现实世界中的难题是什么样的，人工智能的工作或失败会对人们的生活产生重大影响？什么样的问题不断出现？

我可以很容易地在书架上摆满关于这个问题的讨论；没有办法看到这个领域中的每一个有趣的问题，甚至是其中的大多数问题。但我会举六个例子，我发现这些例子帮助我思考了很多其他问题，而不是因为他们给了我正确的答案，而是因为他们帮助我问了正确的问题。

#### 一。乘客和行人

一辆自动驾驶汽车正驶过一座狭窄的桥，这时一个孩子突然从桥前冲了出来。现在停车已经太晚了；汽车所能做的就是向前冲，撞到孩子，或者转弯，把自己和乘客送进下面湍急的河里。该怎么办？

我从这个问题开始，是因为在过去的几年里，这个问题在公开场合被讨论了很多，而且讨论往往非常聪明，并且展示了我们真正需要问的问题。

首先，对整个问题有一个很大的警告：这个问题在实践中很不重要，因为自动驾驶汽车的全部要点是，它们不会在一开始就陷入这种情况。孩子很少突然出现；大多数情况下，要么是司机开得太快，自己的反应无法处理从他们能看到的障碍物后面跳出来的孩子，要么是司机分心了，不知为什么，直到太晚才注意到孩子。这两种情况都是自动驾驶者不会有问题的：一次查看周围的所有信号，连续几个小时，而不会感到无聊或分心。像这样的情况会变得越来越罕见，而这就是拯救生命的来源。

但是“几乎从来没有”和“从来没有”不是一回事，我们必须接受有时会发生这种情况。当它发生时，汽车该怎么办？它应该优先考虑乘客的生活还是行人的生活？

这不是一个技术问题：这是一个政策问题，在上面的形式中，它被归结为简单的核心。作为一个社会，我们可以在任何一个答案（或任何组合）上达成一致，我们可以对汽车进行编程来实现这一点。如果我们不喜欢这个答案，我们可以改变它。

有一个很大的不同于我们今天生活的世界。如果你问人们在这种情况下会做什么，他们会给出各种各样的答案，并用各种各样的“取决于”来警告他们。事实上，我们不想做这个决定，我们当然也不想公开承认，如果我们的决定是为了保护我们自己，而不是为了孩子。当人们真的处于这样的情况下，他们的反应最终遍布地图。

从文化上讲，我们有一个答案：在当下最热的时刻，在你看到即将到来的灾难和灾难发生的那一瞬间，我们认识到我们不能做出理性的决定。我们最终会让司机对他们的决定负责，并承认这是不可避免的，不管他们做出什么决定。（尽管我们可能会让他们对在最后一刻之前做出的决定（比如超速或酒后驾车）负更多责任。）

对于一辆自动驾驶汽车，我们没有这个选择；在事故发生前的几年里，程序实际上有一个空间，它现在问我们：“当这种情况发生时，我该怎么办？我应该如何权衡乘客的风险和行人的风险？”

它会照我们说的做。编写计算机程序的任务要求我们对希望它做出的决定非常诚实。当这些决定影响到整个社会的时候，就像在这种情况下一样，这意味着作为一个社会，我们面临着同样艰难的选择。

#### 2。文雅小说

机器学习的模型有一个非常恶劣的习惯：他们会学习数据显示的内容，然后告诉你他们学到了什么。他们固执地拒绝学习“我们所希望的世界”，或“我们所希望的世界”，除非我们明确地向他们解释这是什么，即使我们喜欢假装我们没有做这样的事情。

2016年年中，高中生Kabir Alli尝试在谷歌上搜索“三个白人青少年”和“三个黑人青少年”，结果比你想象的还要糟糕。



“三个白人青少年”出现了有魅力的、运动型青少年的库存照片；“三个黑人青少年”出现了关于三个黑人青少年被捕的新闻报道中的马克杯照片。（如今，无论是搜索还是搜索，大多都会出现有关这一事件的新闻报道）

这里发生的事情并不是谷歌算法的偏见：而是底层数据的偏见。这种特殊的偏见是“看不见的白人”和媒体在报道中的偏见的结合：如果三个白人青少年因犯罪而被捕，不仅新闻媒体不太可能展示他们的头像，而且他们也不太可能把他们称为“白人青少年”。事实上，几乎只有一次青少年群体被明确地被贴上“白色”标签的是库存摄影目录。但如果三个黑人青少年被捕，你可以指望这句话会在新闻报道中大量出现。

许多人对这些结果感到震惊，因为他们似乎与我们的“后种族”社会的国家观念大相径庭。（请记住，这是在2016年年中）但基本数据非常清楚：当人们在媒体上用高质量的图片说“三个黑人青少年”时，他们几乎总是把他们说成罪犯，当他们谈论“三个白人青少年”时，他们几乎总是在做股票摄影广告。

事实上，这些偏见确实存在于我们的社会中，它们几乎反映在你所看到的任何数据上。在美国，如果你的数据没有显示出某种种族差异，那你一定是做错了什么。如果您试图通过不让race作为模型的输入来手动“忽略race”，那么它会从后门进入：例如，某人的邮政编码和收入会非常精确地预测他们的种族。一个看到这些而不是种族的ML模型，被要求预测一些实际上与我们社会中的种族有关的东西，将很快发现这是它的“最佳规则”

人工智能模型向我们举着一面镜子；当我们真的不想诚实时，他们不明白。如果我们提前告诉他们如何对我们撒谎，他们只会告诉我们有礼貌的小说。

这种诚实会迫使你非常坦率。最近的一个很好的例子是在一篇关于“这是一个非常流行的名为word2vec的ML模型，它学习了英语单词 - 的各种含义之间的关系，例如“国王是人，这篇论文的作者发现它包含了很多社会偏见的例子：例如，它也会说“计算机程序员是男人，家庭主妇是女人。”这篇论文是关于他们为消除这种偏见而提出的一种技术。

对于这篇文章的普通读者来说，不明显的是，没有自动消除偏见的方法。他们的过程是相当合理的：首先，他们分析了Word2VEC模型，发现对字沿他/她轴急剧分裂。接下来，他们让一群人来确定哪一对代表了有意义的分裂（例如，“男孩对男人就像女孩对女人一样”），哪一对代表了社会偏见。最后，他们应用数学技术从整体上减去模型的偏差，留下一个改进的模型。

这是一项很好的工作，但重要的是要认识到，在这个 - 确定哪些男性/女性分裂应该被移除的关键步骤 - 是人类的决定，而不是自动的过程。它要求人们直截了当地说出他们认为哪些分裂是自然的，哪些不是自然的。此外，原始模型得出这些分裂是有原因的；它来自对世界各地数百万书面文本的分析。最初的word2vec模型准确地捕捉到了人们的偏见；经过清理的模型准确地捕捉到了评分者对于哪些偏见应该被消除的偏好。

这突出的风险是，“当我们混淆了什么是什么和应该是什么时会发生什么”。如果我们想用原始模型来研究人们的感知和行为，那么它是合适的；如果我们想用它来产生新的行为，并向他人传达某种意图，那么修改后的模型是合适的。如果说修改后的模型更准确地反映了世界是什么，那是错误的；如果说因为世界是某种方式，所以也应该是这种方式，那也是错误的。毕竟，任何一种模型的目的都是为了做出决定。决定和行动完全是关于我们希望世界是什么样的；如果不是这样，我们就什么都不会做。

#### 三。大猩猩事件

在2015七月，当我是谷歌社会努力的技术领导者（包括照片）时，我收到了来自谷歌的一位同事的紧急信息：我们的照片索引系统已经和他有很好的理由。

我大声发誓之后，立即做出的反应是传呼球队，而这不是我们认为可以的。该团队迅速采取行动，并禁用了冒犯性的描述，以及其他一些潜在的风险，直到他们能够解决根本问题。

许多人怀疑这个问题和六年前引起的问题是同一个问题：“面孔”的训练数据完全由白人组成。这也是我们首先怀疑的事情，但我们很快就把它从名单上划掉了：培训数据包括各种种族和肤色的人。

实际发生的是三个微妙问题的交集。

第一个问题是人脸识别很难。在我们看来，不同的人看起来如此鲜明地不同，正是因为我们大脑中有很大一部分只专注于识别人的面孔；我们花了数百万年的时间来进化工具，而不是其他东西。但是如果你比较两张不同的脸和两张不同的椅子有多不同，你会发现，即使是在不同的物种中，两张脸的相似性也比你想象的要大得多。

事实上，我们发现这种虫子并不是孤立的：这个系统也容易把白脸误认为是狗和海豹。

这就涉及到第二个问题，这是问题的真正核心：ML系统在他们的领域非常聪明，但对更广阔的世界一无所知，除非他们接受了教育。当我们试图思考不同的图片可以被识别为不同的物体的所有方式时，这个人工智能不仅仅是关于面孔，没有人想向它解释黑人被比作猿而失去人性的漫长历史。正是这种背景使得这个错误如此严重和有害，而将某人的幼童误认为海豹只会很有趣。

这个问题没有简单的答案。在处理涉及人类的问题时，错误的代价通常与极其微妙的文化问题联系在一起。与其说是很难解释，不如说是很难事先想到它们：快，给我列出可能出现在武器图片周围的最高文化敏感度！

这个问题不仅仅表现在人工智能中：当人们被要求跨文化做出价值判断时，它也会表现出来。这方面的一个特殊挑战是在网上发现骚扰和虐待行为。这类问题几乎完全由人类而不是人工智能来处理，因为制定规则是极其困难的，即使是人类也可以用来判断这些事情。我花了一年半的时间在谷歌开发这样的规则，并认为这是我所面临的最大的智力挑战之一。举一个很简单的例子：人们常说“好吧，一个明显的规则是，如果你说n\*\*\*\*r，那就不好了。”我向你提出挑战，要求你把这个规则应用到（1）几乎任何一首Jay-Z的歌，（2）Langston Hughes的诗，（3）Chris Rock的，（4）如果他在白人面前表演的话，同样的规则观众，（5）如果特德·纽金特一字不差地对他的一位观众表演了同样的节目，并对发生的事情给出了连贯的解释。这是可能的；这远非简单。而这仅仅是五个例子，涉及到出版、编辑、创作的作品，甚至不是正常的对话。

即使有一队人制定规则，而不是人工智能，强制执行规则，文化障碍也是一个巨大的问题。印度的评论者不一定会有美国种族诽谤的文化背景，美国的评论者也不会有印度的文化背景。但是全世界的文化数量是巨大的：你如何用一种任何人都可以学习的方式来表达这些想法？

教训是：系统中最危险的风险往往不是来自系统内部的问题，而是来自系统与更广阔世界互动的意想不到的方式。我们还没有一个好办法来处理这个问题。

（大猩猩事件中的第三个问题是摄影中的种族主义问题。从商业电影的第一天起，颜色和图像校准的标准就包括“标准化模型的图片”。这些模型都是纯白色的，直到20世纪70年代那部电影无法准确捕捉到深色木材的棕色色调！尽管现代的颜色校准标准更加多样化，但我们对于构成“好图像”的标准仍然压倒性地偏向于白脸，而不是黑脸。因此，用手机摄像头拍摄的白人业余照片效果相当不错，但黑人特别是深色皮肤的人的业余照片往往曝光不足。人脸被简化成模糊的棕色斑点，眼睛有时还有嘴，这对于图像识别算法来说是很难理解的。摄影导演艾娃·伯克夫斯基最近接受了一次极好的采访

#### 四。不幸的是，人工智能会照你说的做

“电脑帮我装上了/我希望他们能把它卖了。/它永远不会做我想做的事，而只做我告诉它的事

人工智能的一个重要用途是帮助人类做出更好的决定：不是直接操作某个执行器，而是告诉一个人它推荐的东西，从而更好地装备他们做出正确的选择。当选择具有高风险时，这是最有价值的，但是真正影响长期结果的因素对该领域的人类来说并不是很明显。事实上，如果没有明显有用的信息，人类可能很容易根据自己的无意识偏见而不是真实数据采取行动。这就是为什么许多法院开始使用自动化的“风险评估”作为其量刑指南的一部分。

现代风险评估是ML模型，其任务是预测一个人在未来再次犯罪的可能性。在一个地区的法院历史的完整的语料库训练，它可以形成一个令人惊讶的谁是和不是一个风险的良好画面。

如果你已经仔细阅读到目前为止，你可能已经发现了一些方法，这可能是可怕的，可怕的，错误的。2016年的一份报告显示，这正是全国发生的事情。

COMPAS系统的设计者遵循了最佳实践，该系统由佛罗里达州布劳沃德县使用。他们确保他们的训练数据没有人为地被群体偏见，例如确保所有种族的人都有平等的训练数据。他们注意确保race不是他们的模型可以访问的输入特性之一。只有一个问题：他们的模型无法预测他们认为的预测结果。

量刑风险评估模型应该问的问题是，“这个人将来犯重罪的概率有多大，这是你现在给他们的量刑的函数？“这将考虑到人和判决本身对他们未来生活的影响：它会永远监禁他们吗？释放他们，却没有机会得到一份正式的工作？

它被训练回答“谁更有可能被定罪”，然后被问到“谁更有可能犯罪”，而没有人注意到这是两个完全不同的问题。

但我们并没有一种魔法灯在每次有人犯罪时都会熄灭，我们当然也没有培训过这样的例子：同一个人同时被判两个不同的刑罚，结果却有两种不同的方式。因此，COMPAS模型是在一个真实的、无法获得的数据的代理上进行训练的：考虑到我们在判刑时所知道的一个人的信息，这个人被判有罪的概率是多少？或者用“这两个人中哪一个将来最有可能被判有罪”来比喻两个人？”

如果你对美国的政治有任何了解，你可以马上回答这个问题：“黑人！“黑人比白人更容易犯同样的罪行，因此，一个研究数据的ML模型，完全忽略其他一切，总是预测黑人被告在未来更有可能被判另一项罪行，事实上，这个模型预测得相当准确。

但是这个模型被训练的目的并不是模型被用来做什么。它被训练回答“谁更有可能被定罪”，然后被问到“谁更有可能犯罪”，而没有人注意到这是两个完全不同的问题。

（康帕斯没有把种族作为一个明确的输入没有区别：在我们很多地方，住房是非常隔离的，在布劳沃德县，所以知道某人的地址和知道他们的种族一样好。）

显然这里有很多问题。一是法院对人工智能模式过于重视，将其作为量刑决定的直接因素，跳过了人的判断，其可信度远远超过任何模式所能保证的。（最近，一个很好的经验法则是，对人的严重后果的决定应该由人来进行理智的检查，并且应该有人的超驰机制可用。）当然，还有一个问题，这暴露了潜在的系统性种族主义：黑人更有可能被逮捕并被判同样的罪行。

但这里有一个特定于ML的问题，值得注意：您想要测量的数量和您可以测量的数量之间通常存在差异。当这些不同时，您的ML模型将能够很好地预测您测量的数量，而不是它作为代理的数量。在信任您的模型之前，您需要非常仔细地分析它们的相似和不同之处。

#### 5个。人是理性的动物

在机器学习的讨论中有一个新的流行语：“解释权”。这个概念是，如果ML被用来做任何有意义的决定，人们有权理解这些决定是如何做出的。

直觉上，这似乎是显而易见且有价值的 - 然而，当提到ML专业人士时，他们的脸会变色，并试图解释所要求的实际上是不可能的。这是为什么？

首先，我们应该理解为什么做这件事很难；其次，更重要的是，我们应该理解为什么我们期望它容易做，以及为什么这种期望是错误的。第三，我们可以看看我们能做什么。

早些时候，我将ML模型描述为包含数百到数百万个刻度盘。这不符合真实模型的复杂性。例如，现代人每次输入一个字母。这意味着，模型必须在阅读了多少字母之后，表达其对文本理解状态的条件，以及每一个连续的下一个字母可能如何影响其对含义的解释。（而且它很有效；有了英语和西班牙语这样的语言对，它的表现和人类一样好！）

对于模型遇到的任何情况，它对它所做的唯一的“解释”是“嗯，下面的1000个变量都处于这些状态，然后我看到了字母‘c’，我知道这会改变用户根据下面的多项式谈论狗的概率……”

这不仅是你无法理解的：它也无法理解的毫升研究人员。调试ML系统是该领域中最困难的问题之一，因为在任何给定时间检查变量的单个状态告诉你关于模型的大致相同，因为测量人的神经电位会告诉你他们晚餐吃了什么。

然而，接下来的第二部分，我们总是觉得我们可以解释自己的决定，人们（尤其是监管者）一直在期待这种解释。“我之所以将这项抵押贷款的利率定为7.25%，是因为他们的FICO得分中值，”他们预计会说，“如果他们从益百利获得的FICO得分高出35分，利率就会下降到7.15%。”或者，“我建议我们雇用这个人，因为他们在我们的采访中对机器学习的解释很清楚。”

但认知心理学或行为心理学中的每个人都知道一个黑暗的秘密：所有这些解释都是无稽之谈。我们对是否喜欢某人的决定是在谈话的最初几秒钟内做出的，可能会受到一些看似随机的东西的影响。无意识的偏见充斥着我们的思想，即使我们没有意识到它们。认知偏差是当今心理学研究中最大的（也是我最感兴趣的）分支之一。

事实证明，人们擅长的并不是解释他们是如何做出决定的：而是为他们的决定提出一个合理的解释。有时这是完全无辜的：例如，我们确定了一些在决策过程中对我们来说很重要的事实（“我喜欢汽车的颜色”）并专注于此，而忽略了对我们来说可能很重要但却看不见的事情。（“我继父有一辆掀背车。我恨他。”）它也有更深层次的动机：通过解释我们做了什么或不想要什么来解决认知上的不和谐（“葡萄可能是酸的，无论如何”），或者避免对我们可能不想承认的事情想得太近。（“第一个候选人听起来就像我毕业时一样。那个女人很好，但她感觉不一样……她不适合和我一起工作。”）

如果我们期望ML系统为他们的决定提供实际的解释，我们将遇到很多麻烦，就像我们要求人类解释他们自己决定的实际基础一样：他们所知道的并不比我们多。

但当我们要求解释时，我们真正感兴趣的是哪些事实既突出（因为改变它们会实质性地改变结果）又可变（因为对它们的改变值得讨论）。例如，“有人给你看了这篇招聘启事；如果你住在西边十英里，你会看到这篇招聘启事”在某些情况下可能很有趣，但“有人给你看了这篇招聘启事；如果你是一只鸸鹋，你会看到一个装着木尔加种子的容器”则不是。

当提供给ML系统的反馈的轴时，该信息特别有用：例如，通过向人们展示一些显著的和可变的项目，他们可以提供对这些项目的更正，并提供更新的数据。

产生这种解释的数学技术正在积极发展中，但你应该意识到其中有一些非平凡的挑战。例如，大多数这些技术都是建立在第二个“解释性”ML模型的基础上的，该模型不太准确，只适用于某些给定输入（您自己的）的微小变化，更易于理解，但基于与所描述的“主”ML模型完全不同的原则的输入。（这是因为只有少数几种ML模型，如决策树，是人们完全可以理解的，而在许多实际应用中最有用的模型，如神经网络，显然是不可理解的。）这意味着，如果你试图给系统反馈说“不，改变这个变量！“就解释模型而言，可能根本没有明显的办法将其转化为主要模型的投入。然而，如果你给人们一个解释工具，他们也会要求有权用同样的语言合理地、但不可行地改变它。

人类处理这个问题的方法是在大脑中拥有一种极其普遍的智能，它可以处理各种各样的概念。你可以告诉它，当涉及到种族历史时，它应该注意图像识别，因为同一个系统可以理解这两个概念。在人工智能领域，我们还不能做到这一点。

#### 6。人工智能归根结底是一种工具

在讨论人工智能伦理时，很难不举大家最喜欢的例子：人工智能杀手无人机。这些飞机在高空飞行，只有一台电脑能帮助他们完成在保护平民生命的同时杀死敌方叛乱分子的任务……除非他们决定任务需要一些“附带损害”，正如委婉说法所说。

人们对这种装置的恐惧是理所当然的，如果他们听到更多关于那些永远面临死亡威胁的人突然从晴朗的天空中出来的故事，他们会更加害怕。

人工智能是这场对话的一部分，但它的中心作用比我们想象的要小。大型无人机与有人驾驶飞机的不同之处在于，它们的飞行员可以在数千英里之外，不受伤害。人工智能自动驾驶仪的改进意味着，一个无人机操作员很快就可以驾驶的不是一架飞机，而是其中的一架。最终，大型无人机舰队99%的时间都可以完全自主驾驶，只有在需要做出重要决定时才会召唤人类。这将为更大的无人机舰队或成本更低的无人机空军提供可能性，使空中轰炸人民的权力民主化。

在这个故事的另一个版本中，人类可能完全脱离了“杀戮链”，即是否发射武器的决定过程。（大多数西方军队已经非常清楚地表明，他们不打算做这样的事情，因为这显然是愚蠢的。但一支处于极端状态的军队可能很容易做到这一点，如果没有其他恐怖因素的话，它可能会制造出未知数量的飞机四处飞行，随意杀人，而且我们可能预计未来会有更多的军队拥有无人机。）现在我们可能会问，谁对完全由机器人决定的杀人负有道德责任？

这个问题比我们最初想象的要简单和复杂。如果有人用石头砸到另一个人的头上，我们责怪的是那个人，而不是石头。如果他们投掷长矛，即使长矛“在自己的力量下”飞行了一段时间，我们也不会想到责怪它。即使他们构建了一个复杂的死亡陷阱，印第安纳琼斯式的，意志行为是人的，这个问题只会变得模棱两可，中间演员可以自己决定。

简单是因为这个问题远非新问题。军事纪律的主要目的是建立一支战斗力，在战斗中不要试图过于自主地思考。在其军队是欧洲体系的后裔的国家，征募和非征募军官的作用是执行计划；委任军官的作用是决定执行哪些计划。因此，从理论上讲，决策责任完全由军官承担，而军官之间根据军衔、指挥区域等明确划分责任范围，决定了谁对任何给定的命令负有最终责任。

虽然在实践中，这往往是相当模糊的，原则是我们已经理解了几千年，而人工智能没有添加任何新的画面。即使在他们最大的决策能力和自主性下，他们仍然会融入到这个讨论中，而我们离他们还有几十年的时间，实际上他们有足够的自主性，可以让谈话甚至开始接近我们为围绕人的讨论所确立的水平。

也许这是人工智能伦理的最后一个重要教训：我们在人工智能方面面临的许多问题，仅仅是我们过去所面临的问题，是由于技术的某些变化而凸显出来的。在我们现有的世界中寻找相似的问题常常是有价值的，以帮助我们理解我们如何对待看似新的问题。

#### 我们从这里去哪里？

我们还可以讨论许多其他问题，其中许多问题对我们这个社会来说是非常紧迫的。但我希望上面的例子和解释能给你一些背景知识，让你了解事情的对与错，以及人工智能系统中的许多道德风险来自何方。

这些很少是新问题；相反，向计算机解释我们的欲望的正式过程——一个没有文化背景或没有能力推断我们不说的话的人的最终情况——迫使我们以通常不习惯的方式明确表达。无论是提前数年做出生死攸关的决定，而不是拖到关键时刻才做出决定，还是对我们的社会现状进行长时间的、认真的审视，并且非常明确地说明我们想要保留哪些部分，想要改变哪些部分，人工智能都会把我们推到舒适之外礼貌小说的领域，进入一个我们必须非常明确地讨论事情的世界。

Every one of these problems existed long before AI; AI just made us talk about them in a new way. That might not be easy, but the honesty it forces on us may be the most valuable gift our new technology can give us.