# 机器学习中的伦理学

在一个阳光不太充足的日子里，我和哈尼·塞吉博士（实际上）坐下来讨论机器学习中的伦理问题。哈尼出生并在伊朗长大，是美国山景研究所的一名研究科学家。在加入谷歌大脑之前，哈尼在西雅图做了两年的研究科学家。她于2015年春季获得南加州大学博士学位。她的论文题目是。



哈尼2010年从伊朗最负盛名的工程学院谢里夫理工大学（Sharif University of Technology）获得电气工程学士和硕士学位后移居美国。

在把我介绍给她可爱的波斯猫米莎后，我们就跳了进去。另外，我和哈尼用波斯语交谈。下面我已经转录和翻译了我们的对话，哈尼审查了准确性。

罗亚·帕克扎德：哈尼，用你自己的话说，机器学习的公平性是什么？

哈尼·塞吉：有很多例子和定义。但让我从一个非常著名的例子开始：刑事司法系统中的算法偏见。在这种情况下，如果我们不成比例地将非裔美国人犯罪的数据输入犯罪预测模型，那么我们的模型的预测当然会偏向黑人社区。这是一个歪斜的样本。另一方面，如果少数群体在我们的数据样本中的代表性不足，例如妇女或变性人，那么，我们可以再次预期，我们的模型会有偏差。

罗亚：有没有解决这个问题的数学方法？

哈尼：某种程度上是的。有三种不同的方法：

1. 模型校正的后处理。这意味着，我们校准分类器参数，使其对敏感特征的所有子组（如种族、性别等）具有相同的接受率。
2. 数据重采样以删除倾斜的样本。但是，由于许多原因，收集更多的数据并不容易，有时会给个人带来问题。
3. 因果推理：我们在因果图中捕捉到不同的路径，从而得到相同的观测数据。这基本上意味着建模可能的因素，例如性别、种族和其他敏感特征，以确保它们的影响被捕获，而不会直接影响结果变量。

罗亚：在决策算法的公平性建模方面有哪些挑战？

哈尼：一个非常重要的问题是缺乏对“公平”的具体定义，我可以告诉你，在公平的定义上，有很多定义，有时研究小组也不在同一页。当你没有一个清晰的定义，那么你如何才能正确地建模呢？

另一个问题是社会科学家和人工智能研究人员之间需要合作。你知道，你不能指望人工智能研究人员自己对公平有一个清晰的认识。我们不仅需要社会科学界的人与我们合作来定义这些词，而且我们还需要将这种合作一直持续到产品研发的最后。

“一个非常重要的问题是缺乏对公平的具体定义。”

但需要注意的是，人工智能研究人员和社会科学家之间的一些合作已经开始。例如，康奈尔大学（Cornell University）的Solon Barocas和加州大学伯克利分校（UC Berkeley）的Moritz Hardt一直在与社会科学家积极合作，研究定义和建模公平的问题。

罗亚：但他们是学者。他们预计不会满足公司生产的最后期限，也不会提供超实际和财务上可行的解决方案。他们可能不会受到我们在大型科技公司所看到的人工智能竞争的压力。你认为我们如何才能鼓励数据科学家和工程师意识到这些问题，因为他们可能面临严峻的截止日期压力？

哈尼：嗯，对他们来说，应该有一些道德和公平方面的培训。

罗亚：如何确保这些培训有效？我自己也参加了一些公司规定的在线45分钟职业道德测验。我承认我点击了所有的答案，直到我得到了正确的答案，没有真正的训练！

哈尼：我明白。这是一个我无法回答的问题，但我也很想知道如何做到这一点。但我知道，关于偏见和公平，只要我们的培训是以某人讲授社会性别或种族偏见的基本知识的形式进行的，这种培训就不太可能有效。当您在没有给出任何解决方案或解决方法的情况下就这些问题向我们发出警告时，该培训如何才能完全有效？

我也认为首先我们应该有一个实用的指导方针。目前，在机器学习的伦理学领域还没有任何全面的指导方针（实际上是可以实施的）。我给你举个例子：差别隐私。这意味着确保您所拥有的数据不会泄露数据所属人员的身份。它有一个明确的数学定义。因此，例如，在Netflix大奖赛上，为了建立一个推荐系统，我们作为研究人员，能够识别出他们的数据集侵犯了差异隐私。但为了公平起见，我们目前没有这个，因此，它不容易被教导、监督和审计。

罗亚：你是伊朗人，你完成了电气工程的学士和硕士学位。你也在这个行业工作。你如何看待伊朗的人工智能研究现状？

哈尼：嗯，我最近在伊朗，有机会和那里的一些人工智能研究人员交谈。有很多优秀的初创公司主要致力于开发与我们在这里使用的一切类似的应用程序。但是在人工智能的研究方面，进展是落后的。机器学习本身还是一个相对较新的领域。造成这种差距的一个原因可能是美国对获取某些研究论文的制裁，更重要的是，对云计算平台的制裁。

罗亚：我相信你比我更清楚，女性在科技行业的代表性不足，女性离开科技行业的比例也很高。我们如何解决这两个问题？

哈尼：关于缺乏代表性，我认为我们应该鼓励早年在STEM领域工作的女孩。谷歌和Twitter目前正在采取一些重大举措。例如，他们邀请年轻女孩进入校园，给她们一些小项目，激励她们把机器学习看作是一个能够解决许多现实世界问题的领域。

拥有强有力的榜样也很重要。有时这种缺乏代表性会导致对妇女缺乏自信。他们可能会因为没有被倾听和智力上的尊重而感到沮丧。最终，他们可能会离开战场。所以，有坚强的女性榜样也很重要。我童年的榜样之一是玛丽·居里。在ML中，我也有很多优秀的榜样，比如Anima Anandkumar和Jennifer Chayes。我从他们坚强的个性和自信中学到了很多。

罗亚：最后一个问题：你担心人工智能的未来吗？

哈尼：我不相信把艾未未想象成魔鬼。我们是老师，我们可以控制我们所教的内容。\*\*\*我相信机器学习有很大的潜力让我们的生活更轻松，从简单的任务，如天气预报和人工智能助理，到自动驾驶汽车。但我确实认为，我们还有很多工作要做，使人工智能达到任何接近人类水平的东西。例如，你可以让你的人工智能回答你的问题，但你还不能指望它完全合理。仍然有许多难以回答的挑战，这是使这个领域最有趣的我！

We wrapped up here. This is the first conversation of my interview series for my newsletter [Humane AI](https://www.royapakzad.co/newsletter/). I will continue talking with both policy and technical experts in the field of ethics of AI in future installments. Tune in to know their opinions about many issues including cybersecurity and AI, private sector approaches in addressing such ethical challenges, Human rights and AI for social good, and much more. To subscribe to the newsletter, click [here](https://royapakzad.us17.list-manage.com/subscribe?u=9138308bb26620c53a0881c20&id=8152ed9f0c).

You can find Dr. Hanie Sedghi on [Twitter](https://twitter.com/HanieSedghi?lang=en) and [LinkedIn](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwjylvP-sePYAhUV6WMKHQCEC08QFggzMAE&url=https://www.linkedin.com/in/hanie-sedghi-71bb2582&usg=AOvVaw0YEx3cTOS3g83S7Bs38Baa).