



普适性人工智能伦理的研究范围与方法

 演讲者 / Vincent C. Müller

【学者介绍】

埃尔朗根-纽伦堡大学 (FAU) 亚历山大·冯·洪堡教席教授, 并担任该校人工智能哲学研究中心 (PAIR) 主任; 埃因霍温理工大学访问教授; 欧洲认知系统协会主席; 人工智能哲学学会主席; 欧洲机器人技术联盟“伦理、法律与经济社会议题”专家组主席。

非常感谢主办方的邀请, 很荣幸大家愿意倾听我的观点。

我对我们能在这些事情上展开合作持乐观态度, 而且我被要求讲得相对宽泛一些。请现场的业内人士多包涵。我觉得, 如果您是这个领域的专家, 可能不会从我的分享中听到太多新鲜内容。但另一方面, 我今天想表达的是, 我对一些事情感到担忧。或许我们可以一起努力解决这些问题, 看看会有什么结果。

下面我先简要介绍一下我接下来要讲的内容。在伦理领域, 就像在其他任何行业一样, 都存在着供给和需求。我觉得这里也存在着不合理之处, 而这与我们没有妥善思考这个领域的范围和方法有关。换句话说, 就是哪些属于这个领域, 哪些不属于, 以及我们该如何开展研究。我会稍微解释一下我的看法以及可能的解决思路。我在这里引入的术语, 也就是标题中的“普世性人工

智能伦理”(universal AI ethics)，我认为这是一个重要的概念，之后我会适时进行解释。

这里有个小小的“健康警告”。你知道，有时候在网上看视频，会看到有人做一些非常危险的事情，比如骑着摩托车在山里狂飙之类的，然后视频里会提醒“请勿在家模仿”。我现在的情况有点类似。意思是说，如果我的学生过来跟我说，“我想做一个关于伦理研究方法的演讲”，我会告诉他。那不是严肃的科学研究，这个题目太大了，难以驾驭。还是做些更具体、更严谨的研究吧。但我已经获得了终身教职，有一份稳定的工作，不会被解雇，所以有时候我可以允许自己做这样的事情。不过对学生来说，这可不是个好主意。

好了，下面开始分析问题。我认为在人工智能伦理领域存在一些问题。我会先从一些负面的情况说起，然后再试着解释如何摆脱这些困境。

正如大家所知，到处都有人在说伦理至关重要。这不仅仅是在我们这样的会议上，在政治会议、各种讨论等场合也都是如此。几年前情况并非如此。我从事这方面的研究已经有 20 年左右了，在大部分时间里，人们都不愿意听这些。这在当时根本不是什么问题。但与此同时，如果你去问一位人工智能伦理专家，这个领域包含什么、主要议题是什么、主要论点有哪些、目前的研究进展如何，他们可能会陷入困境。我不知道你们有没有听说过“冒名顶替综合症(impostor syndrome)”，我就有这种感觉。就是当你身处某个职位，却觉得自己并不真正具备胜任这个职位的资格。所以，当有人说“您是人工智能伦理领域的顶尖专家”，我会想，“这位顶尖专家肯定是别人吧。”当然，从某种意义上说，这种想法很愚蠢，因为如果这个领域真的有专家，我至少应该算是其中之一。但当人们说“您应该知道答案”时，我还是会感到担忧，接下来我会试着解释一下这种担忧。

我认为，这种担忧源于两个相关方面的不一致。一方面是需求。正如大家所知，人工智能发展非常迅速，变化巨大。我认为人工智能正在带来且将要带来的改变是前所未有的。我很高兴看到中国人比欧洲人更能意识到这一点。欧洲人比较保守，觉得目前一切都还不错，应该慢慢来，不要破坏现状等等。显然，这种态度是很危险的。我认为我们正面临着一场重大的变革。这不仅仅是又一项新技术的出现，其影响的规模是不同的，更像是一场工业革命。我觉得这对哲学来说是一个机遇，因为纵观哲学发展史，在社会发生重大变革的时期，哲学往往会有所突破。哲学并不是脱离社会独自发展的，它们之间是相互影响的。你无法想象，如果没有当时的技术变革，会有启蒙运动；如果没有工业革命，会有马克思的思想；如果没有科学革命，会有维特根斯坦的哲学。所以我认为现在也会出现类似的情况。

我很喜欢亨利基·辛格的一句话。他当然不是哲学家，而是一位政治家，在某些方面还存在争议，但无论如何，他说过“潜在的主导技术正在寻找一种指导性的哲学”。我之所以喜欢这句话，

显然是因为它听起来好像政治家们需要哲学。那我可以说，“这不就是我嘛”。或许你们也需要类似的东西。我觉得这里面有一些有趣的现象值得我们去探究。

我们来看看供给方面的情况。我认为我们是被这种需求驱动的。2025 年的伦理研究与 2015 年的不同之处在于，现在的需求大得多，而且这是一个新兴的领域。早在 21 世纪初，就已经有人开始撰写关于人工智能伦理的文章了。2012 年，Nick Bostrom 和我组织了首届人工智能伦理会议，从哲学的角度来看，这并不是很久以前的事，相当近期。我在 2020 年发表了第一篇关于人工智能伦理的综述文章，差不多在同一时间，还有其他一些人也发表了相关文章。可以说，2020 年左右，这个领域才成为一个相对成熟的学术学科，我们可以明确其基本议题，人们可以以此为基础进行学习，教师也可以开设相关课程，讲解基本内容。所以说，这个领域还相当新。

第三个问题是我现在经常遇到的。好像每个人都能自称是人工智能伦理专家。不知道你们有没有注意到，无论你走到哪里，都会有很多人说“人工智能伦理啊，我懂”。很多来自不同领域的人都认为自己在这个领域有专长，这在某种程度上与他们各自不同的专业背景有关。

如果你来自人工智能这个技术领域，你看待问题的方式会和我大相径庭。如果你来自政府政策部门，你看待问题时会想，“我要制定政策，该如何制定呢？”这与人工智能本身所存在的问题完全不同，所以，什么算是问题，取决于你来自哪个领域；什么算是重要的问题，也取决于你来自哪个领域。举个例子，我们都知道人工智能系统会学习到偏见，其中之一就是从人类身上学到种族主义。因为人类存在种族主义倾向，所以人工智能系统也会学到。这当然会引发各种各样的问题，人们也写了很多关于这方面的文章，引起了广泛关注。我曾经想过，为什么会这样呢？原因当然是，这个领域的很多研究者都来自美国，而在美国，种族主义是一个重大的社会问题。但在其他地方，虽然在某种程度上也存在种族主义问题，但并没有那么严重。所以，什么东西会被视为问题，很大程度上取决于你所处的环境。

还有一些人对“奇点”（即存在性风险）很感兴趣。看看这些人，他们大多来自非常富裕、非常安全的环境。如果你坐在牛津大学一所富有的学院里，可能会觉得“奇点”或许是个问题。但如果你在尼日利亚，可能会有不同的看法，因为那里的人们连饭都吃不饱，你会觉得那比“奇点”更重要。因此，我认为我们需要思考一下人工智能伦理的范围和研究方法。

现在，我来说说积极的方面。

我认为，人工智能伦理不同于其他应用伦理学。如果你去看看传统的哲学系，会发现哲学有很多分支，比如形而上学、认识论，而伦理学是其中之一。还有一些人研究应用伦理学，即针对特定领域的伦理问题，比如医学伦理学、生命伦理学、战争伦理学等等。所以，当人们听说有人在研究人工智能伦理时，会说“你在做应用伦理学研究啊。因为你在针对一个新的、前所未有的

领域做应用伦理研究。这通常被称为“领域伦理学”，即针对人类不同活动领域的伦理学。但我认为，这样看待人工智能伦理并不恰当。我以前也这么认为，但现在觉得自己错了，因为我发现人工智能伦理并没有明确的边界。

研究医学伦理学，其范围相对容易界定，有医疗活动和非医疗活动之分，大致可以划出一个范围。但人工智能伦理则不同。我认为原因如下：首先，人工智能，顾名思义，与智能有关。而智能本身并不局限于特定领域，它是人类普遍具有的特征。你可以说某人在数学或其他领域更聪明，但原则上，智能涉及方方面面。计算也有类似的特性，原则上具有普遍性。正如“丘奇-图灵论题”所说，原则上，能被某台计算机计算的东西，也能被任何一台计算机计算。我认为正是这两个特性导致人工智能伦理没有明确的边界。

显然，人工智能伦理与人工智能技术有关，但这种技术可能涉及任何领域，这也是我们难以把握它的原因之一。其他技术，比如汽车制造，都有其特定的应用范围，你能清楚地知道什么属于该领域，什么不属于。但人工智能似乎渗透到了各个领域。因此，我认为在人工智能伦理研究中，我们需要考虑到这一特点。我们无法将人工智能伦理装进一个精致的小盒子里，虽然那样会很方便，但我觉得这是不可能的。它更像是一个巨大的盒子，原则上可能包含一切。所以，我认为这是一个具有普遍性的领域。相应地，我们需要一种能够理解一切的伦理学。

现在，你们可能明白为什么当有人说我是人工智能伦理专家时，我会有“冒名顶替综合症（impostor syndrome）”了。因为这意味着我需要成为伦理学、人工智能以及其他所有领域的专家，这显然是不可能的。所以，我显然需要与其他专家交流合作。比如，我需要和人工智能领域的专家交流，也需要和社会科学、人文科学领域的专家交流。有些专家研究人工智能如何影响政府和民主制度，我对此知之甚少，而研究这方面的政治学家则是专家，所以如果我想研究这个领域，就应该和他们合作。如果我想研究医学领域的人工智能伦理，就应该和医生合作，以此类推。

因此，我们最终会拥有不同类型的专家，并且这个领域也会出现专业化分工，这一趋势已经显现。现在有人会说，“我研究这方面的人工智能伦理，不研究那方面的。”我认为这是一个新现象。

我们还需要摒弃哲学研究中一些常见的观念，这更多是针对哲学家而言的。在哲学中，我们经常区分规范性问题和伦理问题。我们认为，规范性问题涉及事物“应该如何”，而伦理问题则涉及“应该如何做才符合伦理”。还有一些规范性问题并不属于伦理范畴。例如，“应该如何饮食”是一个规范性问题，但不属于伦理问题。吃饭时把食物洒得满地都是，这不雅观，但并非不道德。但我发现在人工智能伦理研究中，这种区分并不是特别有用。

如果你和工程师交流，会发现他们想成为优秀的工程师。优秀的工程师意味着能建造出不会倒塌的桥梁，能告诉你这座桥能承受 20 吨重的卡车，能使用多少年，以及需要如何维护等等。

他们是这方面的专家，这很好。优秀的工程师能告诉你这些特性，或许还会说“这座桥在没有地震的情况下是安全的”，或者“它不会污染环境”，以及其他一些你可能关心的更广泛的特性。这些虽然不是传统意义上的伦理问题，但我认为对我们来说很重要，因此应该纳入人工智能伦理的研究范围。

我还认为，在伦理研究中，我们不能只谈论风险。目前，人们普遍认为，我们可以从两个方面来思考我们的义务：一是产生积极的结果，二是以正确的方式产生积极的结果。例如，我可能通过对你撒谎来达到某个积极的目的，我相信你们实际上也这么做过，比如出于礼貌的谎言。当别人问“你觉得我的裙子怎么样”，你可能会说“太漂亮了”，你这样做是为了产生积极的结果，让对方开心，但你使用的方式是撒谎。因此，我们在判断时会兼顾这两方面：一方面看结果，或许会认为这样做没问题；但另一方面，会觉得你使用的方式并不恰当，如果能不撒谎就达到目的，那就更好了。所以，我认为这两个方面都很重要。风险只是其中之一，但不能涵盖所有内容。

下面我来建议一下我们可以如何开展研究。研究伦理问题时，我们会发现有些问题反复出现。我们可以找出这些问题是什么。例如，偏见问题、隐私问题、人类自主性问题等等。这个列表还可以继续延长。然后，我们可以看看人工智能在各个领域的应用，比如自动驾驶、人脸识别、人工智能助手等等。这样就形成了一个矩阵，你可以看到在某个技术领域存在哪些伦理问题。如果有人对某个特定技术领域感兴趣，就可以纵向研究，看看该领域存在哪些问题。虽然并非所有领域都存在所有问题，但至少能有一个大致的框架。

当然，如果你是一名伦理研究者，你可能会说“我是研究隐私问题的专家”。而隐私问题存在于各种场景中。例如，自动驾驶就涉及隐私问题，因为它会收集数据。或许这就是一种有条理地看待问题的方式。

英语中有句俗语：“切蛋糕的方法有很多种。”一块蛋糕，你可以这样切，也可以那样切，可以切成很多小块，也可以切成少数大块。我在这里的建议是，按照伦理问题来划分，也就是“切蛋糕”的方式是按伦理问题来。我也可以这样写这篇文章：先讲自动驾驶中的伦理问题，再讲人工智能助手中的伦理问题等等，这是另一种“切蛋糕”的方式，但蛋糕还是同一个。

对此我还有一点补充。正如我一开始所说，哲学的进步往往在社会发生巨大变革的时期和领域。我一直在想，或许现在就是这样一个时期。这让我困惑了一段时间，因为人们认为像我这样的人，口袋里揣着一套伦理知识。不，我用伦理知识去研究问题，而伦理理论本身是会发生变化的。我并非把一套固定的理论装在口袋里，到处套用去解决问题。实际情况是，当我运用理论时，理论也会随之改变。工具在使用过程中会发生变化，或许我可以借此改进工具。或许可以把这看作一个循环过程。比如说，我提出一个理论。例如，很多人（不是我）对意识感兴趣，他们可能

会说“这是我的意识理论”，然后用这个理论去研究某个事物，看看是否适用。然后他们可能会说，“这个系统可能具有意识，但出于其他原因，我不想这么说”。那么，或许他们就需要回过头去修改理论，这样就能更好地进行哲学研究。我认为这其中是有希望的。我觉得我们不仅能为人工智能领域做出贡献，也能为哲学本身做出贡献，这可能是一件好事。

以上就是我想解释的，为什么我认为人工智能伦理领域存在“供给”与“需求”的不一致，以及或许可以通过明确其普世范围和相应的研究方法来解决这些问题，希望能让大家看到这个领域取得进展的希望。

谢谢大家。