【人工智能伦理专题】

从碳基伦理到硅基伦理

——人工智能时代的伦理学浅论

蓝　江*

〔摘要〕 人工智能技术的迅速发展及其在日常生活中的广泛应用，日益向传统伦理学提出了新的问题，即是否有可能超越人类中心的视角，重构人工智能时代的伦理学。事实上，从机器学习到深度学习，人工智能的实现依赖于大数据的运行，而大数据需要创造出一种与人类的生命环世界平行的数据环世界。与生命环世界中将复杂的情境还原为人与其他存在物的关系不同，在数据环世界中，所有的情境都被还原为良序的数据交换。因此，存在着一种不同于以人为中心的碳基伦理学的硅基伦理学，硅基伦理学的基础是数据，而数据和信息交换的善的标准则在于尽可能在系统中降低交换汇总的不确定性。

〔关键词〕 人工智能　碳基伦理　硅基伦理　伦理学

〔中图分类号〕B82－057　〔文献标识码〕A　〔文章编号〕1007－1539(2020)05－0036－09

DOI:10.13904/j.cnki.1007-1539.2020.05.005

当看美剧《西部世界》时，我们会看到这样的剧情，一个平常道貌岸然的绅士，在进入到《西部世界》的机器人乐园时，会对那些和人类长相无异的机器人做出最残忍的事情，人类最恶劣的本质在其中体现得淋漓尽致。实际上，对于熟悉伦理学理论的人来说，人类本身的这种兽性并不是不存在的，只是在进入文明社会之后，由于现代伦理学和道德哲学的制约，人与人的平等和互动的人类学机制将这种恶劣的兽性本能压抑了下来，这种伦理学机制使人类在面对同类（甚至包括猫、狗、兔子等亚同类）主体的时候，不会作出极端丑恶的行为。但是《西部世界》这样的影片带给我们的思考是，一旦 AI 机器人不被视为与我们具有同等地位的伦理主体，我们是否有足够的权利对它们施暴。《西部世界》尽管是一个极端的例子，但一些更为显著的例子出现在我们生活的周围，例如在一些互联网游戏中，游戏角色可以对 NPC（游戏中的非玩家角色）施暴，甚至虐杀，这种针对 NPC 的暴力是否应当被视为伦理问题？

问题也可以从相反的角度来看。在针对ISIS的军事行动中，美军曾经使用 MQ－9 军用无人机对恐怖分子进行定点清除。MQ－9 无人机当然是一种武器，是一种具有 AI 功能的新型武器，在一定意义上，我们不能将无人机视为传统意义上的阿帕奇武装直升机或 F－35 这样的战斗机，因为在战场上，MQ－9 的航行和射击指令完全由人工智能掌控，人类指挥官只给无人机提供对象的地点，剩下的所有决策都是由 AI 做出的。但是，如果 AI 做出的判断，在该地点射杀

＊ 作者简介：蓝　江，南京大学哲学系教授、博士生导师；南京大学马克思主义社会理论研究中心研究员，江苏省青年社科英才（江苏南京　210046）。

基金项目：国家社会科学基金重大项目（18ZDA017）。

的不是恐怖分子而是平民，谁应该承担伦理上的责任？另外，AI为了完成任务，做出的破坏行为，如为了射杀恐怖分子，破坏里面仍然有居民的民房，在伦理上能否被人们接受？即便不在战争上，AI作为伦理主体的问题仍然存在，例如在Facebook和Twitter等应用软件上，都有AI对用户的账号进行监控的情况，AI会判断一个用户的发言是否违反规定，在极端情况下，会将某个用户判定为机器人用户，并封掉这个账号。倘若这个账号的真实用户是一个真正的人类用户，那么AI账号封掉并未真正违反规定的人类账户的做法是否正当？

所有这些问题给伦理学提出的一个根本问题是，传统意义上建立在人类中心主义或者仅仅将人类作为基本行为者的伦理学正在受到巨大的挑战，这种挑战是由于另一种智慧行为体的出现造成的，随着机器学习、深度学习、云计算、边缘计算等新技术的发展，随着5G等通信技术日益成为新信息时代的基础设施，我们越来越感到AI正在从科幻小说中的对象逐渐成为我们生活中与我们产生行为互动的对象，面对这样的对象，我们与它们之间的交往和互动仍然需要一定的行为法则的约束，但是这种行为法则已经不能在纯粹的人与人之间的伦理框架下来思考，这势必意味着我们需要思考一种面向人工智能时代的伦理学，而这种伦理学最基本的内涵已经不再是人类中心或人与人之间的互动模式，AI的活动已经成为这个巨大的社会关系网络中不可或缺的一环。因此，思考人工智能时代的伦理学或道德哲学，已经成为今天学者们必须面对的问题。

一、人工智能时代的电车问题

我们可以从一个经典的伦理问题开始讨论，这就是电车问题。

在伦理学中，电车问题已经广为人知。迄今为止，电车问题已经有了诸多变种，但电车问题的原初版本起源于菲利帕·富特（Philippa Foot）早期的一篇论文《堕胎问题和双重效果论》（*The Problem of Abortion and the Doctrine of the Double Effect*）①。富特之所以以电车问题来举例，其目的就是为了解释在堕胎中出现的伦理难题。实际上，富特举出了很多类似的难题，如医生能否为了挽救另外五个人来杀死一个身患绝症的病人，移植他的器官。一个失控的飞行员是选择撞向更多的人还是撞向更少的人等，但电车问题无疑成为富特讨论中最著名的案例，我们看看富特自己的假设：

> 尽可能举一个类似的例子，这个例子也是假设的：一个运行中的有轨电车的司机，能做的事情就是将电车从一个轨道导向另一个轨道，有一条轨道上有五个正在作业的工人，而另一条轨道上只有一个人，这辆电车行驶上的轨道上的人会必然死亡。[1](23)

富特之所以提出电车问题，正是为了解释她所提出的“双重效果论”，用富特的话来说，“双重效果论基于一种区别，即一个人可以预见其自愿行为的结果与这个人在严格意义上的意图之间的区别”[1](20)。在电车案例中，这种双重效果论体现为一个矛盾：一方面是为了拯救一条铁轨上的五个人的动机，另一方面是实现该动机的手段是将电车导向另一段铁轨，放任那段铁轨上的另一个人的死亡。这样，拯救五个人的生命是以放任另一个人的死亡为代价的。在“器官移植”案例中，同样是牺牲了一个人的生命（尽管这个被牺牲的人的生命已经所剩无几），拯救了另外五个人的生命。在富特的案例中，存在着一个做出伦理判断的主体，在电车案例中是电车司机，在器官移植案例中是医生。在轨道上的六个人只是作为需要做出伦理判断的电车司机的条件，这是一个选言型结构：要么电车司机不改变电车行动的方向，任由五个人死亡；要么改变电车方向，牺牲另一条轨道上的一个人。对此，富特给出的解释是：“从人类的伦理直觉角度，为了保全更多的生

① 菲利帕·富特的这篇论文最早发表于1967年11月5日的《牛津评论》（*Oxford Review*）上，是现在被视为最早提出电车问题的论文。

命而实施无害且必要的手段来导致另一个人的死亡，且我们可以预知另一个人的死亡在该手段下是必然的，这种做法是错误的。”[1](27) 富特的答案并不是一种功利主义的回答，而是一种直觉主义的回答，也就是说，富特认为在放任人死亡和主动导致人死亡之间存在着区别①。这里虽然人的数字很重要，在富特自己的案例中是 5 人对 1 人，但如果电车游戏变成了数人数的游戏就会导致伦理上的悲剧。假如说，一个铁轨上是 31 人，另一个铁轨上是 30 人，我们是否还会支持从 31 人的铁轨转向 30 人的铁轨。因此，在富特看来，伦理的主要原则并不是数量上的比较，而是在放任故意和主动故意之间的区别。不过，在后来的许多讨论电车问题的版本中，更多的人会倾向于选择功利主义式的计算，即将电车导向人数更少的一边，从而保存更多人的性命。为了在富特的直觉主义和功利主义之间作出区别，后来的茱蒂丝·J. 汤姆森(Judith J. Thomson)给出了电车问题的另一个版本，即“旁观者版本”。

富特的电车问题的原版有一个很大的问题，就是驾驶电车的司机实际上始终存在着责任，无论他是走向五个人的轨道，还是走向一个人的轨道，都无法免责。也就是说，在后果评价体系中，很有可能会以他导致死亡的人数来对他的责任进行追究，在这种情况下，电车司机选择更少人数的主动故意便是合情合理的。或许也正是出于后果论的考虑，更多的人选择了功利主义的方案，即选择主动伤害另一条轨道上的一个人比伤害原先轨道上的五个人在伦理原则上更为优先。倘若如此，富特在放任故意和主动故意之间作出的区别，或者说她用来针对讨论堕胎问题的双重效果论实际上是失败的，原因在于，后果上的数字比较要比动机考察更为优先。为了摆脱这种困境，茱蒂丝·J. 汤姆森给出了电车问题的一个变种：

> 这个案例我可以称之为旁观者三项选择。旁观者可以用两种方式来扳动开关。假如扳到右边，电车将会转到右侧岔道，杀死一位工人。扳到左边，电车会转向左侧岔道，而旁观者自己就站在左边，会把他自己撞死。或者，他什么也不做，任由那五个工人死亡。[2](364)

在汤姆森的版本中，大家更关心的是她给出了一个自我牺牲选项。但实际上，更值得关注的内容并不是旁观者的自我牺牲，即向左扳动开关，而是作为伦理判断主体的旁观者与富特案例中的电车司机有着完全不同的伦理境遇。在富特的案例中，无论电车司机如何选择，他都有一定的伦理责任，对于这个责任，需要从结果来做评判。但在汤姆森的案例中，情况发生了微妙的变化，旁观者并不是直接作为伦理主体出现的，也就是说，唯有当旁观者触动了扳道的开关，他才能成为伦理主体。简言之，旁观者的放任行为是可以在伦理上免责的，这是他与电车司机最大的不同。在这种情况下，旁观者的放任故意(不去触动扳道开关)和主动故意(触动开关，或让电车撞向那一位工人，或作出自我牺牲)之间便出现了巨大的差别。在结果论看来，旁观者的放任故意是没有责任的，但他的主动故意会导致一个人的死亡(无论这个人是一位工人还是他自己)。在不考察汤姆森的自我牺牲的案例的情况下，实际上，在旁观者情形中，放任故意比主动故意更为可取，在这个讨论中，人数的差别进一步被边缘化了，因为对于旁观者来说，评价他的行动的结果不是纯粹以死亡的人数作为标准的。

电车问题已经成为今天最为经典的伦理问题，在这个案例中，直觉主义、动机论、结果论、规范伦理学、情境伦理学都可以给出不同的答案，但案例中的孰是孰非、熟善熟恶并不是伦理学关注的重点，而是电车问题是否存在着一个最终的根本原则，这个原则能够成为绝对的衡量标准。现代伦理学的发展告诉我们，这个最终原则是不

① 实际上，器官移植案例更能体现富特所谓的“放任故意”和“主动故意”的双重效果论之间的矛盾，医生不去救另外五个人，是放任故意，但杀掉一个人来用他的器官救活另外五个人则是主动故意，而主动故意造成的死亡，在富特的伦理学中，是不被允许的。

存在的，正如谢利·卡根所说："假设我们已经找到了最终原则Q，这就结束了吗？我们已经解开电车难题了吗？远非如此！因为很可能的情况是，一旦原则Q所依据的那些观念和差异特征被识别出来，我们却并不准备接受它们。"[3](198)实际上，在卡根看来，电车问题的伦理学问题更多的是偶然性，而不是绝对的价值性判断。

不过，在今天，我们可以给出电车问题的另一个变种，即一种人工智能的变种。这个变种依赖于茱蒂丝·J. 汤姆森的旁观者情形：

> 在今天的人工智能时代，电车遇到的岔路口是一个由人工智能主控的扳道开关。电车是无人驾驶的，里面没有司机。而人工智能的开关只能决定是往左开还是往右开。其他的条件不变，即一条道上有五名工人，而另一条道上有一名工人。而这六名工人的性命取决于人工智能扳道开关。

之所以给出电车难题的人工智能变种，正是因为我们之前提出的伦理判断的主体全部是人类，无论是电车司机还是一个旁观者，他们都具有伦理判断的主体性，也同时承担着这种伦理判断的义务和责任。现在的问题是，这个人工智能的扳道开关是否具有伦理判断的主体性，这个扳道开关本身没有任何能力让失控的电车停下来，它唯一能做的是一个二选一的选择，同时扳道开关根据数据收集，已经知道两侧的铁轨上的人员情况，在这种情况下，人工智能会做出如何判断？

事实上，用电车问题作为案例，并不是要假定一个极端情况作出虚拟的推演，在不久的将来，甚至可以说在今天，人工智能的伦理问题正在日益成为我们生活中的现实问题。例如，在人工智能的无人驾驶中，一个突然出现的横穿马路的行人让汽车躲闪不及，从而撞向了路边的小店，导致小店营业员的死亡；或者撞向了桥梁，使里面的乘客溺水而亡。对于人工智能的应用，如无人驾驶、无人机、智能家居等，我们不能只假设便利与和谐的情况，在理论上，我们必须假设一些特殊的失控的情况，来研究和探讨人工智能的伦理。我们在这里的假设并不首先是人工智能存在对人类的恶意，而是在同时涉及人类生命的案例中，人工智能会如何抉择，它们的伦理思考会与人类完全一致吗？我们是否可以直接将富特和汤姆森关于电车问题讨论的结论直接移植到人工智能的案例中来？

二、从机器学习到深度学习：超越人类中心主义

对于这样的问题，我们显然不能仅仅从现在的以人类为中心的伦理学做出回答。在回答这样的问题之前，我们首先需要了解人工智能究竟是如何来进行思考和决策的。

1997年IBM的电脑"深蓝"与当时的国际象棋冠军加里·卡斯帕罗夫进行了一场举世瞩目的对弈，在六局比赛中，"深蓝"以3胜2负1平的战绩战胜了棋王卡斯帕罗夫。"深蓝"的胜利代表着计算机开始在某个特定的领域中超越了人类。但是，"深蓝"并不是严格意义上的人工智能，"深蓝"之所以取胜，并不是因为它拥有了和人类一样的思维能力，而是因为软件工程师给"深蓝"提供了大量的棋谱和大量真实的历史对弈的记录，并推演了多种变化的可能性。也就是说，"深蓝"的胜利仅仅只能视为机器存储能力和计算能力的胜利，"深蓝"在储存的棋谱和对弈数据中提取每一步相关的数据，计算棋盘上各种变化的可能性，这种快速调用数据和计算推演的能力，是包括卡斯帕罗夫在内的真实的人类所不具备的，在这个意义上，卡斯帕罗夫输给了计算的存储容量和计算速度，而不是输给了人工智能。

后来由谷歌公司开发的人工智能AlphaGo，完全不是这样的情形。在战胜李世石和柯洁的对弈中，AlphaGo已经具有了一种特殊的人工智能的能力，即机器学习。在"深蓝"时代，机器对弈建立在人类对弈的经验基础上，比如说，在输入了大量人类真正对弈的棋谱上，"深蓝"的下棋步骤仍然是人类可以理解的，因为它所下的每一步虽然带有计算机快速计算的痕迹，但是它所走出的步骤，并没有真正脱离人类的思维。而AlphaGo并不是在既定的人类对弈经验上来进行推演和决策的，也就是说，AlphaGo是一个从零基础开始培育的学习的人工智能，它通过自己

的数据提取和积累，完成了属于 AlphaGo 自己的推理逻辑。在与柯洁的对弈中，AlphaGo 显然下出了一些根本不可能是人类棋手下出的步骤，即便柯洁在之后多次复盘，也无法理解 AlphaGo 的某几个步骤是如何思考、如何落子的。于是，我们遇到了今天在人工智能领域中十分重要的概念：机器学习。

实际上，机器学习从一开始就不是以模仿人类思维来设定的。1982 年，约翰·霍普金斯大学的年轻科学家谢伊诺斯基（Sejnowski）和加州大学圣迭戈分校的研究生杰弗里·辛顿（Geoffrey Hinton）在一场夏季学术研讨会上相遇了，他们关心的是一种“神经—激活”的工程。谢伊诺斯基原先的专业是物理学，而他现在已经将主要的精力放在神经生物学上，而辛顿在加州大学所学习的是认识心理学。这次会面让两位未来在人工智能领域声名显赫的科学家感觉到传统认知科学和计算机领域中的陈旧，他们希望通过人工的“神经元”的联网互动实现一种全新的智能方式。谢伊诺斯基和辛顿合作，开发了一个命名为“Nettalk”的程序网络，这是一种初级的“人工神经网络”，它采用了 300 个被他们称之为“神经元”的模拟电路来捕捉单词，生成语音输出，这 300 个神经元具有 18000 个人工“突触”[4](147)。在神经生物学中，突触是每一个神经元与其他神经元相互链接的结构，它们通过一个生物电脉冲将一定的信息与另一个神经元的信息链接起来。而谢伊诺斯基和辛顿的“人工神经网络”的“突触”实际上是一个数据交换系统，每一个“突触”都从周围环境中提出数据，并在人工神经网络中进行信息合成和分析，形成属于“人工神经网络”自己的数据，其进行数据分析的基础也是这些数据。

什么是机器学习？按照比较正式的定义，机器学习是“让计算机修改或适应它们的行为（无论这些行为是做出决策，还是控制一个机器人），这样，它们的行为会越来越精确，而行为的精确性是这样来测定的，即既定行为能否做出正确的反映”[5](4)。譬如，我们与一个具有机器学习的算法的计算机一起下棋。一开始，这台计算机没有任何数据，我们可以很轻松地在对弈中取胜，在多盘对弈之后，计算机的水平基本上和玩家旗鼓相当了，到达一定的盘数之后，我们已经基本上无法战胜计算机了。事实上，在每一个下棋过程中，计算机都在进行机器学习；在每一盘对弈中，计算机都提取了数据，并在数据的基础上给予最优化分析，最后做出相应的决策和行为。在达到足够多次的对弈之后，计算机的对弈行为已经达到了很高的精确值，我们作为棋手，已经无法与计算机的推演相提并论。这就是一个典型的机器学习过程。同样的案例还有，一些从事机器学习的研究者给一台计算机不断展示猫的图片，然后让计算机自己来理解什么是猫。在刚开始的时候，计算机给出的抽象非常不精确，但在给计算机展示了上千张猫的图片之后，计算机就可以根据自己的理解，描绘出一只它抽象出来的猫的形象，这个形象可能没有真实的猫与之对应，但是在我们人类看来，这种猫的形象已经与我们理解的猫几乎没有太大差别。

这种机器学习与早期的图灵机式的 IF—THEN 结构有着很大的区别，IF—THEN 是条件式命令，是一种从上到下的规定形式，在以往计算机的程序设计中，程序设计者占据的是一个至高地位，需要洞悉全局，考虑到所有条件的可能性与对应的结果。之前的电车问题就是一个典型的 IF—THEN 结构，所有对电车问题的伦理讨论，无论是最开始的富特，还是汤姆森的变形，以及迈克尔·桑德尔在公开课上对电车问题的分析，都意味着电车问题的讨论者站在一个至高的地位上，俯视着大地上正在发生的电车事故。然而，真正的伦理问题不应该是这个角度，实际上伦理行为者和伦理评判者处在同一个层面上。这样，当我们面对一个人工智能时代的电车问题时，实际上考察的角度会完全不同于那种处于高高在上的上帝视角的伦理评判者的地位。

机器学习的本质是基层的数据互联，并在这些基层人工神经元基础上提取数据，加以抽象，上升一个层次，在抽象的层次上，会进一步抽象，得出更高的层次。比如从猫的图片中，机器学习可以抽象出智能对猫的理解，同时在更高的层次

上，机器可以尝试着将生物与静态的环境分离开来，这是一种梯度的结构，在谢伊诺斯基的定义中，这是一种深度学习（deeping learning）。浅层的机器学习帮助从基层神经元数据中提取了抽象概念，最终，这些进一步抽象出来的概念服务于一个更高层次，用于更复杂的操作。譬如，在无人驾驶的情境中，完成驾驶任务的智能不仅仅需要辨析周遭环境中哪些是人，哪些是动物，哪些是路标，在更高的层次上，智能需要将生命体、活动的物体与静态的环境分离出来，将地面上的路标、指示灯、路牌与一般性的文字分离出来，分离出来的这些数据关系会进一步用于处理更高层次的内容，如看到左转的地面路标和指示牌要准备左转，看到红灯需要停止下来，看到人行道上的行人需要停止下来。这些具体命令的执行，建立在浅层的机器学习的纵深的深度，深度学习表达出类似于多层次的从下至上的结构，它所执行的命令不是来自某个上帝的命令，更不是某个绝对的伦理法则，而是来自从最底层的数据收集上不断在深度上抽象而来的学习过程。

我们在这里需要注意一个问题，人工智能的深度学习建立在大量数据采集的基础上，也就是说，所有机器学习的基础是数据。正如谢伊诺斯基指出的："让深度学习得以腾飞的是大数据。就在不久之前，收集太字节（terabyte，即兆兆字节）还需要一整排计算机。如今，我们完全可以在一个记忆棒中存储太字节的数据。互联网中心如今可以储存拍字节（petatype，1000太字节）……如果没有互联网提供的大量的图像和其他标记数据，我们几乎不可能训练出能进行真正深度学习的网络。"[6](164) 至此，我们可以针对从机器学习到深度学习的人工智能在伦理学上的意义给出以下一些概括。

第一，机器学习、深度学习或人工神经网络的人工智能基础是大数据，每一个层次上的学习都是以海量级的数据抽象为基础的，这是一种自下而上的认知学习结构。可以说，在机器学习之前，没有任何先入为主的概念来统领机器的认识，它们的体系是建立在对海量级数据的深度学习的基础上的。这样，在电车问题的案例中，决定人工智能扳道开关是否起作用的，并不是某种从上至下的命令，而是人工智能建立在海量数据基础上的判断。

第二，无论AlphaGo的对弈还是无人驾驶和无人机技术的实现，机器学习的人工神经网络都已经远远超越了人类思维的轨道。我们不能用我们自己的知识与判断来理解人工智能的决定，正如从柯洁和李世石的角度，无法理解AlphaGo的每一步棋招一样。在电车问题的案例中，至关重要的问题是，对于人工智能来说，不存在直觉主义，也不存在后果论，更不存在先验性的道德规范来约束人工智能的判断。在这种情形下，人工智能的判断并不是既往人类伦理思维所能把握的对象。

第三，对于具体情境，人工智能的理解只能建立在数据的基础上，简言之，任何复杂的情境只有还原为数据，对于人工智能才具有意义，即便是伦理学问题也是如此。在无人驾驶的案例中，人工智能看到的不是具体的场景，而是被数据化的图像界面，人工智能的作用是在这个图像界面上进行布尔代数式的分析：根据一系列的函数，进行分类甄别分析，最后做出判断。

显然，对于电车案例中的伦理问题也是如此，轨道上的五个人和另一轨道上的一个人，以及整个铁轨的路线图、整个铁轨系统上的所有列车的运行情况、失控列车本身的状况等是人工智能面对的对象，在这里，很有可能出现一种情形，即在人工智能的运算中，轨道上的人（无论是五个人还是一个人）不是人工智能首要考虑的对象，而具体存在的人的情况，恰恰是伦理学首要考虑的问题。为什么会如此呢？人工智能的机器学习不会首先面对一个伦理问题，因为在复杂的系统中，伦理体系只是所有体系的一部分。在具体案例中，铁轨上的人工智能扳道开关首先考察的是整个铁轨运行图和列车运行状况的处置，而不是进行伦理抉择。在收集的所有信息面前，扳道开关首先面对的是如何不影响整个铁道系统的运行，而不是考虑哪边铁轨上的人更少。

因此，我们在这里可以梳理一下常规伦理学与人工智能视阈中电车问题的差异。无论是伦

理学还是人工智能，都对复杂的情境做出了还原式简化，因为无论是做出伦理判断的人，还是作为人工神经网络的智能，都不可能处理一个环境中的所有信息，其结果是，伦理学将所有复杂情境还原为抽象的人（在富特的案例中，重点也是不同轨道上人的数量以及做出判断的主体的动机），而在人工智能的情况下，复杂情形被还原为对应数据，在这些数据的基础上的机器学习抽象地决定了人工智能的判断基础。因此，人的问题（无论是结果还是动机）均不是人工智能面对的主要问题。从这些分析我们可以看出，传统的伦理学在人工智能时代遇到了一个巨大的瓶颈，那种从上至下的道德律的模式，那种以人类为中心进行动机考察和结果算计的模式，在人工智能时代或许会遭遇前所未有的冲击。也正是因为如此，我们是否需要建立一种新的伦理学，一种超越传统的人类中心主义，将人的价值放在绝对不可动摇地位上的伦理学？如果存在一种人工智能时代的伦理学，它会是什么样的呢？

三、数据环世界下的硅基伦理学

尽管在整个伦理学发展史上，伦理学常常被界定为关于善与恶的学问，但这种善恶之学在绝大多数时候指向的是人的行为本身，即一种实践性的学问。尽管在不同的时代，人们对伦理学的认识千差万别，但总体来说，伦理学离不开人与人之间的关系，它总是试图讨论什么样的行为是正当的，什么样的关系是正义的，什么样的品质是善的这样一些问题。正如雅克·蒂洛在他的《伦理学与生活》一书中十分明确地指出的："道德基本上是讨论人的问题，讨论人同其他存在物（包括人和非人）的关系如何。道德讨论人如何对待其他存在物，以促进共同的福利、发展、创造性和价值，力求扬善抑恶、扶正祛邪。"[7](28) 尽管在学科范畴上伦理学已经发展出生态伦理学、动物伦理学甚至基因伦理学等分支，但这些分支在根本上并没有真正取代雅克·蒂洛关于伦理学的定义，即在这些伦理学范畴中，我们看到的还是人与自然、人与动物、人与基因等的关系。这样，人的关系始终在伦理学中占据着十分中心的地位，这种以人类为中心的伦理学是衡量各种伦理现象的尺度，我们究竟如何才能超越人类中心主义的方式来设想人工智能时代的伦理学呢？

在现实生活中，伦理学的案例无疑是按照人的标准来进行，也就是说，无论多么复杂的情形，我们只需要将其还原为一种人的分析。在科尔伯格的著名的海因茨偷药案例中，海因茨的两难问题实际上是他愿意为拯救他妻子的生命去偷药，还是尊重法律的要求，放任其妻子的死亡。这里的冲突其实不是伦理的内部冲突，而是一种抽象的法律和人的生命的冲突。在伦理学限度内，人的生命的拯救显然是优先于对抽象法律的恪守的。经过这样的还原，伦理学的法则是并不难掌握的。但是在富特给出的移植器官的案例中，刻意剥夺一个病人的生命来拯救另外五个人的生命是不被允许的，因为这类情境被还原为另一种伦理原则，即作出行为的伦理主体的动机是否存在故意，不能故意去伤害一个人的生命成为伦理判断的标准。

但是，人工智能时代的情况有所不同，我们看到的往往不是一个具体生活中的案例，新的案例可能出现在网络平台中。比如说，在网上购物的时候，与我进行交流的是一个人工智能的机器人，我在网络平台上对这个对话机器人以最侮辱性的言辞进行辱骂攻击是否违背伦理原则？如果对方是一个真正的人，在伦理学上，对对方进行辱骂显然是不允许的。但对方并不是人，而是一个智能对话机器人，我们是否可以进行辱骂？同样的情形出现在《西部世界》的电视剧中，西部乐园的机器人都被事先植入了永远不能反抗人类的绝对律令，因此，人类玩家对这些机器人的任意施暴行为都是允许的，在这种情况下，这些具有人类外表的机器人只是一个器物，而不是具有与人类同等人格身份的伦理存在物。那么，我作为参与主体与对话机器人之间存在的关系是否构成一种伦理关系？还有更为复杂的情形。例如在著名的网游《王者荣耀》中，经常出现的是 5 对 5 的战斗配对，当你加入战斗并与其他玩家并肩作战的时候，实际上你没有办法判断参与这类游戏、与你一起战斗的玩家是否是一个真正的

人，但是在战斗中你必须相信他是一个人类玩家，能够与你进行合作，并肩作战①。在这个时候，如果仅仅按照以人类为中心的伦理学来考察，情形会变得十分复杂。我们面对的是一个不确定的对象，这个对象不像那些明确作为机器人的NPC，我们更多时候将他们判定为人类玩家，而实际上它们可能是人工智能玩家，那么这种我们误认为是人与人之间的关系实际上是人与人工智能之间的关系，甚至会出现人工智能与人工智能之间的关系，在这种情况下，蒂洛笔下的用来调节人与人、人与其他存在物的道德，是否还适用？

或许我们可以换一个角度来思考人工智能时代的伦理问题。海德格尔在《存在与时间》中曾经使用过一个有趣的概念：周围世界（Umwelt）。海德格尔说："如今人们常说'人有他的环境（周围世界）'……因此此在本质上以'在之中'这种方式存在着，所以它能够明确地揭示从周围世界方面来照面的存在者，能够知道它们利用它们，能够有'世界'。"[8](75) 海德格尔的这个周围世界不同于真正的自然世界，在一定程度上，周围世界是依靠此在彼此的打交道（Umgang）而形成的，也就是说，这个空间是人作为此在构筑的空间。有趣的是，并不是海德格尔首先使用了Umwelt这个词，首先使用这个词的是德国的一位生物学家尤克斯考尔（Uexküll）②，其实环世界（Umwelt）就是生物通过自己的感知和行为构筑的一个将自己与真正的自然世界隔离开来的世界，这个世界和我们与周围环境打交道有关系，这样一条鱼所构建的环世界，是不同于人类本身构筑的环世界的。在这个意义上，人类世界的伦理行为是无法用来简单理解鱼类的相濡以沫的现象的，这也正是庄子的"子非鱼，焉知鱼之乐"的含义。我们是否可以做这样的推理，所谓的伦理，即在人类打交道构筑的环世界中所约定的规范和善恶的标准，这种尺度和善恶标准仅仅只在人类所构筑的环世界中有效，例如，人与人之间的伦常不能移植到鱼类的关系之中。伦理学实际上是专属于人类生物环世界的法则，不能将这种法则作为普遍性的自然法则推演到人类环世界之外的领域。

如果伦理法则是人类的环世界的法则的推论成立的话，我们可以进一步推论，在人工智能时代，可能出现一种全新的环世界，这种环世界不再被还原为人与人或人与其他存在物的关系，而是直接被还原为某种带有标签的数据。在人与人工智能的关系中，人工智能需要将具体情形还原为一组数据，并对其加以抽象和分析，才能做出行为上的判断和决策。也就是说，在人工智能时代，海德格尔意义上的打交道的方式从直接的此在关系变成了数据之间的交换关系。只有一切都变成了数据，在人工智能那里才可以被理解。严格来说，根本不是我们自己的肉体在与人工智能打交道，与人工智能的学习机制打交道的是一种数据化的自我，一种虚拟的自我，或者说一种虚体（virtual body）。正如本纳德·哈库特（Benard Harcourt）所说："今天的数据构成了一种新的虚拟身份，一种虚拟的自我，现在这种虚拟自我比起我们自身来说更加触手可及，更加权威，更加显著，更加稳固，也更加可靠。"[9](1) 所以，我们可以这样假定，在我们的身体性存在之外，还存在着一个数字化的虚拟自我，而与人工智能发生关系的恰恰是这种数字化的虚体，这样，存在着一个与我们生命性环世界平行的数字化环世界，我们可以称之为"数据环世界"（Data—Umwelt）。这个数据环世界并没有取代生命性环世界，而是以另一种模式存在着，而人工智能建立在高度数据化的世界之上，因此，我们与人工智能的任何关系都是数字环世界之中的关系。

① 实际上，腾讯游戏在游戏中加入了人工智能，这些人工智能可能是在进行机器学习，也可能是在平衡游戏的力量，让游戏更具有娱乐感。

② 在我的另一篇文章中，我将尤克斯考尔的Umwelt翻译为环世界，以区别于陈嘉映先生翻译的海德格尔的"周围世界"。后文中，在不涉及海德格尔的用语时，Umwelt一律翻译为环世界。可以参见蓝江：《环世界、虚体与神圣人》，《探索与争鸣》2019年第4期。

由于生命性环世界是由有机的生命体构成的，我们可以称之为碳基伦理学（碳基伦理学中包含了人与人、人与自然、人与动物之间的关系），但是数据环世界中存在着另一个法则，也存在着另一种伦理学，由于在数据环世界中，以半导体为基础的芯片和数据交换构成了其主要的打交道的方式，所以数据环世界的伦理学是一种硅基伦理学。

硅基伦理学必然有着与碳基伦理学不同的规范。在碳基伦理学中，人与周遭存在物的关系构成了伦理的基本法则，而硅基伦理学的基本关系是数据交换，人工智能正是在这种数据交换的关系上形成的，而人工智能所能理解的任何法则（包括伦理法则），都建立在数据交换的基础上。英国哲学家卢恰诺·弗洛里迪（Luciano Floridi）给出了适用于数据环世界的伦理学的一个不错的思考，即“只有当一个行动在其实施过程中从不产生任何形而上学熵，它才是无条件地可赞同的”[10](102)。这里所使用的熵（entropy）的概念，实际上是信息论的开创者香农从热力学中引入到信息学当中的，它代表着信息的增多从而引出的不确定性的增大。根据弗洛里迪的说法，在数据环世界中，最重要的德性是尽可能保持系统状态的稳定，避免熵的增加。对于人工智能来说，能够稳定地搜集数据，实现与其他虚体之间的数据交换就是稳定的伦理状态，相反，不确定性因素的增加，使人工智能无法做出判断，就是熵的增长，这种情况意味着数据环世界中的恶。

回到人工智能中的电车问题。在硅基伦理学之下，重点已经不是两侧的轨道究竟有多少人，或者哪边轨道上的生命更值得保留的问题，因为人工智能考虑的信息要更为广泛，要考虑是否会造成不确定性。例如，在 5 人施工的轨道是主轨道，会造成后续的拥堵，并给整个铁路系统带来不确定性，这个是人工智能需要首先排除的问题，在这个问题上，人工智能考虑的是系统的熵减。同样，在一个网络平台上，对不良信息的删除甚至封号，也是为了降低熵带来的不确定性，在这个意义上，某个账号是真实的人类账号还是一个人工智能的账号，已经不再重要。由此可见，硅基伦理学的基础是良序的数据交换，而导致熵增的行为会被数据环世界判断为恶的行为，反之则为善的行为。

最后，必须指出的是，虽然存在了数据环世界的硅基伦理学，但这并不意味着以生命体为核心的碳基伦理学的消失。同时，尽管在硅基伦理学中人的生命不构成最高的绝对价值，但是其仍然占据着举足轻重的地位。在很长一段时期，硅基伦理学和碳基伦理学互相有交叉，碳基伦理学的价值仍然存在，并直接影响到硅基伦理学。同时硅基伦理学也在很大程度上冲击了碳基伦理学的基本体系。或许在我们今天看来还十分严肃的电车问题，在数据环世界会变成一个非伦理的问题，好比在一个虚拟游戏中，讨论电车是走向有五个人的轨道还是有一个人的轨道变成了一种纯粹的模拟场景。而在这些特殊的数据环世界中，我们需要从新的角度重新来思考一种超越人类中心主义的伦理学。

参考文献

[1] Philippa Foot, *Virtues and Vices and Other Essays in Moral Philosophy*, Oxford: Oxford University Press, 1978.

[2] Judith J. Thomson, “Turning the Trolley”, *Philosophy & Public Affairs*, 2008, 36(4).

[3] 〔美〕埃里克·拉科夫斯基. 电车难题之谜[M]. 常云云，译. 北京：北京大学出版社，2018.

[4] 〔美〕马尔科夫. 人工智能简史[M]. 郭雪，译. 杭州：浙江人民出版社，2017.

[5] Stephen Marsland, *Machine Learning: An Algorithmic Perspective*, Boca Raton: Taylor & Francis Group, 2014.

[6] Terrence J. Sejnowski, *Deep Learning Revolution*, Cambridge, MC: The MIT Press, 2018.

[7] 〔美〕雅克·蒂洛，基思·克拉斯曼. 伦理学与生活[M]. 程立显，刘建，等，译. 北京：世界图书出版公司，2008.

[8] 〔德〕海德格尔. 存在与时间（修订译本）[M]. 陈嘉映，译. 北京：商务印书馆，2018.

[9] Bernard Harcourt, *Exposed: Desire and Disobedience in the Digital Age*, Cambridge: Harvard University Press, 2015.

[10] 〔英〕卢恰诺·弗洛里迪. 信息伦理学[M]. 薛平，译. 上海：上海译文出版社，2018.

责任编辑：陈　菊