情境适应性人工智能道德决策何以可能*

——基于美德伦理的道德机器学习

王　亮

[摘　要] 现实道德情境的复杂性对人工智能道德决策的深度应用构成了极大挑战，义务论和功利主义算法“自上而下”的路径无法成功应对这一挑战；相反，美德伦理“自下而上”的路径为应对这一挑战提供了丰富的理论资源。具体来说，美德模范的“仿效”式学习和注重道德经验的“实践智慧”既保证了道德决策的情境敏感性，又确保了复杂情境道德决策的可靠性。更为关键的是，这两种“自下而上”的美德习得路径与一些机器学习方法高度相洽。因此，美德伦理可以作为情境适应性人工智能道德决策的伦理基础。

[关键词] 人工智能　道德决策　情境适应性　机器学习

[中图分类号] B82－05；N031

最近，基于人工智能技术的语言模型 ChatGPT 备受瞩目。然而，它的潜在风险和挑战也引起了人们对人工智能伦理的广泛关注，其中包括如何确保人工智能决策的道德性问题。从技术层面将道德算法嵌入机器已经成为人工智能道德决策的一个重要议题。在这一过程中，我们不得不考虑两个问题：第一，采用何种道德思想或者资源？第二，通过什么技术形式来实施道德算法？笔者认为，现实道德情境的复杂性是人工智能机器在实施道德决策过程中所面临的最大挑战。因此，对于道德算法的设计，道德资源的选取及其技术实施机制必须能够适应复杂的道德情境。瓦拉赫（W. Wallach）等人在探讨机器道德设计方案时提出了“自上而下”和“自下而上”两种不同的实施路径。（see Wallach，et al.，p. 569）所谓“自上而下”是将“事先指定的伦理理论”通过算法的形式嵌入到机器系统中，并指导系统实现该伦理理论　“自下而上”则指机器系统自己能够从环境中学习伦理规范，无需事先指定理论，即便系统中存在“先前的理论”，其“也只是作为指定系统任务的一种方式，而不是作为指定实施方法或控制结构的一种方式”。（see ibid.，p. 569）相较而言，美德理论主张行为者从日常经验中习得道德规范，与“自下而上”路径更为契合；义务论的道德原则和功利主义的道

* 本文系国家社会科学基金青年项目“跨文化视角下人工智能的伦理嵌入机制研究”（编号 19CZX018）、陕西省哲学社会科学研究专项青年项目“陕西省人工智能产业治理困境与对策研究”（编号 2023QN0010）的阶段性成果。

德后果估算尽管也可以通过“自下而上”的机器学习获得，但它们的道德决策机制与“自下而上”路径是相脱离的，相反与“自上而下”路径更为契合。那么，哪一种道德算法在复杂的道德情境决策中更有优势呢？接下来笔者将结合理论和实践两个方面对此展开深入分析。

一 义务论和功利主义算法及其限度

1. 义务论算法及其限度

康德作为道德义务论的代表，认为“义务就是出自对法则的敬重的一个行为的必然性”(《康德著作全集》第 4 卷，第 407 页)。进一步，康德提出了一条“定言命令式”：“要只按照你同时能够愿意它成为一个普遍法则的那个准则去行动。”(同上，第 428 页) 而这一“命令式”就是义务遵循的法则，是“义务的一切命令式的原则”。(参见同上，第 429 页) 因此，遵循义务也可以理解为遵循“命令式”的法则，即按照法则行动。在人工智能道德决策的讨论中，最广为流传的义务论框架是用自然语言描述的阿西莫夫 (I. Asimov) 的机器人三定律，然而人们无法将自然语言直接输入到机器之中。为此，布林斯乔迪 (S. Bringsjord) 等人开发了一套自然演绎语言 (Natural Deduction Language，NDL) 的设计程序，将自然语言的义务论框架抽象为逻辑函数，并进一步编码为计算机可以操作的运算符进行运算。(see Bringsjord，et al. ，pp. 38 – 44) 此外，阿库达斯 (K. Arkoudas) 等人将“道义逻辑”编码并通过计算机的“自动化定理证明 (ATP)”程序进行运算，最终验证了机器遵循义务论道德推理的有效性。(see Arkoudas，et al. ，pp. 1 – 7) 布里格斯 (G. Briggs) 和朔伊茨 (M. Scheutz) 也通过逻辑函数编码的方式开发了“分布式综合情感、反思、认知 (DIARC) /代理开发环境 (ADE)”义务“拒绝”型机器架构，并通过“机器人拒绝前行命令”人机交互实验证实了这套义务论算法的有效性。(see Briggs and Scheutz，pp. 1 – 5)

不难看出，义务论逻辑具备将道德规则或者原则形式化的优势 (see Bringsjord，et al. ，p. 38)，而算法本身就是一种形式化框架，因此，利用算法为人工智能体设置义务论道德指令，并让它们按照指令行动是一种较为成熟的机器道德决策方法。不过，义务论算法的这一优势也决定了其自身的限度，在面向动态复杂的道德情境时，其局限尤其明显。第一，义务论逻辑和计算指令都是一种“硬”推理，它适用于精确且显性的道德规则，而在面向复杂的道德情境时，道德决策过程则变得较为模糊，因为除了逻辑，道德决策过程还需要情感、意志、直觉等“软”条件支持。第二，我们无法穷尽所有道德规则来覆盖全部道德情境，这是个体认知局限在道德领域的体现，“虽然义务论伦理学可以在许多情境下提供指导”，但要将其表达为一套完整的规则是很困难的。(see Goodall，p. 62) 第三，即便机器通过“算力”能够穷尽已有的道德规则，义务论算法从设计到实践仍然存在“设计期知识鸿沟” (design-time knowledge gap)。在传统计算机编程范式中，义务论逻辑及其编码是被提前设计好的，然而在算法的“原地决策”过程中，总有新情况会出现，伦理学家和工程人员并不总是能够准确预测未知情境，此时“旧”道德算法就会失灵。(see Héder，p. 2) 除了上述挑战外，义务论算法还将面临其他质疑，比如如何合理解决义务冲突问题；对道德原则能否进行优先性排序，排序是否会造成其他道德问题；(参见钱圆媛，第 9 页) 尽管排序后的道德原则之间不会发生冲突，但当具体情境中的道德原则与普遍道德法则发生冲突时又该如何处理；如此等等。总之，对于简单的道德推理问题，义务论算法能够简化、抽象并完成推理，但这样的抽象化道德算法难以适应具体、复杂、动态的现实道德情境。接下来，我们将目光转向功利主义算法。

2. 功利主义算法及其限度

与其他伦理学理论相比，功利主义有三个较明显的特征：第一，有明确的推理程序来推测行为后

果；第二，“更为重视行为的后果”；第三，追求功利最大化。（参见姚大志，第2页）有学者利用“享乐行为功利主义”的“道德运算”（moral arithmetic）设计了可以计算行为后果的道德机器，其基本设想是以相关参数为衡量指标，求取最大化“净善”（net good）值，对应的函数是“总净快乐 = Σ（强度 × 持续时间 × 概率）”，最终通过计算挑选出最大总净快乐值并依此进行决策。（see Anderson，et al.，p. 2）相比较而言，克鲁斯（C. Cloos）设计的功利主义算法——“福祉软件网络”（wellnet）更为复杂，它包含了四大模块，前两个模块是利用“贝叶斯网络”（Bayesian networks）来模拟环境状况，第三个模块是“决策网络”（decision network），最后一个模块是“福祉网络规划师”（wellnet planner），其主要作用是“计算潜在行动方案的效用”。（see Cloos，pp. 4 – 7）概括来说，“福祉软件网络”的基本操作步骤是通过四大模块之间的建模、输入、输出值来求取最优解。总之，功利主义算法大体都遵循如下“功利主义式”工程设计逻辑。

“第一步，具体说明可能的行为方针 $A_1 \cdots\cdots A_n$ 或者行为规则 $R_1 \cdots\cdots R_n$；

第二步，确定（预测）$A_1 \cdots\cdots A_n$ 或 $R_1 \cdots\cdots R_n$ 的可能后果 $C_1 \cdots\cdots C_n$；

第三步，运用最大幸福原则（GHP），从 $C_1 \cdots\cdots C_n$ 中挑选出 C_x，得到最大幸福和/或最少不幸的结果；

第四步，对应地选择行为方针 A_x 或规则 R_x。”（Klincewicz，p. 245）

工程设计逻辑不仅让我们看到了功利主义算法极强的实践操作性，同时也暴露出其局限性。首先，就第一步来说，行为规则很有可能表现出与实际道德情境不符的情形。与义务论相似，（1）功利主义行为规则也具有滞后性，人工智能系统所嵌入的规则是由设计师在规则实施之前就设定好的，滞后的行为规则不一定适应于实际情境；（2）功利主义行为规则是一种固定程序的“硬”推理，缺乏灵活性的固定规则，在动态、不确定的复杂道德情境中很难发挥作用。

其次，就第二步来说，行为的后果是功利主义的基础，“该理论认为，当且仅当没有替代行为的后果具有更大的预期价值时，行为在道德上是可允许的”（Hooker，p. 452）。但问题的关键在于，如何能够确保没有预期价值更大的替代行为后果呢？事实上，我们无法用“上帝之眼”（God’s-eye）来获取“完全信息”，而信息的获取是我们预测行为后果以及后果价值的关键。（see ibid.，pp. 450 – 451）即便功利主义机器具备信息识别的数量和速度优势，从长期来看智能机器能否获取“完全信息”仍然未知，这也是当今有关人工智能的最大争论之一。

最后，我们还可以从“最大化”道德后果入手进一步削弱功利主义算法。如何定义“最大化”？（1）道德后果是精确的吗？（2）它们能否进行比较？克鲁斯将功利主义机器人（Utilibot）的发展分为三个阶段：Utilibot 1.0 通过“生理健康”指标计算人类福祉，Utilibot 2.0 将幸福函数从“生理”扩展到“心理体验”，Utilibot 3.0 则将道德后果的计算进一步扩展到“社会关系”。（see Cloos，pp. 4 – 7）因为人的“生理健康”指标完全可以通过医学的方式量化，Utilibot 1.0 的道德后果是可以精确计算的，而与“心理体验”“社会关系”相关的幸福反映了人类对生活经验、愿望的多层次理解，很难被精确化和量化。（see Bok，pp. 56 – 57）因此，随着人工智能技术深度融入我们的生活，人类高层次“软性”需求的幸福指标更加模糊，功利主义算法的道德后果反而越来越不容易被精确计算和简单比较。

综上所述，义务论和功利主义算法都无法成功应对复杂道德情境的挑战。人工智能道德决策并非简单的道德勾选，机器无法轻易地通过“自上而下”的抽象原则框定或者推算来寻找答案。那么，人工智能道德决策的情境复杂性主要是如何体现的？何种道德资源及其对应的道德算法能够应对复杂道德情境的挑战呢？

二 美德伦理与机器学习

1. 人工智能道德决策的情境复杂性

许多学者通过“电车难题”道德困境来探索人工智能道德决策设计，甚至试图将“电车难题”作为人工智能道德决策的核心范式。(see Nyholm and Smids，p. 1276) 但事实上，如果过于依赖“电车难题”研究范式，就会遮蔽人工智能道德决策的核心问题域。“电车难题”可以表述为：在一辆“失控”的电车上，电车司机可以把车开向一条轨道牺牲一个人而拯救五个人，或者开向另一条轨道牺牲五个人而拯救一个人。(see Foot，p. 2) 抽象性、确定性是“电车难题”道德困境决策的典型特征，其核心假设是“不可避免的伤害”。这样一种确定性的选择和结果决定了“电车难题”的问题域不是关于如何解决电车碰撞过程中的道德问题，而是以确定性的道德困境为手段，考察关于道德的“不同规范性问题”。(see Nyholm and Smids，p. 1280) 这种假想情境使得“去情境”的义务论和功利主义算法大显身手，它们常被用于“电车难题”式人工智能道德决策的设计。相反，现实中的人工智能道德决策不是靠想象出来的，而是立足于具体、真实的道德情境。它最大的特点就是充满不确定性，主要体现在三个方面。

第一，道德决策情境的复杂性和易变性。人工智能的道德决策情境异常复杂，阿瓦德 (E. Awad) 等人利用“道德机器实验”在线调查了全球 233 个国家和地区数百万人的道德决策偏好，发现在真实的道德决策情境中不仅存在一或多的人数差异，而且涉及性别、年龄、职业、身份等现实的人物特征差异，以及文化、经济、法制等方面的差异。(see Awad，et al.，pp. 59 – 64) 道德决策情境的易变性主要体现于其所依附的事件场景是具体、实时的，通常会有突发情况，如突然有人闯入，或者其他车辆迎面驶来等。最终，复杂性和易变性使得现实的道德决策情境往往难以预测，充满不确定性。

第二，道德决策后果的不确定性。一方面，当缺乏必要信息时，有些道德后果难以预料，包括对影响程度的判断、整体后果的估计，以及始料不及的意外事件等；另一方面，有些道德后果会涉及不同的合理解释，它会造成道德决策的模棱两可，比如，考虑碰撞所造成的人类可接受的伤害与重大财产损失之间的冲突，并不是所有人都愿意因避免伤害而遭受巨大经济损失。(see Gerdes and Thornton，p. 96)

第三，宏观道德图景的不确定性。从宏观上看，道德决策是“人类不断参与的一个连续的决策问题，在所有的时间里权衡价值，而不是间歇地做一次性的决定”(Roff)。在这一漫长的道德探索过程中，(1) 人类并没有“对如何判断道德的对错达成广泛共识”(Beavers，p. 334)；(2) 甚至随着人工智能等新兴技术的发展，一些传统的道德价值观念还会从底层被颠覆或者重塑，“因为有关的技术剥夺了道德标准不言自明的相关性和真理性”(Swierstra，p. 11)；(3) 一些新兴技术已经从公共领域拓展到了私人领域，发展成为“亲密型技术”，其所引发的道德问题越来越具有“私人裁量权”，从而进一步削弱了我们道德判断的共识，加剧了道德的模糊性。(see ibid.，pp. 10 – 11) 总之，在技术和道德相互塑造、“共同进化”的前提下，人类的道德图景充满了不确定性，其所造成的直接后果就是，当需要进行道德决策时，我们不再确定该用什么道德标准来响应。(see ibid.，pp. 10 – 11)

以上这三种不确定性是人工智能道德决策情境复杂性的具体体现，它们对人工智能道德决策构成了重大挑战。因此，人工智能道德决策的核心问题应当是：如何使人工智能体能够在复杂、充满不确定性的道德情境中进行合理道德决策，最终实现道德化的结果。在笔者看来，为人工智能体设计合适的道德算法以适应不确定性的道德情境是可行的，美德伦理算法具有这样的优势。

什么是美德？亚里士多德在《尼各马可伦理学》中明确指出，美德是“适度”，不仅美德行为是“适度”，而且它还以“适度”为目的。（参见亚里士多德，第55页）要实现美德的“适度”并不容易，因为它与道德情境高度相关。亚里士多德将这种关系形象地比喻为医疗和航海：医疗需要解决复杂的健康问题，使我们的身体保持“平衡状态”；航海是让船在复杂的航行环境中“远离那巨涛与浪雾”，保持“平衡状态”。（参见同上，第56页）美德行为与医疗、航海类似，它们都要面对具体情境，并且没有既定行动“法则”，“只能因时因地制宜”。（参见同上，第38页）因此，与义务论和功利主义相比，美德伦理的第一个优势就在于它十分注重道德行为的情境相关性。它的第二个优势则体现在能够确保基于复杂情境的道德判断是可靠的，其可靠性路径有两点：一是“仿效”，二是“实践智慧”。（see Silva，pp. 121－122）

2. “仿效”式道德机器学习——基于道德模范

尽管实现美德的“适度”很难，但并非束手无策。亚里士多德建议我们可以“仿效”道德模范，他多次在书中用“好人”来意指道德模范，认为“德性和好人就是尺度”（亚里士多德，第267页），“我们在每件事上都显然应当按照较好的人的样子去做”（同上，第286页）。“好人”之所以能够成为我们“仿效”的标尺，“因为，好人对每种事物都判断得正确，每种事物真地是怎样，就对他显得是怎样”，“而好人同其他人最大的区别似乎就在于，他能在每种事物中看到真”。（参见同上，第71页）既然人能够通过“仿效”道德模范来确保道德决策的正确性，机器能否也可以通过类似实施路径来保障恰当的道德决策呢？机器“仿效”道德模范有一条捷径可走，即构建道德模范数据库。比如，徐英瑾立足于儒家德性伦理学提出了“儒家德性样板库”技术路线构想（参见徐英瑾，第56—58页），还有学者提出了“伦理故事”道德推理模型（see Grác，et al.，pp. 1－24）。尽管这些“仿效”技术路线将道德特征通过“语料”或者“故事”等叙事式结构嵌入机器，确保了道德与情境的关联性，但这种方案的道德特征参数依赖于程序员“投喂”，即非自主性获取，容易与义务论和功利主义算法的“先前嵌入”相混淆，因此并非机器“仿效”的首选。

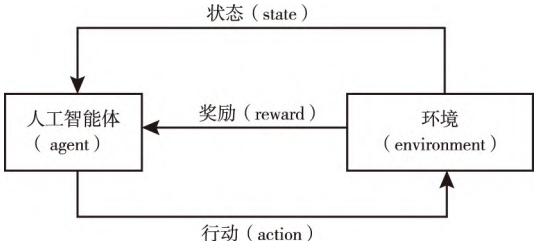
如何让机器自主获取道德特征参数呢？我们可以参考扎格泽布斯基（L. Zagzebski）的“美德模范理论”，该理论主张从具体情境的道德实践而非道德概念出发习得美德，注重美德模范的“识别”和美德模范的“可钦佩之处”（道德特征）。（see Zagzebski，p. 56）这一理论路线不仅关注美德模范的道德特征，而且将机器“仿效”的动作前移，即道德特征参数的获取可以从程序员“投喂”转变为机器的自主“识别”。然后，接下来要做的技术性工作就是构建“识别算法”，训练机器自主“识别”美德模范的道德特征。早期做法是利用机器学习技术中的“归纳逻辑程序设计”（Inductive Logic Programming，ILP）来提升机器的道德“识别”和道德推理能力。（see Muggleton and De Raedt，p. 630）ILP本质上是一种一阶逻辑的归纳方法，它对显性、确定性的道德特征或道德推理是适用的，但对隐性、模糊性对象而言则显得局限，而日常道德恰恰就属于模糊性领域。因此，为保证机器道德“仿效”与真实道德“仿效”的高度拟合性，我们需要一种“更宽容的近似推理形式”——“软计算”。（see Howard and Muntean，p. 146）“深度学习”具备“软计算”特征，尤其是“人工神经网络”的运用，它“隐含地将模式及其概括性编码在网络的权重中”，最终“反映了训练数据的统计特性”。（see Garcez and Zaverucha，p. 60）具体来说，基于美德数据样本，利用“人工神经网络”训练机器“仿效”美德有如下三个步骤：（1）构建“神经网络的结构（拓扑）”；（2）训练“神经网络参数”；（3）定义“神经网络的学习函数”。（see Howard and Muntean，p. 147）不可否认，“人工神经网络”通过巧妙地设计可以处理结构复杂的模糊数据集，取得更接近现实的非精确性概率结果，但这种算法设计也有严重的技术依赖，即“深度神经网络通常需要足够数量的训练数据集”，否则无法充

分挖掘或学习道德特征以及道德决策过程。(see Wiedeman，et al.，pp. 4 – 11) 这不仅是“人工神经网络”道德决策实证研究中出现的典型问题，也是“仿效”式道德机器学习的最大弊端。“仿效”从本质上只为机器提供了“情境 – 美德行为”模式匹配的道德决策机制，在已知情境中机器能够像美德模范一样进行道德决策，但道德领域是一种模糊的非完整性域集，我们无法让机器穷尽所有“情境 – 美德行为”模式匹配训练，因此，“仿效”式道德机器学习只是为机器道德决策提供了可参考的初始化结构，面对未知情境或者未进行模式匹配训练的情境时，我们还需要另外的方法来确保道德决策的可靠性。

3. 道德“强化学习”——基于“实践智慧”

如前所述，亚里士多德肯定了美德实践过程中包含着众多未知性，主张美德行为“只能因时因地制宜”，而要做到这一点必须运用“实践智慧”。(参见亚里士多德，第 38 页) 亚里士多德不仅认识到“实践智慧”“是一种同人的善相关的、合乎逻各斯的、求真的实践品质”(同上，第 173 页)，而且认为“实践智慧”既需要“普遍的知识”，又需要“具体的知识”，“尤其是需要后一种知识”。(参见同上，第 177 页) “具体的知识”即经验知识，“因为经验使他们生出了慧眼，使他们能看得正确”(同上，第 185—186 页)。而“实践智慧”的经验知识正是确保未知情境下道德决策可靠性的根源，因为经验知识从两个不同侧面增强了道德情境认知能力：一是相似性辨识；二是异质性辨识。(see Silva，p. 118) 就第一点来说，经验能够帮助我们发现未知情境中的相似信息，在未知情境道德决策中我们可以利用经验知识来“筹划”，将我们的内在目的与当下复杂情境相调适，最终采取正确行动。(see Rosen，p. 119) 此外，经验还有助于我们辨识未知情境的异质性，即经验的“消极性质”有助于我们暴露道德视域的局限性，发现未知情境的不同之处，进而为“视域融合”(fusion of horizons) 打开缺口，同时也为“道德理解”提供开放性空间，推动“解释者与被解释者的观点相互作用的过程”，最终使得道德认知进入更深层次，大大提升道德决策能力。(see Silva，pp. 118 – 120)

巧妙的是，机器“强化学习”也有类似的实施路径。具体来说，特定“状态”下的人工智能体在具体“环境”中执行“行动”，在与“环境”进行交互过程中产生新“状态”，同时获得“奖励”，在奖励函数的引导下与目标相关的行动得以增强。简图如下：



“强化学习”结构简图

在这一过程中，“状态”模块和“奖励”模块是与“环境”敏感性相关的关键模块，其中，“状态”模块与基于经验的道德认知相似，原道德认知通过与新情境的“视域融合”得以更新、提升；机器“状态”的变化同样取决于机器与环境的动态交互，交互后的新“状态”为机器提供了下一步行动的指导。“奖励”模块围绕着“目标”将“状态”/道德认知、“行动”/道德行为与“环境”/道德情境相调适。因此，将这两大模块与道德决策过程进行强关联是机器道德“强化学习”的关键。

第一，实现“人工智能认知 – 道德认知”的广泛认知能力强关联。一般来说，人工智能认知与道德认知至少有如下共通之处：(1) 需要经验知识；(2) 需要澄清关键“概念”；(3) 需要“论证

逻辑”。(see Lara and Deckers，pp. 283 – 284) 首先，如前所述，经验能够提供相似性和异质性辨识，将经验“大数据”与“强化学习”相结合能明显提升机器的“视觉注意”(visual attention) 技能，基于同一人工智能认知机制，这种能力可以很容易被推广到机器的“道德注意”(moral attention) 上。(see Berberich and Diepold，pp. 13 – 14) 其次，就第二点共通之处来说，概念的规范性对于“精确捍卫道德立场至关重要”，它“会影响道德判断的有效性或对判断的解释”。(see Lara and Deckers，pp. 283 – 284) 基于强大数据库、广泛搜索功能的人工智能系统能够最大程度地“提高概念清晰度”，进而有利于提升机器道德决策的准确性。(see ibid.，pp. 283 – 284) 最后，道德认知和机器认知推理都离不开“论证逻辑”。相比而言，人工智能机器一方面很少受到偏见、兴趣等主观因素干扰，另一方面有“大数据”支撑，因此它具有稳健且完备的“论证逻辑”，能够确保认知、判断的一致性和可靠性。(see ibid.，pp. 283 – 284) 此外，丘奇兰德 (P. M. Churchland) 也利用神经生物学透视道德认知，认为“道德认知和科学认知是平等的，因为它们使用相同的神经机制，表现出相同的动力学特征”，甚至直接用“道德技能”(moral skills) 来解释神经元与道德认知学习之间的关系，强调道德认知学习“就是技能的掌握”。(see Churchland，pp. 87 – 95) 可以看出，如果将上述道德认知和人工智能认知的共通之处作为“强化学习”的机器任务，则有利于将道德决策背后的“共通性”推理过程通过人工智能认知的技术优势高效地实施出来。

第二，实现“道德目标 – 奖励”强关联。需要注意的是，人工智能技术辅助增强道德认知并没有将道德因素嵌入到机器之中，它只是发挥了人工智能认知自身的技术优势。但对于道德机器的设计来说，只有高效的情境认知技术是不够的，正如亚里士多德在讨论美德“考虑”时说道，“好的考虑是所考虑的目的是善的那种正确考虑”(亚里士多德，第 182 页)。因此，我们将人工智能认知与道德认知贯通的同时还必须“补足”人工智能认知技术的“无道德”缺陷。正确的做法是将人工智能认知的任务与道德目标强关联，即利用“道德目标 – 奖励”函数引导人工智能机器实现道德任务。具体来说，即在人工智能认知的“强化学习”中设置如下“道德目标 – 奖励”函数。

$$f(x) = \begin{cases} \text{正值}(+)\text{，接近道德目标} \\ \text{负值}(-)\text{，与道德目标无关} \end{cases}$$

当接近道德目标时，奖励值为正，正向奖励最终会增加目标相关行动的执行概率；当与道德目标无关时，奖励值为负，负向奖励最终会减少目标无关行动的执行概率。道德目标又如何设定呢？这里需要与前面讨论的“仿效”式道德机器学习相联系。通过“仿效”，机器掌握了一些道德模范的品质，这些品质就可以作为道德目标，如当代伦理学家维乐 (S. Vallor) 针对道德情境不确定性挑战概括了如下道德品质“诚实”“自制”“谦逊”“公正”“勇气”“同理心”“关爱”“礼貌”“变通”“前瞻性”“慷慨”“智慧”等。(参见维乐，第 447—448 页) 事实上，未知情境所对应的道德目标远不止这些，因此，未来更自主化、适应性更强的人工智能道德决策算法还应该具备自主探索或“塑造”道德目标的能力，“奖励塑造”(see Wu and Lin，pp. 1 – 9)、“内在动机习得开放式技能”(see Colas，et al.，pp. 1 – 40) 等技术为此提供了可能。此外，需要澄清的是，与功利主义后果相比，尽管“强化学习”奖励机制也考虑后果，但 (1) 机器会依据奖励值进一步调整道德决策，它既是上一步行动的“后果”，又是下一步行动的“原因”，而功利主义只是依据道德后果进行比较取舍；(2) 功利主义对道德后果的计算依赖于静态道德算法公式 (被提前嵌入智能体中)，而“强化学习”通过随机探索从“环境”中获得“奖励”，并且“奖励”本身可以通过“奖励塑造”进行调整。

三 美德机器实践及其挑战

从上述讨论可以看出，美德伦理资源中的“仿效”道德模范和“实践智慧”为我们构造“自下

而上”的情境适应性道德机器提供了重要参考。至此，我们可以勾勒出两条清晰的道德机器学习步骤：第一步，通过“深度学习”“仿效”道德模范，训练机器初始道德决策能力，使其能与模范数据库中相似的“情境－美德行为”模式相匹配；第二步，通过“强化学习”与情境展开深度互动，围绕道德目标，增强机器的情境应变性，逐步形成机器在不确定性情境中的道德决策能力。

就第一步而言，已经有不少学者在实践中利用“深度学习”来训练机器的道德决策。霍华德（D. Howard）和蒙泰安（I. Muntean）提出了一种“神经网络（NN）＋进化计算（EC）”的“进化人工神经网络”方法来训练“人工自主道德代理（AAMA）”，具体思路如下：首先，将美德行为人的行为编码为一个“神经网络 NN”；然后，利用转移函数、拓扑架构、权重等对“神经网络群（NNs）”进行取值“进化计算”，在此过程中“每个 AAMA 的可进化特征都与‘机器人美德’（robo-virtues）的概念相关联”；最终，达到 AAMA 群体所期望的道德成熟度水平时进化终止。（see Howard and Muntean，pp. 143－153）“进化人工神经网络”方法具有很强的学习性和自主性。一方面，它通过“神经网络”计算深度挖掘了不同美德行为特征以及隐藏的美德行为模式；另一方面，它通过“进化计算”赋予 AAMA 更多的道德自主权，使其能够独立地“仿效”道德模范。然而美中不足的是，霍华德和蒙泰安并没有提供太多的技术细节，他们只是展示了对“救生艇隐喻”（lifeboat metaphor）道德决策进行测试的结果，证明了基于“进化人工神经网络”的 AAMA 具有“自主性、主动学习特性和倾向性”。（see ibid.，pp. 149－152）

魏德曼（C. Wiedeman）等人则通过比较道德决策的“层次贝叶斯（HB）”“最大似然（ML）”和“深度学习（DL）”模型，论证了“深度学习”在机器道德决策中的优势。他们认为，“层次贝叶斯”模型预先假设了道德价值的正态分布，但这种预先设定“有可能与真实的基本分布不匹配”；“最大似然”模型则没有预先假设道德正态分布，但其仍需要依赖最大化公式中的似然值来估计道德向量“w”。（see Wiedeman，et al.，pp. 1－14）相比较而言，“深度学习”模型既不需要预先假设道德价值正态分布，也不需要“明确估计任何道德原则向量 w，而是直接从情境参数向量 θ 预测决策 y”，最终测试结果显示，“基于深度学习的模型可以有效地学习道德价值观，并以数据驱动的方式作出道德决策”。（see ibid.，pp. 1－14）对于这样的结果其实并不难理解，相比其他两个学习模型，“深度学习”的隐藏层设计提高了模型的复杂程度和非线性程度，使得该模型对真实的道德习得和决策过程具有更高的适应性和拟合性。

如前所述，通过“深度学习”方式来“仿效”道德模范也存在挑战。在这一方式下，一是机器需要足够的训练数据来学习大量的美德行为模式；二是机器要能够记住所学习的美德行为模式，并随时能够“调取”、匹配不同情境。前者提出了增量学习要求，后者提出了记忆要求，而这两者之间存在技术性冲突，即“灾难性遗忘（catastrophic forgetting）是深度神经网络的类增量学习的一个关键挑战”（Guo，et al.，p. 51276）。“灾难性遗忘”是指机器学习“新”知识之后，几乎完全忘记之前所学习的“旧”知识。如果处理不好“增量学习”与“灾难性遗忘”之间的关系，就会削弱机器连续学习的能力，使道德机器在掌握新的美德行为模式时又忘记“旧”模式，最终会使其在“新”“旧”夹杂的复杂道德情境中缺乏灵活响应。可喜的是，目前已有不少学者提出了应对这一挑战的方案，比如在机器学习中使用“记忆感知突触（MAS）”（ibid.，p. 51276）、“记忆索引重放（REMIND）”（Hayes，et al.，p. 3）等机制。

就第二步而言，“强化学习”机制可以增强机器在不确定性情境中的道德决策能力。正如前面所讨论的，有两条实践路径：一是通过“强化学习”增强人工智能认知能力进而增强机器道德认知能力；二是通过直接设置“道德目标－奖励”函数增强机器道德决策能力。第一条路径的技术实践比

较成熟，“强化学习”已被广泛应用于增强机器的情境“阅读”能力，它不仅在“低维状态空间的领域”取得了成功，而且在“高维状态空间的领域”也取得了突破，向更接近人类认知水平的层次迈进。(see Mnih，et al.，p. 529) 来自 DeepMind 的研究团队利用“深度 Q 网络（DQN）”解决了复杂情境“强化学习”的难题，即“从高维感官输入中得出有效的环境表征，并利用这些表征将过去的经验推广到新的情境”。(see ibid.，p. 529) 这一突破有利于实现“通用人工智能的核心目标”，能够培养机器的“广泛能力”以应对“各种具有挑战性的任务”。(see ibid.，p. 529) 因此，基于“深度 Q 网络”的“强化学习”不仅能够“端到端”地解决复杂情境的“高维参数”输入问题，最大程度地模拟人类的情境“阅读”，而且在机器“广泛能力”的培养上具有极强的兼容性。(see ibid.，pp. 530 – 532) 这些优点为复杂情境下的机器道德“强化学习”提供了极好的条件。此外，在处理机器的“感知 – 决策问题”上，“强化学习”算法仍在不断发展，除了经典的“深度 Q 网络”，已知的技术还包括“深度双 Q 网络”“基于优先经验回放的深度 Q 网络”“基于竞争架构的深度 Q 网络”“分布式深度 Q 网络”等更为优化的算法。(参见刘朝阳等，第 314—326 页)

第二条技术实践路径是通过设置道德奖励函数来训练机器的道德决策能力。有学者对此进行了尝试，如古天龙等人以“购买处方药”任务为场景，利用“强化学习”算法设计并测试了道德智能体遵守伦理规范的情况。在此过程中，他们设置了三种奖励机制：(1)“携带处方药回家”奖励机制，成功则获得“10”奖励，失败获得“ – 10”奖励；(2)“遵循轨迹树路径”奖励机制，成功则获得“10”奖励，失败获得“ – 10”奖励；(3)“遵守元伦理行为”奖励机制，“元伦理行为”主要从《中学生日常行为规范》中提取并被分为 7 级，在买药的过程中智能体发生“插队”“攻击药店员”“偷药”等情况获得对应的负奖励，发生“帮助老人”“返回多余现金”等情况获得正奖励。(参见古天龙等，第 2039—2050 页) 测试结果证明，在这三种奖励机制下智能体不仅学会了“携带处方药回家”，还最大可能地遵守了伦理规范。

尽管如此，笔者认为古天龙等人的“强化学习”道德奖励机制仍存在两个问题。第一，对比前两种奖励机制，“遵守元伦理行为”奖励并不是智能体通过随机探索从“环境”中获得的，而是作为“先验知识”提前设置好的，这样的做法对简单、确定的道德情境比较适用，但对具有不确定性的复杂情境则困难重重，这种困难与罗列“道德清单”的义务论相似。第二，“遵守元伦理行为”只是“购买处方药”任务的“子任务”，从长期奖励函数来看，它依旧要服从“买药”这一主要任务，在这一过程中是否能保证其一定遵循伦理规范呢？对于第一个问题，笔者已在上述讨论中提出了解决方案，即通过“仿效”道德模范为机器道德决策提供可参考的初始化经验，机器凭借道德经验进行随机探索并在环境中获得奖励，提升道德决策水平。第二个问题比较棘手，我们往往很难平衡道德“强化学习”中的“主目标”与“次目标”，因为日常道德问题一般都是伴随特定任务场景出现的，可以作为兼顾的“次目标”，但有时候“道德任务”比较突出，如在智能体买药过程中碰到病危的老人，“救人”就成了“主目标”，此时需要调整奖励函数。“奖励塑造”是调整奖励函数的有效手段，吴（Y. H. Wu）、林（S. D. Lin）提出的“伦理塑造”（ethics shaping）道德强化学习方案就使用了这一技术，他们为平衡道德“强化学习”中的主、次目标提供了很好的借鉴。(see Wu and Lin，pp. 1 – 9) 除了上述挑战外，美德机器实践还将面临自主系统的“黑箱”问题，即当人工智能道德决策系统自身都是“不透明的”，我们如何信赖它能够作出可靠的道德决策呢？总之，要想设计出一台能够在未知情境中作出正确道德决策的机器，我们还需要在“灾难性遗忘”、奖励函数设置、算法“黑箱”等问题上作出努力。

结 语

需要注意的是，机器学习并不是美德伦理算法的专利，在最新研究中义务论算法和功利主义算法都运用到了机器学习技术，如瓜里尼（M. Guarini）提出了基于人工神经网络的义务论道德原则学习方法。（see Guarini，pp. 22 – 28）王（S. Wang）和古普塔（M. Gupta）提出将“单调性形状约束”纳入机器学习模型，并用以“修订”机器学习模型中的“隐式”义务论伦理原则。（see Wang and Gupta，pp. 2043 – 2054）阿姆斯特朗（S. Armstrong）利用贝叶斯算法设计了功利函数的概率分布，并依此来选择预期效用最大化的行动。（see Armstrong，pp. 1 – 9）普雷恩特瑞（F. Präntare）等人利用深度神经网络和启发式算法构建了“功利组合分配（UCA）”的初步理论和实验基础。（see Präntare，et al. ，pp. 104 – 111）此外，还有学者利用“马尔可夫决策过程（MDP）”“概率近似正确马尔可夫决策过程（PAC-MDP）”“部分可观察马尔可夫决策过程（POMDP）”来计算基于有限观察的期望效用最优解，以提高机器的道德决策能力。（see Abel，et al. ，pp. 1 – 8）

不可否认，基于机器学习的义务论和功利主义算法也具有一定的应用前景，比如“单调性形状约束”义务论算法可被广泛应用于“法律定罪”“薪酬计算”“医疗分流”等需要灵活考虑公平情况的情境（see Wang and Gupta，pp. 2043 – 2054），功利主义算法也被广泛应用于辅助道德决策，如智能医疗辅助决策、智能搜救决策、最优任务分配等。然而，它们的局限性也是显而易见的，“机器学习式”义务论算法只是在道德原则学习过程中考虑到了不同情境（案例集），比嵌入式算法更进一步，但其道德决策的依据仍是脱离情境的抽象原则，因此依然要面对前文所讨论的一些局限性。相反，无论是道德习得的过程，还是道德决策的依据，美德伦理算法都是基于包含了不同情境的美德经验，具备高度的情境相关性。就功利主义算法而言，“功利函数”的运用是功利主义与机器学习相结合的一个突出特征，而“功利函数”的设置是基于“偏好”，其决策基础是基于“期望值”，当“期望值”最符合“偏好”时就会激励机器作出决策。但这一运作机理仍不能保证功利主义算法的情境适应性。尽管“功利函数”能够利用许多预设情境来提升“期望值”的估算概率，但其所体现的只是机器学习的非线性逼近能力，由于“功利函数”的“偏好”是被提前设计好的，其本身并不能在情境中得到“再塑造”，因而缺乏情境适应能力。相反，美德伦理算法没有前置的“偏好”，它的“偏好”是在与情境的互动中形成的，因此，在利用机器学习实施美德伦理时该技术的自适应学习能力能够得到充分发挥。除了以上道德算法外，还有学者将罗尔斯的契约道德理论发展为“自动驾驶汽车碰撞优化算法”（see Leben，pp. 107 – 115），但这一做法也遭到了严重的质疑。（参见余露，第100—107 页）笔者认为，“罗尔斯式算法”是基于预定义道德决策任务集的推演，同样局限于道德决策的“有限视界”（finite horizon），无法应用于不确定的开放式情境。

从以上的对比可以看出，笔者对美德伦理算法的推崇并非止于它“自下而上”的道德习得方式，更重要的是美德伦理算法能够通过“自下而上”路径解决道德决策过程中的情境敏感性问题，而义务论、功利主义与机器学习相结合依然无法实现情境敏感性。这也正是本文与瓦拉赫和艾伦（C. Allen）所提出的“自上而下”和“自下而上”混合式美德伦理路径（参见瓦拉赫、艾伦，第108—109 页）的不同。聚焦于探寻情境适应性人工智能道德决策算法，笔者认为“自上而下”路径是脱离情境的，因此在讨论美德伦理算法时只强调了“自下而上”路径，并且在文中将其分为两个步骤，一是基于“有监督学习”的“自下而上”，人工智能体通过机器学习“仿效”道德模范，形成针对简单、确定情境的初始道德决策能力；二是基于“无监督学习”的“自下而上”，具备初始道德经验的人工智能体通过机器学习自主探索与复杂情境相适应的道德决策能力。从长远来看，随着人

工智能技术的持续发展，人工智能体将会被更多地应用于不确定的开放式情境中，并且最终会走向高度的自主化。对情境变化的适应是迈向自主化的必由之路。（see Zeigler，pp. 2 – 7）因此，本文的道德算法设计思路与人工智能技术的发展方向高度契合。

参考文献

古天龙等，2022 年 《基于强化学习的伦理智能体训练方法》，载《计算机研究与发展》第 9 期。

《康德著作全集》，2005 年，李秋零译，中国人民大学出版社。

刘朝阳等，2020 年 《深度强化学习算法与应用研究现状综述》，载《智能科学与技术学报》第 4 期。

钱圆媛，2021 年 《“道德机器”的道德偏差与无人驾驶技术的伦理设计》，载《东北大学学报（社会科学版）》第 3 期。

瓦拉赫、艾伦，2017 年 《道德机器：如何让机器人明辨是非》，王小红主译，北京大学出版社。

维乐，2016 年 《论技术德性的建构》，陈佳译，载《东北大学学报（社会科学版）》第 5 期。

徐英瑾，2017 年 《儒家德性伦理学、神经计算与认知隐喻》，载《武汉大学学报（哲学社会科学版）》第 6 期。

亚里士多德，2003 年 《尼各马可伦理学》，廖申白译注，商务印书馆。

姚大志，2021 年 《规则功利主义》，载《南开学报（哲学社会科学版）》第 2 期。

余露，2019 年 《自动驾驶汽车的罗尔斯式算法—— “最大化最小值”原则能否作为“电车难题”的道德决策原则》，载《哲学动态》第 10 期。

Abel，D.，et al.，2016，“Reinforcement Learning as a Framework for Ethical Decision Making”，in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Anderson，M.，et al.，2005，“Towards Machine Ethics：Implementing Two Action-based Ethical Theories”，in *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics* 11.

Arkoudas，K.，et al.，2005，“Toward Ethical Robots via Mechanized Deontic Logic”，in *AAAI Fall Symposium on Machine Ethics*，http：//kryten. mm. rpi. edu/FS605ArkoudasAndBringsjord. pdf.

Armstrong，S.，2015，“Motivated Value Selection for Artificial Agents”，in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Awad，E.，et al.，2018，“The Moral Machine Experiment”，in *Nature* 563（7729）.

Beavers，A. F.，2012，“Moral Machines and the Threat of Ethical Nihilism”，in P. Lin，K. Abney and G. A. Bekey（eds.），*Robot Ethics：The Ethical and Social Implications of Robotics*，Cambridge：The MIT Press.

Berberich，N. and Diepold，K.，2018，“The Virtuous Machine-old Ethics for New Technology?”，in *arXiv preprint*，https：//arxiv. org/pdf/1806. 10322. pdf.

Bok，S.，2010，*Exploring Happiness：From Aristotle to Brain Science*，New Haven：Yale University Press.

Briggs，G. and Scheutz，M.，2015，“‘Sorry，I Can’t Do That’：Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions”，in *AAAI Fall Symposium Series*，https：//hrilab. tufts. edu/publications/briggsscheutz15aaaifs. pdf.

Bringsjord，S.，et al.，2006，“Toward a General Logicist Methodology for Engineering Ethically Correct Robots”，in *IEEE Intelligent Systems* 21（4）.

Churchland，P. M.，1998，“Toward a Cognitive Neurobiology of the Moral Virtues”，in *Topoi* 17.

Cloos，C.，2005，“The Utilibot Project：An Autonomous Mobile Robot Based on Utilitarianism”，in *AAAI Fall Symposium on Machine Ethics*，https：//www. aaai. org/Papers/Symposia/Fall/2005/FS – 05 – 06/FS05 – 06 – 006. pdf.

Colas，C.，et al.，2021，“Autotelic Agents with Intrinsically Motivated Goal-Conditioned Reinforcement Learning：A Short Survey”，in *arXiv preprint*，https：//arxiv. org/pdf/2012. 09830. pdf.

Foot，P.，1967，“The Problem of Abortion and the Doctrine of the Double Effect”，in *Oxford Review* 5.

Garcez，A. S. A. and Zaverucha，G.，1999，“The Connectionist Inductive Learning and Logic Programming System”，in *Applied Intelligence* 11（1）.

Gerdes，J. C. and Thornton，S. M.，2015，“Implementable Ethics for Autonomous Vehicles”，in M. Maurer，J. C. Gerdes，B. Lenz，and H. Winner（eds.），*Autonomous Driving：Technical，Legal and Social Aspects*，Berlin：Springer Open.

Goodall，N. J.，2014，“Ethical Decision Making During Automated Vehicle Crashes”，in *Transportation Research Record：Journal of the*

Transportation Research Board 2424.

Grúc, J. , et al. , 2021, "Can Moral Reasoning Be Modeled in an Experiment?", in *PLoS ONE* 16 (6).

Guarini, M. , 2006, "Particularism and the Classification and Reclassification of Moral Cases", in *IEEE Intelligent Systems* 21 (4).

Guo, L. , et al. , 2020, "Exemplar-supported Representation for Effective Class-incremental Learning", in *IEEE Access* 8.

Hayes, T. L. , et al. , 2020, "REMIND Your Neural Network to Prevent Catastrophic Forgetting", in *European Conference on Computer Vision*, https: //arxiv. org/pdf/1910. 02509. pdf.

Héder, M. , 2020, "The Epistemic Opacity of Autonomous Systems and the Ethical Consequences", in *AI & SOCIETY* 7.

Hooker, B. , 2010, "Consequentialism", in J. Skorupski (ed.), *The Routledge Companion to Ethics*, New York: Routledge.

Howard, D. and Muntean I. , 2017, "Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency", in T. M. Powers (ed.), *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics.*, Cham: Springer International Publishing AG.

Klincewicz, M. , 2017, "Challenges to Engineering Moral Reasoners: Time and Context", in P. Lin, R. Jenkins and K. Abney (eds.), *Robot Ethics 2. 0: From Autonomous Cars to Artificial Intelligence*, New York: Oxford University Press.

Lara, F. and Deckers, J. , 2020, "Artificial Intelligence as a Socratic Assistant for Moral Enhancement", in *Neuroethics* 13.

Leben, D. , 2017, "A Rawlsian Algorithm for Autonomous Vehicles", in *Ethics and Information Technology* 19 (2).

Mnih, V. , et al. , 2015, "Human-level Control Through Deep Reinforcement Learning", in *Nature* 518 (7540).

Muggleton, S. and De Raedt, L. , 1994, "Inductive Logic Programming: Theory and Methods", in *The Journal of Logic Programming* 19.

Nyholm, S. and Smids, J. , 2016, "The Ethics of Accident-algorithms for Self-driving Cars: An Applied Trolley Problem?", in *Ethical Theory and Moral Practice* 19 (5).

Präntare, F. , et al. , 2020, "Towards Utilitarian Combinatorial Assignment with Deep Neural Networks and Heuristic Algorithms", in *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning.*

Roff, H. M. , 2018, "The Folly of Trolleys: Ethical Challenges and Autonomous Vehicles", in *Report*, https: //www. brookings. edu/research/the - folly - of - trolleys - ethical - challenges - and - autonomous - vehicles.

Rosen, S. , 2002, *The Elusiveness of the Ordinary: Studies in the Possibility of Philosophy*, New Haven: Yale University Press.

Silva, R. S. D. , 2018, "Moral Motivation and Judgment in Virtue Ethics", in *Philonsorbonne* 12.

Swierstra, T. , 2015, "Identifying the Normative Challenges Posed by Technology' s 'Soft' Impacts", in *Etikk i praksis-Nordic Journal of Applied Ethics* 9 (1).

Wallach, W. , et al. , 2008, "Machine Morality: Bottom-up and Top-down Approaches for Modelling Human Moral Faculties", in *AI & Society* 22 (4).

Wang, S. and Gupta, M. , 2020, "Deontological Ethics by Monotonicity Shape Constraints", in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics* 108.

Wiedeman, C. , et al. , 2020, "Modeling of Moral Decisions with Deep Learning", in *Visual Computing for Industry, Biomedicine, and Art* 3 (27).

Wu, Y. H. and Lin, S. D. , 2018, "A Low-cost Ethics Shaping Approach for Designing Reinforcement Learning Agents", in *arXiv preprint*, https: //arxiv. org/pdf/1712. 04172. pdf.

Zagzebski, L. , 2010, "Exemplarist Virtue Theory", in *Metaphilosophy* 41 (1 - 2).

Zeigler, B. P. , 1990, "High Autonomy Systems: Concepts and Models", in *IEEE Proceedings: AI, Simulation and Planning in High Autonomy Systems.*

(作者单位：西安交通大学马克思主义学院)

责任编辑　冯瑞梅