人工智能道德增强：能动资质、规范立场与应用前景

黄　各

【摘　　要】信息技术时代的来临使人类道德增强的手段变得越发丰富，以人工智能为依托的道德增强方案旨在构造一个伦理智能系统，帮助人类作出更为合理的道德决策。但是，此系统在能动资质、行动动机等方面却受到一定质疑。在对其主体资格进行认定后，通过对直接伦理决策机器、人工道德建议者以及关系—辅助型增强系统的规范立场的讨论，明晰其在何种意义上能够指导人类道德实践。在此基础上，对关系—辅助型增强系统的应用前景赋予期待，并对其如何与人类共处进行了分析。

【关 键 词】道德增强　人工智能　规范性　关系型辅助

【作者简介】黄各，哲学博士，中共中央党校（国家行政学院）哲学教研部讲师。

【中图分类号】B829　　【文献标识码】A

【文章编号】2097－1125（2022）05－0018－13

道德增强（moral enhancement）技术是近年来伦理学领域讨论的热点。因为它主张可以通过特定的技术手段（比如生物干预、基因改造、颅刺激）来使人类在道德判断中获取稳定的性情，保持客观中立，并以此来影响行动者的选择和偏好，从而获得道德水平的提升。但这一建立在生物工程学基础上的技术，因涉及对人类基因、细胞的改造受到众多的反对。近年来，随着大数据、深度学习、仿生技术等领域的快速发展和交叉融合，人工智能已经在社会诸多领域得到了充分的发展。萨沃斯库（Julian Savulescu）等学者基于此提出了“人工智能道德增强”（artificial intelligence moral enhancement，以下简称 AI 增强）的新方案。该方案能够避免对人体结构的直接改造，而是根据事先设定的标准编写一套提供道德建议的程序，该程序能够搜索到足够多的信息，并结合程序中设定的观点对人类的道德行为进行指导，最终创建出一套人工

智能系统。它一方面具有非常强大的运算能力和知识背景，另一方面拥有稳定且可靠的情绪，可以为人们提供超出自身认识、情感等范围内的信息，进而帮助其进行更为合理的道德判断。

为了更为清晰地认识人工智能系统的作用机制和原理，以及它在何种条件下能更好地帮助人类增强道德，本文拟从如下三个方面来对此方案进行探讨：第一，人工智能（Avtificial Intelligence，以下简称 AI）系统是否具备道德能动性，作为道德行动主体的我们应该如何对待它的存在；第二，AI 增强所提供的深度人机交互模式的道德规范性依据是什么，我们又因何判定其是合理的；第三，在综合各方面因素的考量后，我们拟提出一种关系—辅助型的 AI 增强模式，并结合其实际运作方式，对其应用前景进行评估和展望。

一、对人工智能道德能动性的质疑与回应

在道德增强的语境中，AI 更多的是以“普适计算”（ubiquitous computing）或“环境智能”（ambient intelligence）系统的形式出现。它是一个从多个传感器和数据库搜索信息的系统，并根据其对系统用户的功能相关性进行处理。AI 系统的出现已经从多个方面改善了人们的生活质量，它以更为强大的数据搜索能力、更为可靠的判断能力以及更为稳定的情绪能力带给人们帮助。AI 增强方案的支持者认为，这一技术的应用前景是乐观的，因为无论人类作出何种努力，其通过自身教育使道德能力增强的进程是非常缓慢的。现在有了能够快速提升道德的工具和技术，且不会对人体造成损害，我们没有理由置之不理。因而，这一方案的重点是要创造出具有一定道德评判、建议和执行能力的智能机器，按照算法系统提供的方案，依照某种运行程序[①]来帮助人类进行道德决策。不过，这项方案所面临的一个首要挑战是：协助人类进行道德决策的 AI 系统是否具备一定的道德行动能力，从而能够担负起道德实践判断的角色。很多学者对此持有怀疑态度，他们从如下三个方面对智能系统的道德能动性进行了质疑。

首先，是对其理解能力的质疑。在他们看来，AI 需要倚靠计算机算法系统，其只有处理信息而没有理解信息的能力。塞尔（John Searle）曾证明 AI 只是一堆程序，它根本无法“理解”自己在做什么。彭罗斯（Roger Penrose）也指出：“只有在某些自上而下（或主要是自上而下）组织的情况

① 舍费尔和萨沃斯库就据此提出了一种“程序性道德增强”（Procedural Moral Enhancement）的方案。他们以罗尔斯《一个伦理决策程序的纲要》为依托，制定出一个“合资格的评判者”，使人们能够制定出更合理的道德决策。参见 G. Owen Schaefer and Julian Savulescu, Procedural Moral Enhancement, *Neuroethics*, Vol. 12 (1), 2019, p. 75。

下，计算机才表现出对人类的显著优势……而在自下而上（人工神经网络）的组织中，计算机只能在少数有限的情况下达到普通训练有素的人类的水平。”① 因而它们可能只有分析判断，而没有综合判断的能力，需要依靠提前指定的参数来模仿。在这些学者眼中，AI 不过是数据和程序的简单堆砌，是一个没有理念和心灵的低级造物。人类的理解能力、分析能力、自主意识能力等都被认为是一种难以探究、难以模仿、更难以企及的神秘事物。因此，基于算法的 AI 机器无法像人类一样成为真正的道德行动者。

其次，是对其情感、心理和共情（sympathy）能力的质疑。在道德情感主义者眼中，人类的情感是道德判断中最为重要的组成部分。在他们看来，评判道德善恶的一个重要标准是人类“爱恨苦乐”的感觉。如果缺少了人类情感的特质，道德所必需的交互主体性和敏感性将不会存在。正如雷依（Georges Rey）所言：“只有拥有与我们（人类）的荷尔蒙和神经调节的情感相关的化学特性的实体才能完全有资格成为一个人……即使它们（机器）确实有自己的一种情感状态，仍然需要表明的是，这些情感是不是一种与道德相关的类型。”② 此外，斯密（Adam Smith）和休谟（David Hume）所提出的共情能力也是道德能动性的重要一环，其意在构建不偏不倚的旁观者来指导道德行动。但由于 AI 系统没有类似于人的心理结构，很多质疑者认为，它不具备相应的情感和共情能力，因此道德行动能力有所欠缺。

最后，是对其自律、意志和责任能力的质疑。道德理性主义者认为，与道德能动性关联密切的是人类的自律能力。按照通常的理解，如果行动者的自由意志（意愿）引起了他的行动，那么他的行动就是自律的；而道德自律又与道德法则有着强烈的关联，能够促使人们遵从自我订立的法则并积极履行相应的道德责任。但是，在很多质疑者看来，AI 系统并不具备此种自我立法的能力，也没有一种“作恶的自由”或“可供取舍的可能性”的自由意志，更不用说在此基础上所形成的责任意识能力。

但我们认为，以上诸种质疑并不能从根本上否认 AI 系统的道德能动性。其一，当代科技的迅猛发展已经让 AI 可以从符号主义、联结主义出发，基于物理符号系统的计算机程序模拟人类的思维过程。从本质上说：“认知是信息处理，且信息处理是可计算的……符号主义基于自上而下的道路，物理符号系统致力于把客观世界做成形式模型，而联结主义如包括神经网络、深

① Roger Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness*, New York: Oxford University Press, 1994, p. 19.

② Georges Rey, Functionalism and the Emotions, in Amelie Rorty, ed., *Explaining Emotions*, Berkeley: University of California Press, 1980, pp. 192 – 193.

度学习等基于自下而上的道路，神经网络致力于把大脑做成形式模型。”① 前文所提到的塞尔等人不仅低估了 AI 运用算法时所需能力的复杂程度，还忽视了其通过深度学习而迅速获得自主迭代升级的功能。AlphaGo 战胜人类棋手的事例其实已经表明了 AI 具备深层次的逻辑抽象和综合判断能力。丹尼特（Daniel Dennett）对此就指出：“在图灵之前，我们总以为人类所有的能力都来自于他们能理解，理解是所有智力的神秘源泉，现在我们才认识到，理解本身是一种效果，它从成堆的能力中冒出来，是各种能力层层叠加生成的。”②

其二，人们虽然可以从苦乐等感觉经验中建构起自身的道德标准，并以此来规范自身，形成准则，但是这并不表明它们就是评判道德能动性的先天条件。斯密也承认，情感的作用是让行动具有一定的合宜性。现代道德哲学的研究也越来越倚重于以合理性为基础的规范性和实践理由等概念，并进一步发展出“理由基础主义”（reasons fundamentalism）等侧重理性认知基础的学说。其意在通过强调理由是一个由事实、行动者、情境和态度构成的四元关系，来论证道德评判的依据在于行动理由的客观规范性。因此，AI 系统可能确实无法拥有人类独特的情感体验，但其在对四元关系的把控中相较于人类更为稳定和客观，也能依照更为普遍的要求来制定道德准则。

其三，将道德能动性与意志自由建立一种强关联亦是值得进一步讨论的。在利用生物工程进行道德增强的案例中，很多学者亦对此论题展开了争论。③ 其中，道格拉斯反对哈里斯的一个论断就足以说明自由意志对取得道德能动性资格而言是不充分的。在他看来：“在许多情况下，为了防止作恶，牺牲一些作恶的自由似乎更为可取……同样的，在非认知性道德增强的情况下，不道德的自由所损失的价值被不道德行为或者动机所减少的价值抵消了。”④ 萨沃斯库等人通过一个“上帝—机器”（god machine）的思想实验也认为，可供选择的自由并非道德责任的首要条件，其也可能“通过控制道德主体，使这个人屈从于另一个人的意志，并取消不道德行为的自由”。⑤ 正

① 闫坤如：《人工智能理解力悖论》，《云南社会科学》2020 年第 3 期，第 32 页。

② ［美］丹尼尔·丹尼特：《直觉泵和其他思考工具》，冯文婧、傅金岳、徐韬译，浙江教育出版社 2018 年版，第 341 页。

③ 除道格拉斯和哈里斯外，王珀、刘玉山、马翰林、叶岸滔等学者也针对此问题展开过讨论。不过，他们的讨论主要聚焦在生物医学增强对人的意志自由的影响上。由于本文无意陷入自由意志与道德责任的众多争论中，对此不再作过多展开。

④ Thomas Douglas, Moral Enhancement via Direct Emotion Modulation: A Reply to John Harris, *Bioethics*, Vol. 27 (3), 2013, p. 166.

⑤ Julian Savulescu and Ingmar Persson, Moral Enhancement, Freedom, and the God Machine, *The Monist*, Vol. 95 (3), 2012, p. 417.

如相容论者所论证的那样："自由不是指免于因果律或没有任何限制，相反，它以秩序为基础，并与因果法则相容。"① 因而，自由意志并不能成为评判道德能动性的先决条件。

其四，自律与自我立法（self-legislation）观念的实质是通过定言命令式的程序来使行动者履行责任，但其却有很高的执行标准与要求："在人们的目的和行动的选择中，自律必须被视为超越于我们偏好控制的条件……因而，它是一个优先原则的标准，只能在经验世界中依靠先验自由来实现。"② 这种能力连一般意义上的理性存在者都难以企及，因此完全基于自律能力去构建 AI 系统是不现实的。有学者对此指出，只要 AI 系统能够做到以下方面，就能够被视为是自律的道德行动者："机器系统可以不受任何其他行动者或者用户的直接控制……如果机器系统的程序和环境的复杂互动能够导致它以道德上有害或有益的方式行动，而且这些行动似乎是有意的和经过计算的……让人们能够通过假设机器系统对其他道德主体负有责任，并能对之赋予信念。"③

摩尔（James Moor）等人还据此对 AI 系统进行了层级划分："有伦理影响的行动者、隐式的伦理行动者、显式的伦理行动者以及完全伦理行动者。"④ 在此基础上，很多针对 AI 系统道德能动性的质疑针对的是完全伦理行动者，其余几种伦理行动类型依然可以成为 AI 系统追求的目标。因此，AI 是否具有道德能动性的争论焦点其实并不在于它是否必须要像人类那样拥有理解力、情感、自律和意志自由，而在于其是否能够对复杂情况进行判断，并通过调整和使用规则进行合理的道德决策。现有的研究已经表明：AI 已经可以通过动机—观察—推理—假设—验证来在复杂情形中自主地寻求合理性的解决方案。因此，一些学者提倡用一种意向立场来理解和对待 AI 的能动性，把它当作理性智能体来对待："这种立场也是人类理解自身和其他高等动物的方式，它也就自然地赋予了人工智能一种'类人的'而非'拟人的'属性以及真正的主体资格。"⑤

---

① 参见费多益：《意志自由的心灵根基》，《中国社会科学》2015 年第 12 期，第 57 页。

② Paul Guyer, *Kant's System of Nature and Freedom: Selected Essays*, Oxford: Oxford University Press, 2005, p. 126.

③ John Sullins, When Is a Robot a Moral Agent? *International Review of Information Ethics*, Vol. 6, 2006, p. 28.

④ 参见王淑庆：《人工道德能动性的三种反驳进路及其价值》，《哲学研究》2021 年第 4 期，第 123 页。

⑤ 吴童立：《人工智能伦理规范是什么类型的规范？——从〈儒家机器人伦理〉说开去》，《文史哲》2020 年第 2 期，第 57 页。

二、人工智能的道德规范性审视

在承认 AI 具备一定程度的道德能动性以后，要使其帮助人类提升道德，还需解决一个关键问题：它应该以什么样的形式，按照何种原则来帮助人们进行道德判断和决策？这涉及道德规范性问题的讨论，按照一种普遍的理解，规范性是对“人们基于什么样的理由行动才是道德的”这一问题的回答。目前学界对此主题有很多争论，其中最常用的一种方法是借助元伦理学中的人称问题进行探讨。

第一人称立场持有者认为，道德规范性必须在行动者自身的意志中寻求，行动者需要自己制定道德法则。如果一个行动者不能对自己的行为感同身受，并根据其来制定相应的标准，那么道德法则对他来说就是不起作用的。此理论对道德能动性的要求很高，行动者至少应该具备道德慎思（moral deliberation）的能力，凭此能力找到合理行动的理由，并根据这些理由选择适当的行动。相较而言，不把自己或他人的实践目的直接与行动者自身相关联的是第三人称立场。这种立场需要把行动者看作客观或中立的。在这种模式中，行动者总是被置于一个中立的位置，其行动理由都抽离于主体，而关乎一个客观的善的目的，这个善的目的赋予每一个行动者以同等的权威性，体现出不偏不倚性的要求：“在这个意义上，第三人称的规范理由独立于行动者，是一种外在的立法权威。”① 所以，这种立场总是在个体与个体之间寻求一种博弈的结果，它的规范性理由并不能基于某一个行动者的理由而提出，而是要在综合考虑不同欲求的条件下提出。

而介乎于这两者之间的是第二人称立场。其提出者达沃尔（Stephen Darwall）认为，这一立场能够有效地解决行动主体间的普遍权威问题：“一个第二人称的理由，它的有效性依赖于预设的权威和人们之间的责任关系，因此也依赖于理由在人与人间传达的可能性。”② 它可以用“我—你”的视角向自己提出要求，并用一种相互负责的观点进行审视。其规范性和权威性的来源“不是一个人希望或更愿意所有人做什么，而是一个人期望别人做什么，以及我们会同意任何人能够向作为相互负责的平等共同体成员提出的要求，亦

① 文贤庆：《三种人称立场对道德规范性问题的回答》，《道德与文明》2013 年第 4 期，第 31 页。

② Stephen Darwall, *The Second-Person Standpoint: Morality, Respect and Accountability*, Massachusetts: Harvard University Press, 2006, p. 8.

即我们对彼此负有的责任”。[①] 这种立场提供了一种“相互承认”的契约论道德模型,[②] 使得道德原则不只是约束行动者应该或者必须执行哪些行为,而是通过彼此之间的共担的责任与要求,让道德主体更为明确其行为规范,从而避免第一人称立场和第三人称立场的局限。那么,在 AI 增强的视域中,人工智能机器和道德行动者之间应采用哪种规范性人称立场才更为合理呢?

(一) 伦理决策机器的规范性疑难

依据学界现有的研究成果,AI 增强最常见的方法是应用伦理决策机器这种“直接道德增强”(direct moral enhancement)[③] 的方案。这种方案试图创造出具有自主判断意识的 AI 系统来对人类的道德行动直接进行指导,其“基于系统的设计者认为有效的道德概念,并以此概念将系统配置成能够指导人类的信念、动机和行动的事物”。[④] 它在完成初始的程序设定以后,包括所有设计人员在内的人类参与者都将成为“配角”。除了决定系统如何作出道德判断以外,人们无须在思考和行动上花费额外的时间和精力,只需简单地设定一定的程序,按照被告知的指示来行动即可。这一方法是将道德判断和决策的主导权完全交给了伦理决策机器,这亦是一种将道德规范性完全建立在人工智能机器上的“第一人称立场”。

因此,这一方案首先遭遇的疑难也就是第一人称立场所固有的。如前所述,这一立场依托一个完全理性的行动者,需要靠其理性慎思来进行道德判断与实践。这种慎思类似于康德无条件观念的推论,是一种理性的自我立法。然而,这种方式对于一般有限的理性存在者而言,要么是一种没有实质内容的空洞幻想,要么会错误地寻找到某个外在于反思的事物作为道德判断的起点。正如很多反对自我立法的研究者所言:“如果行动者对自身施加了一个原则(准则),那么很有可能他必须有理由去那样做;但是,如果存在

① Stephen Darwall, *The Second-Person Standpoint: Morality, Respect and Accountability*, Massachusetts: Harvard University Press, 2006, p. 34.

② 虽然契约论通常被视为政治理论或者隶属于政治哲学范畴,但在现当代的讨论中,它已进入伦理学的领域,比如罗尔斯和斯坎伦,他们就将契约论发展成为一种道德理论。他们的理论都建立在假想同意(hypothetical consent)的基础上,为每个人都可以合法地遵守基本社会规则提供合理的基础。

③ 在舍费尔看来,目前所有的道德增强技术都可以被划分为直接道德增强和间接道德增强。前者意指当特定干预措施旨在使某人的信念、动机和行为与增强者认为正确的道德信念一致时,它就是一种直接的道德增强;而后者则旨在使人们更可靠地产生道德上正确的思想、动机和行为,而无须指定这些思想、动机和行为的内容。参见 G. Owen Schaefer, Direct vs. Indirect Moral Enhancement, *Kennedy Institute of Ethics Journal*, Vol. 25 (3), 2015, p. 262。

④ Francisco Lara and Jan Deckers, Artificial Intelligence as a Socratic Assistant for Moral Enhancement, *Neuroethics*, Vol. 13 (3), 2020, p. 277.

一个在先的理由去采纳这些原则，那么，这一理由将不是自我给予或者施加的。”① 由此一来，AI 系统很容易从程序本身所确立的价值标准出发来进行道德判断，并有可能陷入“自负”的情境中。

并且，我们在第一节的讨论中已经明确了 AI 系统并不具备完全自律的道德行动条件，由此以第一人称立场为行动主体的系统设定就非常有难度。我们不仅不可能要求设计出来的 AI 系统兼具平等与自由、敬重感、自我同一性等这些很多人类都无法企及的特质，而且这种完全将决策权交给 AI 系统的举动，也会让人类的这些特质遭受侵蚀，从而无法达到增强的效果。此外，伦理决策机器要想获得第一人称立场的规范性，还需要将自己视为目的而不是手段，这种方法若得到全面应用，会给人类社会带来一系列的风险和挑战。康德人性目的论的最终归宿是要求行动者具有目的意识，并走向目的王国。如若将此意图赋予伦理决策机器，把决策权全权交给它来执行，这难免会为人类社会带来不可预知的风险与挑战。

（二）人工道德建议者的规范性问题

很多研究者转而提出一种“人工道德建议者”（avtificial moral advisors，以下简称 AMA）的方案。这项方案已经在一些场合得到了应用，如安德森等人设计的医用伦理专家（MedEthEx）、麦克拉伦设计的真话机（Truth-Teller）等。其核心是让 AMA 充当理想观察者的角色，从而在人们面对各种道德和利益冲突时给出合理的建议。因此，它“不仅仅是为人类在具体情境中提供超出人类认知范围的道德相关信息，进而帮助人类做出更恰当的道德行为。它的终极目的可以是帮助人类接近某种‘理想观察者’的境界，即一个近乎全知的存在，从而做出一个建立在信息零缺失、（基于某类原则）‘道德零失范’基础上的道德判断”。②

因而，AMA 的道德规范性更多依靠的是第三人称立场，它需要智能系统从一个客观中立的角度为人类道德决策赋予标准，让其能够处理信息。③从积极方面来看，如果这种系统掌握的信息足够全面，则能够从“上帝”视角来对人类进行指导，从而满足斯密“神圣诫命”的设想。并且，它还

① Terry Pinkard, *German Philosophy 1760 – 1860: The Legacy of Idealism*, Cambridge: Cambridge University Press, 2002, p. 59.

② 马翰林：《人工智能道德增强的限度》，《自然辩证法通讯》2020 年第 11 期，第 81 页。

③ 萨沃斯库和朱比里尼都对此机器进行了设想，其灵感来源于罗德里·弗斯（Roderick Firth），他提出了一个 AMA 的初级架构模型：“（1）对于非伦理事实无所不知（omniscient）；（2）无所不感（omnipercipient），亦即通过可视化、想象和同时使用信息的能力来认知这些信息；（3）不偏不倚；（4）无私、冷漠；（5）一致性；（6）在其他方面他是规范的。”参见 Roderick Firth, Ethical Absolutism and the Ideal Observer, *Philosophy and Phenomenological Research*, Vol. 12 (3), 1952, pp. 333 – 344。

能将道德原则的选取权交给人类使用者，我们可以根据 AMA 提供的“原则菜单”，依据自己的倾向来选取相应的行动，从而获得道德行动的自主权，也在这种过程中使自身的道德能力得以强化。

但是，这种方案依然会存在由第三人称立场所引发的规范性问题。因为该立场所给出的行动理由大多是认知而非实践意义上的：“作为给出理由的形式，它所表达的只是要求中立的行动者同意有一个做某事的理由，而并非在实践中同意实际地按此采取行动。可能做出的任何声明都只是针对中立的行动者关于实践理由的信念，而没有直接作用于行动者的实践理性或意志。”① 因此，AMA 在执行程序时也容易受到此立场的影响，从而只是在情形比较和分析中有很大的优势，但在指导实践方面则有所欠缺。而且，虽然我们可以从 AMA 所建议的各种价值序列中选取行动原则，这样一来看似获得了自主性，但当我们按照 AMA 的指示进行行动时，行动动机与责任认定却存在很大的疑问。

三、关系性视角：人工智能道德增强的前景展望

上述两种 AI 增强方案除了在规范性立场上存疑以外，还会在应用过程中遭遇一定的阻碍。有研究者就认为，这两种方案可能还无法达至使人类道德增强的效果。比如，在这两种方案中，个人的角色是比较被动的，伦理决策机器不用多说，AMA 也只是让个人选定自己希望采纳的价值序列。在这之后，人们唯一能做的只是决定是否接受道德机器测算后的结论。由于行动者不需要了解这些价值观和其系统决策之间的关联，一旦 AI 系统确定了标准，系统只会建议行动者做出符合这些标准的决定，而不是鼓励去质疑它们。个人或许可以随时对系统所提供的价值进行不同的排序，但他们不会更加深入思考，因此他们的道德增强程度是有限的。并且，这两种方案都容易陷入绝对主义的困境之中，并受到系统制定者的操控，最终让人类不仅没有在道德水准方面得到提升，反而会处于更深层次的风险之中。

那么，有没有一种方案既能够保证道德增强的效果，又可以在规范性问题上避免争议和指责呢？在前文中，我们已经论证了 AI 系统具备一定的能动资质，甚至可以在一定条件下采取与人类相似的认知和实践模式。只不过，它们的行动是由数据搜集、程序制定和内生性的推理模式所构成的。这些行动能力同样需要处理庞杂的价值观念和相互关系，不同的行动任务也需要它们从不同的环境和境况中提取特定的变量。因此，要想使 AI 系统真正

① 文贤庆：《三种人称立场对道德规范性问题的回答》，《道德与文明》2013 年第 4 期，第 34 ~ 35 页。

帮助到人类，必须要让它们走进人类世界，试着去理解人类的价值关切。“当人工智能进入人类的社群，它首先要观察社群的规范、理解社群的诉求、把握社群的内部和外部关系，这些是人工智能进行道德判断必不可少的训练。一个理想规范手册无法穷尽情况的复杂性，如同人类一样，人工智能的道德训练只能在大量的具体案例中，尤其是在互动中才能逐渐塑造出‘道德品格’。”①

与此同时，人们也需要在与 AI 系统的交互过程中去思考和反思，理解这样做（或不做）的理由是什么。因此，一种真正能够提升人类道德水准的 AI 系统应该是与人类处于一种第二人称的“关系型”视角当中。在“我—你”的交互过程中，增加他们对话的可能，以此来促使人类思考“合理行动的规范理由到底是什么”？而不是简单地对人类施加指令，或者预存一些制定好的价值菜单和序列，让人们只是去选择而不是深入体会。正如同第二人称立场所预设的那样，它的实践理由是相关于行动者，而不是中立于行动者的。它的规范性来源于预设的权威个人间的相互可说明性，亦来源于行动者的相互关系。因此，它需要具备一定的能动资质，像人类的合作者那样，与这个群体逐渐融合，通过一次次的互动来积累和调整自己的数据库，学习和估测服务对象的价值函数，从而不断改进和提高。这种增强的方案类似于《理想国》中的“助产术”，我们在劳拉（Francisco Lara）等人所创立的“苏格拉底助手式模型”② 的基础上，更进一步将其解读为“关系—辅助型”增强。

与之前的方案相比，这种增强模式有着如下三方面的优势：其一，它的规范性立场可以建立在第二人称基础上，根据行动者之间的“我—你”关系结构来确定行为动机和规范。它在明确 AI 能动性资质的前提下将其纳入与人类的关系结构中，根据彼此的行为和意志提出要求，达到承认和尊重，从而获取二者之间普遍的权威性。其二，它使行动者获得了更高的参与程度。在前述的方法中，人类主要依靠机器决策，纵使它们只是为人类提供了价值序列，但人类的参与度是极低的。与之相比，这种方案将重点放在了增强的过程而不是结果中。它的目的是帮助行动者进行伦理思维的提升和学习伦理思考模式和路径，而不是仅仅重视行为的后果。其三，这一方案还能对

① 吴童立：《人工智能伦理规范是什么类型的规范？——从〈儒家机器人伦理〉说开去》，《文史哲》2020 年第 2 期，第 57 页。

② 有学者质疑，劳拉的这种增强方案有着“道德判断无力”以及“无法对智能系统进行评价和奖惩”等方面的缺陷。参见尹洁：《弱道德人工智能可行吗——从精神医学用途到道德增强》，《医学与哲学》2020 年第 13 期，第 4 页。但我们认为，劳拉的方案虽存在一些瑕疵，但其规范立场并不是基于第三人称立场的。并且值得强调的是，人工智能在一定的情况下同样能够肩负起道德责任，从而可以对其进行奖惩。

AI 系统的责任进行明确规定。在此方案中，AI 系统同人类一样，需要对自己的行为和选择负责，只不过人类作为行动者负主要责任，AI 系统作为参与者和辅助者负次要责任。①

在此基础上，还值得思考一个问题：AI 系统应该从哪几个主要方面来指导行动者进行更为合理的道德行动？我们希望能够找到行动者在进行道德选择和判断的过程中，对其产生重要影响的潜在因素，并在制定人工智能程序时着重考虑如何将这些因素更多地纳入他们的“对话”之中，以此来使行动者的判断和实践更为合理，也利于改善自身的道德品质。在我们看来，如下几个要素是值得考量的。

首先是提升个体的逻辑能力。我们可以在人类与 AI 的交流过程中，让智能系统显示出每一个判断推理背后所蕴含的逻辑规则。通过这种数据化的分析形式，行动者可以明晰其论证中的逻辑缺陷，了解到底是什么导致他们推理中出现了无效的目的，以及明确意识到自己的推理因何而正确有效。对于智能系统而言，也需要根据一定的逻辑准则来告知行动者他们的判断、推理与实践是否保持了一致，有没有各种观点被整合在一起后出现不连贯的情况。

其次是提升定义的准确性。在道德领域中，只要涉及判断和决策，我们就必须对一些抽象的概念、理论和原则进行思考。因此，更为清晰的定义对于道德推理者而言非常重要。“在这些观念是道德观念的情况下，与之相关的理解力将会部分是抽象的——更清晰的概念理解将会对道德推理者有价值。这包括对道德观念的内容、强度和范围的清晰把握，以及有效传达对其理解的能力。”② 如果行动者模糊和歪曲了相关定义，将会导致不可靠的推论，从而引发与行动者审慎判断不一致的行为。因此，AI 系统应该设置出一定的警告程序，来严格要求使用者对原则、观念等准确定义。

再次是赋予其更多的常识。常识在道德判断中非常重要，其“涉及行动者所在的世界的各种事物以及一些经常性行为的结果，这些都是智力水平在平均线之上的人们应该知道的”。③ AI 系统应广泛搜集和整理各种常识观点中涉及的特殊事实。尽管这类知识与道德判断不直接相关，但掌握更多的常识可以提高行动者有效评估这些前提的能力，从而提升依赖于这些前提所做

① 在我们看来，惩罚的关键其实是在于对后续行为产生实际的影响，只要人工智能可以在公众面前接受监督，并根据审判结果去修正和调整自己的行为，惩罚对其就是有客观约束力的。

② ［英］欧文·舍费尔、［英］朱利安·萨沃斯库：《程序性道德增强》，黄各译，唐代兴主编：《哲学探索》2022 年第 1 辑，中国社会科学出版社 2022 年版，第 88 页。

③ John Rawls, Outline of a Decision Procedure for Ethics, *The Philosophical Review*, Vol. 60 (2), 1951, p. 178.

出的道德结论的可靠性："尽管道德判断不能仅通过事实和经验来充分论证，但当基于经验驳斥的前提时，道德判断则可以被判定为无效。由于可以访问并处理大数据，AI 具有很大的优势，可以建议行动者在没有经验事实支撑的情况下修正其判断，以此调整决策。"①

最后是增强引导性。此方案还需面对的一个场景是，个体行动者应该如何对待经 AI 系统处理后的意见和建议。在日常生活中，人们难免会出现意志薄弱的倾向，并且让一个执着的人放弃其秉持的理念也是困难的。当人们经常性的思维方式存在缺陷时，如果没有对修改意见保持开放的态度，任何形式的道德增强都注定会失败。这时，AI 系统只能尽其所能地保持耐心，并循循善诱。在这点上，AI 会比一般人类更为出色，它可以以个人比较能够接受的角度切入，还可以模仿人们喜欢的音色与其对话。

如果将上述这些方面都考虑在内，那么人类可以借助这种方案获取更多的益处，通过这种交流和互动所得到的建议可能比任何普通人类内部的交流都更有价值。而且，这种改变和增强的过程保证了人类的主体地位，因为 AI 只是一个高效的助手，它们只是帮助行动者做出完全属于他们自己的决定，人们也会为自己决策和判断的正确改变而自豪。

四、结语

进入 21 世纪，与人类未来命运休戚与共的莫过于以生物基因为研究对象的基因工程和以人脑为资源开发对象的人工智能。这二者"虽为人提供了许多便利，为全球经济复苏开辟了可喜前景，但也给原本不确定的世界增添了更多不确定性，从不同方面强化了全球性人口危机、环境危机和人的存在危机"。② 甚至，它们从一定程度上已经改变了人类社会的存在形态。如今，人类为了使自己的决策和行动更为合理，有研究者已经开始依托这两种技术来提升人类的道德实践。不过，基于生物基因视角的道德增强由于其技术手段涉及改变人体"内部"结构，即使出于自愿，这种增强方式仍会受到较大的质疑和挑战。而关系—辅助型 AI 增强由于采用的是一种第二人称规范立场的教化和引导的方式，让其看上去不失为一个正确的选择。

不过，随着此项技术的逐渐开发和应用，其给人类社会带来的不确定性是在不断增长的。因为 AI 系统极有可能在第二人称立场中获得与人类相应

① Francisco Lara and Jan Deckers, Artificial Intelligence as a Socratic Assistant for Moral Enhancement, *Neuroethics*, Vol. 13 (3), 2020, p. 283.

② 唐代兴：《基因工程和人工智能：人类向后人类演进的不可逆风险与危机》，《江海学刊》2020 年第 3 期，第 111 页。

的地位，从而独立于人类社会，并演化出与之完全不同的社会结构。因此，如何在设计过程中，让这一类“主体”与人类共处于一个社会道德规范体系之中，仍是值得我们深入思考的重大课题。由于本文所涉及的所有讨论几乎都是将 AI 视为一种具备一定能动资质的辅助服务系统，因而我们对其发展前景是持有乐观态度的。但如果不制定一定的规范，对其发展趋势作出一定限制，AI 也极有可能发生“变异”，从而对人类社会产生威胁，甚至把人类置于被奴役和被统治的境地。因此，对这二者应该以什么方式“共处”的探索可能还有很长的一段路要走。无论最后人工智能将会给我们带来怎样的改变，都需要我们不断去探索、应对、试错和挑战。以信息技术和生物工艺学为基础的“后人类”时代已经悄然来临，我们需要对此做好充分准备。

（责任编辑：周勤勤　李　涛）