哲学研究

人工智能的道德判断及其伦理建议

王银春*

［摘　要］学术界关于AI的道德判断主要存在三种不同的立场与观点。然而，对AI进行道德判断需要区分两个问题：其一是AI本身的道德评价问题；其二是AI研发与应用后果的善恶评价问题。二者既存在差别，又紧密相联。前者解决的关键在于AI道德判断的前提性规定，即AI是否真正意义上的人格实体与道德主体问题的讨论，而后者解决的关键仍在于人类自身。只有在人类与AI的关系等进行批判性反思的基础上才能对上述问题作出道德判断。具体而言，当前对AI应采取以下审慎态度：一是在大力研发和使用“弱AI”同时，应适当限制“强AI和超强AI”的研发和使用，尤其应当限制甚至禁止“杀人机器人”的研发与使用；二是国际社会、行业协会等应为不同发展阶段AI的研发与应用制定公约，确立行业发展伦理标准，并敦促相关法律法规出台，通过道德调整与法律规制引导AI健康发展；三是通过道德教育培养科技人员的道德自觉意识与社会责任意识。科技人员应当遵守科技伦理与职业道德，履行“通告和预防义务”。

［关键词］人工智能；心灵能力；理解能力；科技伦理；道德判断

2017年8月21日，人工智能（Artificial Intelligence，以下简称AI）公司的116位创始人和首席执行官们，给联合国写了一封公开信，希望禁止“杀手机器人（即自动化武器，如无人机、无人战车、哨兵机器人、无人驾驶军舰等）”的研发与使用。他们认为，这些机器人有成为恐怖武器的可能。独裁者和恐怖分子有可能利用这些武器对无辜人群使用，或以其他不可取的方式使用。然而，2017年10月15日，英国政府在发布的《英国发展AI》报告中指出，AI具有改善教育医疗保健、提高生产力的巨大潜力，可为英国释放6300亿英镑的经济活力，到2035年，AI可为英国经济增加额外收入8140亿美元（约合6300亿英镑）。该报告从数据、技术、研究以及政策的开放与投入等方面，对AI发展提出了18条建议，意欲使英国成为AI的世界领导者。那么，AI到底是造福人类的利器，还是毁掉世界的终极恶魔？对AI的善恶到底该如何评价？我们应当以何种态度对待AI的发展？

*哲学博士后，东华大学马克思主义学院讲师，200051。本文是国家社科基金项目“资本逻辑视域下的技术正义研究”（15BZX034）以及东华大学现代化与文明发展研究基地研究阶段性成果。

一、学术界有关AI道德判断的争论

AI的道德判断需要区分两个问题：其一，AI本身的道德评价问题；其二，AI研发与应用后果的善恶评价问题。目前有关AI的善恶评价可谓此起彼伏，莫衷一是。总体而言，学术界主要存在以下三种立场与观点：

第一种乐观主义立场。持此种立场的专家学者认为，AI只是一项手段与工具，本身无所谓“善与恶”“好与坏”，关键在于使用它的人类本身，对AI的未来发展前景持乐观主义态度，总体而言，AI的研发与广泛应用对人类发展利大于弊，能产生巨大的经济效益与社会效益。中科院自动化所复杂系统管理与控制国家重点实验室主任王飞跃研究员是此种立场与观点的代表，他认为AI在未来会拥有自己的第三个轴心时代。这个轴心时代是从哥德尔开始，现在是在哲学上突破，再是科学上突破，然后是技术上的突破。针对当前AI的广泛应用所带来失业危机的担忧，他乐观地认为，人们当前的工作正是依赖机器来提供的，人类向无用阶级的转变其实是社会的进步。一般而言，乐观主义立场大多是由一些与AI研发与应用相关，或出于自身利益考虑的AI界人士，或出于对科学技术盲目崇拜的科学主义者所坚持的，其缺陷在于片面地、孤立地看待AI的积极方面，比如能够产生巨大的经济与社会效益，重构包括金融、医疗、教育、交通等几乎所有的行业，从而推动人类生活方式的整体变革。他们有意或无意地忽视或者掩盖AI的消极作用，比如杀人机器人的诞生将会给人类带来安全威胁以及人类过度依赖AI文明可能造成人类文明的退化问题，等等。

第二种中立主义立场。持此种立场的专家学者承认AI本身存在“作恶”的可能，它的研发与应用对人类具有潜在威胁，可能带来严重后果，但基于一些理由仍大力支持发展AI技术。第一条支持理由，AI目前尚处于发展的初级阶段，其危害远远不够强大，所以不必过分担忧。例如，AI领域的全球领衔分析师汤姆·奥斯汀表示，霍金等“彻底开发AI会导致人类彻底毁灭”的警告是“非常愚蠢”的，其理由在“现在AI还很低级”。第二条支持理由，认为“人造的东西不可能超过人”，这一观点源于某种宗教情怀，即“造物主一定比所造之物高明”，所以不用杞人忧天。令人讽刺的是，本该最具无神论精神的科学主义者，却从宗教信念中寻求思想资源。第三条支持理由，认为“人类可以设定AI的道德标准”，但是从未对“AI必然会服从人类设定的道德标准”提供过有效的论证。还有人认为只要对AI进行道德教育，就可以保证他们一心向善，全心全意为人类服务。然而，问题是道德教育如何能够防止AI朝不道德方向发展。

第三种悲观主义立场。持此种立场的专家学者认为，AI不再是工具地位，自身具有生命意识与学习能力，在道德上具有“作恶”的两种可能：一种是AI的强大威力可能引发“人类作恶”，另一种是AI自身具有“作恶”的能力，而且人类对AI的“作恶”无法应对，最终将使人类走向虚无与毁灭。因此，他们表达了AI将来有可能失控或危害人类的担忧。2015年初，史蒂芬·霍金、比尔·盖茨和埃隆·马克斯等人签署了一封公开信，呼吁控制AI开发。马克斯认为AI会“唤出恶魔”，比核武器对人类的威胁还大。霍金则明确断言：彻底开发AI可能导致人类灭亡。上海交通大学江晓原教授也认为AI存在三个层面的威胁：首先，近期威胁。其一，AI开始大批取代蓝领工人和下层白领。如果达到某个临界点，社会就有可能发生动荡。其二，军用AI盲目研发。这一研发存在风险，一方面军用AI有可能直接操控武器，一旦失控，后果难以设想；另一方面即使没有失控，研发军用AI杀人武器，存在潜在风险。其次，远期威胁。AI对人类的反叛给人类所带来的威胁。最后，终极威胁。所有依赖AI的文明，因其极度单调无趣，人类因之失去生存意志，变得孱弱、颓废、没落而奄奄一息，最后终将走向灭亡。持悲观主义立场的人认为，技术并非使人类获得解放的道路，它“并非通过控制自然而从自然中解放出来，而是对于

自然和人本身的破坏，不断谋杀生物的过程将最终导致总体毁灭”[①]。上述观点的确具有合理之处，但有些观点也有待商榷，比如AI会造成失业问题，应当辩证视之。从人类历史进程来看，科技进步往往是伴随着失业同时发生的，四次工业革命皆是如此，现在AI也不例外，这是必然会发生的现象；而且，科技进步可以创造新的就业岗位，产生新的就业需求。随着AI的迅速发展，将会对AI人才产生大量需求，但目前此类人才非常匮乏。英国在其最新发布的AI发展报告中，建议政府、产业界和学术界打破成见，共同努力，扩大参与，支持AI硕士、博士学位人才的培养、鼓励发展AI MOOC和持续专业发展课程、设立图灵AI奖学金等具体措施，大力培养AI人才。反驳这一观点的学者认为，普通白领人才有可能通过培训，继续从事与AI相关的行业，但失业的大多数为蓝领工人，无法在AI创造的新就业岗位中找到匹配的工作。但是，AI会使很多人从劳动中解放出来，成为有闲阶级，则意味着对服务业的需求会更大，相应地创造的就业岗位会更丰富多元，蓝领工人也可以通过培训实现再就业。

总之，以上站在不同立场上提出的观点及其辩护都有一定合理性，也存在诸多缺陷，其中最大的缺陷在于对AI进行道德判断的关键问题没有解决，即未对AI本身善恶评价问题，与AI的研发与应用所带来的后果的善恶评价问题进行区分。后一问题解决的关键还是在人类自身。但是前一问题的解决，我们就不能基于现有的伦理道德框架对其进行评判，而应对传统的科技伦理进行批判性反思。人们通常认为，科技伦理是人类为了避免或减少科技滥用给自己带来的“负面”作用，从而为人们的科技行为制定出一套规范和一定的限制，以限制其发展的限度。这种观点只是简单粗暴地从既定的伦理道德立场出发，对科技行为的“是非与善恶”进行道德评判，并向科技提出规范性的呼吁或指令，而不对伦理道德体系本身进行反思、修正与发展。事实上，科技发展的许多重大问题，“已经不是传统的伦理道德框架能够提供某种有价值的答案，而需要对传统的伦理观念及其前提性规定进行批判反思”[②]的基础上，才能做出有效的回答。具体到AI的道德评判问题，则需要对“AI是否是真正意义上的人格实体与道德实体”进行批判性反思，而不是在某些细枝末节问题上纠缠不清，否则对问题的实质把握及其根本性解决是无所助益的。

二、AI是否真正意义上的人格实体与道德主体

2017年10月26日，在沙特阿拉伯首都利雅得举行的“未来投资倡议”大会上，“女性”机器人索菲娅被授予沙特公民身份，她也因此成为史上首个获得公民身份的机器人。这是在法律意义上首次承认机器人的公民身份，但是否也意味着AI已经是真正意义上的人格实体与道德主体？事实上，AI自诞生之日，就存在与之有不解之缘的问题：AI或机器人能否成为认识的主体？机器的形式化过程与人的思维在本质上是相同的吗？如果承认AI有智慧，这种智慧能够超过人类吗？[③]几十年过去了，相似的问题仍然在不断地被讨论，但是国内外学术界仍未对此问题做出具有充分说服力的解释与论证。在此我们尝试对此问题的解决提供一种参考思路：首先确定什么是人？其次，确定什么是AI？最后对AI的本质与人的本质进行比较判断？如果是，则确证了AI是真正意义上的人格实体与道德主体；如果不是，那他们之间的区别在哪里？

Flawell在“概念发展”回顾中开宗明义地断言，“为‘概念’术语寻找令人满意的定义是词典编撰者们的噩梦”[④]。《牛津英语词典》中将“概念”界定为“同一类事物的观念”。当然，这离概念的确切定

---

①［德］卡尔·雅思贝尔斯：《历史的起源与目标》，魏楚雄、俞新天译，北京：华夏出版社，1989年，第132页。

②李德顺：《沉思科技伦理的挑战》，《哲学动态》2000年第10期。

③［英］玛格丽特·博登：《人工智能哲学》，刘西瑞、王汉琦译，上海：上海译文出版社，2001年，第4页。

④J. Flavell, “Concept Development”, in P. H. Mussen (ed.) *Carmichael's Manual of Child Psychology*, New York: Wiley vol.1, 1970, pp.2—3.

义仍有一定的距离。Medin在近期“概念及其结构”一文中，将概念的定义向推进了一步，认为“作为观念的概念包括与之相关的所有的独特性”①，如果有新的异议提出“概念”被证明是不充分的，它只能是因为我们仍未构建出令人满意的概念构成理论。直到我们能够发现类似于人类、动物、植物、正义、真理以及概念自身等“概念和范畴的本质问题”，我们才能期待一种令人满意的“概念”定义。换言之，认识论的目标就在于发现“概念的本质属性”②。那么，AI是否是真正意义上的“人”的问题，就可以转换成“人的概念”与“AI概念”是否具有同一性的问题，或者“人的本质属性”与“AI的本质属性”是否具有同一性的问题。

《汉字源流词典》中将“人的概念”或“人的本质属性”界定为“能制造并使用工具的高等动物”③。当然，不同学科有着不同定义，不同语境下也有不同的引申意义。然而，在一般哲学意义上将人定义为“自由意志的存在”④。苏格拉底式的“人的概念”是指“一个对理性问题能给予理性回答的存在物，是‘有责任的（responsible）’存在物，成为一个道德实体”。⑤由此可见，苏格拉底意义上的人，理性与道德才是人的本质属性。但这个定义并未包括全部领域，它是不充分的，犯了以偏概全的谬误。因为“经验领域的丰富性、微妙性、多样性和多面性才是人之为人的特性”。⑥人类生存的基本要素是矛盾，人根本没有“本性”——没有单一或同质的存在，人是存在与非存在的奇怪混合。⑦因此，不同的哲学家开始从不同的经验事实来定义人的概念，尼采公开赞扬权力意志，弗洛伊德突出性欲本能，马克思则推崇经济本能。然而，这种从各自经验领域或学科研究角度出发对人的概念探究，与其说在寻求人的概念的普遍性，不如说在寻求一种特殊性。这种诸多特殊性的混合使人的概念开始“去中心化”、“去权威化”，更加呈现出混乱复杂状态，这种混乱直接导致了人对自我的认识迷失方向。卡西尔试图从诸多经验性的定义中找到人充分的“共同属性”、“本质属性”，他虽然承认“理性”是人的固有属性，但认为“人不再生活在一个单纯的物理宇宙之中，而是生活在一个符号宇宙之中，语言、神话、艺术和宗教则是这个符号宇宙的各部分”，因此，他提出应当把人定义为“符号的动物（animal symbolicum）”来取代把人定义为“理性的动物”。⑧符号化思维与符号化行为是人类生活的典型性特征，而且人类文化的全部发展都依赖于这些符号系统。

接下来，我们再考察一下AI的本质是什么？简而言之，AI即为“机器对心灵的解释”，属于认知科学范畴。如果说克隆人是对人身体的“基因复制”，那么AI则可视为对人心灵的“机械复制”，但AI并不是独立的“复制”技术，而是结合各个行业的大数据应用到各个具体任务中的一系列技术。在AI内部存在两大分支：一大分支是由图灵、冯·诺伊曼规定的以符号逻辑为基础的算法系统的传统AI，另一大分支是建立在统计分布规律之上的并行分布式系统，包括对大脑神经网络的模拟联结论AI。它在很大程度上弥补了传统AI的不足，具有容错能力和较强的学习功能。⑨AI与物理、化学等自然科学有着根本的区别，它的研究对象是大脑思维活动，是对与“存在、物质、客观”等相区别的“意识、精神、主观”等活动的研究。⑩那么，AI“思维”运行机制到底是怎样的？任何一个事物进入意识，都要借助于某种形式化过程。在这个意义上，人类思维的过程就是形式化过程，即一切被感知的事物，和由他们组合而成的复杂事物，以及被意识到的精神活动，都以符号和其他某种形式在意识中形成对应物（图形、声音等代号形式），即构成意识的基本材料，这些东西总是表现为一定形式的，我们思维活动就是

---

①D. L. Medin, “Structural principles in categorization”, in T. J. Tighe & B. E. Shepp (eds.) *Perception, Cognition and Development: Interaction Analyses*, Hillsdale, N. J.: Lawrence Erlbaum, 1983, p.1469.

②S. Shanker, *Wittgenstein's Remarks on the Foundations of AI*, London & New York: Routlege,1998, p.187.

③谷衍奎:《汉字源流字典》，北京:华夏出版社，2003年，第7页。

④《辞海》，上海:上海辞书出版社，1989年，第344页。

⑤⑥⑦⑧［德］卡西尔:《人论》，甘阳译，上海:上海译文出版社，2004年，第9、15、16、33页。

⑨⑩［英］玛格丽特·博登:《人工智能哲学》，第6、7页。

通过对这种素材的组织而完成的。那么，这就意味着“形式化的界限即为思维的界限”。无论是传统AI，还是联结论AI，都是源自麦卡洛克和皮茨“神经活动内在概念的逻辑演算”的设想，即根据事先规定好的运行方式，将某一过程形式化，并建立一种算法，将这一过程描述出来。任何事物，只要能够形式化，就可以由计算机来完成，其逆反推论也成立：任何不能形式化的事物，计算机都无法实现。因此，在这个意义上，“形式化的界限就是计算机的界限。”[①]在大多数学家看来，思维是在所有事物中最适于建立计算机模型的心理活动，而动机和情感是另外一类事物，人们对通过计算机模型来模拟或解释心灵普遍持怀疑态度。因此，在这个意义上，我们问“AI是否是真正意义上的人格实体和道德实体”？实质上是在问“AI是否有心灵能力、情感能力以及理解能力”？

心灵主义（Mentalism）是关于心灵问题研究最重要流派，它可以追溯到他的鼻祖笛卡尔，其著名的“我思故我在”的命题，已确立为近代唯心论的基本原则。在心灵主义看来，“心灵”是一种事物区别任何其他事物的名称，是我们身体的组成部分，是拥有者的私有物。人只能觉知到自我的心灵，不能觉知到他人的心灵，称之为“他心难题”[②]。他人的心灵是否真的存在，“我”无法确定，因之，“我”无法确定“你”与“我”是否为平等的、独立的个体。华东师范大学颜青山教授以此作为立论基础，推论出“我”也无法觉知到AI“心灵”，从而提出“机心难题”，所以不能证实亦不能证伪AI“心灵”的存在。那么，我们可以据此得出结论：无法证实亦无法证伪AI是否是真正具有心灵能力与理解能力。在心灵主义看来，心灵是人类特有的本质属性，因此可以进一步作出推论：无法证实或证伪AI是否是真正意义上的人格实体和道德主体。

一直以来，意向性（intentionality）[③]被视为AI与人最关键的本质区别之一。AI也能完成人较为高级的情感性思维活动，比如弹钢琴、写作、跳舞等，甚至有时候AI比人完成得更好，但是他们二者之间唯一的差别在于人有意向性，而AI没有。在大多数情况下，人的任一行为活动是在某种意识的引导下完成的，无意识的情况寥寥无几。而AI的任一行动都只是某种机械的运行过程，它的运行必须存在一个先决条件，即由人从外部输入其始发动力，即根据程序员所编写的程序发出的指令，或者程序员所发出的指令所产生的某个指令运行，并不是由AI自发产生的指令，这是一种机械式的思维活动，没有主观上的动机与目的，有的只是机械的思维与行为的模仿。我们经常赞扬AI具有“不知疲劳、不为情感所动、不会出现粗心错误、不受外界干扰”等优点，而这一切恰恰是不具备主观性、意向性的产物，比如“心绪不佳、感情冲动、注意力分散”正是意向性的产物。[④]即使AI出现偏离程序所预先设计的行为产生失控的情况，“一项程序所进行的许多推理链可能异常复杂，以致人无法遵从；此外，程序还有可能会接受意料之外的信息（来自电传、照相机，或传声器），所以，AI可能会使我们无比惊讶”[⑤]，这也是因为程序或算法出现纰漏，而非AI主观故意为之。因此，无论AI如何强大、如何智慧，都无从改变其所处的工具地位。AI是具有机械性思维能力的机器，不是具有道德地位的道德主体，也就意味着AI本身不能享有道德权利与义务，也不能成为承担道德责任的主体。当无人驾驶汽车失控的时候，所造成的人、财、物的损失，责任不应由无人驾驶汽车承担，而应由无人驾驶汽车公司，或者设计无人驾驶汽车的程序员承担道德责任和法律责任。当机器人杀人的时候，所带来的严重后果，责任同样应当由机器人公司或者设计机器人的程序员承担，而非机器人本身。

然而，客观唯物主义认为人的意识仅仅是大脑的物理、化学活动的宏观表现，“思维是大脑这台机

---

①④［英］玛格丽特·博登：《人工智能哲学》，第9、11页。

②笛卡尔主义原则：我们自身存在的自明性是坚不可摧、无懈可击的，近代哲学肇始于此原则，但是心理学知识的进展几乎根本无法证实笛卡尔主义原则，但在本文中仍然将这一原则作为心灵主义的立论基础展开论证。参见卡西尔：《人论》，第3页。

③意向性简而言之是指人的意志的指向性，是心理状态的某种特征，由于这种特征，心理状态指向或是涉及世界中的客体和事物状态。例如，信念、欲望和意图等都是意向性状态。参见［英］玛格丽特·博登：《人工智能哲学》，第7页。

⑤［美］亚当·库珀、杰西卡·库珀：《社会科学百科全书》，上海：上海译文出版社，1989年，第40页。

器的产物”[①]，意识在宏观上看来是不可思议的独特现象，其本质仍是大脑近千亿神经细胞（神经元）互动的电、化学活动。塞尔认为，根据直觉判断，神经蛋白可以生成意向性，而金属和硅则不能。[②]因此，客观唯物主义者认为人类一旦合成“神经蛋白”这种特殊物质，AI就可以生成意向性思维。对AI研究具有奠基性贡献的图灵也深信，不出意外，在未来某个时刻，机器将能够进行思维。施米德胡贝教授同样认为，不久的将来，人类将能够制造出基于神经网络的AI机器人，富有好奇心，具有学习、计划与推理能力，善于解决问题。第一个阶段是创造类动物的AI，下一个阶段将是类人的AI，而且进程会非常快，一旦成功，所有的文明都会改变，一切都会改变。“在先进技术的时代中，知识的增长，连同机器统治的扩大，看来是限制了人的潜力，尽管同时也充实了他。”[③]换言之，AI在未来某个时刻达到某种临界状态会觉醒，可以自组织、自我迭代，并呈现出人的某种意向性与目的性，甚至会超越人类。那么，在这个时候AI就会成为真正意义上的人格实体与道德实体。

综上所述，我们无法证实亦无法证伪AI是真正意义上的人格实体与道德主体。然而，根据客观唯物主义的假定，人类一旦合成能够产生意向性思维的神经蛋白，就能制造出类动物的AI，甚至是类人的AI。根据人权理论，无论何种方式出生的人，都应当具有平等的人权。[④]那么到这个时候，人与AI的关系将发生根本性转变，将从之前“主奴关系”转化为“平等主体关系”，甚至存在转化为“奴主关系”的可能。但是，至于未来是否能够合成产生具有意向性的思维的神经蛋白，未来AI能否具有心灵能力、理解能力以及情感能力，则是一个不确定性的可能性问题，需要未来科技发展的结果来检验。因此，对于AI本身的善恶、是非等道德判断的问题，这仍是一个“悬而未决”的问题，仍是一个开放有待继续讨论的问题。但是，我们能够确定的是，现阶段的AI仍然属于机械范畴、物质范畴、工具范畴，是机器对人类思维、行为的解释与模仿。这意味着现阶段的AI（弱AI）并不是真正意义上的人格实体与道德实体。因此，现阶段，我们仍需要将AI作为器物来对待，它与人之间的关系也并非平等主体之间的关系，仍然属于“主奴关系”。鉴于此，我们需要将AI本身的善恶问题悬置起来，明确意识到现阶段AI能够为我们提供什么，也要明确意识到他的局限及其潜在危险性，只有这样才能巧妙地避开既迷信AI又憎恨AI的谬误。

三、AI研发与应用的伦理建议

根据对AI道德判断的前提性问题，即“AI是否是真正意义上的人格实体与道德实体”进行批判性反思所得出的结论，现阶段，我们对待AI的基本立场仍是“人类中心主义”原则，主张用一种审慎态度来对待AI，既不将AI视为洪水猛兽，也不毫无保留地拥抱赞美AI，主张在不同阶段、不同层次用不同态度与方法来对待AI，并从伦理学的维度提出三个层面的具体建议。

其一，应当大力研发与使用“弱AI”，限制“强AI与超强AI”的研发与使用，尤其应当限制甚至禁止“杀人机器人”的研发与使用。正如文艺复兴时期的宣言：技术进步是不会停止的，人类只能适应这不可阻挡的进步。在海德格尔看来技术既不是技术的东西，也不是达到某种目的的手段，而是一种“去蔽”的方式，现代技术框架的本质并不是由人的主体性膨胀所致，而是来自一种不可抗拒的东西——“去蔽的命运”[⑤]，在这个意义上，任何试图以道德或人性观念来“控制科技进步”的企图都是徒劳无功的，因此，我们必须正视AI技术的发展。然而，几乎所有的技术都有利有弊，AI技术亦是如此。在专业

①② [英]玛格丽特·博登：《人工智能哲学》，第7、8页。
③ [德]卡尔·雅斯贝斯：《时代的精神状况》，王德峰译，上海：上海译文出版社，2013年，第17页。
④高兆明：《克隆人人格与权利问题研究——兼与甘绍平先生商榷》，《南京师大学报》（社会科学版）2005年第4期。
⑤ [德]海德格尔：《技术与转折》，弗林恩译，昆特尔·奈斯克出版社，1988年，第20页。

人士那里，人工智通常被分为四个级别：第一级别，没有学习能力（智能程度也可以相当高）；第二级别，具有固定的学习能力，而且习得的技能会趋于一个定值；第三级别，具有固定的学习能力，技能增长没有极限，但因为它的学习能力是事先设定的，所以被认为不会成长为超级AI；第四级别，具有无限的学习能力，可以成长为超级AI。[①]前两个级别通常被称之为“弱AI”，后两个级别分别被称之为“强AI”以及“超强AI”。超强AI可以“像人一样按照同样的步骤解决问题”[②]。“弱AI”对个体的日常生活，社会进步以及人类发展具有积极意义，尤其能产生巨大的社会和经济效益，应当大力发展。

“强或者超强AI”发展给未来社会带来的可能弊端：人的自然力下降，以及人的“虚无化”。虚无化意味着“人类自然性的萎缩和沦丧，而这恰恰构成个体自由的限度，甚至人类作为一个类存在的限度”，[③]可能加速“绝对的虚无主义”的到来，可能使世界成为一个“没有人的世界”，[④]而成为一个“机器人的世界”，人类文明也因之被智能文明所替代。“杀人机器人”正式名称为“致命自主武器系统”，它无需人类的帮助就可以确定目标、发动攻击并清除目标。因之就会衍生出一系列根本性的伦理问题：什么样的杀人行动是合乎伦理和道义的行动？如何区分与评价杀人行动的善与恶、对与错？谁对杀人行动的后果承担责任？杀人机器人虽然可以降低军人的风险，但却会因为这些“不知疲倦的战争机器”可能导致武装冲突变成无休止的战争。同时，杀人机器人的程序编程中无法具体区分平民、受伤、无作战能力和被俘的军人，而且杀人机器人没有怜悯、悔恨、愧疚等情感，无法掌握好战争中的适度问题，因此有可能带来更多更大的不必要的伤害。更为危险的是，杀人机器人一旦被暴君和恐怖分子所利用，将会变成人类的一场彻底灾难。因此应当限制“强或超强AI”的研发，尤其限制或禁止研发“杀人机器人”。

其二，国际社会、行业协会应当通过商谈达成道德共识，区分现阶段的AI（弱AI）与未来可能发展的AI（强AI），制定不同的研发与使用公约，确立行业发展伦理标准，并敦促相关法律法规出台，通过道德调整与法律规制引导AI健康发展。现阶段的AI（弱AI）作为工具价值与技术价值，本身并无危险，危险只存在人自身之中，可能使有益的、可控制的技术向有害的、无法控制的技术转变。因此，弱AI是中立的，具有两重性，本身没有目的，它就在一切善恶之彼岸，既有可能有助于人类的幸福，又有可能给人类带来灾难。正因如此，它才需要加强引导，用伦理道德与法律来规范科学技术发展，制定现阶段AI（弱AI）的研发与使用的伦理规范，确保“弱AI”为人类福祉而服务。

类动物AI，甚至类人AI（强AI）一旦研发成功，这类“强AI”本身及其后果的善恶需要重新审视与判断。但是，在现阶段我们仍需对其存在的风险提高警惕。江晓原教授在谈论安全问题的时候，有一个非常有价值的观点，安全并非纯粹的客观问题，它还需要主体参与构建，只有主体感到安全了，否则“安全”的建构尚未完成。因为AI的广泛运用，尤其“杀人机器人”的诞生，给人们带来的安全感的隐忧，是一种潜在的安全隐患，是不能在短时间内就能得到验证的安全威胁。“你如果‘主观上’缺乏安全感，那你的身心就会在‘客观上’受到影响和伤害”[⑤]。因此，为了防止风险失控，我们需要提前做好风险管控，所以就需要国际社会以及利益相关群体达成道德共识，在此基础之上制定不同层级的伦理道德规范。首先，联合国尤其是安理会常任理事国应采取务实的行动，从宏观层面共同规定AI研发与应用限度、界域，并根据AI发展的实际情况适时作出调整；其次，AI应当成立行业协会，通过行业内部协商共同制定AI行业发展的伦理标准，发挥行业协会的自律功能，并确立AI行业退出机制；最后，尽快敦促AI研发与应用的法律法规出台，发挥法律的震慑与惩罚功能，健全AI的他律机制，为AI的健康发

①江晓原：《警惕人工智能的五个“烟雾弹”》，《解放日报》2016年3月22日。
②S. Shanker, *Wittgenstein's Remarks on the Foundations of AI*, p.38.
③S. Shanker, *Wittgenstein's Remarks on the Foundations of AI*, pp.40—43.
④孙周兴：《虚拟与虚无——技术时代的人类生活》，《探索与争鸣》2016年第3期。
⑤江晓原：《当代科学争议中的四个原则问题》，《上海科技报》2013年11月8日。

展提供法律支撑。

其三，通过道德教育培养科技人员的道德自觉意识与社会责任意识。科技人员应当遵守科技伦理与职业道德，履行“通告和预防义务”。科学研究的学术价值与社会价值密切相关，有的直接服务于“企业的经济兴趣，甚至是政治上的需求和决策”①。学术价值与社会价值之间的关系存在三种情况，其一，学术价值与社会价值相吻合，内在具有一致性，比如杂交水稻对于解决人的温饱问题。其二，学术价值与社会价值在一定的条件下相吻合，比如核武器研究，如果将核武器作为保卫国家安全手段，对于其他国家或群体起到震慑这个意义上而言，其学术价值与社会价值是一致的；但是将核武器作为侵略其他国家或族群的工具，进行人类的暴力大屠杀，在这个意义上，其学术价值与社会价值是相背离的。其三，学术价值与社会价值完全相违背，比如利用人体进行细菌武器的研发，不仅违背社会价值，而且违背人类基本人权。总之，无论是从研究手段、研究目的、研究过程还是研究成果来看，科技活动不再是价值中立的行动，而是“同其他的人类行为一样受制于普遍的道德准则与规范”②。鉴于此，应当通过常态化的道德教育活动，培养科技人员的道德自觉意识，使他们对自身研究行为及其结果负有强烈的责任意识，一旦发现科研活动的社会应用会给社会带来弊大于利的后果时，就应当停止研究进程，同时，他们还负有独特的“通告与预防责任”，将其研究项目向有关当局或媒体进行通报。③在AI领域，科学家负有同样责任与义务，事实上，马斯克联名116名AI专家向联合国发出呼吁，禁止研发“杀人机器人”诞生，就是科学家们自觉遵循科技伦理、履行“通告和预防义务”的行为。

（**责任编辑**：杨嵘均）

Moral Judgement on Artificial Intelligence and Some Ethical Suggestions

WANG Yin-chun

Abstract: There are mainly three types of understanding of the moral judgement on artificial intelligence (AI). In this paper, we argue that the moral judgement of AI should be based on the distinction of two issues, i.e. the judgement on AI itself and the evaluation of AI development and application. These are two independent issues but at the same time they are closely connected with each other: The former concerns the premise of moral judgement on AI, i.e. whether or not AI is an entity with personality and a morality subject in a real sense; the latter's solution, however, depends totally on the human beings themselves. Moral judgments can be made only on the critical reflection on the relations between human beings and AI systematically. Specifically, prudent attitudes towards AI should be taken. First, 'weak AI' should be developed and applied vigorously, while 'strong AI' and 'super strong AI' should be restricted, and in particular the development of 'human killing robots' should be restricted and prohibited. Second, international communities and industry associations should formulate norms guiding the development and application of AI at different stages, establish ethical standards and relevant laws and regulations governing the development of AI, hence a healthy development of AI. Third, the technicians should cultivate the consciousness of morality and social responsibility, follow the scientific and technological ethics and professional ethics, and carry out the obligation of 'notification and precaution.'

Key words: artificial intelligence; mental competence; comprehension ability; scientific and technological ethics; moral judgement

①②③甘绍平：《科技伦理：一个有争议的课题》，《哲学动态》2000年第10期。