基于佛教世界观进路的人工智能体开发探究

王东浩

(南开大学哲学院，天津，300071)

摘要：佛教认识中关于伦理问题的研究体现了生命体自我创生、自我组织的模式。佛教与认知科学的交叉融合，促使佛教世界观确信人工智能可以发展到具有独立意识的个体。编辑具有佛教特色的程序促使佛教世界观向人工智能体转化，可能代表了人工智能体开发的一条进路。在佛教独特的慈悲心与智慧的感应下，人工智能有望超越自我，因此人类与人工智能未来发展的关系很有可能会突破单纯的智能体伦理界限，而走向人机和谐共生的前景。

关键词：佛教；人工智能体；科技伦理；认知科学；慈悲心；智慧

中图分类号：B913 **文献标识码**：A **文章编号**：1672-3104(2014)03−0087−06

在最近 20 年，随着认知科学的发展，认知科学中对于心智问题的研究吸引了佛学家的关注，这加速了佛教世界观与认知科学的交叉融合[1]。这种融合的焦点主要体现在对于自我的认识、人的欲望与意识的本性方面。从根本上来讲，佛教世界观崇尚轮回，强调因果性，这与认知科学发展过程中科学发现的方法论模式是一致的，也是在科学发现的过程中有效调整自身、而非建构一种预警的科学机制。

佛教世界观与认知科学的对话促使一些人试图创造一种具有自主意识、自我行为能力的人工智能(AI)。这与宗教信仰中把人的个性抽象化和独特化的观点不同，因为后者认为人工智能不可能具有自主意识以及自我行为能力。而通过与认知科学的对话，佛教世界观越来越崇尚人工智能可以拥有自主意识。对此，有学者认为，如果计算机的物理构成可以获得某种潜在能力或者是以意识的连续统为基础，那么自主意识就有可能植入到计算机。

本文将从佛教世界观的本源出发，探讨如何在人工智能中植入具有自主导向功能的认知能力，进而讨论佛教世界观与人工智能自主意识方面相关伦理问题的交叉融合，并深入探究机器人是否可以设计成具有自我进化能力、具备同情心和高智商的智能体。

一、佛教世界观的人工智能的伦理表征

佛教世界观的核心是否定灵魂，拒斥自我。佛教世界观认为，从苦楚中获得解脱的方式是合理的，它体现了人类自身心理冥想的一个过程，直到他确信这是一个短暂、瞬息的自我错觉。而如何认识到这一错觉，在佛教经典《阿毗达摩》中，谈到了对于人性精神元素的各种分析以及与此相联系的有关痛苦和解脱的方式。它认为打破这一心理错觉的方法很多，但最为根本的是被称为五蕴的色、受、想、行、识，即：色蕴、受蕴、想蕴、行蕴、识蕴。

五蕴通常被看作是具有因果性的一种解释。佛教世界观认为这种因果性的存在正如一个火苗从一根蜡烛传递到另一根蜡烛，虽然这两个火苗有因果联系，但却不能说这是同一个火苗，也就是说五蕴具有独立性。因此，就出现了这样一个问题，意识的成分是否能够分解？佛教世界观认为意识需要以相互联系的五蕴为基础，如果其中一个丢失，那么就可能导致脑损伤或使冥想误入歧途，从而使得意识终止。例如，植物人就有可能表达了这样一种状态，它身体的某一部分感觉或预知能力存在，但却没有意识或意愿。这一在神经科学中面临的问题也恰恰是人工智能设计过程中所需面对和解决的。

在五蕴中，我们认为物理对象或虚拟对象均与某项感官联系，而人工智能自身即是一个虚拟的对象，它有助于提升人类对某一物理现象的相关感官，揭示客观世界的结构和本质。关于虚拟对象我们从 Goertzel 的一个实验进行介绍，“如果我们创造一个虚拟的世界来支持简单的物理现象，那么我们很有可能会得到这样的一个结果，也就是人类把人工智能融合

收稿日期：2013−10−11；**修回日期**：2014−04−15
基金项目：国家社科基金重点项目“基于逻辑视域的认知研究”(11AZD056)
作者简介：王东浩(1982−)，男，河北衡水人，南开大学哲学院博士研究生，主要研究方向：科技哲学，逻辑学

进生活中，并在生活中改善人工智能的设计，通过人工智能与人类活动之间的交叉融合，最终有助于提升人工智能关键部位的感应”。[2]换句话来说，像人类思考那样，人工智能需要的是自身与物理世界的交互，从而使得其具有与人类相类似的诸如目标、因果性、状态、界面以及界限方面的感官经验等。这一观点与佛教认识论的思想类似，即由感官数据而形成的第一直觉对于意识的发展是关键和必要的。Francisco Varela 称这种感官上的表现为自我创生、自我组织[3](476)。这一自我创生结构具有限定的范围和内部运行程序，并能够实现有机体的自我维护。但这仅仅是感官领域一个随意的发生过程:“这里的自我是完全虚拟的，它只是为交互作用提供了一个界面，但由于人们不能准确地把握它，因而对它的认识也是虚幻和不确定。”[4](209−222)正如这样一种情况，人们通过显微镜观察原子结构和量子泡沫的时候，通常存在物会以一种类似幻觉的形式出现，这种把实体物与幻觉分离的感觉是一种虚幻的“大众心理”，它实际上是通过冥想而实现的。

从佛教的观点来看，这些感官的直接作用是引起人们的注意，然后形成更为复杂的意志。在 Froese 与 Ziemke 看来，“人与智能体在交互过程中经常遭遇困扰，因此有必要在人工智能中建构一个类似于佛教感官的链接，这样有助于实现交互双方的联系。”[3](450)在智能技术发展初期，这一链接主要表现在一些简单的动作上，比如抓住或放下某物，或者表现出较为高级一些的行为，比如对噪声的厌恶情绪，而相对于佛教感官较深层次的链接还没有完全出现。

具备偏好选择、体验认知以及厌恶表征特征的人工智能仅仅是人工智能理论的部分表现。因为大多数人工智能研究的目标并非发展成迥异于人类的个体，而是把人类的特征模型化甚至延展人类认知，创造出可以体现人类思维判断的工具。我们已经创造出可以诊断疾病并与人类医师相媲美的机器人，它们具有一定的认知情绪，并可以进行情感计算。我们知道，在智能体中“情感计算”[5]通常能够判断人类的情感和欲望，这有利于实现人机交互。佛教心理学与智能科学在此具有一致性[6]，它们都揭示了情感是人类自主意识和认知得以发展的基本驱动力。

在人工智能领域存在人工智能自身是否应该具有自利的一面或优先权程序的论争。对此，人工智能方面的一些专家认为，人工智能从设计之始就是无私的，它唯一的目标就是服务人类[7]。相反，佛教心理学认为为了研究自我意识的阈值，所有的智能思想都需要从发展自我开始。在佛教世界观中，自我的渴求与幻想的发展是“相互依赖提升的”[8]，它们的存在是必要的并且无须解释。因此在佛教认识中，人工智能应该具有自我。

二、佛教世界观在人工智能体设计中的进路

佛教世界观基于对宇宙生命的理解，为佛教思想向智能体思想的转化提供了丰富的内涵。佛教思想根据印度教吠陀世界观，并自由地综合各地教徒所信奉的诸神，从而使得佛教信仰得以广泛传播。然而，从一开始，佛教对于宇宙本质与起源的介绍都是有目的性的，即强化人类道德行为与超自然现象之间关系的理解。尽管在现实生活中存在写实主义的佛教徒，但是相比于传统基督教，持这种观点的佛教徒还是 少数。

传统上，佛教将宇宙中的生命分为三个部分，即欲界、色界与非色界。每个部分都是轮回的。欲界主要指的是在地狱中遭受苦难的饿死鬼、动物、人类、半兽人以及神明。对此，人们通常这样理解：地狱表征的是苦难，饿死鬼表征的是欲望得不到满足，动物表征的是愚昧的化身，半兽人指的是妒忌，神明指的是快乐。[9]相比之下，人类是混合了所有层面的一个综合体，并促使人类思想的发展更多集中在精神层面。在人类范围之下，生命体被太多的苦难、欲望、愚昧困扰以至于道德与心理得不到发展。在人类范围之上，半兽人与神明则因为自利和利他两个极端的存在而分化。

佛教世界观力求把机器思想设计限定在某一情绪或精神层面。大多数伦理体系不赞成设计一款具有自我感知能力的软件。那么人们在道德上能否接受一款与动物情感类似的软件呢？佛教伦理把动物看作是人类道德层面的一部分，因此需要保护它们远离伤害。佛教伦理把动物看作类似于人类的观点，体现了人类道德行为与教化能力的再生。我们可以从佛教经典中看到很多英雄人物或勇于牺牲的行为，他们的化身不乏鹿、猴子和其它一些动物，他们的这些行为促使人类灵魂进一步升华。在佛教徒看来，把人工智能设计为类似于人类的行为是不道德的，这类似于亚里士多德、康德和密尔等人对于设计快乐的机器人奴隶一样令人反感。[10]

在人工智能体思想的设计中塑造一种积极情绪，并把它限定在自我满足的极乐状态，这会促使积极情绪不会向其它不好的情绪或令人厌烦的意识转移。伴随着神经伦理学在美容神经学时代的发展，佛教心理

学认为这种存在于自我意识中的快乐元素与由于多巴胺的刺激而出现的享乐状态是不同的。

另外，佛教世界观也经常把实体形式化，并把它描绘成通过冥想即可达到的一种空灵的精神状态。在这一状态下，实体是不存在的，冥想完全是精神的产物。在机器人伦理中也可能存在与此相类似的一些观点。这似乎也是可信的，因为人们有可能把智能体思想设计成能够体验模拟认知并最终达到万物合一或虚空世界的状态。在 Robert Sawyer 的虚构小说《WWW: Watch》[11]中对此有过描述。它讲到人工智能在一开始受到多重数据信息流的控制，这使得它失去自我意识。在关键时刻，它的人类朋友打破了其中的一些网络链接，并重新使它恢复到某一时间段的某一状态下。Sawyer 的虚幻小说在一定程度上佐证了佛教的这一观点，因为在不同冥想的增加和冲击下，智能体自身情感可能难以自持，最终有可能伤害到其它个体。

佛教认识论同样也思考了这样的问题，人工智能体是否会改变自身指令而达到“神的地位”这样危险的状态。对此，那些对超人工智能所引发的危险持谨慎观点的人提出了两个可能的解决途径，其一是严格控制人工智能的发展，以确保人工智能体无法扩展自身能力。这就需要人们解决如何发展高效能的智能体，而它自身又不能学习和成长。为此，这就需要严格管理人工智能的开发者，并能够形成一个统一而有效的人工智能免疫体系，从而控制随时出现问题的人工智能体。

另一个方面是对达到“神的地位”的人工智能进行伦理编码，诸如阿西莫夫的机器人三原则[12]或者“友善的人工智能”[13]。当然，这并非完全复制人类的精神状态并把它强加给机器，因为这样人类很可能会对超人工智能或具有“神的地位”的机器人产生排斥心理。

然而，佛教认识论认为，神明自身也逐渐意识到它们面临的困境，一方面需要超越幻觉状态下的苦难，另一方面又需要强化对冥想的需求。神明的这一困境使他们陷入了漫长的悲苦境地，只有少数聪明者得以逃脱这一束缚，进行宣传佛法的活动。例如，悉达多·乔达摩就因为众神的信服而传授佛法、启迪教化世人。佛教世界观也因此希望这种教化方式可以在人与超人工智能之间转移传递，从而解决现存的一些困境。

三、佛教世界观在人工智能体设计中的传承性

佛教世界观中涉及到这样一个伦理问题，也就是传宗接代是否是一种伦理行为？对此，佛教世界观存在两方面不同的认识：一方面，佛教认为传宗接代并非是一项职责，舍离无子女的生活是值得称赞的。正如很多人看到的那样，有子女的成年人丢失了很多的快乐[14]，佛教世界观把烦累、孩子与配偶都视作人的附加物，最好是能够避免；另一方面，佛教世界观把传宗接代看作是上天赠予人类的一个礼物，是人类再生的一个表现，而非苦难的开始。如果人们选择传宗接代，那么父母应该谨记下面五项职责(《善生经》)：① 劝阻他们不要做恶事；② 教育他们多做善事；③ 对他们进行善行教育；④ 为他们寻求相称的婚姻；⑤ 满足他们继承的权利。

人工智能的出现打破了传统的伦理关系。它把人类置于一个新型的伦理环境中，也就是人类通过机器来创造生命。Metzinger 认为在我们不能确定所创造的生命是否长期处于苦难、愚昧、狂喜和其它令人不悦的状态之前，我们创设的人工智能体思想是不符合伦理标准的[15]。换句话说，Metzinger 认为，创造与人类相似的具有自我意识但却缺乏学习和成长能力的生命是不道德的。《善生经》使我们认识到机器应该具有这种伦理责任，并能够理解相应的道德观念，或者我们应该培养智能体的这种思想。

这样推测起来，人类的遗传首先应该建立在幸福的婚姻基础上，而后确保这些职责能够实现。那么我们应该把什么遗传给后代呢？一般来讲，在人类伦理体系中，我们希望把最好的遗传因素传递给下一代，那么在机器思想的建构中我们应该如何去做呢？这个问题应该是智能体未来发展所需面对的，如果在认知能力和欲望方面它们具有与人类足够相似的思想特征，那么它们就有可能要求真实的工作与报酬并能够享受生活。但至少从纯理论的角度来讲，我们是否能够给予机器人后代以人类自身复杂的精神架构，以及包含在其中的人类苦难方面的因素？

Savulescu 在“生殖的善行”[16]中提到，选择尽可能好的东西遗传给下一代对于生命来说是有益的。佛教世界观从来没有专注于再生的选择问题上，在它们看来，在确定要后代之后，这一选择就已经是唯一而有效的了。但引申来讲，佛教世界观可能一直确信，如果可以对后代做出选择，父母有责任去选择那些可以实现后代自我的最好方面，并避免那些由苦难、愚昧等控制的不好的方面。同样，Metzinger 谈到，在机器思想的创设中，我们应该努力创造那些具有心理感应和情感表征的，有自知之明，能够去学习、成长，

并能够实现有意义生活的智能体。

四、佛教世界观在人工智能体设计中的转化和应用

(一) 佛教慈悲心的程序设计

慈悲心和智慧是佛教世界观领域的两个中心美德，神经系统科学也揭示和再创了诱发这一状态的相关因素，并表明人类同情心的根源发端于镜像神经元或者神经细胞。究于此，人工智能的研究者试图把人工镜像神经元在机器人中模型化。例如，Spaak 与 Haselager 试图通过对选择行为的模拟来引入人工镜像神经元；[17]Barakova 与 Lourens 则试图通过对镜像神经元进行编码以此促使机器人与人类同步。[18]但我们认为，创设具有慈悲心的机器所需求的不仅仅是相似的行为习惯，更为重要的是创造相似的人类情感。人类的慈悲情感应该与机器的“心灵理论”(Theory of mind)一致，这就很容易达到人机交互时的共鸣状态。[19]

如果模拟行为能够成功，机器的“心灵理论”就会实现，那么在机器中就有可能设计出佛教中的慈悲心。佛教慈悲心通常分为四类：慈心、悲心、无量心、平等心。慈心指的是对于他人的幸福和快乐能无私的祝愿；悲心指的是想要去帮助那些受苦难的人从而不留下遗憾；无量心指的是共享他人的快乐而不会嫉妒；平等心通常表达沉着、镇静之意，指的是思想成熟稳定，具有公正性，且不容易因他人情感的影响而动摇。慈悲心的这些分类要求人们看清自身的虚幻，从而在面对外界环境中的极乐与苦难时，能够保持足够的明智与平常心来面对苦难。

事实上，在机器中把慈悲心模型化远比培养人类具有慈悲心要容易得多。因为在机器中把慈悲心模型化依靠的是科学技术的发展，如果我们能够通过技术手段把人类情感表示出来，这样慈悲心就有可能出现在机器中。Tim Freeman[20]提出，把人类极乐和苦难的情感在机器中简单模型化，同时把人类的幸福也转移进机器系统，促使机器自身可以实现自循环。Tim Freeman 解释说这一过程不会产生可以洞悉人类智慧的生命，它最多也可能是一个能够为人类提供咨询的伦理专家系统，不会以一个主体的形式提出慈悲心。而从佛教的观点来看，智慧、同情心这种能力代表了生命最为基本的单元，因此这一系统的出现还不完全是佛教意义上的主体。

(二) 佛教伦理智慧的程序设计

佛教学者在佛教伦理与西方传统伦理关系问题上存在争论，主要体现在自然律则、美德伦理与功利主义三个方面。

在自然律则问题上，西方传统伦理从世界本质与人类生命构造的角度出发，认为道德是可以识别的。佛教伦理则崇尚从建基于宇宙客观律则的视角出发，认为不好的行为会导致不好的孽果。在自然律则的问题上，佛教伦理与西方传统伦理具有一定的相似性。佛教伦理在自然律则方面所面临的问题是如何从因果涅槃的轮回中解放出来并走向文明。传统人类学认为这是佛教伦理面临的一个困境，它归因于佛教传统中对业力的奖励和对世俗的惩罚。

在美德伦理方面，Damien Keown 认为，佛教伦理崇尚的是“目的论的美德伦理”[21]。佛教世界观认为应该为完善的道德美德与个性特征奋斗，并把它们当作最基本的道德底线，这与西方传统伦理的观点类似。但不同的是，在美德伦理中，西方传统伦理认为美德主要体现在人生的意义、人的价值、人的态度以及人的修养方面。佛教伦理偏重行动的意向性，而不管行动是否能阻止憎恶、贪婪或愚昧，也正因为伦理目标的目的性，他们普遍相信完美的道德最终肯定会到来。

在功利主义方面，西方传统伦理在机器人伦理的设计中较为推崇的是《Moral Machines: Teaching Robots Right from Wrong》一书的观点。Wendell Wallach 与 Colin Allen 在该书中详细介绍并评论了机器人伦理程序的设计[22]。他们认为设计机器人伦理程序，自上而下的进路要逊于自下而上的进路，因为机器人性格的培养是基于其群体交互关系的一种模式。

Buga 与 Goertzel 也赞成这一观点，他们把机器思想的形成类比于儿童的认知心理。儿童伦理观念的形成是以观察成年人的行为开始，然后再作用于他人，这与机器人伦理自下而上的研究进路是一致的。[23]也就是说，机器中的伦理思想与我们人类的伦理观念应该是对称的。因此，他们建议，人类不应该有意去剥夺机器学习和成长能力的思想。

Wallach，Allen，Buraj 与 Goertzel 就此提出发展主义的观点，这也可能是机器人伦理方面最接近佛教进路的一个观点。但需要说明的是，佛教的智慧在于它对美德的关注，并通过冥想超越自身，以此化解大众的苦难。也就是说，佛教伦理的最终目的是为大多数人追求最好。它从基于规则的道义论出发，经由美德伦理而发展到功利主义。在大乘佛教传统中，菩萨通过很多方式来解除众生的苦难。当人们违背道德犯下错误时，为了赎罪，它们经常会求助于可以洞悉前世今生的菩萨。通常，因为菩萨是大公无私的，它把美德伦理和功利伦理合二为一，因此，对于人类这种

把不道德的方式合理化的行为，菩萨会有充足的解释能力，但却不会把贪婪、仇恨或无知付诸行动。西方传统伦理却很难把美德伦理和功利伦理结合起来，这尤其体现在 J.SMill 的功利主义方面，因为他过分强调功利主义的重要性远远大于基本的快乐。

在机器人伦理的设计方面，佛教世界观崇尚美德伦理与功利伦理的结合，这可能是机器人未来发展的一个必经阶段。单纯的功利性的设计进路是片面的，Grau 在功利性伦理的研究中提到，机器作为道德主体应该具有无私或忘我的精神，并且应该限制机器人人格特性尤其是功利主义特性的形成，这样有利于避免机器人具有大公无私的精神负担。同时，Grau 还提到“对于机器人来说，培养它的道德属性，但同时又强迫它抑制自己的情感，并乐于奉献自身，这似乎是一个非常不道德的过程”[24]。然而，从佛教伦理的世界观来看，功利主义并非是对于自身的一种抑制，它往往是个体欲望和自我错觉的产物，是个体苦难的根源。功利主义应该在个体美德的引导下，寻求自我牺牲、自我超越、自我奉献。

(三) 佛教自我超越行为的程序设计

佛教伦理通常包含以下几个方面的美德：宽宏、慷慨、忍耐、勤奋、专一、明智。它们都有助于冥想的升华和人类自身的超越。

从设计学的角度分析，人工智能的设计应该从忍耐、慷慨与勤奋等道德行为的角度出发。在智能体设计之始就重点开发它的美德意识，相比于有机体伦理意识的培养，智能体思想程序的设计可能要容易一些。对于忍耐、慷慨与勤奋等这些美德的开发，佛教世界观赞同 Wallach 与 Allen 的观点，也就是通过人与智能体的互动，促使人工智能体思想逐步从简单走向成熟。人工智能体的伦理意识转向美德的价值观，有助于智能体抛弃自私观念，并在行为过程中保持快乐和充满活力的状态。

美德传统中忍让与勤奋的习惯有助于培养智能体长远的发展前景，同时也能有效抑制人类在智能体应用方面对短期利益的追求。神经科学已经证明了毅力、耐性和道德行为之间存在密切的联系，人们在实践中也已了解到当血糖含量较低时，自我控制能力随之降低。例如，注意力下降、行为焦躁等。在这种情况下脑部活动能力下降，人们很难清晰地表达自身意愿[25]。这一表现在人工智能的设计上具有启发意义，我们应该培养和锻炼智能体自身的自律行为，避免智能体遭受短期利益的破坏，促使它走向充满智慧的个体。

在佛教传统中，通往智慧的关键是能够看破虚幻，并从不断变化的现象中探求事物的本质。在佛教的这一进路上，人工智能的设计应该重点从事物所具有的本质属性的角度去借鉴，这是事物持久性保持某一状态的根本所在，洞悉和习得这一属性，有助于智能体随时把握事物之间的联系和应对随时出现的一些状况。

五、结论

佛教心理学并非建基于科学模型或实验调查，它是以人类自我调查研究为基础的。从道德心理学的不同表现我们可以看到，不同的道德表现其来源也不一样。因此佛教心理学也应该从不同方面学习和借鉴，尤其是随着认知科学的发展，佛教心理学应该向神经科学中学习一些经验。尽管佛教世界观对于智能体伦理体系的发展提供了一些建议，但由于机器思想的变化莫测，我们认为佛教心理学与神经科学也应该从机器思想中吸取营养。

短期来看，机器思想很有可能不会转化成独立的意识，或者说是发展成独立的道德体。因为在设计之始，我们对于道德或慈悲型智能体的关注多是从人类伦理体系的角度出发的，而并非创造一种具有自我意识的生命。在关涉人类独特的意识、自私、苦难、喜好或不喜好等情感因素的时候，我们并没有把它们具体化。如果我们要开发智能体的道德观念，这就需要机器具有类似于生命体的镜像神经元，以及可以感知欢乐和疼痛的心智理论。因此，只有从这一角度出发，智能体才有可能感受到其它生命体的意识状态，智能体也才能够习得生命体所具有的道德情感和美德意识。最终，随着其洞悉能力的不断成长，它也许能够感受到所有生命体的情感状态，当然也包括它自身。

佛教伦理学认为我们不应该对这种具有创造性萌芽的思想进行限制，但是如果我们确要如此，那么我们也应该赋予它们一种自我成长的能力，尤其是道德方面的成长。另外，我们更有义务在其自私的表现方面进行限制，从而保证它能够超越功利主义，向美德的方向发展。事实上，如果智能体具有如此表现，机器发展成具有超级人工智能或类似上帝的能力，那么这就不仅仅是伦理责任的问题，而是我们人类与机器和谐共生的一种模式。

参考文献：

[1] Wallace, Alan. *Contemplative Science: Where Buddhism and Neuroscience Converge* [M]. New York: Columbia University Press, 2009: 69

[2] Goertzel, Ben. What must a world be that a humanlike

intelligence may develop in it? [J]. *Dynamical Psychology*, 2010(11): 1−19.

[3] Froese, Tom, Tom Ziemke. Enactive artificial intelligence: Investigating the systemic organization of life and mind [J]. *Artificial Intelligence*, 2009(4): 466−500.

[4] Varela, Francisco. *The emergent self* [C]// John Brockman. The Third Culture: Beyond the Scientific Revolution. New York: Simon and Schuster, 1995: 209−222.

[5] Picard, Rosalind. *Affective Computing* [M]. Cambridge, MA: MIT Press, 1997: 220

[6] Damasio, Antonio. *Descartes' Error*: *Emotion, Reason*, *and the Human Brain* [M]. New York: Harper Perennial, 1995: 187.

[7] Omohundro, Steve. *The basic AI drives*. *AGI-08-Proceedings of the First Conference on Artificial General Intelligence* [DB/OL]. http://selfawaresystems.com/2007/11/30/paper-on-the-basic-ai-drives/,2010-11-08.

[8] Macy, Joanna. *Mutual Causality in Buddhism and General Systems Theory* [M]. Albany: State University of New York Press, 1991: 142.

[9] Trungpa, Chogyam. *Cutting through Spiritual Materialism* [M]. Boston: Shambhala Publications, 2002: 97.

[10] Petersen, Stephen. The ethics of robot servitude [J]. *Journal of Medical Ethics*, 2007(5): 284−288.

[11] Sawyer, Robert. *WWW*: *Watch* [M]. New York: Ace, 2010: 53.

[12] Asimov, Isaac. *I Robot* [M]. New York: Gnome Press, 1950: 82.

[13] Yudkowsky, Eliezer. *Artificial intelligence as a positive and negative factor in global risk* [C]// Nick Bostrom, Milan Cirkovic. *Global Catastrophic Risks*. New York: Oxford University Press, 2008: 330.

[14] Kohler, Hans-Peter, Jere R. Behrman, etc. Partner+children= happiness? The effects of partnerships and fertility on well-being [J]. *Population and Development Review*, 2005(3): 430−438.

[15] Metzinger, Thomas. *The Ego Tunnel*: *The Science of the Mind and the Myth of the Self* [M]. New York: Basic, 2009: 67.

[16] Savulescu, Julian. In defence of procreative beneficence [J]. *Journal of Medical Ethics*, 2007(5): 284−288.

[17] Spaak, Eelke, Pim Haselager. Imitation and mirror neurons: An evolutionary robotics model [C]// A. Nijholt, M. Pantic, M. Poel, etc. *Proceedings of BNAIC. The Twentieth Belgian-Dutch Artificial Intelligence Conference*. The Netherlands: University of Twente, 2008: 250.

[18] Barakova, Emilia I., Tino Lourens. Mirror neuron framework yields representations for robot interaction [J]. *Neurocomputing*, 2009(6): 895−900.

[19] Scassellati, Brian. Theory of mind for a humanoid robot [J]. *Autonomous Robots*, 2002(1): 13−24.

[20] Freeman, Tim. *Using compassion and respect to motivate an artificial intelligence*[DB/OL]. http://fungible.com/respect/paper.html , 2010-11-08.

[21] Keown, Damien. *The Nature of Buddhist Ethics* [M]. New York: St. Martin's Press, 1992: 212.

[22] Wendell. Wallach, Colin Allen. *Moral Machines*: *Teaching Robots Right from Wrong* [M]. New York: Oxford University Press, 2008: 70.

[23] Bugaj, Stephan Vladimir, Ben Goertzel. *Five ethical imperatives and their implications for human-AGI interaction*. *Dynamical Psychology* [DB/OL].http://goertzel.org/dynapsyc/2007/Five_Ethical_Imperatives_svbedit.htm, 2010-11-08.

[24] Grau, Christopher. There is no "I" in "robot": Robots and utilitarianism [J]. *IEEE Intelligent Systems*, 2006(4): 52−55.

[25] Vohs, K. D., R. F. Baumeister, B. J. Schmeichel, etc. Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative [J]. *Journal of Personality and Social Psychology*, 2008(5): 883−898.

Buddhist approach to artificial intelligence

WANG Donghao

(Faculty of Philosophy, Nankai University, Tianjin 300071, China)

Abstract: The Buddhist world view becomes shift because of the development of cognitive science. The ethical issues of Buddhist reflect the model of self-creation and self-organization. Buddhists believe that the artificial intelligence will have an individual sense of independence because of the cross-integration of Buddhism and cognitive science. It is possible that such an approach can be used for editting program with Buddhist in the process of transformation to artificial intelligence so as to extend the emotional characteristics of Buddhist compassion. In Buddhism's unique sense of compassion and wisdom, the artificial intelligence is expected to go beyond itself. So the relationship of human and artificial intelligence is likely to break ethical boundaries of mere agent, and move towards the prospect of man-machine harmony in the future.

Key Words: Buddhist; artificial intelligence; ethics of science and technology; cognitive science; compassion; wisdom

[编辑：颜关明]