# 以人工智慧应对人工智能的威胁

**Developing Artificial Wisdom to Deal With the Threat of Artificial Intelligence**

汪凤炎 /WANG Fengyan　魏新东 /WEI Xindong

（南京师范大学道德教育研究所暨心理学院，江苏南京，210097）
(Institute of Moral Education Research, School of Psychology, Nanjing Normal University, Nanjing, Jiangsu, 210097)

**摘　要：**近年来人工智能的飞速发展引发了人们对其安全问题的关注与讨论。现有研究更多地从预防强人工智能、引导弱人工智能的角度提出应对方案，却很少有直面强人工智能的有效方案。本文从“智慧的德才一体理论”中获得灵感，主张将人工智能升级为人工智慧来应对这一威胁，以图灵智慧测验来检测是否达到人工智慧，并将人工智慧分为弱人工智慧与强人工智慧两种类型。最后探讨了人工智慧实现的具体路径。

**关键词：**智慧　人工智能　人工智慧　图灵智慧测验

**Abstract:** In recent years, with the rapid development of artificial intelligence (AI), its security problems have attracted many people’s attentions and provoked much discussions. However, more attention, in existing studies, was paid to preventing strong AI or guiding weak AI rather than directly facing the real threat from strong AI. The latest theory of wisdom of integration of morality and cleverness has shed light on this problem by advocating that AI should be upgraded as artificial wisdom. Only in the way of equipping such software with the corresponding hardware, and making it pass the Turing wisdom test successfully, can the AI be upgraded as artificial wisdom with the capability of integrating morality and cleverness. There are two types of artificial wisdom. Weak artificial wisdom is controlled by human beings, while strong artificial wisdom is not. According to this approach, specific ways of turning AI into artificial wisdom are discussed.

**Key Words:** Wisdom; Artificial intelligence; Artificial wisdom; Turing wisdom test

**中图分类号：**N0　　**文献标识码：**A　　DOI:10.15994/j.1000-0763.2018.04.002

人工智能（Artificial Intelligence, 以下简称AI）是研究并开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。[1] 近年来，随着计算机硬件和软件的突飞猛进，大数据的不断生成与拓展，以及深度学习算法的发展，AI取得一系列重大突破。其中围棋一直被视为人类智力的最后堡垒，如今却被人工智能攻破，AlphaGo横扫围棋界以李世石与柯洁为代表的世界顶尖高手，最近，最新版AlphaGo Zero不依赖人类经验，通过强化学习（reinforcement learning）与自我对弈（self-play），[2] 训练几天击败AlphaGo，引发了对人工智能安全问题的关注与讨论。[3] 霍金等人认为AI的全面发展可能会导致人类的灭绝。[4] 人类应该如何应对AI的威胁？国内学者从安全学的视角探讨预防AI不受控的情形发生，解决方案包含内部进路与外部进路：内部进路包括伦理设计、限定AI的应用范围及限制AI的自主程度和智能水平等；外部进路主要指依靠政府部门的监管及AI科学家的责任意识等。[5] 戴维斯（J. Davies）认为当前最为迫切的问题是如何管控好AI已经带来的现实问题，因为AI发展出意识、不受人类控制并不表明其发展出伤害人类的能力。[6] 鉴于AI在医院、金融等多领域的应用广泛，

**基金项目：**教育部人文社会科学重点研究基地2016年度重大项目（项目编号：16JJD880026）。
**收稿日期：**2017年11月18日
**作者简介：**汪凤炎（1970–）男，江苏南京人，南京师范大学心理学院教授，研究方向为中国文化心理学与智慧心理学。Email: fywangjx8069@163.com
魏新东（1991–）男，江苏宿迁人，南京师范大学心理学院心理学博士生，研究方向为中国文化心理学与智慧心理学。Email: weixindong@icloud.com

人们就需要对这一技术可能造成的对伦理道德、社会、文化以及政策制定的影响进行评估。克劳福德（K. Crawford）与加洛（R. Calo）提出AI的社会系统分析（Social-systems analyses），认为人们应该利用好哲学及法律等学科的研究成果，综合考察AI对社会文化与人们生活的影响，以应对AI带来的现实挑战。[7]上述学者从不同学科视角对AI可能带来或已经带来的威胁提出了应对方案，对今后AI设计、开发具有一定的指导作用，不过，它们或仅是主要针对弱人工智能提出的方案，或是从哲学层面提出的设想，可操作性弱。本文依据心理学对智慧的最新研究成果，尤其是根据汪凤炎提出的“智慧的德才一体理论”,主张将AI升级为“人工智慧”（Artificial Wisdom），提出用“图灵智慧测验”来判断AI是否升级为人工智慧，并以是否受人类控制为界将人工智慧分为强、弱两种类型。最后对人工智慧的具体实现途径进行探讨，以期通过一揽子解决方案，彻底消除各类水平的人工智能所可能存在的风险。

## 一、什么是人工智慧

依据智慧的德才一体理论，[8]-[10]可对人工智慧作出如下定义：AI一旦具有德才一体的性能，就升级为人工智慧。当人工智慧面临某种复杂问题解决情境时，就能让其适时产生下列行为：在“善”的算法或原则引导与激发下，及时运用其聪明才智去正确认知和理解所面临的复杂问题，进而采用正确、新颖（常常能给人灵活与巧妙的印象）、且最好能合乎伦理道德规范的手段或方法高效率地解决问题，并保证其行动结果不但不会损害他人和社会的正当权益，还能长久地增进他人和社会的福祉。那么，如何判断AI变成了人工智慧？ AI先驱图灵于1950年在其所著《计算机器与智能》一文里设计出图灵测验（Turing Test），用以检测机器是否具有了智能：测试人在与被测试者（一个人与一台机器）隔开的情况下，通过一些装置（如键盘）向被测试者随意提问。如果被测试者超过30%的答复不能使测试人确认出哪个是人、哪个是机器的回答，那么这台机器就通过了测试，并被认为具有智能。[11]参照图灵测验可以设计出图灵智慧测验，也叫“汪氏智慧测验”：测试人在与被测试者（从人类中推选一个公认的智慧者和一台机器）隔开的情况下,通过一些装置（如电脑）向被测试者提问。问过一些问题后，如果被测试者的智慧性作答中超过30%的答复不能使测试人确认出哪个是智慧者、哪个是机器的回答，那么这台机器就通过了测试，并被认为具有人工智慧。

人工智慧是以AI为基础，两者有三个方面的联系：（1）两者都表现为一种功能主义。功能主义认为心灵是一种功能，心灵在作用于主体的外部刺激和行为反应之间起一定的因果或功能作用。[12]AI功能主义体现在强调人脑功能与计算机功能相似，推向极端就是认为人脑不过是一台计算机，人的心灵不过是一种程序。[13]这与当代主流认知心理学相符，即将人脑视为一种信息加工系统。心理学家虽一贯重视研究心理的生物基础，在具身认知思想的影响下，[14]更加重视考察心理的生物神经学基础。体现在智慧心理学领域，一些人也试图揭示智慧的生理机制，目的之一是为了验证某种智慧理论，[15]不过，限于科技发展水平和研究范式，至今也没有弄清智慧在大脑中的生成机制。因此想要人工智慧通过模拟人脑产生智慧的机制来达成在目前仍不现实，目前人工智慧也体现为一种功能主义，即用计算机程序展现智慧。（2）两者都是一定的物理符号系统。纽维尔（A. Newell）与西蒙（H. Simon）认为一个物理符号系统对于展现智能具有必要和充分的手段：[16]一方面，任何一个展现智能的系统归根结底都能够被分析为一个物理符号系统；另一方面，任何一个物理符号系统只要具有足够的组织规模和适当的组织形式，都会展现出智能。心理学家对智慧大都采取一种兼顾认知与善的观点来探求智慧的成分与结构，如“智慧的德才一体理论”就将智慧视为“德”与“才”的统一。这种成分与结构的细化为用物理符号系统表征智慧提供了可能，而用机器与程序来实现人工智慧，则表明用物理符号系统来刻画又是一种必然。（3）两者判定方式相似：AI以通过图灵测试为标准，而人工智慧则以通过图灵智慧测试为标准。

AI与人工智慧的区别主要是因智能与智慧之间有本质差异所致。首先，AI可脱离人类而存在，人工智慧则不行。对于智能，动物可以有，[17]机器也可以有。莱格（S. Legg）与赫特（M. Hutter）提出“普遍智能”（Universal Intelligence）：智能是主体在一个广阔的环境中达成目标的能力。[18]这一定义基本涵盖自然界中所有行为主体的智能，

为人工智能提供了理论支撑。虽然这里提出人工智慧，但并不代表机器具有所谓“机器智慧”，而只是人类智慧在机器上的延伸。因为它诞生的主要动机是应对AI的威胁，想让AI行为符合人类的价值观念，并且智慧主要来源于后天对知识的学习与转化，所以智慧本身就具有一定的文化属性，不能脱离人类社会。其次，人工智慧中有“善”，而AI则是中性的。西方学者对智力（intelligence，即智能）的研究偏向于价值中立，认为智力是中性的概念，并无善恶之分，[19]智慧则是良好品德与聪明才智的合金。人工智慧依据智慧的德才一体理论，其中“德”即为“善”，而主流AI界依然坚持传统的技术中立论。[20]最后，AI与人工智慧对问题的解决方式不同。面对一个问题情境时，如果AI能够解决，其给出的解决方案往往是中性的或是最有效率的，但绝不会优先考虑善的解决方案；与此不同，假若人工智慧能够解决，那么，人工智慧一定会给出最善的解决方案，尤其是当有多种解决方案可供选择时更是如此。

## 二、人工智慧的类型

在AI发展初期，塞尔（J. Searle）首次将其分为弱AI和强AI，认为前者可以作为研究心灵的工具，对心智活动进行模拟；后者不仅仅是人类研究心灵的一种工具，被恰当程序设计的强AI等价于人类心灵。（[13]，p.417）郝格兰德（J. Haugeland）从研究进路的角度将AI分为“老而妙人工智能”（Good Old-Fashioned Artificial Intelligence，GOFAI）与“新式人工智能”（New-fangled Artificial Intelligence，NFAI），前者认为人类智能在很大程度上就是对物理符号的一种机械操作，或至少可以被分析为这类操作；后者用来表示不能被GOFAI囊括的各种研究进路，其典型为“人工神经元网络”（Artificial Neural Network，ANN）模型，也叫“联结主义”进路。[21]21世纪以来，随着来自社会各界的推动，一些其它领域学者对强AI可能导致的后果进行深入思考，波斯特洛姆（N. Bostrom）提出超强AI概念，认为超强AI在所有领域远远超过人类，从而会给人类带来存在性危险。[22]以上研究者从不同的角度对AI进行分类，限于研究旨趣，这里在借鉴塞尔和波斯特洛姆的人工智能分类思想的基础上，以是否受人类控制为界限，将AI分为弱AI和强AI两大类：凡是在人类制定的规则范围内行动、无法自定行动规则的AI，都属弱AI；反之，完全摆脱人类的控制且能够自定行动规则的AI，就是强AI。以弱AI能处理的任务种类为标准，又可将弱AI分为单任务弱AI与多任务弱AI，前者指只能处理一种任务的弱AI，后者指可以处理两种或两种以上任务的弱AI，它们之间的区别仅在任务种类的数量上存在差异。同样，强AI又可分为单任务、多任务以及通用型强AI，单任务与多任务与上述弱AI中相对应的含义相同，通用型强AI则指可处理更多甚至所有的任务，不过这里处理任务的能力与多任务或单任务强AI的能力并不是简单的量的差异，而是质的区别。

相对应可以将人工智慧也分为两大类：受人类控制的弱人工智慧与不受人类控制的强人工智慧。而弱人工智慧又可以分为单任务与多任务两类，强人工智慧又可分为单任务、多任务及通用型强人工智慧。弱人工智慧本质上依旧是人类的工具，而强人工智慧不受人类掌控。相对于人类智慧，强人工智慧主要有两个优势：（1）人类智慧会随着智慧者的死亡而在某种程度上丢失，这类智慧的延续依赖后来者对其生平言论、发明、发明或作品等的学习，由于言论、发明、发明或作品等的不完备性、默会知识会随个体的死亡而消失以及后来者受制于自身的知识经验、兴趣和时代的局限性等，因此这类智慧不可能得到完整且无偏差的还原；而强人工智慧因为拥有无限的“寿命”，所以它的智慧一旦生成，一般不会消失；退一步讲，即使遭到毁坏，其所生成的智慧依然可以通过“拷贝、复制粘贴”的方式放置到其它人工智慧中，从而得到完整的保存与延续。（2）人类智慧通常只是特定的一个或几个领域中的智慧，属于领域特殊性智慧（domain-specific wisdom），强人工智慧则可在几乎各个领域中展现出智慧，属于全知全能智慧（overall wisdom）。[23]

## 三、弱人工智慧的实现

人工智慧实现的基本进路为：在AI的软件系统中内置蕴含德才一体性质的软件，此软件一旦生成，除非重回原厂经公认的智慧团队对其升级或当AI生成强人工智慧后可自行升级外，任何人、任何病毒以及AI或人工智慧自身都无法对其降级、篡改或删除，也无法让其处于“沉默状态”而不“工作”，并配置相应的硬件设备，使其顺利通过图灵

智慧测验。弱、强AW的实现分别以弱、强AI为基础，不过具体路径存在一定的差异。

研究者针对现有弱AI提出的“AI的社会系统分析”、（[7], p.313）“道德机器”（moral machine）[24]或是倡导为现有AI编写伦理代码（[6]，p.291）等措施本质上与弱人工智慧中的“德才一体”软件所起的作用相同，都是在弱AI完成任务的过程中考虑社会与道德因素。弱人工智慧中的伦理道德设定可分为两类，一类是面向整个系统的道德法则，无论是单任务还是多任务系统，在任何时刻都应该遵守的法则，这一类法则理应体现“人类中心主义”，著名科幻小说家阿西莫夫提出的“机器人三法则”就符合这一要求，具体内容为:（1）机器人不能伤害人类，也不能在人类受伤害时袖手旁观;（2）在不违反第一条法则的前提下，机器人应该服从人类的一切命令;（3）在不违反前面两条法则的前提下，机器人应确保自身的安全。[25]另一类是面向特定任务的道德规范，以“自动驾驶”这一弱AI为例，当自动驾驶汽车不得已面对类似“电车两难”这一道德困境时，它应如何抉择？这里就需要针对具体的情境，设定具体的道德规范，这里规范的前提是以不违反第一类面向系统的法则为基础。

除设定具体的法则与规范外，要体现出智慧，“德才一体”软件还要为弱AI所面临的任务设定一个可以体现德才一体的目标。对于任务本身，可以分为“目标界定精确”（well-defined goals）任务和目标无法精确界定任务，前者包括具体的规则与精确的目标，例如棋类游戏；后者由于目标本身涵盖广泛无法对其精确界定，例如“识别一只狗”，由于狗的外形多变，品种丰富，无法找到一个可以精确量化的标准。对于目标精确任务而言，通过强化学习，可以不借助人类经验，例如AlphaGo Zero仅仅通过“自我对弈”的方式就在围棋上达到超人水平，而对于目标无法精确界定任务而言，则离不开人类经验的参与。人工智慧主要面对的是复杂任务，一般而言不是目标可以精确界定的简单任务，具体而言，依据德才一体理论，所谓能够体现“德才一体”的目标，就是指在高效率完成任务以及不损害相关人员以及社会正当权益的基础上，增进他们的福祉。由于这一目标的“模糊性”，就使得“德才一体”软件还要为相关任务提供必要的人类经验，即相应知识与智慧案例。这些知识与智慧案例，一方面要保证生成方案的可行性，另一方面要体现创造性。与法则和规范在弱人工智慧中所扮演的角色类似，可行性的知识是面向系统的，而能够体现创造性的知识则是面向具体任务的。可行性以常识性知识来保证，包括“朴素物理学宣言”[26]中计算化的人类日常物理学知识与社会文化常识，有“深度学习教父”之称的欣顿（G. Hinton）预测在不远的将来将实现“具有常识”的计算机系统。[27]可行性知识同时也起到让弱人工智慧先行判断任务能否完成的作用，若不能则停止对任务的加工，反之则进入下一步，提取能够体现创造性的相关知识及案例。一般来说，对单任务人工智慧只要赋予其该领域的智慧案例与相关知识即可，而对于多任务的人工智慧而言，除了要赋予其各个相关领域的智慧案例与知识外，还要考虑到不同任务之间的交叉领域的相关案例与知识。对于案例的选择，则在智慧的德才一体理论的指导下，遵循心理测验的标准模式，让多位评价者进行综合评价，在此基础上进行筛选，以避免软件设计者个人或单个团队因素的影响。另外，虽然有大数据技术保证案例数量，考虑到日新月异的社会，软件在面对新问题时依然存在无法解决或不是智慧解决的可能，因此有必要对软件的解决方案与具体任务成果进行反馈性评估，通过评估结果来不断调整软件对案例的认知，以达到“训练”软件的目的。无论是弱AI还是弱人工智慧，它们的规则与目标由人类来设定，相关知识由人类赋予，行为结果由人类来评估，体现出两者完全在可控范围内，并不会给人类带来存在性威胁。

## 四、强人工智慧的实现

虽然对强AI能否实现一直存在争议，但目前一些AI研究者从哲学、未来学、神经科学等角度向我们论证了其实现的可能性，例如徐英瑾认为在维特根斯坦哲学的启发与指导下，结合“非公理化推理系统”（Non-Axiomatic Reasoning System）可以开发出不同形式的通用智能系统，实现强AI；[28]库兹韦尔（R. Kurzweil）认为2045年到达奇点，到那时非生物智能在这一年会10亿倍于今天的人类智能；[29]众多神经科学家及机器学习专家在“皮层神经网络机器智能”（Machine Intelligence from Cortical networks, MICrONS）项目上，即主要绘制啮齿类动物大脑皮层结构与功能图谱，已取得突破性进展，为“下一代人工智能”提供理论计算构件的原理。[30]霍金等人对AI的担忧实质上是对强AI的担忧，下面就在强AI基础上来具体探讨强人工智慧的实现。

简单通过模拟弱AW的路径来实现强AW是不可行的。一方面，虽然强、弱人工智慧的最终目标都是“德才一体”地完成任务，但是并不能由人类来为强人工智慧设定目标，主要因为人类对问题的考量并不一定比强AI全面，所以其设定的目标也不一定是最佳的。最好是让强人工智慧自行生成目标。另一方面，为“德才一体”软件中“灌输”智慧案例也是不必要的，一是因为强AI可以通过某种手段自行获取这些必要的知识；二是人类并不能阻止强AI自行获取其它知识或案例，这其中就可能包括愚蠢案例。

无论是让强AI生成“德才一体”的目标，还是让其能主动获取与学习智慧而不是愚蠢案例，本质上就是赋予强AI道德判断能力，解决强AI的道德性问题。目前将人类道德属性赋予AI主要包括从上而下与从下而上两种路径，前者指将一种道德规则转化为某种算法用来指导AI的行为；后者指模仿人类道德发展模式，通过不断学习来具备道德判断能力。[31] 应以自下而上的路径来设计强人工智慧中的“德才一体”软件，一方面因为人类的发展性与当前知识的局限性，无法为强人工智慧构建出一个永恒不变并且对人类长期有益的道德规则，另一方面虽然强AI足够强大，能够自定规则并且脱离人类掌控，但它并不能摆脱自然法则的约束，其能力的发展必然也要遵循一个从无到有、从低到高的阶段。自下而上的路径本质上就是将道德规则从“是”的层面转化为“应当”，弱人工智慧是在“是”层面上接受具体的道德法则与规范的限定，而强人工智慧则变为在“应当”的层面上对道德进行自主的认知加工。那么机器能否实现这种转化呢？换句话说，对于“绝对律令”的执行能否还原为一定的物理符号系统？如果否认这种可能性，等于认为“应然”的发生可以绕开“实然”的基础，这显然不符合常识，（[28]，p.93）因为道德律令的产生，离不开大脑的物质基础，对其的执行更离不开个体与环境的交互。柯尔伯格（L. Kohlberg）等人对道德发展的研究展示了这一转化过程，柯尔伯格运用道德两难法，经过一系列的实证研究，将个体的道德发展划分为“三水平六阶段”：前习俗水平、习俗水平、后习俗水平及服从与惩罚阶段、相对的功利主义阶段、人际和谐或好孩子阶段、维护权威或秩序阶段、社会契约阶段、普遍伦理原则阶段。前习俗水平根据行为直接后果和自身利害关系判断好坏是非，习俗水平根据行为是否符合他人愿望，是否有利于维持习俗秩序进行道德判断，后习俗水平指能摆脱外在因素，着重根据个人自愿选择的标准进行道德判断。[32] 这种最终发展为“普遍伦理原则”的过程实质就是由“是”转化为“应当”的过程。

必须确定的是，强人工智慧所拥有的最低道德水平至少要达到柯氏所讲的习俗水平的第二阶段，即“遵守法规取向”。并且，这里的“法规”不是指某个国家或地区所制定和认可的法规，而是指对人类都有利的法规。但柯尔伯格的理论主要考虑的是道德中的“公正”，忽视了“关爱”与“宽恕”。吉利根（C. Gilligan）通过实证研究论证关爱取向与公正取向是人类两种不同但是同样重要的道德价值取向，并提出关爱道德取向三水平：自我生存定向、善良、非暴力道德。恩赖特盖尔（R. D. Enright）等人通过对宽恕的研究发现，类似于柯尔伯格道德发展阶段，宽恕理由也可以对应分为六个阶段：报复性宽恕、归还和补偿性宽恕、预期的宽恕、合法的宽恕、和谐需要的宽恕、爱的宽恕。[33] 结合关爱与宽恕取向的研究，在强人工智慧的最低道德发展水平上可定为，在遵守法规的前提下，关心人类尤其是关心人类的情感，合法地宽恕人类，相应在可以预见的第六阶段上有一个“第七阶段”，即遵循普遍伦理原则的前提下，无条件地宽恕、关爱和公正地对待人类中善良的群体与个体。

强人工智慧与弱人工智慧所遵守的法规有以下两点不同：（1）因为弱人工智慧中的法规由设计者领导的智慧团队制定，其内容一方面受制于设计团队的背景与经验，另一方面还要考虑到所在国家的法律法规；而强人工智慧的法规因为要面向全人类，因此就需要类似联合国这样的国际组织共同商议制定。（2）弱人工智慧中的法规设定后，除非经由专业的智慧团队人为改动，其自身只能遵守，无法取消，亦无法自行升级；强人工智慧的法规只是最低水平，其未来可能会在“遵守法规取向”阶段保持稳定，或走向柯氏理论中的最高阶段以及可预见的“第七阶段”，但不能倒退或取消。

依据不确定性原理（uncertainty principle），避免产生确定性效应。现在无法确定的是强人工智慧最终发展的道德水平，目前可以预计到的是“第七阶段”，不能确定往后是否还会有更高的阶段。未来无法确定的相关事件有二：（1）可能发生有人或强人工智慧自身试图篡改所处的最低道德阶段状态，即取消这一设定或让其倒退到低阶段。（2）

随着人类的发展可能发生现阶段“法规”与那时人类整体利益相冲突的情况。对前一种情况的应对措施是启动“德才一体”软件中的自毁装置，使其毁灭；对后一种情况的应对措施是应保留发展到柯氏后习俗水平或更高水平的强人工智慧，对于仍然处于第四阶段的强人工智慧依旧启动自毁装置，并保证日后生产设计的强人工智慧中的“法规”适应那时的人类整体利益。这样便可有效解决强人工智能带来的威胁。

## ［参 考 文 献］


[1] 王晓阳. 人工智能能否超越人类智能 [J]. 自然辩证法研究, 2015, 31（7）: 104–110.

[2] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. 'Mastering the Game of Go Without Human Knowledge'[J]. *Nature*, 2017, 550(7676): 354–359.

[3] 朱滢. 怎样面对来自人工智能的威胁？[J]. 心理与行为研究, 2017, 15（1）: 2.

[4] 翟振明、彭晓芸. “强人工智能”将如何改变世界—人工智能的技术飞跃与应用伦理前瞻 [J]. 人民论坛·学术前沿, 2016,（7）: 22–33.

[5] 杜严勇. 人工智能安全问题及其解决进路 [J]. 哲学动态, 2016, 9: 99–104.

[6] Davies, J. 'Program Good Ethics into Artificial Intelligence'[J]. *Nature*, 2016, 538: 291.

[7] Crawford, K., Calo, R. 'There is a Blind Sport in AI Research'[J]. *Nature*, 2016, 538: 311–313.

[8] 郑红、汪凤炎. 论智慧的本质、类型与培育方法 [J]. 江西教育科研, 2007, 5: 10–13.

[9] 汪凤炎、郑红. 智慧心理学的理论探索与应用研究 [M]. 上海：上海教育出版社, 2014, 189.

[10] 汪凤炎、郑红. 品德与才智一体：智慧的本质与范畴 [J]. 南京社会科学, 2015, 3: 127–133.

[11] Turing, A. M. 'Computing Machinery and Intelligence'[J]. *Mind*, 1950, 59(236): 433–460.

[12] 唐热风. 论功能主义 [J]. 自然辩证法通讯, 1997,（1）: 6–12.

[13] Searle, J. 'Minds, Brains, and Program'[J]. *Behavioral and Brain Sciences*, 1980, 3: 417–458.

[14] 叶浩生、麻彦坤、杨文登. 身体与认知表征：见解与分歧 [J]. 心理学报, 2018, 50（4）: 462–472.

[15] Sanders, J. D., Jeste, D. V. 'Neurobiological Basis of Personal Wisdom'[A], Ferrari, M., Weststrate, N. M. (Eds) *The Scientific Study of Personal Wisdom: From Contemplative Traditions to Neuroscience*[C], New York: Springer, 2013, 99–112.

[16] Newell, A., Simon, H. 'Computer Science as Empirical Iquiry: Symbols and Search' [A], Haugeland, J. (Eds) *Mind Design II: Philosophy Psychology Artificial Intelligence*[C], London: The MIT Press, 1976, 35–66.

[17] Zetall, T. R. 'Animal Intelligence' [A], Sternberg, R. J. (Eds) *Handbook of Intelligence*[C], New York: Cambridge University Press, 2000, 197–215.

[18] Legg, S., Hutter, M. 'Universal Intelligence: A Definition of Machine Intelligence'[J]. *Minds and Machines*, 2007, 17(4): 391–444.

[19] Sternberg, R. J. 'A Balance Theory of Wisdom'[J]. *Review of General Psychology*, 1998, 2(4): 347–365.

[20] 洪小文. 我们需要什么样的机器人 [J]. 中国计算机学会通讯, 2014, 10（11）: 50–54.

[21] Haugeland, J. 'What is Mind Design?' [A], Haugeland, J. (Eds) *Mind DesignII: Philosophy Psychology Artificial Intelligence*[C], London: The MIT Press, 1997, 1–28.

[22] 波斯特洛姆. 超级智能：路线图、危险性与应对策略 [M]. 张体伟、张玉青译, 北京：中信出版社, 2015, 143.

[23] 汪凤炎、傅绪荣. “智慧”：德才一体的综合心理素质 [N]. 中国社会科学报, 2017–10–30（6）.

[24] Wallach, W., Allen, C. *Moral Machine: Teaching Robots Right from Wrong*[M]. New York: Oxford University Press, 2009, 10.

[25] Murphy, R., Woods, D. D. 'Beyond Asimov: The Three Laws of Responsible Robotics'[J]. *IEEE Intelligent Systems*, 24(4): 14–20.

[26] Hayes, P. 'The Naïve Physics Manifesto' [A], Boden, M. (Eds) *The Philosophy of Artificial Intelligence*[C], New York: Oxford University Press, 1979, 248–280.

[27] Lecun, Y., Bengio, Y., Hinton, G. (2015). 'Deep Learning'[J]. *Nature*, 2015, 521(7553): 436–444.

[28] 徐英瑾. 心智、语言和机器——维特根斯坦哲学和人工智能科学的对话 [M]. 北京：人民出版社, 2013, 427.

[29] 库兹韦尔. 奇点临近 [M]. 李庆诚、董振华、田源译, 北京：机械工业出版社, 2012, 85–122.

[30] Underwood, E. 'Barcoding the Brain'[J]. *Science*, 2016, 351(6275): 799–800.

[31] 王东浩. 机器人伦理问题研究 [D]. 南开大学, 2012.

[32] 汪凤炎、燕良轼、郑红. 教育心理学新编 [M]. 广州：暨南大学出版社, 2016, 190–191.

[33] Enright, R. D., Santos, M. J., Almabuk, R. 'The Adolescent as Forgiver'[J]. *Journal of Adolescence*, 1989, 12(1): 95–110.


［责任编辑　李斌　赵超］