第21卷 第1期 东北大学学报(社会科学版) Vol. 21, No. 1
2019年 1月 Journal of Northeastern University (Social Science) Jan. 2019

doi: 10.15936/j.cnki.1008-3758.2019.01.003

# 人工智能威胁论:逻辑考察与哲学辨析

李 帅
(南开大学 哲学院, 天津 300350)

**摘 要**: 近来年,人工智能发展异常迅猛,引发了一些学者的担忧,他们认为人工智能在不久的将来可能会威胁人类生存。从哲学和逻辑的角度考察这个议题,试图构造一个溯因推理论证驳斥这一论题。如果人工智能会威胁人类生存,那么它必然拥有情感认知能力。获得这种能力可能要借助全脑仿真技术,而这项技术的实现过程中会遇到"指数爆炸"问题。威胁论的论证基于归纳推理,不具备必然的可靠性。但如果仅将威胁论视为一个信念,进行一场类似"帕斯卡赌"的赌局,考虑到信念成真的巨大代价,有必要采取一些防范措施。

**关 键 词**: 人工智能; 威胁; 溯因推理; 帕斯卡赌
**中图分类号**: N 031 **文献标志码**: A **文章编号**: 1008-3758(2019)01-0014-06

# Threat of Artificial Intelligence: Logical Investigation and Philosophical Analysis

*LI Shuai*
(College of Philosophy, Nankai University, Tianjin 300350, China)

**Abstract**: In recent years, with the rapid development of artificial intelligence, some scholars are concerned that in the near future, artificial intelligence may threaten human survival. This issue is examined from the philosophical and logical point of view so as to construct an argument of abduction. If artificial intelligence threatens human survival in the future, it must have emotional cognition. To acquire this ability, whole brain simulation is likely to be relied upon and the realization of this technology may face the "exponential explosion". The current argumentation of this hypothesis is based on induction and does not have the necessary reliability. However, if the thesis of threat is only treated as a belief whereby a gambling similar to the "Pascal's wager" is conducted, risks of a belief becoming the reality can be huge. Thus, it is necessary to take some precautions.

**Key words**: artificial intelligence; threat; abduction; Pascal's wager

中德两国正在共同打造具有跨时代意义的"工业4.0",该项目号称"第四次工业革命",主导方向是智能制造,推动制造业向智能化转型。人工智能是未来工业转型的重要助推器,不仅在工业制造领域扮演重要角色,而且必将成为革新我们生活方式的未来趋势,全方位地重塑我们的生活。人工智能发展速度令人惊叹,很多人工智能专家系统在一些领域已经极大地超越了人类,比如我们所熟知的AlphaGo完胜世界顶尖围棋选手,它的升级版AlphaZero在短时间内就超越了上一代,并"精通"各种棋类。更令人匪夷所思的是,DeepMind推出了能够从事"科研"的

收稿日期: 2018-06-20
基金项目: 国家社会科学基金重大资助项目(15ZDB018); 教育部人文社会科学研究资助项目(18YJC72040001); 天津市社会科学重点资助项目(TJZX17-001)。
作者简介: 李 帅(1991-),男,湖北荆州人,南开大学博士研究生,主要从事逻辑学、科学哲学研究。

AlphaFold，通过基因序列，成功预测出蛋白质的3D形状。人工智能的迅猛发展引发了一些学者的担忧，他们担心人工智能如果以这样的惊人速度发展下去，是否会产生情感认知因素，更具体地说，会不会产生不友好的动机。像科幻电影《终结者》中的情节那般，人工智能奴役、毁灭人类。我们可以通过构造一个驳斥人工智能威胁论的溯因推理结构，证明人工智能威胁论信念是基于不可靠的归纳论证；而以哲学的视角看，可以将人工智能威胁论视为一种信念。鉴于该信念达成之后的灾难性后果，有必要采取适当的防范策略。

## 一、通向乌托邦还是奴役之路

我们目前无法确定人工智能是否是一个潘多拉之盒，但它必定是一个黑箱，至少现在无法预测人工智能的走向。根据当前的发展态势，有部分学者对人工智能的前景持积极乐观的态度，他们认为人工智能会实现技术的乌托邦，人将得到最大限度的解放。他们指出，人工智能所有的知识都是人类“喂给”的，其自身无法创造新的明述知识或命题性知识，而且人工智能的学习能力基于算法而不是社会交往[1]。所以，人工智能只是人类器官的投影和延伸，它们不会出现自主意识，更不会出现社群组织，不会对人类产生危害[2]。一种心智考古学的观点认为，一个能够自主地威胁人类的人工智能体首先必须是一个具有自治情感的生命系统，智能只是心智的一个子集，没有情感系统的引导，智能就不会发挥作用。因此人工智能体不会对人类产生生存威胁[3]。还有一部分人则对人工智能发展到高阶水平的可能后果充满担忧，他们预言人工智能将是人类的终结者。国内学者以江晓原为代表，他认为人工智能近期的威胁会导致大量工人失业，以及军事化用途所带来的世界格局震荡。长远的威胁就是人工智能的失控和反叛[4-5]。我们把这种论调称为人工智能威胁论。

近年来，人工智能威胁论的支持者越来越多。其中不乏一些著名科学家和知名人物，如史蒂芬·霍金（Stephen Hawking）、迈克斯·泰格马克（Max Tegmark）和伊隆·马斯克（Elon Musk）等。马斯克曾多次在推特上发长文警告，人工智能可能是人类有史以来最为严峻的生存威胁。霍金指出，人工智能的崛起要么是最好的，要么就是最糟糕的。

一些研究中心或智库开始研究人工智能可能带来的风险，牛津大学的“人类未来研究所”是目前最具代表性的技术风险研究机构，下设“人工智能政策研究中心”。泰格马克牵头在麻省理工学院成立了类似的机构：“生命未来研究所”。剑桥大学成立了“利弗休姆未来智能研究中心”。北京大学于2018年11月成立了“北京大学人类未来研究中心”……。此外，许多科技企业纷纷成立人工智能研究部门，探讨人工智能的未来前景。

当谈到人工智能威胁论的时候，我们到底针对的是人工智能的哪些方面或哪个阶段。学界粗略地将人工智能分为三个阶段：弱人工智能（artificial narrow intelligence）、强人工智能（artificial general intelligence）和超级人工智能（artificial superintelligence）。弱人工智能只擅长某一具体领域，强人工智能适用范围要更大，能与人类智能比肩；而超级人工智能是机器智能的顶峰，牛津大学“人类未来研究所”所长、人工智能思想家尼克·波斯特洛姆（Nick Bostrom）把超级人工智能定义为“在几乎所有领域远远超过人类的认知能力”[6]29。人工智能拥有巨大的潜力，它可以让人类永生，抑或彻底灭绝。所以当我们言及人工智能威胁论时，主要针对的是超级人工智能。这种威胁论主要基于以下三个前提。

第一个前提认为，最先被创造出来的超级人工智能与包括人类智能在内的所有其他形式的智能相比，拥有决定性的战略优势。这种优势足以使超级人工智能控制并塑造地球上所有智能生命的未来。如果超级人工智能是友好良善的，那么我们无须担心。但根据波斯特洛姆的说法，我们没有任何理由认为它是友好的。第二个前提指出，智能水平和善行之间没有必然关系。但智能水平与其终级目标有着一定的相关性：几乎任何水平的智力都与它的最终目标相一致。波斯特洛姆构想的“曲别针最大化”（paperclip maximiser）思想实验生动地阐释了这一点，如果超级人工智能确定一个目标，就会最大化地实现这个目标。假如它的目标是制造曲别针，那它会尽可能地制造更多的曲别针[6]153。第三个前提指出，超级人工智能很可能设定一些与人类利益无关的目标。比如它可能会执着于获取广泛的资源，其中可能包括人类赖以生存的资源。

就已发表的相关文献来看，波斯特洛姆的专

著《超级智能:路线图、危险性与应对策略》代表了这一趋势的典范。波斯特洛姆在书中详细论述了超级人工智能得以实现的几种形式、可能存在的风险、应对措施等。泰格马克的大部头著作《生命3.0:人工智能时代而为人的意义》呈现了一幅专业物理学家对人工智能的展望图景。书中对人类的未来作出了最大胆的想象,构造了对未来生命、科技、文化和社会的全新理解。泰格马克追问人类到底能发展到什么极限?人类到底要走向哪里?"生命3.0"体现了"生命"系统从1.0到2.0,再到3.0的更迭。生命1.0是简单生物阶段;生命2.0是文化阶段,人类就是生命2.0的代表;生命3.0则是科技时代,系统不断地升级自己的软件和硬件,直接跳跃漫长的自然进化,堪称智能系统全面觉醒的阶段。比起有血肉的生命之躯,届时便产生了新的"生命"[7]27。泰格马克担忧新"生命"会与旧生命发生冲突。这些担忧是杞人忧天的无稽之谈,还是有着坚实的经验基础?我们接下来做一番考察。

## 二、驳斥人工智能威胁论的溯因推理论证

我们可以尝试着构造一个溯因推理结构反驳人工智能威胁论。这个论证结构是这样的:如果威胁论者们认为人工智能在未来会对人类生存构成致命威胁,那么它必然要拥有善恶、喜恶这样的认知情感因素,它甚至可能会伪装。我们进一步追问,人工智能何以拥有像"动机"这样的情感认知功能,我们选定一种可能的实现途径:全脑仿真。这项技术的实现又依赖智能设备运算能力的"指数爆炸"假设;而该假设目前的成功基于归纳,不具备必然可靠性。

1. 预测人工智能善恶的归纳论证

有学者对人工智能威胁论提出质疑,他们对波斯特洛姆的观点提出了一个明显的反驳:为了确保人工智能的安全使用,将人工智能应用于实践之前,必须对其进行严格测试,验证它是否友好。这样做的目的是将人工智能限制在一个安全可控的环境中,并反复测试它的友好属性和安全性。建造一个模拟真实场景的"沙盒",当我们在实验环境中做了足够多的测试,验证它是友好、合作、负责任之后,才会把它从"沙盒"中释放出来[6]145。在一个高度仿真的实验环境中,反复经验检测,如果发现人工智能看起来没有威胁,那么我们就有理由相信它是友好的。这里的推理模式属于典型的归纳论证。

波斯特洛姆试图回击这个论证,他提出了"背叛转折"(treacherous turn)概念[6]147。即超级智能化的人工智能会运用策略,它在实施计划的同时,还会预测人类的反应。我们从人与人之间的日常交流中可以发现,人们为了达成目标,有时会欺骗他人。例如,某人可以假装对他同事的尖端研究缺乏兴趣,希望同事会信赖他,向他披露所有的研究细节,然后他就能窃取同事的成果,据为己有。超级人工智能是否也会做同样的事呢?难道它们没有意识到,我们正在测试它们?难道它们就不会采取假装友好的策略来走出困境吗?

一个人工智能体看似完全友好,与人类合作并保护人类的利益,但它实际上可能筹备着威胁人类生存的长远计划。在欺骗我们陷入虚假的安全感之后,超级人工智能可能会触发背叛转折。所以,质疑人工智能威胁论的归纳论证在"背叛转折"原则下失效。然而,该原则预设了一个前提:人工智能系统具备很强的自主性,甚至能够产生坏的"动机"。那么,人工智能可否发展出诸如"动机"这样的情感要素呢?

2. 人工智能的"动机"论何以成立

针对超级人工智能"动机"论的质疑,波斯特洛姆予以了回应。他认为:"只要具有足够的智能升级能力,所有其他智能就都在系统的间接方式范围内,因为系统能够开发所需的新的认知模块和技能。"[6]112如此一来,波斯特洛姆就把人类的各种认知能力视为智能系统内化的功能,随着智能升级,这些功能便随之开发出来。波斯特洛姆并没有直面质疑,没有解释超级人工智能为何会产生"动机"。其实,这里涉及一个更为复杂的问题,就是人工智能是否会拥有意识,如果拥有意识,是依靠何种方式实现的。倘若我们不首先说明人工智能的"意识"来源及可能性,那么讨论人工智能的动机,必是漫天猜想,没有理据可循。

波斯特洛姆曾提出过超级人工智能的五种可能实现形式:人工智能、全脑仿真、生物认知、人机交互、网络和组织。其中,生物认知基于这样一个观念:生物技术的进步可能会直接控制人类遗传学和神经生物学,而无须实施人类优生繁衍计划。简言之,就是通过生物技术对人体进行无限改造,以提升人类智能。人机交互与网络和组织也是类

似的原理，人机交互意图将机器智能和人脑互联，大幅度提升人脑的能力，网络和组织则企图通过技术手段把人脑连接在一起，形成强大的集体智能[6]37-52。这三种途径都强调对人类智能的增强和提升，属于人类增强的范畴，因为在这样的智能形式中，人脑依旧是主体，增强的是人类智能，而非人工的智能体。这样的超级智能必然会保有人类的各种认知能力，不在我们讨论范围内。

所以，严格意义上说，超级人工智能只有两种可能的实现途径：人工智能和全脑仿真。人工智能途径就是现在所采用的主流方式，依赖算法和硬件的提升。人工智能在“情商”的学习过程中，效率有余，效果不足。微软的人工智能“微软小冰”在 2017 年出版了一本所谓的题名为《阳光失了玻璃窗》的“诗集”。我们来看其中的一首标题为“黄昏里来了一碗茶”的“现代诗”：黄昏里来了一碗茶/回家一齐看/嘴里的妻子已失去了/让野火的人们/风景如风车里一碗茶凉/是少年的故事/回家一年的时候/我猜我也一例有敌骑的呼声晌[8]14。这里的诗歌“意象”更多的是句型与词汇随机混搭生成的“后现代风格”，不能展现人类诗歌才有的情感张力。故而有人批评人工智能有智商而无情商，有智能而无智慧。智能系统要想发展出自主意识，仅靠算法无法达成，可能需要借助全脑仿真技术。这项技术可能完全模拟出被模仿对象的大脑活动，然后加以原理化，但这种技术建立在强大的硬件基础上。接下来又产生了一个新的问题，硬件升级速度能否跟上人工智能的发展？

3.“指数爆炸”的迷雾

其实全脑仿真有一个前提，即大脑本质上是“计算”的，并且能够被模拟。这里涉及到诸多哲学议题，我们不做深究，暂且假定该前提成立。波斯特洛姆认为全脑仿真不需要理论上的突破，但需要一些非常先进的技术。他认为实现这项技术须满足三个先决条件：扫描、翻译和模拟。其中扫描需要有足够分辨率和相关检测性能的高通量显微镜，模拟则需要强大的硬件。以上两项先决条件能否满足，很大程度上取决我们能否制造出性能优越的硬件设备。

众所周知，硬件迭代速度依赖“指数爆炸”假设。“指数爆炸”原指指数函数的爆炸性增长。此后借用到计算机领域，一般与“摩尔定律”同义，表示计算机的计算能力将在固定的周期内以指数形式提升。从第一台现代意义上的计算机诞生至今，计算机运算能力的发展符合“指数爆炸”。我们就拿目前的超级计算机领域来说，超算的运算能力大约每 14 个月提升一倍。超级计算机代表着目前计算机设备性能的最高水平。2018 年 6 月，IBM 公司助力美国田纳西州橡树岭国家实验室推出了一台名为“Summit”的超级计算机，计算峰值惊人，每秒可执行 $2^{1\,017}$ 次运算，其性能远超我国研发的神威·太湖。倘若超级计算机的发展一直符合“摩尔定律”，那么似乎就能满足全脑仿真技术在硬件上的要求。但这个推理模式是典型的归纳论证，基于目前的发展水平来推断未来发展速度依旧保持不变。更为严峻的是，“摩尔定律”已经遇到瓶颈，芯片制造已经接近物理极限，经济成本不降反升。由此，我们认为人工智能威胁论论题站不住脚。

对人工智能威胁论的批评还有很多，其中凯文·凯利(Kevin Kelly)认为我们把智能误解为可以无限增长的空间，反对算力和智能之间存在正相关关系。至于人工智能最终发展到像电影《超验骇客》那般无所不能的地步，他将这种过分夸大、毫无根据的观点称为“意淫主义”(thinkism)[9]。我们充其量只能将人工智能威胁论视为一种信念，我们可以选择相信它为真，但却很难辩护其为真。

## 三、人工智能威胁论与帕斯卡赌

倘若我们将人工智能威胁论仅仅视做一种信念，会产生积极的启示作用吗？有些人会嗤之以鼻，因为该信念带有某种神秘的宗教色彩。还有一些学术评论员批评未来学家雷·库茨韦尔(Ray Kurzweil)提出的所谓“奇点”临近，因为他们的信念体系是信仰主义的，即基于信仰或缺乏合理的论证基础[10]。

我们可以尝试将人工智能威胁论信念与上帝存在的信念作一番类比。后者是宗教学领域中争论的焦点之一，前者可以看做是科学领域中的预言。这两个信念都有一个共同的特点，被证明为真的概率极低，并且假如我们先前拒斥该信念，但后来证明为真，其后果是毁灭性的。比如说，我们无法证明上帝存在，因为根据现有的科学理论体系，假想一位有形体有人格的全能神似乎是一件很荒谬的事情。倘若上帝不存在，我们在世间纵

情享乐,不信奉任何教义,那相信上帝存在与否于我们都没有影响。但如果上帝存在,不信上帝存在者虽然享受了短暂的快乐,死后则会坠入地狱,遭受无穷无尽的折磨。既然如此,我们是选择相信上帝存在呢,抑或拒斥?这就是帕斯卡诉诸信仰的论证上帝存在之路径。帕斯卡将相信上帝存在与否视为一场赌博,是赌博就会有输赢,虽然赌徒不知道赌局的结果如何,但他知道输赢的奖励和惩罚,也就是"赌注",所以赌徒会权衡利弊,然后决定选择是否相信上帝存在。类似地,我们可以将帕斯卡赌应用于人工智能威胁论。显然,人工智能威胁论信念到底有多大概率为真,我们无法预测,但倘若该信念成真,那么后果是无法估量的。正如波斯特洛姆所定义的"存在风险"(existential risk)威胁着地球上智慧生命的根本福祉,它具有潜在的"毁灭性的"或"地狱般的"影响[11]。波斯特洛姆的末日预言暗示了超级聪明的人工智能可能会给人类带来灭顶之灾。因此,波斯特洛姆认为,即便现有的种种证据都不利于人工智能威胁论,其发生的总体概率很低,但背叛转折依旧不容忽视。

诚然,在相信上帝存在的承诺和人工智能威胁论的承诺之间存在一些明显的区别。相信上帝存在的承诺是一种朴素的宗教信念,是一种对现实的终极本质和原因的信仰,信仰者可以诉诸非理性的方式;而人工智能威胁论的信念是对某项科技发展后果的信念,这种信念基于一定的经验基础,基于对现有科技成就的评估和预测[12]。宗教信念与科学信念不能划等号,此处类比只是较弱层面上的。

我们通过更深入地考察"背叛转折",可以发掘人工智能威胁论信念的一些实践意义和认知意义。超级人工智能可能会设定对人类不友好的目标,可能会欺骗我们,用邪恶的方式来达成目标。必定会有人批评波斯特洛姆的"背叛转折",其观点违反我们的直觉。但正是这种反常观念,才会迫使我们反思科学成功的合理性,重新评估我们习惯性地使用的各种归纳推理。肯定"背叛转折"就要质疑科学取得成功所依赖的主要推理工具,此举难以让人接受,因为在人工智能领域中,归纳推理不仅能够解决复杂情况,而且还相当可靠。但我们试着转换视角,其实人工智能威胁论和休谟的怀疑论有着相似的旨趣。在休谟的经验怀疑主义看来,我们不能根据我们迄今为止吃的面包有营养,就推断出接下来吃的面包也会有营养。在波斯特洛姆看来,我们无论在广度、深度,抑或接触时间上,与人工智能的互动极为有限,所以我们无法断言人工智能未来依旧安全,更何况目前无人智能驾驶系统事故频出。在某种程度上,我们能够理解休谟问题背后的深刻哲理,在具体实践中,却容易忽视归纳推理的固有缺陷。对归纳合理性的审视,促使我们思考人工智能威胁论信念折射出的深刻蕴义,提防"背叛转折"。

即使是那些希望我们认真对待人工智能风险的人也会争辩说,人工智能威胁论的一些观点似乎危言耸听。他们指出,超级人工智能可能会有不可控风险,但也有巨大的益处。几乎没有人会怀疑人工智能的社会变革力量,我们现在更需要关注的是如何理解、应对人工智能的潜在风险。像波斯特洛姆、泰格马克等等这样的威胁论者和未来预言家免不了充斥着天马行空的想象,但他们以这种方式呈现出人工智能的可能威胁,提醒我们在充分发展人工智能时,警惕"背叛转折"风险,彰显了人工智能威胁论的认知与实践意义。

## 四、规避人工智能潜在风险及其挑战

我们正处于弱人工智能阶段,针对人工智能的一些建议性政策主要集中在司法审判、大数据隐私、机器人伦理等具体的应用问题上,很少有预防人工智能发展到较高阶段威胁人类生存的策略。

人工智能价值观的加载问题是目前学术界讨论较多的议题。即如果我们给人工智能系统设定一套代码形式的价值观,可否确保人工智能不会偏离正常的发展轨迹。学者们之所以尤为关注该议题,是因为我们通过监视超级人工智能的程序运行预测其动机的方式行不通。正如波斯特洛姆所言:一个不友好的人工智能可能会变得足够聪明,能够意识到隐藏自己的一些能力会获得更大的收益。它可能不会显露进步,并故意放弃一些难度较大的测试,以避免在获得决定性的战略优势之前引起恐慌。程序员试图通过秘密监视人工智能的源代码和它的内部工作来防范这种可能性,但是一个足够聪明的人工智能会意识到它可能被监视,并相应地调整它的应对策略。监视超级人工智能行为的路径走不通,从源头入手,似乎

更为行之有效。霍金一语道出了要害："尽管人工智能的短期影响取决于控制人工智能的人，它的长期影响取决于人工智能到底能否受到控制。"[13]让人工智能戴上像阿西莫夫构想的机器人三定律式的"紧箍咒"，是最为高效便捷的处理方式。

然而，该进路面临如下三重困境。首先是理论层面。所谓的"价值观"由非常含混的观念集合而成，我们何以确保我们要设立的价值观必定是合理不悖的。退一步说，倘若我们可以确保价值观是一致的、没有矛盾的，那么我们需要给人工智能系统输入什么价值观，以及智能系统能否"理解"这些加载的价值观？如果让人工智能系统回答"电车难题"，会出现什么样的情形呢？我们需要给人工智能系统设定什么样的价值标准？[14]这一系列追问都是令人懊恼的哲学和伦理学问题。其次是技术层面。以现在的技术水平，把伦理准则嵌入人工智能系统为时尚早。要想实现伦理准则的嵌入，需要靠在设计计算力和机器智能的大规模应用中推动[15]。技术层面难以实现，应用层面亦是困难重重。我们设定一个具体的应用情景，假如我们给人工智能系统设立一个密尔的功利主义原则："确保人类总体快乐最大化"。如果它认为快乐就是刺激负责快乐的中枢神经，那么它极有可能创造出《黑客帝国》里弱化版的"缸中之脑"。纵而览之，虽然人工智能价值观嵌入看似一劳永逸，实际上面临着理论、技术和应用三面难题，还存在许多难以克服的问题。

其实，我们也可以借鉴其他高风险技术的发展经验，如基因编辑技术、克隆技术和 NBIC 聚合技术等新兴技术。虽然"贺建奎事件"让基因编辑技术蒙上阴影，但总体上还是朝着安全平稳的势头发展。一个重要的原因就是各国政府和整个学术共同体在这些领域制定了规范研究的准则和公约。以克隆技术为例，鉴于克隆人技术特别是生殖性克隆可能会引发严重的社会、伦理、道德、宗教和法律问题，所以联合国在 2002 年制定了《禁止生殖性克隆人国际公约》，许多国家分别制定了禁止生殖性克隆和治疗性克隆的法律。类似地，各国在人工智能领域竞争时，合作和沟通也必不可少，有效减少人工智能研发过程中的草率和盲动行为。居安思危，防微杜渐，是应对将要到来的人工智能时代所应持有的合理立场。

随着人工智能发展的深入，与大众生活日渐交融，必定会出现新的情况。根据形势的发展，还需要修订或增加新的规约和法律条文。谁也不知道新兴技术最终把我们引向何方，但无论如何，在面对人工智能这项重大课题时，多方协作、共同应对、群策群力、小心谨慎是预防人工智能"背叛转折"的有效法宝。

## 参考文献：


[1] 王礼鑫. 马克思主义新认识论与人工智能——人工智能不是威胁人类文明的科技之火[J]. 自然辩证法通讯，2018，40(4)：15-19.

[2] 黄欣荣. 人工智能与人类未来[J]. 新疆师范大学学报(哲学社会科学版)，2018，39(4)：101-110.

[3] 李恒威，王昊晟. 人工智能威胁与心智考古学[J]. 西南民族大学学报(人文社科版)，2017，38(12)：76-83.

[4] 江晓原. 人工智能：威胁人类文明的科技之火[J]. 探索与争鸣，2017(10)：18-21.

[5] 江晓原. 人工智能的危险前景[J]. 编辑学刊，2015(5)：40-41.

[6] 尼克·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.

[7] Tegmark M. Life 3.0：Being Human in the Age of Artificial Intelligence [M]. New York：Alfred A. Knopf，2017.

[8] 微软小冰. 阳光失了玻璃窗[M]. 北京：北京联合出版公司，2017.

[9] Kelly K. The Myth of a Superhuman AI[EB/OL]. (2017-04-05)[2018-06-16]. https：//www.wired.com/2017/04/the-myth-of-a-superhuman-ai/.

[10] Yampolskiy R. Leakproofing the Singularity Artificial Intelligence Confinement Problem [J]. Journal of Consciousness Studies，2012，19(1/2)：194-214.

[11] Bostrom N. Existential Risk Prevention as Global Priority [J]. Global Policy，2013，4(1)：15-31.

[12] Danaher J. Why AI Doomsayers Are Like Sceptical Theists and Why It Matters[J]. Minds and Machines，2015，25(3)：231-246.

[13] 姚人杰. 人工智能对人类构成威胁吗？[J]. 世界科学，2015(4)：58-59.

[14] 李帅. 防范人工智能潜在威胁[N]. 中国社会科学报，2018-09-04(7).

[15] 邓小铁. 智能系统中人类伦理嵌入的计算挑战[J]. 科学与社会，2018(1)：1-13.


（责任编辑：李新根）