人工智能风险性刍议

王治东

摘 要：当前，人工智能技术的发展一路高歌猛进，尤其是阿尔法狗围棋（AlphaGo）以 4:1 的比分击败前世界围棋第一人李世石之后，人们对人工智能的发展比较乐观，当然对人工智能的风险性也产生巨大的担忧。那么，人工智能将带给我们一个什么样的未来？人工智能的风险性是否存在？风险性生成的机制是什么？相互之间的关联度如何？人工智能风险防范的边界在哪里？这些问题亟待不同视角的探讨，哲学在其中有其特有使命。

关键词：人工智能；技术风险；人机边界

中图分类号：B80 文献标志码：A 文章编号：2095-0047（2017）05-0031-11

1956 年人工智能（artificial intelligence）概念在美国达特茅斯大学的研讨会上被正式提出，标志着人工智能学科的诞生。“顾名思义，人工智能就是人造智能，目前的人工智能是指用电子计算机模拟或实现的智能。同时作为学科，人工智能研究的是如何使机器（计算机）具有智能的科学和技术，特别是人类智能如何在计算机上实现或再现的科学或技术。”①随着人工智能技术的发展，人工智能被进一步划分为“弱人工智能”和“强人工智能”。“就弱人工智能而言，计算机在心灵研究中的主要价值是为我们提供一个强有力的工具；就强人工智能而言，计算机不只是研究心灵的工具，更确切地说，带有正确程序的计算机其实就是一个心灵。”②就目前而言，弱人工智能技术已经基本实现，以计算机为载体的人工智能技术在自动化工业中发

作者简介：王治东，东华大学马克思主义学院教授。

基金项目：本文为国家社科基金项目“资本逻辑视域下的技术正义”（课题编号：15BZX034）、东华大学预研究重点课题“元哲学视角下人工自然哲学研究”的阶段性成果。

① 渥维克：《机器的征途》，李碧等译，呼和浩特：内蒙古人民出版社 1998 年版，第 1—2 页。

② 廉师友：《人工智能技术导论》，西安：西安电子科技大学出版社 2000 年版，第 1 页。

挥了巨大作用，“我们可以通过各种自动化装置取代人的躯体活动”①，人类的生产效率因此得到极大的提升。人类一直朝强人工智能的道路上强劲迈进，如果人工智能广泛应用，未来社会是否有更多的风险？风险何在？如果有风险，产生风险的机制是什么？人工智能与人之间的边界或者禁区在哪里？本文尝试从哲学的角度进行初步的探讨。

一、风险因子分析：人工智能风险何在？

乌尔里希·贝克与安东尼·吉登斯的风险理论开启了对技术风险问题的关注。乌尔里希·贝克在 20 世纪 90 年代提出“风险社会”概念，认为科技发展在促进社会进步的同时，也对生态环境甚至人自身造成威胁。“在风险社会中，风险已经代替物质匮乏，成为社会和政治议题关注的中心。”② 贝克认为，当前社会是一个充满各种风险的社会，政治、经济、文化、科技、生产、贸易等各个领域都存在诸多风险，而技术风险无疑是其中影响最为深远的风险类型。吉登斯从现代性的视角出发，提出现代社会的风险形式是一种人类制造出来的风险，“‘人造风险’于人类而言是最大的威胁，它起因于人类对科学、技术不加限制地推进”③。

（一）人工智能与一般技术风险的区别

风险意味着危险的可能性，也是目的与结果之间的不确定性，是危险的概率指标。技术的风险性首先表现为技术的不确定性。技术的不确定性有多种表现形式，技术使用后果的不确定性是技术不确定性的主要方面。技术风险也主要来源于此。国内学者对技术风险问题的认识已经比较成体系，代表性的观点如下：从技术风险的属性来看，技术风险既具有客观实在性，也具有主观建构性；从技术风险的生成来源来看，技术风险既是技术自身的内在属性，亦是人的行为结果；从风险性后果来看，风险事件逐年增多、破坏性不断增强、不可预测性日趋复杂、风险控制愈加困难等。

技术风险的另一个说法是墨菲法则，那就是，如果事情有变坏的可能性，不管这种可能性有多小，它迟早都会发生。人工智能技术也是如此，如果人们担心某种情况发生，那么它就有发生的可能性，因为风险是一种可能性的存在。人工智能技术风险问题既与一般技术风险具有同源性和同构性，但也有很大的区别性。

技术在很大程度上都是作为它者的存在，一般性技术在很大程度上都是外在化的风险，如环境风险、生态风险、经济风险等。“由于技术与社会因素的相互作用，

① 杜文静：《人工智能的发展及其极限》，载《重庆工学院学报（社会科学版）》2007 年第 1 期。
② 乌尔里希·贝克：《风险社会》，何博闻译，南京：译林出版社 2004 年版，第 15—19 页。
③ 安东尼·吉登斯：《现代性的后果》，田禾译，南京：译林出版社 2000 年版，第 115 页。

因此，在风险社会中，风险都会从技术风险自我转换为经济风险、市场风险、健康风险、政治风险等。”①

但人工智能技术却不能简单地作为它者存在，除了外在的风险之外，人工智能技术很大程度上是内在化的风险，那就是人的存在性地位的挑战风险以及人与物边界复杂性的风险。内在化风险不是物质层面的风险，而是一种精神上的冲击风险，是基于人的自我认识和认同的风险。因此人工智能技术的风险因子不仅仅在经济维度、环境维度，而且在于人机边界的厘定，以及人机之间竞争关系的形成方面。

在此方面，很多人工智能事件都引起人工智能取代人的担忧。自 1997 年电脑“深蓝”战胜国际象棋冠军加里·凯斯帕罗夫 19 年之后，在 2016 年 3 月 9 日—15 日，由谷歌 DeepMind 研发的神经网络围棋智能程序 AlphaGo 以 4:1 的比分击败前世界围棋第一人李世石。2017 年 1 月 6 日江苏卫视《最强大脑》上演了一场精彩的人机对决，这次的战场不再是围棋，而是人脸识别。据悉，“‘百度大脑’已建成超大规模的神经网络，拥有万亿级的参数、千亿样本、千亿特征训练，能模拟人脑的工作机制。百度大脑智商如今已经有了超前的发展，在一些能力上甚至超越了人类”②。“小度”对战人类大脑名人堂选手，上演人机大战，在图像和语音识别三场比赛中，以 2 胜 1 平的战绩胜出。2016 年 11 月百度无人车已经能够在全开放的道路上实现无人驾驶。当前快递捡货机器人已经大规模投入快递行业。2016 年富士康公司在昆山基地裁员 6 万人，用 4 万台机器人取代人力。基于以上事实，很多人认为：人工智能取代人类的时代已经到来，敌托邦式构想即将成为现实。并且通过几场“人机大战”，普通大众开始表现出对人工智能风险性问题的强烈关注。强人工智能技术尽管还没能实现，但从这场 AlphaGo 围棋大战中，让人似乎看到未来人工智能超越人类的可能，因为人工智能的三大基础——算法、计算平台、大数据——已经日渐成熟。南京大学林德宏教授曾指出：“电脑不仅能模拟人的逻辑思维，还可以模拟形象思维、模糊思维、辩证思维，人工智能将来可能全面超过人脑智能。”③ 人工智能风险性考虑，主要是基于人工智能对人类的可能性超越。这是一种内在性的风险，是人工智能之于人的关系性的风险。

（二）人工智能风险的表现形式

“工程师和技术专家倾向于把技术风险界定为可能的物理伤害或者厄运的年平均律，哲学家和其他人文主义者认为技术风险无法定量，它包含了较之物理伤害更为广泛的道德内容。”④ 有学者直接认为：“‘风险’包括两部分，一部分是物理性的，

① 芭芭拉·亚当等：《风险社会及其超越》，赵彦东等译，北京：北京出版社 2005 年版，第 334 页。
② http：//tech.qq.com/a/20170107/001226.htm.
③ 林德宏：《“技术化生存”与人的“非人化”》，载《江苏社会科学》2000 年第 4 期。
④ 李三虎：《职业责任还是共同价值——工程伦理问题的整体辨释》，载《工程研究》2004 年第 1 期。

更为实际有形的、可被量化的危险，即技术性的风险；而另一部分是由心理认知建构的危险，即感知的风险（perception of risk）。”① 人工智能风险同样包含这两个层面：一个是客观现实性的物理层面，一个是主观认知性的心理层面。在人工智能技术大规模运用之前，很大程度上风险的认识来自主观认知的心理层面。在人工智能发展过程中，人工智能（类人）与人（人类）之间的关系一般经历三个阶段：首先是模仿关系阶段，人工智能首先基于对人的模仿，使机器初步具有人的智能；二是合作关系阶段，人工智能协助人类完成大量的工作，体现出人工智能强大的利人性；三是竞争关系（取代关系）甚至是僭越关系阶段，是人工智能大规模广泛应用情况下出现人工智能与人之间的依赖、竞争、控制等复杂的关系情况。

人工智能在大规模应用后，潜在的风险性主要有以下几种表现形式：一是人工智能技术的发展将（至少暂时性地）导致未来失业率的大幅度提升。现代工业中，弱人工智能技术已经能够替代人类，从事一般性的体力劳动生产，未来人类的部分脑力劳动也必将被人工智能技术所取代。因此，对未来人类可能面临巨大失业风险的担忧不无道理。二是人工智能的发展使人类遗忘人工智能技术。也就是说，人类将越来越依赖机器的“智能性”，而忽视其“人工性”，这将导致人类与机器的关系转换成人类与“类人”的关系，人类很可能对机器产生类人情感，甚至产生类人的依赖感。一旦人类将机器视为同类，必然带来相应的伦理问题。如性爱机器人如果大规模应用，将使婚姻生育等问题变得复杂，人的两性关系以及很多伦理问题都会相应而来。三是未来机器人不仅具备类人思想，还可能具备类人的形态，人类在与机器人的日常交互中，如果将机器人视作同类，机器人能否获得与人类等同的合法地位，人与机器人之间的关系如何界定，这也是复杂的问题。以人工智能技术为核心的机器（至少部分性地）超越人脑，存在威胁人类主体性地位的可能。依托强人工智能技术的机器一旦具备甚至超越人类智慧，机器很可能反过来支配人类，这将对人类存在性（主体性）造成巨大的威胁。

当然，上面都是人工智能作为它者的存在与人之间的关系的风险。但还有更复杂的情况，2017 年 3 月 28 日，特斯拉创始人马斯克成立公司致力于研究“神经织网”技术，将微小脑电极植入人脑，直接上传和下载想法。在此之前，后现代哲学家哈拉维提出赛博格的概念，是人与机器的杂合。这种以智能植入方式，将人与机器联机，人与机器的边界何在？对人类未来的影响是积极的还是消极的？人对未来终极问题的思考对人类的心灵造成巨大的困扰，这种主观认知性的心理层面的风险并不弱于客观现实性的物理层面。

① 转引自曾繁旭、戴佳、王宇琦：《技术风险 VS 感知风险：传播过程与风险社会放大》，载《现代传播》2015 年第 3 期。

二、人工智能风险形成机制分析：一种现象学的视角

人工智能风险目前更多地体现在主观认知性的心理层面，是人们对人工智能发展的一种担忧，哲学的思考大有用武之地，其中现象学更具解释力。

（一）从外在模仿到内在超越：人工智能技术的放大效应

人工智能多是以独立的形式对人的模仿甚至超越。“行为的自动化（自主化），是人工智能与人类其他早期科技最大的不同。人工智能系统已经可以在不需要人类控制或者监督的情况下，自动驾驶汽车或者起草一份投资组合协议。”① 与一般技术一样，人工智能技术之于人有两个层面：一是机器操作代替人的劳动，使人从繁重而复杂的劳动生产中解放出来，让人获得更多的自由空间。二是人工智能取代人类智能，人类受控于机器，人类主体的存在性地位丧失。技术发展呈现完全相反的两种进路，这是由技术二律背反的特性决定的，技术具有“物质性与非物质性、自然性与反自然性、目的性与反目的性、确定性与非确定性、连续性与非连续性、自组织与他组织”② 等特性。

技术还有一个内在属性就是具有放大性功能。技术放大功能是技术内在结构的属性，是技术模仿人类功能并对人类能力的放大，它完全内置于技术结构中。“人—技术—世界”的结构模型是现象学的基本模型，表达了人是通过技术来感知世界的，人与世界的关系具有了技术的中介性。例如，在梅洛-庞蒂所举的盲人与手杖的例子中，盲人对方位的感知是通过手杖获得的，手杖成为连接盲人与空间方位的转换中介，扩展了盲人的空间感。在这里，技术通过转化人类的知觉，扩展了人类的身体能力。“只有通过使用技术，我的身体能力才能得到提升和放大。这种提升和放大是通过距离、速度，或者其他任何借助技术改变我的能力的方式实现的。”③ 人类对技术无限放大性的追求也是现代技术发展潜在的动力，也是技术风险生成的根源。而技术的放大效应既是内置于技术内核的结构性特征，也是人类目的性的现实要求。在目的性结构中，技术是表达人的意愿的载体，人工智能技术就是放大人类的意愿，在某种程度上可以代替人的意愿。当一个中介完全把人的意愿变成中介的意愿时，人工智能的本质得以实现，技术的放大效应达到最大化。但人的意愿可以被机器表达时，人的可替代性也逐步完成，人也失去了自我。技术便有可

① 马修·U. 谢勒：《监管人工智能系统：风险、挑战、能力和策略》，曹建峰、李金磊译，载《信息安全与信息保密》2017 年第 3 期。

② 王治东：《相反与相成：从二律背反看技术特性》，载《科学技术与辩证法》2007 年第 5 期。

③ 唐·伊德：《技术与生活世界——从伊甸园到尘世》，韩连庆译，北京：北京大学出版社 2012 年版，第 75 页。

能朝向背离人类预期的方向发展，技术风险由此生成。

在前人工智能技术时代，技术只是对人类“外在能力”的模仿与扩展，即使像计算机、通信网络等复杂技术也是以一种复合的方式扩展人类的各项技能。但人工智能技术却内嵌了对人类“内在能力”的模仿，对人脑智慧的模拟。这一技术特性使人工智能技术具备了挑战人类智慧的能力。千百年来，人类自诩因具备“非凡的”智慧而凌驾于世间万物，人的存在地位被认为具有优先性。康德“人为自然界立法”的论断，更是把人的主体性地位推到了极致。一旦人工智能技术被无限发展、放大，具备甚至超越人脑机能，人类对技术的“统治权”将丧失，人类的存在性地位也将被推翻。尽管就目前而言，人类对人工智能技术的研发仍处于较低水平，但人工智能表现出的“类人性”特征，已经不似过去技术对人脑机制的单向度模拟。特别地，AlphaGo 在面对突发状况时表现出的“随机应对”能力，远远超出开赛前人类的预估。我们似乎看到人工智能正在从对人类“智”的超越，转向对“慧”的模拟，这种风险越来越大。

（二）从它者性到自主性的循环：人工智能技术矛盾性的存在

早期技术就是作为一种工具性的存在，也是一种它者的存在。但发展技术的潜在动力就是不断让技术自动化程度越来越高，越来越自主。技术的自主性发展表现为，“技术追求自身的轨道，越来越独立于人类，这意味着人类参与技术性生产活动越来越少。”[①] 人工智能技术的自主化程度取决于人工智能的“类人性”。也就是说，人工智能越趋近于人类智能，技术的自主性也就越可能实现。就人类预期而言，人工智能技术的发展是自主性不断提升的过程，也是使更多的人从日常劳作中解脱出来，获得更多自由的过程。但当技术发展到具有人一样的智能时，技术在新的起点上成为一个它者。因为技术发展的不确定性使技术既有“利人性”也有“反人性”。这两种看似相反的特性是一个问题的两个方面，智能技术将这两种特性又进一步放大。人工智能技术的“利人性”是技术自主性的彰显。但也正是基于人工智能的“类人性”特点，使达到自主化奇点的技术可能出现“反人性”倾向。

技术的“反人性”表现为它者性的生成，技术它者性是技术发展违背人类预期的结果。理论上，当技术成为一个完全自主、独立的个体时，它将不依附于人且存在于人类世界之外，技术相对地成为它者。在伊德看来，“我们与技术的关系并不都是指示性的；我们也可以（同样是主动的）将技术作为准对象，甚至是准它者”[②]。技术的（准）它者性可以表示为：人类 → 技术（世界）。“它者”一词本身暗含着人类对技术完全对象化的担忧，这种担忧在海德格尔看来由技术的“集置”特

① Jacques Ellul, *The Technological Society*, New York: Alfred A. Knopf, 1964, p.134.

② 唐·伊德：《让事物“说话”：后现象学与技术科学》，韩连庆译，北京：北京大学出版社 2008 年版，第 57 页。

性决定，“集置（Ge-stell）意味着那种摆置（Stellen）的聚集者，这种摆置摆置着人，也即促逼着人，使人以订造方式把现实当作持存物来解蔽”①。事实上，人工智能技术的发展趋势，就是在不断提高技术较之于人的它者地位。

在实际应用中，人工智能技术的“反人性”倾向会以它者的形式呈现。“技术还是使事物呈现的手段。在故障情形中发生的负面特性又恢复了。当具身处境中的技术出现故障了，或者当诠释学处境中的仪器失效了，留下来的就是一个强迫接受的、并因此是负面派生的对象。”② 在伊德的技术体系中，尤其在具身关系和诠释学关系中，技术（科学仪器）通过故障或失效导致技术它者的呈现。技术在承载人与世界的关联中，本应该抽身隐去，但却以故障或失效的方式显现自身，重新回到人类知觉当中，必然阻断人与世界的顺畅联系。本来通过技术实现的人对世界切近的感知，转换成人对（失效了的）技术的感知。这时，（失效了的）技术的它者性仅仅表现为感知的对象性。同样，人工智能技术同样也存在技术失效的可能，但这种失效不是以故障而是以一种脱离人类掌控的方式成为它者。人工智能技术的失效不仅会转换人类知觉，更为严重的是，一旦技术在现实中摆脱人类控制，自主化进程将以故障的方式偏离预定轨道继续运行，技术的“反人性”开始显现，技术它者由此形成，技术的自主性成为它者的“帮凶”。人工智能技术的风险在于经历了“它者性—自主性—它者性”过程之后，这种风险结构被进一步放大。

现实中需要不断通过技术发明和技术改造提高人工智能技术的自主性，但又不得不防范人工智能技术的它者性。由此，人工智能技术自主性和它者性便成为技术发展过程中的一种冲突。

三、“人类”与“类人”界限：人工智能技术的禁区何在?

从概念可以看出“人工智能”由两部分组成：一是人工，二是智能。相对于人工智能，人在某种程度上就是一种天然智能。“准确定义人工智能，困难不在于定义‘人工’（artifiality），而在于‘智能’（intelligence）一词在概念上的模糊性。因为人类是得到广泛承认的拥有智能的唯一实体，所以任何关于智能的定义都毫无疑问要跟人类的特征相关。”③ 人与人工智能之间的界限有两个维度的比较很重要：第一个维度是知、情、意、行四个基本特征；第二个维度是人的自然属性和社会属性两个方面。

（一）关于知、情、意、行的边界问题

人是知、情、意、行的统一。随着人工智能技术的不断发展，人工智能将趋近

① 李霞玲：《海德格尔存在论科学技术思想研究》，武汉：武汉大学出版社 2012 年版，第 82 页。
② 唐·伊德：《技术与生活世界——从伊甸园到尘世》，第 99 页。
③ 马修·U. 谢勒：《监管人工智能系统：风险、挑战、能力和策略》。

于人类智能，承载人工智能技术的机器也将具有更多“类人性”。这种“类人性”不仅表现为机器对人类外在形态的模仿，更表现在机器对人类“知、情、意、行”的内在模拟。智能首先要学会语义分析，能够读懂指令，如 2016 年 4 月刷屏爆红的“贤二机器僧”由北京龙泉寺会同人工智能专家共同打造，在最初阶段有效回答问题率 20%—30%，但人工智能强大的学习能力是一般计算机系统无法相比的，人工智能具有学习能力，逐渐增加的信息变成知识，继而形成知识库，通过知识库形成机器人大脑，进而形成能够与人进行有效交流的智能系统，随着访问量的增加，贤二机器僧的数据库相应增加，有效回答率达到 80% 左右。这样与人交流的人工智能，让人感觉不到是与一台机器在交流。

当然，人工智能也是有禁区的，这种禁区，首先来自技术不能逾越的禁区。当前，可计算性是人的逻辑判断部分，而情感支配的思维是无法被计算的，因此人工智能对个体人的超越还是存在困难的。但如果像马斯克公司一样，通过人工智能去解读人脑的思维，上传和下载功能能够实现，人工智能能够读懂人，这样的冲击风险对人而言，其风险具有更大的加强性。当然能够解读人类思维也仅仅是一种识别和解读，人与人工智能之间还有一道重要的区别，就是自我意识的区别。“一个种的全部特征，种的类特征就在于生命活动的性质，而人的类特征就是自由有意识的活动。”① 人是有意识的存在，意识总是关于某物的意识，同时也是作为承载者关于“我”的意识，意向性不仅指向作为对象的某物，同时也自反式地指向自身。只有在行动之前首先意识到处境中的对象与“我”不同，人类活动才具有目的性，人类才能“有目的”地进行物质创造和生产劳动。而人的目的性或者说人类需求是社会进步发展的最大动力。人工智能如果有了“我”的概念和意识，不仅是对人的模拟，而且具有了人的核心内核。在这个层面而言，人工智能就在个体上可以成为另一个物种的“人”。

当然，如果人工智能技术一旦具备类人意识，它将首先关注到自身的价值意义，即存在的合法性。而作为对象的人类，将沦落为技术“眼中”的它者。具备类人意识的人工智能对人类智能的超越，将对人类生存构成实质性威胁，到那时，人类生存与技术生存真的会互相威胁，彼此竞争。

（二）关于自然属性和社会属性问题

任何技术都是“自然性和反自然性”的统一。“技术作为人本质力量的对象化有两重属性：一是技术的自然属性，二是技术的社会属性。自然属性是技术能够产生和存在的内在基础，即技术要符合自然规律；技术的社会属性是指技术的人性方

① 《马克思恩格斯选集》（第 1 卷），北京：人民出版社 1995 年版，第 46 页。

面，即技术要符合社会规律。”[①] 人工智能技术也是如此，“人工”是一个前置性概念。“智能”是对人的模仿，在人工智能设定的模仿程序中很大程度上也有社会属性，如军用机器人的战争属性，性爱机器人的性别属性，但这种社会性是单一的属性。恰恰社会属性是人区别于人工智能的核心所在。按照马克思对人本质是类本质和社会本质的论述，起码目前人工智能还不能叫板人本身。“人的本质并不是单个个人所固有的抽象物，在其现实性上，他是一切社会关系的总和。”[②] 人工智能目前是以个体性或者是整体功能性而存在的，不可能具有社会性存在，也就意味着人工智能不能作为一个物种整体具有社会性，而人恰恰具有社会关系性，而这种社会关系是人之为人的根本性存在。“人的本质是人的真正社会关系，人在积极实现自己的本质的过程中创造、生产人的社会关系、社会本质。”[③] 无论人工智能怎么在智能上超越人类，但根植于物种的社会性不是通过可计算获得的。因此这也是人工智能的禁区。如果人类赋予人工智能以社会关系构架，人工智能之间能够做联合，能够选择意识形态，那人类危机也真的不远了。但个人认为，作为人的整体社会建构的文化以及关系，任何人工智能都是无法取代的。

四、结语

科学技术的进步与发展是历史必然，我们终究无法预测人工智能技术究竟能够取得多大的突破，但只要人工智能无法意识到“自我”的存在，它就只能作为工具为人所用，被人所控。关乎人类问题的关键从来也必须从人类自身出发才可能找到解答，无论人工智能技术如何发展，只要人类保持足够的理性，为人工智能技术划定禁区，人类的存在性地位就不可能被超越。尽管人工智能大规模发展会在某种程度上代替人类，让某些人失业，让社会结构发生改变，但人工智能也能创造出新的工作平台和领域，让人在更高的平台上实现人的创造性本质。

当然，有一点需要特别引起注意，对人工智能的应用的限制和立法要提前进入考量范畴，否则如同其他技术一样，如果人工智能技术被别有用心的人滥用，产生的社会问题肯定会超过一般技术的滥用，这一点法学和社会学会大有用武之地。

（责任编辑：肖志珂）

① 王治东：《相反与相成：从二律背反看技术的特性》。
② 《马克思恩格斯选集》（第 1 卷），第 56 页。
③ 《马克思恩格斯全集》（第 42 卷），北京：人民出版社 1979 年版，第 24 页。