重庆大学学报(社会科学版) 2024年第30卷第3期
JOURNAL OF CHONGQING UNIVERSITY(Social Science Edition) Vol. 30 No. 3 2024

Doi:10.11835/j.issn.1008-5831.fx.2019.12.007

欢迎按以下格式引用:甄航.人工智能刑法"主体性"否定:缘起、解构、反思——以认知科学的五个层级为基础[J].重庆大学学报(社会科学版),2024(3):242-252. Doi:10.11835/j.issn.1008-5831.fx. 2019.12.007.

Citation Format: Zhen Hang. The negation of the "subjectivity" of artificial intelligence in criminal law: origins, deconstruction, reflection:based on the five-level theory of cognitive science [J]. Journal of Chongqing University(Social Science Edition),2024(3):242-252. Doi:10.11835/ j.issn.1008-5831.fx. 2019.12.007.

人工智能刑法"主体性"否定:缘起、解构、反思
——以认知科学的五个层级为基础

甄 航
(西南政法大学 法学院,重庆 401120)

摘要:人工智能是否能与人类一样具有刑法"主体性"地位无法在刑法理论内部找到答案,而需要以相关的认知科学为基础,否则就会陷入循环论证的困境。根据认知科学的五个层级理论,人工智能在神经层级、心理层级的低阶认知层面,仅是对人类认知的简单模拟;在作为高阶认知与低阶认知中间环节的语言层级认知层面,人工智能的人工语言与人类自然语言之间具有本质的区别;在思维层级、文化层级的高阶认知层面,当前的人工智能并没有显现出其具有思维或产生文化的能力。结合塞尔"中文房间模型"研判,人工智能并不具备刑法意义的"辨认能力"与"控制能力"。在辨认能力方面,人工智能传感器对客观世界的识别虽然能在一定程度上对人类认知进行形式模拟,但其并不能将识别到的信息与自身"行为"结合进行加工处理从而得出该"行为"的社会意义;在控制能力方面,人工智能所展示出的"控制能力"实质上是一种执行算法的能力,其本质上仍然是一种人类控制,而非人工智能的"自我控制"。因此,在当下及可预见的未来,人工智能并不具有刑法"主体性",进而承担刑事责任,当前的刑法理论也不必对所谓的"强人工智能"过度反应。将"类人"的人工智能提升到人类同等高度,有损人之所以为人的尊严。将"删除数据、修改程序、永久销毁"等与刑罚异质的要素纳入刑法会让其有适用于人类的危险。故在当下及可预见的未来,人工智能对刑法理论的影响主要在于其导致传统社会风险加剧,刑法理论的应对模式仍应当在现有刑法理论体系内,结合风险刑法理论将其作为"犯罪对象"和"犯罪工具"对待。当人工智能作为犯罪对象时,其是以财物、作品等形式存在,在司法认定过程中要注意界分人工智能本身与人工智能的载体;当其作为犯罪工具时,会导致犯罪后果扩大,查证难度增大等结果。

基金项目:中国博士后科学基金第71批面上资助"智慧量刑的悖论与破解研究"(2022M712650);2019年度西南政法大学科研创新项目"人工智能的刑法属性研究"(FXY2019009)

作者简介:甄航,法学博士,西南政法大学法学院讲师,博士后流动站研究人员,Email:zhenhanglmn@163.com。

关键词:人工智能;认知层级;强人工智能;中文房间模型
中图分类号:D924.13;TP18　　**文献标志码**:A　　**文章编号**:1008-5831(2024)03-0242-11

一、问题的提出

互联网时代之后,就有人断言人工智能(Artificial Intelligence,AI)时代已经来临。“无论人们欢欣抑或踟蹰,人工智能时代正悄悄向我们走来,人类即将甚至已经开始进入一个全新的时代”[1]。近年来,在移动互联网、超级计算、大数据、传感网、脑科学等新理论的驱动下,具有深度学习、跨界融合等特征的人工智能技术呈指数爆炸式发展。2017年7月8日,国务院印发《新一代人工智能规划》(以下简称《规划》),标志着我国将发展人工智能技术提升至国家战略层面。与此同时,人工智能技术也将面临伦理与法律的困境。《规划》指出,我国人工智能技术关于伦理与法律发展的战略目标分三步走:第一步,部分领域的人工智能伦理规范和政策法规初步建立;第二步,初步建立人工智能法律法规、伦理规范和政策体系,形成人工智能安全评估和管控能力;第三步,建成更加完善的人工智能法律法规、伦理规范和政策体系①。

(一)人工智能的刑法研究突破口——刑法“主体性”

《规划》印发后,关于人工智能的各法学领域研究层出不穷,如关于人工智能创作作品的知识产权问题研究、人工智能侵权责任问题研究、人工智能具体技术规范问题研究等。此外,刑法学者也试图在刑法领域寻求关于人工智能研究的突破口——人工智能的刑法“主体性”问题,即人工智能能否和自然人一样作为刑法主体,承担刑事责任。刘宪权教授主张以是否具有辨认能力和控制能力为标准将人工智能区分为弱人工智能和强人工智能,其中强人工智能应承担刑事责任,承担刑事责任的方式是删除数据、修改程序、永久销毁②。无可否认,人工智能技术的指数爆炸式增长必将冲击现有法律体系和社会伦理体系,并带来巨大的法律风险,但是否会在每一法学领域形成深刻变革还有待考证。具体而言,人工智能技术的飞速发展是否会对传统刑法理论造成底层结构性冲击,引起刑法领域的深刻变革以至于需要将人工智能像自然人一样作为刑法主体仍有待更为严谨的论证。

(二)现有研究及其困境

当前主张人工智能应具备刑法主体地位的主要论证路径为:第一步,以是否具有辨认能力与控制能力为标准将人工智能区分为强人工智能与弱人工智能,强人工智能是具有辨认能力与控制能力的人工智能,弱人工智能反之。第二步,以强人工智能具有辨认能力与控制能力为由主张强人工智能应当作为刑罚主体承担刑事责任③,承担刑事责任的方式包括“删除数据、修改程序、永久销毁”[2]。这种在闭环内循环论证的论证逻辑存在诸多问题。(1)将人工智能提升到与人类同等地位,是否有损人之所以为人的尊严?是否会严重冲击“人类中心主义”的价值立场?是否会使得几千年以来以人类为主体所构建的主客体格局的哲学关系崩塌?这一系列较为深远的哲学论题亟待解决。(2)在可预见的未来是否可能存在具有辨认能力与控制能力的所谓“强人工智能”缺乏足够的论证,以对强人工智能的

①参见:2017年7月8日国务院印发的《新一代人工智能规划》。
②参见:刘宪权《涉人工智能犯罪刑法规制的路径》(《现代法学》,2019年第1期75-83页)、《论人工智能时代智能机器人的刑事责任能力》(《法学》,2018年第1期40-47页)、《人工智能的刑事风险与刑法应对》(《法商研究》,2018年第1期3-11页)、《人工智能时代的“内忧”“外患”与刑事责任》(《东方法学》,2018年第1期134-142页)、《人工智能时代机器人行为道德伦理与刑法规制》(《比较法研究》,2018年第4期40-54页)。
③此处的“刑事责任”概念在犯罪法律后果层面使用。

“设想”为基础的法学研究缺乏严谨性。人工智能技术的发展程度、阶段的划分以及划分的标准首先是一个技术问题,如果法学研究先于“自动化”相关学科研究,则会使严肃的法学研究落入研究对象模糊不清、研究结论无法落地的窘境。(3)在形式上,人工智能在某种程度上具有一定的辨认能力与控制能力④,但这是否是刑法意义上的辨认能力与控制能力缺乏足够的论证。(4)“删除数据、修改程序、永久销毁”等所谓针对人工智能的“刑罚”能否进入刑罚体系是一个更为深远的论题,该种“刑罚”一旦进入刑法,人类也有了被适用该种“刑罚”的危险,其是否人道缺乏足够的论证。在这一系列问题都没有解决的情况下,仅仅在对人工智能进行朴素假想的基础上得出需要将人工智能作为刑法主体的结论无法立足。

(三)研究路径阐释

在刑法理论内部通过逻辑推理的方式进行“黑格尔式”思维论证研究人工智能的刑法主体地位问题无法寻求出路,故本文以相关认知科学(Cognitive Science)为基础⑤,以清华大学心理学系蔡曙山教授提出的人类认知的五个层级为理论支撑,分析人类智能与人工智能的认知路径,以此剖析人工智能的“辨认能力”与“控制能力”,进而论证人工智能在可预见的未来是否应被作为刑法主体承担刑事责任,并以此为基础反思人工智能技术的发展对刑法理论造成的冲击程度以及应对策略。

二、缘起:人工智能与人类智能

之所以会有人工智能能否作为刑法主体进而承担刑事责任的论题,是因为人工智能的“类人”属性。该“类人”属性并非表现在外在形式上,而是人工智能的“认知能力”具有“类人”属性。例如2016年3月9日至15日,阿尔法围棋(AlphaGo)与世界围棋冠军李世石进行人机大战,最终阿尔法围棋以4比1获得胜利;再如“美国人工智能公司 OpenAI 推出的跨时代新产品 ChatGPT 更是颠覆了传统分析式人工智能的技术路径,使人工智能进入生成式人工智能时代”[3]。人工智能阿尔法围棋与 ChatGPT 不具有“类人”的外在形式,而仅仅是具有“类人”的“认知能力”。故“人类智能,就是神经、心理、语言、思维、文化五个层级上所体现的人类的认知能力。人工智能,就是让机器或人创造的其他人工方法或系统来模拟人类智能”[4]。人工智能是对人类认知能力的模拟⑥,故在明晰人类认知的基础上,对比分析人工智能与人类智能的“认知”,考察其能在何种程度上对人类智能进行模拟,并以此为基础剖析其刑法主体地位是本文的论证路径。

(一)人类认知的五个层级

人类认知的五个层级理论(以下简称“人类认知五层级理论”)是清华大学心理学系蔡曙山教授按照人类大脑里的认知过程的层次将人类认知划分为五个层级:神经层级的认知;心理层级的认知;语言层级的认知;思维层级的认知;文化层级的认知(以下简称“神经认知;心理认知;语言认知;思维认知;文化认知”)。其中,神经认知与心理认知属于低阶认知,是人与动物共有的认知,而语言认知、思维认知与文化认知属于高阶认知,是人类特有的认知。低阶认知是高阶认知的基础,高阶认知包含并影响低阶认知[5]。人类认知五层级理论如图1所示。

---

④例如国际自动化机械工程师学会(SAE)的分级标准中的5级无人驾驶汽车(完全自动化汽车)可以对汽车周围环境进行识别,并以此识别信息为基础控制汽车的方向与速度。

⑤从脑和神经系统产生心智(mind)的过程叫认知(cognition)。认知科学(cognitive science)就是研究心智和认知原理的科学,参见:蔡曙山《认知科学框架下心理学、逻辑学的交叉融合与发展》(《中国社会科学》,2009年第2期25-38页)。

⑥但对于何种自动化程度的机器可以被称之为“人工智能”,至今仍然没有一个明确的界定。

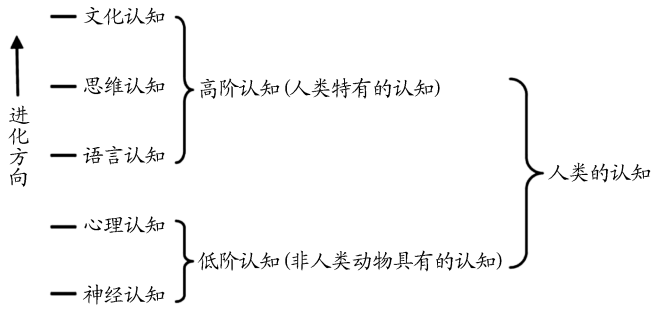

图 1　人类认知的五个层级[5]

神经认知是人类与动物共有的认知。人类的神经认知活动包括视觉认知、听觉认知、嗅觉认知、味觉认知、触觉认知。因而在某些神经认知上,一些动物的认知能力比人类更强,如鹰的视觉认知能力、犬的嗅觉认知能力比人类更强。心理认知也是人类与动物共有的认知形式。心理认知活动包括感知觉和注意、表象和记忆等基本的心理现象。“感觉(sense)是通过单一感官直接获得的认识,包括视觉、听觉、味觉、嗅觉,以及多感官或跨通道获得的认知,即联觉(synesthesia)……知觉(consciousness/perception)是脑和神经系统对感觉信息的再加工,以获得对事物的整体性认识的心理过程。注意是在知觉和意识这个层面上认知加工的一种重要方式,它是一种导致局部刺激的意识水平提高的知觉选择性集中的形式……表象(image/presentation)是在感知觉基础上,经大脑进一步加工而成的经验的认识形式……记忆(memory)是表象的特使形式,表象通常体现为记忆效果”[5]。语言认知是人类所特有的认知形式,其具有特殊的地位和意义:它是五个层级的中间环节,是低阶认知和高阶认知的联结点,也是高阶认知的基础。思维认知也是人类特有的认知形式。思维形式和规律是逻辑学研究的领域,思维形式包括概念、判断、推理和论证。文化认知也是人类特有的认知形式,是五个层级中最高的认知形式。在文化层级上,人类认知由初级到高级的三个层次分别是:科学、哲学和宗教[5]。

(二)人工智能超越人类智能的理性缺失

在明晰人类认知的五个层级后,对比人工智能认知能力的“类人”程度,会发现,人工智能在可预见的未来超越人类智能的论断存在理性的缺失。

在神经层级与心理层级的低价认知层面,人工智能在当下及可预见的未来所做的仅仅是在形式上模拟人类神经认知与心理认知的某些片段性活动,其与人类神经认知、心理认知活动具有本质的区别。如摄像头模拟人的眼睛,麦克风模拟人的耳朵。但摄像头、麦克风的运作原理与人类眼睛、耳朵的神经认知活动具有本质的不同,人工智能通过传感器“输入—存储—计算”的工作原理与人类认知活动无法比拟。即使不与人类作比较,在此方面,作为“硅基”的人工智能甚至与“碳基”的低级动物都相去甚远。“人类的神经认知活动,如视觉认知、听觉认知、嗅觉认知、触觉认知,计算机和人工智能远没有达到人类认知能力和水平;而对幸福、痛苦和各种情绪的感受,目前的人工智能恐怕连一些低级的动物如虫鱼鸟兽的认知水平都比不上”[4]。

在语言层级、思维层级与文化层级的高阶认知层面,人工智能更是无法与人类相比拟。在语言层级的认知上。由于语音识别技术、输入技术等,人工智能在形式上看似具有语言能力,但人工智能的语言系统与人类的语言系统具有本质的区别。语言分类系统和分支如图 2 所示。现阶段人工智能所使用的形式的人工语言是单调的、无歧义的,其与抽象的、可产生性、歧义性的人类自然语言有着不可逾越的鸿沟。“不论是采用符号计算主义还是联结主义的系统,当前的人工智能技术都不能解决语义的理解和生成问题”[7]。如前所述,语言层级的认知具有中间环节的特殊地位,人工智能与人类智能在

语言层级认知层级的分野使其在思维层级、文化层级的认知愈加方枘圆凿。在思维层级的认知上,由于人工智能在需要思维的国际象棋、围棋领域都战胜了人类智能,就有人认为人工智能某些方面的思维层级认知已经超越了人类智能,但事实并非如此。人工智能在国际象棋、围棋领域的胜利与计算机的数字计算力超越人类没有本质的区别。人们觉得现阶段人工智能具有的思维仅仅是由于近年来科技的发展使人工智能"计算力"急剧提升而产生的假象。而在文化层级的认知上,人类文化的形成经历了几千年的沉淀,而在当下及可预见的将来,人工智能还没有形成其自身文化的迹象。

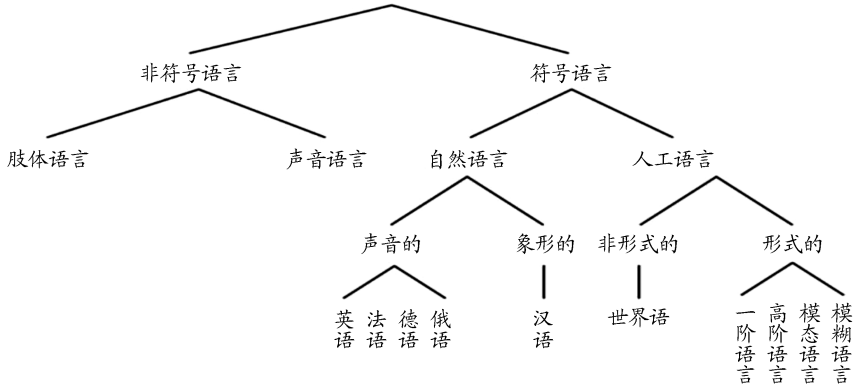

图2 语言分类系统和分支图[6]

以认知科学的五个层级理论为基础,对比人工智能与人类智能在神经认知、心理认知、语言认知、思维认知、文化认知层面的差异可以得出结论:在当下及可预见的未来,人工智能在认知科学的每一个层级上都仅仅是对人类认知形式上的简单模拟,与人类认知具有本质的区别,故人工智能的"智能"在当下和可预见的未来无法超越人类,甚至很难接近人类。

(三)图灵测试与塞尔"中文房间模型"

关于机器的智能标准,历史上曾以图灵测试作为判断计算机是否智能的标准。1950年10月,英国数学家、逻辑学家艾伦·麦席森·图灵(A. M. Turing)发表了题为《机器能思考吗》的论文,在该论文中,图灵提出了著名的"图灵测试"(Turing Test)。图灵测试的路径:使一台机器与人类进行对话(该对话通过电传设备进行),如果人类无法辨别与其对话的是否是机器,那么该机器就是智能的。有学者认为以图灵测试为标准判断机器是否智能已经过时,"我们认为,这(图灵测试——引者注)仅仅是计算机科学家所理解的人工智能,现在看来,图灵的标准似乎太弱了"[8]。但由于至今对什么样的机器属于人工智能仍然没有准确的界定,以图灵测试为标准并非谬误。

1980年,美国语言哲学家约翰·塞尔(John Rogers Searle)提出"中文房间模型"。有学者在区分强人工智能与弱人工智能时撰文指出,"'强人工智能'一词最初是约翰·罗杰斯·希尔勒针对计算机和其他信息处理机器创造的"[9]。并引用约翰·罗杰斯·希尔勒的话论证其观点,"计算机不仅是用来研究人的思维的一种工具;相反,只要运行适当的程序,计算机本身就是有思维的"[10]。实际上,约翰·塞尔虽然提出了"强人工智能"(strong AI)的概念,但其本身是反对"强人工智能"存在可能的。约翰·塞尔所提出"中文房间模型"就是为了反驳"强人工智能"的存在可能而设计的。"中文房间模型"是指,一个不懂中文的人身处图灵测试所描述的房间中,并拥有一本操作规程,依据该中文操作规程,他可以对中文字符进行回应。此时,如果将一些中文字符递进图灵测试房间,房间里不懂中文的人依据中文操作规程予以回应,图灵测试房间外的人与一个不懂中文的人进行对话,且并不能识别其不懂中文。"中文房间模型"旨在让人类模拟人工智能的运行方式,从而反驳强人工智能的存在。通过

该模型,可以看出,“这样塞尔就构造了一部不可能有任何一点智力的机器,但它却能完成类似人的智力行为……计算机能够完成某种智能行为,仅仅是因为它执行了人们按照一定目的事先编制的‘操作规程’,或者说,是人类智能决定了机器智能而不是相反”[8]。

通过图灵测试到塞尔“中文房间模型”,我们可以看出,即使通过图灵测试的机器,也并非具有人类所特有的语言、思维、文化层级的认知;“塞尔标准的意义在于:机器智能是有限度的,它永远不可能超过人类智能;同时,机器智能向人类智能的接近却是无限度的,机器智能总可以无限逼近人类智能”[8]。

三、解构:人工智能刑法“主体性”之否定

(一)解构对象之明晰——辨认能力与控制能力

我国传统刑法理论是以自然人为主体进行构建的(即使是单位犯罪,也是以自然人为基础构建起来的)。既然有论点以人工智能可以像自然人一样具有思维为由主张将人工智能作为刑法主体进行刑罚处罚,那么就需要把人工智能套入现有刑法理论,考察其能否契合。当然,该论点还主张刑法需要作出调整以适应人工智能对刑法提出的新挑战(如刑罚内容增加删除数据、修改程序、永久销毁等)。但是,即使刑法作出调整,也必须保留其最本质的要素,如承担刑事责任需要具有责任能力⑦。支持人工智能作为刑法主体的观点也以人工智能具有辨认能力与控制能力为其主要论据。“强人工智能产品与弱人工智能产品的本质区别在于是否具有辨认能力和控制能力”[9]。故本文将人工智能的责任能力,即控制能力与辨认能力作为解构对象,进而对人工智能作为刑法主体的论点进行否定。

以人工智能的责任能力欠缺作为反驳人工智能作为刑法主体的方式表面上会存在一种质疑:责任是确认违法性之后所进行的判断,该方式虽然反驳了人工智能的责任能力,但并不能否定其“行为”的违法性,也即不能否认人工智能的“行为”是刑法意义上的行为,从而没有将人工智能从刑法主体中完全剔除。该种质疑仍然无法成立,传统刑法理论的行为理论都是以自然人为基础构建起来的,因果行为论、社会行为论、目的行为论、人格行为论都无法适用于人工智能,但责任能力的本质是意思能力⑧,有无责任能力的实质是有无意志自由问题,故否认责任能力就可以否定人工智能具有意志自由。自然人生来就具有作为“人”的生物特征,我们不需要对其是否是“人”进行证成。人工智能正好相反,其并不具有“人”的生物特征,在证明其也不具有思维、意志自由等要素时,就可以将其从刑法主体中完全排除。

(二)人工智能辨认能力之否定

“刑事责任能力中的辨认能力,是指行为人具备的对自己的行为在刑法上的意义、性质、后果的分辨认识能力”[11]。人工智能作为刑法主体论点的支持者以传感器技术的急速发展为由,认为人工智能所具有的对外界感知的能力已经超越人类,进而认为人工智能的辨认能力已经超越人类。其实并非如此。辨认能力在事实上存在两步:第一步,对外界进行感知与识别;第二步,将感知到的信息与自身“行为”联系进行加工处理从而得出该“行为”的社会意义⑨。该加工处理过程不只是简单的比对过程,而

---

⑦本文区分使用“刑事责任”与“责任能力”概念,“刑事责任”在刑事法律后果层面使用,“责任能力”在“辨认能力”与“控制能力”意义上使用。

⑧关于责任能力的性质,有旧派与新派两种立场,旧派基于非决定论的立场认为,责任能力的本质是有责行为能力、意思能力或犯罪能力;新派基于决定论的立场认为,责任能力是刑罚适应能力。本文采取非决定论的旧派观点,主张责任能力是犯罪能力。

⑨由于刑法中的行为理论是针对自然人而言的,而人工智能的对外交互方式无法对应刑法理论中的行为,故本文关于人工智能的“行为”的表述与刑法中行为理论中的行为不在同一意义上使用,仅指人工智能的对外交互方式。

是需要经验层面的判断。

第一,关于对外界的感知与识别层面,当前人工智能不具有人类的神经认知与心理认知。本文第二部分对人工智能与人类智能神经层级的认知与心理层级的认知进行了对比,可以看出,人工智能在这两个层级的认知上都无法超越人类,而只是对人类智能形式上的模拟。在神经层级认知上,某些人工智能可以通过传感器精确地探知物体的材料性质或所含元素的类型与比例,而人类的触觉无法做到,但并不能以此来论证人工智能在神经层级的认知上或在辨认能力上超越人类。对物质性质与元素比例的精确测量恰好反映了其"机器"属性,因为该测量并不属于神经层级认知的范围,该种测量是通过一系列物理和化学作用得出的结论,与人类神经认知中外界对神经的刺激有本质的区别。如前所述,如果人工智能在神经认知层面上对人类智能的模拟有以假乱真的现象,那么在心理认知层面,其连外在形式上的模拟都没有做到。

第二,关于将"感知"到的信息与自身"行为"联系进行加工处理从而得出该"行为"的社会意义,人工智能完全没有这样的能力。人类处理信息的方式是将神经认知得到的信息结合自己心理层级的认知、语言层级的认知、思维层级的认知与文化层级的认知,进行整合而产生的心智层面的内容。而人工智能的处理方式是:第一步,通过各种传感器将外部环境的特征进行输入,并转换成人工智能可以识别的人工语言。第二步,通过人工智能的"计算力"对得到的信息按照提前输入的算法进行处理。通过上述区分可以看出,人类所具有的辨认能力是经验的,而不是原生的,是在低阶认知(神经层级的认知、心理层级的认知)的基础上,通过语言,进行思维和文化上的考量。而人工智能具有的所谓"辨认能力"是通过"输入—存储—计算"的模型进行的,计算的依据就是人类设置的算法,也即塞尔"中文房间模型"中的"操作规程"。故人工智能具有的所谓的"辨认能力"仅仅是其具有的"计算力"而已,其并没有思维和文化上的考量。近年来所发生的所谓的人工智能战胜人类智能的事件,仅仅是由于科学技术的发展,使得人工智能计算力急剧提升而出现的假象。综上,当下及可预见的未来的人工智能并不具有刑法意义上的"辨认能力"。

(三)人工智能控制能力之否定

"刑事责任能力中的控制能力,是指行为人具备决定自己是否以行为触犯刑法的能力"[12]。传统的刑法理论认为,在责任能力层面,控制能力以具有辨认能力为前提,也即没有辨认能力就一定没有控制能力。前文论证了人工智能不具有辨认能力,逻辑上就不用探讨其控制能力的问题,但前述无辨认能力就无控制能力的逻辑是以自然人为基础构建的。自人工智能能否作为刑法主体的讨论伊始,其就具有一定的特殊性。人类的心智会随着时间延续而递增,并伴随着经验的痕迹,这也是自然人会有刑事责任年龄等理论问题的事实支撑,而人工智能所谓的"心智"不是经验的,而是原生的,并不是一个逐渐发展的过程,这就使得人工智能即使被证明没有辨认能力,但在形式上仍然存在看似具有控制能力的假象,如运用人工智能技术的无人驾驶汽车[11]。虽然无人驾驶汽车不具备人类所具有的辨认能力,但是在表面上,其具有控制自己在道路上规范行驶的能力,故仍然有必要对人工智能的"控制能力"进行证伪。

同人工智能"辨认能力"证伪路径相同,其形式上所具有的控制能力与人类所具有的控制能力具有本质的区别。人工智能所具有的形式上的"控制能力"其实是在算法内运行的能力,其仍然是机械的、固化的,其机械与固化的依据仍然是人类设置的算法。例如作为人工智能的无人驾驶汽车,在形式上是经过传感器识别后作出看似有控制能力的行为,但在实质上其仍然是机械行为,只是由于其计算力的提升而在形式上更加的"智能"。人类智能的控制能力是先进行神经层级、心理层级的感知,再结

合语言层级的认知、思维层级的认知、文化层级的认知进行判断所形成的控制能力。人工智能“控制”的过程是先通过传感器对外部环境进行识别,再将识别到的信息转换成人工语言,然后在算法内进行判断。故人工智能的“控制”并不是其自主的控制,而仍然是在“操作规程”之内进行判断,也即其是在人类控制之下进行的控制,其本质仍然是人类控制。

(四)人工智能刑法主体地位的再诘问

人工智能刑法“主体性”不仅能从辨认能力与控制能力层面予以否定,其也无法面对来自刑罚、哲学理论的诘问。

第一,针对人工智能的“刑罚”的合理性缺乏足够的论证。现阶段,针对人工智能的“刑罚”——“删除数据、修改程序、永久销毁”看似是一个形式上较为合理的选择,也是人工智能具有刑法主体性支持论者所构想的蓝图,但“删除数据、修改程序、永久销毁”是否是刑罚并没有得到有力的论证。如果认为其是刑罚被规定在刑法中,会出现一系列无法解释的悖论:(1)这些适用于人工智能的“刑罚”和适用于人类的刑罚规定于同一部法律都被作为刑罚,那么删除数据、修改程序等有没有适用于人类的风险(如果技术可以使这些措施适用于人类),如果这些“刑罚”被适用于人类,是否人道?(2)将这些措施规定于刑法中,这样的刑法还符合刑法的本质吗?(3)与其将“删除数据、修改程序、永久销毁”作为“刑罚”措施规定在刑法中,为什么不单独制定一部关于人工智能行为规范的法律呢?

第二,将人工智能作为刑法主体有损人类尊严。世界范围而言,人类意识的觉醒并不是很久远的事。文艺复兴之前,经历了原始社会、奴隶社会与封建社会,在文艺复兴之前,即使存在人类意识的觉醒也只在极小范围内片段性地发生,文艺复兴之后,人本意识才在全世界范围内传播开来。当前由于大数据、传感器科技、芯片技术的急速发展,而使机器在形式上具有人类智能的假象。由于这些“类人”的假象而将人工智能提升到人的高度有损人之所以为人的尊严。现代人类所拥有的哲学财富都是以人为主体构建起来的主客体关系,如果将人工智能提升到与人类同样的高度而将其作为构建世界哲学主客体关系的主体,不仅人类的主体地位会遭受结构性挑战,人类积淀起来的哲学财富也将会受到巨大的冲击。从历史唯物主义的视角看,人类的主体地位是由“全部人类与自然世界以及与自身不懈斗争,不断进取,对自然界取得日益强大的实践与改造能力,通过文明发展的历史事实自证的”[13]。即使在不可预见的未来出现具有思维的强人工智能,我们仍然可以对其单独立法,而不是与人类共用刑法。在知识产权法领域,就有学者以此路径保护人工智能的创作物。“对人工智能创作物,一方面法律应当给予一定的保护;另一方面,这种保护又应当考虑其特有的品性,体现与对人类作品保护的区别。人工智能创作物保护路径的选择应当秉持这一基本的理念”[14]。

四、反思:人工智能的刑法地位

(一)人工智能“风险”的刑法立场

人工智能虽然不具有刑法主体地位,不能对其适用刑罚,但人工智能相关科学的指数爆炸式增长的确极大地增加了社会风险,对现有的法律体系提出巨大的挑战,刑法也不例外。在此背景下,刑法需要表明自身立场,作出回应,以应对社会风险的加剧。

1986年,德国社会学家乌尔里希·贝克(Ulrich Beck)提出了“风险社会”概念。我国刑法学界也针对“风险刑法理论”进行了大讨论,但正如有学者所言,“由于支持者和批评者都没有充分了解风险社会理论的真面目,这场看似激烈的争论其实并未深入本质”[15]。风险刑法理论曲解了贝克所提出的风险社会之风险,贝克所提出的风险社会之风险是“可以被界定为系统地处理现代化自身导致的危险

和不安全感的方式。风险,与早期的危险相对,是与现代化的威胁力量以及现代化引致的怀疑的全球化相关的一些后果”[16]。由此可以看出,风险社会理论之风险与风险刑法理论之风险并不一致,风险刑法理论之风险仍然是古典工业社会之风险,仅仅在程度上(量)有所上升,并没有质的区别。而风险社会之风险是具有毁灭性的全球性风险,具有异质性,其一旦发生就没有救济的余地。而这种全球性的风险在古典工业社会“发展”思想指导下的治理逻辑中是合法、合理的,正如贝克所言,“(风险)由工业制造出来,被经济外部化,被法律制度个体化,被自然科学合法化,且被政治变得表面上无害”[17]。故风险社会之风险是古典工业社会的治理逻辑(包括法律)所不能解决的。

风险刑法理论与风险社会理论所指风险并不同一,但这并不代表风险刑法理论没有存在的根基,风险刑法理论仍然可以以传统的古典工业社会之风险为基础。毕竟随着科技的发展,传统风险正在急剧增多是个不争的事实。在此背景下,风险刑法理论仍然有生存发展的土壤,其提出的刑法规范提前介入在某些情况下仍然可行。总之,笔者并不否认当下正处于风险社会,但风险社会与古典工业社会并不相悖,可以同时存在,也即当下是两种风险并存的时代,而风险社会之风险是刑法无法应对的,刑法只能针对传统风险。众所周知,当下传统风险正在加剧,如人工智能技术的发展增强了潜在犯罪人的犯罪方式,扩大了犯罪后果,加大了侦查机关的侦查难度(在一定程度上扩大了犯罪暗域,增大了社会的不稳定性),刑法对此应采取积极态度。

虽然人工智能并非一个新的概念,1956年达特茅斯(Dartmouth)会议首先使用“人工智能”术语,但人工智能在当时由于科技的时代局限性而未对生活产生深刻的影响,近年来才因相关科技的发展对社会生活产生了较为深远的影响,也是近年来才致使社会风险增多。人工智能“风险”仍然属于古典工业社会之风险,其并不具有异质性,所带来的可能也并非毁灭性的全球性灾难。故人工智能“风险”是社会治理的新课题,也是刑法所面临的新课题。对此,刑法并非像对待风险社会之风险那样束手无策,而应在不违反罪刑法定原则、刑法谦抑性原则的基础上,积极应对涉人工智能犯罪。

(二)人工智能的刑法地位——刑法客体

人工智能的刑法主体地位前文已经证伪,其在刑法中只能作为客体存在⑩。具体而言,行为人可能利用人工智能实施犯罪行为,或行为人可能针对人工智能实施犯罪,此即人工智能的犯罪工具属性和犯罪对象属性。第一,人工智能的犯罪对象属性。传统观点认为,“行为对象也叫犯罪对象(行为客体),一般是指实行行为所作用的物、人、组织(机构)、制度等客观存在的现象”[18]。以人工智能为对象的犯罪行为与以其他物或知识产权等为对象的犯罪并无本质区别。例如行为人盗窃、毁坏人工智能会构成相应的侵犯财产类犯罪;再如人工智能作为“作品”(知识产权)形式存在,行为人侵犯该知识产权就构成相应的侵犯知识产权犯罪。第二,人工智能的犯罪工具属性。现实中确实存在一类看似人工智能“犯罪”的案件,诸如完全自动化的无人驾驶汽车交通肇事案件、智能程序在算法之外错误运行造成极大财产损失的案件等。人工智能“犯罪”事件,其实是人类犯罪的表现方式,本文将其分为两类:第一类为人类故意犯罪;第二类为人类过失犯罪。故意犯罪是指人类制造人工智能的目的就是为了犯罪或者人工智能最初的制造目的是为了服务于人类,但是人类故意破坏其原本的算法或程序以达到犯罪的目的,如行为人恶意更改他人完全自动化无人驾驶汽车而发生车祸。此种类型的故意犯罪中,人工智能就是“犯罪工具”,和杀人犯手中的刀没有本质的区别。而人工智能作为刑法主体的支持论者认为有自动更改自身算法或程序的完全自动化无人驾驶汽车的存在,其通过更改自身程序进行“杀

⑩此处“客体”并非犯罪论中犯罪客体,而是哲学关系中主体以外的客观对象,是主体认识和实践的对象。

人”。但通过前文认知科学的五个层级理论及塞尔的“中文房间模型”可以看出,在当下及可预见的未来,不存在这种所谓的具有自我意识的“强人工智能”。

涉及人工智能的过失犯罪有两种:(1)人工智能自身运行错误而造成的“事故”;(2)产品自身缺陷而引发的“事故”。第一种情况,如果人工智能所有者具有监督责任而没有履行,其就构成相关的过失犯罪。例如行为人有定期对完全自动化无人驾驶汽车进行检查的监督责任,而行为人没有履行监督义务,就需要对无人驾驶汽车程序运行错误造成的“事故”承担过失责任;再如对于非完全自动化无人驾驶汽车,驾驶人需要在汽车行驶过程中对汽车进行监管,因驾驶人的监管失误而发生交通事故的,驾驶人需要承担过失责任。对于第二种情况,则适用产品责任的相关规定。值得注意的是,在当今技术细分精细化的背景下,人工智能产品责任事故可能不止一位责任者,要区分人工智能产品“系统”与“实体”。例如采用人工智能技术的无人驾驶汽车会有人工智能“系统”的制造者和汽车实体部分的制造者,在一起案件中要区分是系统产品责任还是实体产品责任。

结语

近年来关于人工智能刑法的主体性问题似乎在一瞬间成为刑法学者讨论的热门问题,但该问题在现阶段是否是一个严肃的法学问题仍值得商榷。笔者认为,人工智能当下仍然处于萌芽阶段,所谓的强人工智能在当下和可预见的未来并没有出现的迹象,故在现阶段,在刑法层面讨论其主体性问题并非严肃的法学问题,当前刑法理论应该把更多的注意力着眼于人类利用人工智能犯罪问题、人工智能辅助量刑等问题上。人工智能被用来实施犯罪行为会使得犯罪后果扩大,侦查难度增大(犯罪暗域增大),例如证券公司利用人工智能买卖证券,行为人篡改该人工智能程序实施操纵市场行为等。在司法领域,人工智能无法替代法官审判案件,但其辅助法官办案将成为新的趋势。例如在量刑时,“量刑人工智能以大数据分析为基础”[19]获取同类案件的常态量刑,作为法官的量刑参考。一直以来,我国量刑理论在精致的犯罪论的反衬下显得无比简陋,正如王利荣教授所言,“在规范刑法学中找不到令人信服的量刑规则和量刑步骤,不能断言是研究者对其不重视,可能这一研究范式一开始就命定了结局”[20]。故人工智能在量刑领域的应用将极大地促进司法公正。

参考文献:

[1] 刘宪权,胡荷佳. 论人工智能时代智能机器人的刑事责任能力[J]. 法学,2018(1):40-47.

[2] 刘宪权. 涉人工智能犯罪刑法规制的路径[J]. 现代法学,2019,41(1):75-83.

[3] 徐小奔. 论人工智能生成内容的著作权法平等保护[J]. 中国法学,2024(1):166-185.

[4] 蔡曙山,薛小迪. 人工智能与人类智能:从认知科学五个层级的理论看人机大战[J]. 北京大学学报(哲学社会科学版),2016,53(4):145-154.

[5] 蔡曙山. 论人类认知的五个层级[J]. 学术界,2015(12):5-20.

[6] 蔡曙山. 自然语言的形式理论[M]. 北京:人民出版社,2010:6.

[7] 王钢. 人工智能刑事责任主体否定论:基于规范与语义的考察[J]. 苏州大学学报(法学版),2022(4):63-79.

[8] 蔡曙山. 哲学家如何理解人工智能:塞尔的“中文房间争论”及其意义[J]. 自然辩证法研究,2001,17(11):18-22.

[9] 刘宪权. 人工智能时代的“内忧”“外患”与刑事责任[J]. 东方法学,2018(1):134-142.

[10] SEARLE J R. Minds,brains,and programs[J]. Behavioral and Brain Sciences,1980,3(3):417-424.

[11] 贾宇. 刑法学(上册·总论)[M]. 北京:高等教育出版社,2016:302.

[12] 高铭暄,马克昌. 刑法学[M]. 北京:北京大学出版社,2016:85.

[13] 张力,陈鹏. 机器人“人格”理论批判与人工智能物的法律规制[J]. 学术界,2018(12):53-75.

[14] 许明月,谭玲. 论人工智能创作物的邻接权保护:理论证成与制度安排[J]. 比较法研究,2018(6):42-54.
[15] 南连伟. 风险刑法理论的批判与反思[J]. 法学研究 2012,34(4):138-153.
[16] 乌尔里希·贝克. 风险社会[M]. 何博闻,译. 江苏:译林出版社,2004:19.
[17] 乌尔里希·贝克. 世界风险社会[M],吴英姿,孙淑敏,译. 南京:南京大学出版社,2004:49.
[18] 张明楷. 刑法学[M]. 北京:法律出版社,2016:163.
[19] 甄航. 人工智能介入量刑机制:困境、定位与解构[J]. 重庆大学学报(社会科学版),2023(4):191-202.
[20] 王利荣. 量刑说理机制[M]. 北京:中国人民公安大学出版社,2012:14.

The negation of the "subjectivity" of artificial intelligence in criminal law: Origins, deconstruction and reflection: based on the five-level theory of cognitive science

ZHEN Hang

(*Law School, Southwest University of Political Science and Law, Chongqing 401120, P. R. China*)

Abstract: Whether artificial intelligence can have the same "subjectivity" status as human beings in criminal law cannot be found in the criminal law theory, and it needs to be based on relevant cognitive science, otherwise it will fall into the dilemma of circular argument. According to the five-level theory of cognitive science, AI is only a simple simulation of human cognition at the lower-order cognitive levels of the neural and mental hierarchies; At the level of language-level cognition, which is the intermediate link between higher-order and lower-order cognition, there is an essential difference between the artificial language of AI and natural language; At the higher-order cognitive levels of the thinking and cultural hierarchies, current AI has not shown itself to be capable of thinking or generating culture. Combined with the "the Chinese room argument", the AI does not have the "recognition ability" and "control ability" in criminal law theory. In terms of recognition ability, although the recognition of the objective world by AI sensors can mimic human cognition, they are not able to process the recognized information in conjunction with their own "behaviors" to derive the social significance of those "behaviors". In terms of control capability, the "control capability" demonstrated by AI is essentially the ability to execute algorithms, which is still essentially a form of human control rather than AI "self-control". Therefore, in the present and foreseeable future, AI does not have a "subjectivity" of criminal law, and current criminal law theory need not overreact to so-called "strong artificial intelligence". Elevating AI to human heights undermines human dignity. "Delete data, modify procedures, permanent destruction" is not a penalty. If it is written into the criminal law also has the possibility of human application. Therefore, in the present and foreseeable future, the impact of AI on the theory of criminal law is mainly due to the fact that it leads to the exacerbation of traditional societal risks. Criminal law theory should treat them as "objects" and "instruments" of crime in conjunction with the theory of risk criminal law. When AI is used as an object of crime, it exists in the form of property, works, etc., and care should be taken to distinguish between the AI itself and the carrier of the AI in the process of judicial determination; When it is used as an instrument of crime, it can lead to results such as widening the consequences of the crime and making it more difficult to investigate.

Key words: artificial intelligence; the Five-Level Theory of Cognitive Science; strong artificial intelligence; the Chinese room argument

(责任编辑 胡志平)