理论前沿

人工智能是适格的刑事责任主体吗?

叶良芳

内容提要:肯定“人工智能是适格的刑事责任主体”的主张,是以未来强人工智能的实现为前提的。但如果将智能理解为不仅是编程算法还包括心智意识,则这一主张是不能成立的。设计具有人类心智的人工智能,在根本上是不可能实现的。首先,在认知方面,人工智能仅在单项领域具有较高的认知能力,而在跨界领域却存在难以逾越的技术瓶颈。其次,在意识方面,人工智能既不可能内生自发地生成意识,也不可能外在强行地输入意识,因而不可能具有自由意志。最后,在情感方面,人工智能不可能具有人类的情感动机,无法体验犯罪之乐和刑罚之苦,因而不是适格的受罚主体。

关键词:强人工智能　刑事责任主体　认知　意识　情感

叶良芳,浙江大学光华法学院教授。

一　问题的缘起

“人工智能”这一术语首次出现于1956年夏季在美国达特茅斯学院举办的一次学术研讨会。这次会议的主题是如何开发出像人类一样能够从现有知识中进行自主学习的算法系统。此后,人工智能的发展经历了两次浪潮、两次寒冬的洗礼。近年来,由于大数据、深度学习和神经网络等技术的广泛应用,人工智能研究峰回路转,复为最热门的研究领域,迎来了第三次浪潮。从学科分布来看,不仅数学、计算机科学、信息科学、管理学、统计学、神经生理学、控制论等自然科学先后将人工智能作为研究的重点,而且哲学、伦理学、人类学、社会学、语言学、经济学、法学等社会科学也竞相在各自的研究中增添人工智能的元素。

在这股如火如荼的研究热潮中,刑法学者也不甘落后,纷纷撰文著说,提出不少新锐的见解和主张。与其他学科不同的是,刑法学者的研究主要集中于人工智能的刑事责任主体资格问题,即对于人工智能在设计和编制的程序范围之外独立实施的“犯罪行为”,能否要求其承担相应的刑事责任?对此,存在肯定和否定两种对立的观点。肯定说认为,

刑事责任主体不应局限于人类,强人工智能也可以成为刑事责任主体。智能机器人“一旦通过深入学习,产生自主意识和意志,在设计和编制的程序范围外实施符合犯罪构成的具有严重社会危害性的行为时就应当承担刑事责任”。〔1〕“只要满足刑法的所有相关要求,除人类个体和法人外,一种新型主体也可以归入到现行刑法责任主体的群体中。”〔2〕否定说认为,刑事责任主体只能是自然人和法人,人工智能不能成为独立的刑事责任主体。“人工智能不具有刑法评价意义上的行为认识与控制要素的自主性,也就不具有人的目的理性所支配的可答责基础,欠缺作为刑事责任主体的本质要素。”〔3〕“智能代理缺乏作为人的至关重要的因素:尽管它能够学习并且做出其他人无法预知的决策,但它对于自身的自由并无意识,更遑论将自己视为社会权利义务的承担者。”〔4〕

上述理论之争表面上看仅限于刑法领域,但从深层次看,实际上涉及到人工智能的认知、思维、判断、决策、意识、情感等问题,进而还涉及到人的本质以及人与工具、物质与精神的界分等哲学问题。只有对这些问题进行正本清源式地剖析和探究,才能正确地回答人工智能的刑事责任主体资格问题。然而,目前的否定论对肯定论的反驳却犯了论证方法错误,要么简单地陈述自己的论点而没有予以系统和详细的论证(“不论证的否定论”),要么强调目前人工智能发展的低阶段性而不可能具备承担刑事责任的要件(“摇摆的否定论”),但却承认高阶段的人工智能可以成为独立的刑事责任主体,因而都是一种“不彻底的否定论”。本文坚持“彻底的否定论”,即认为无论是现在还是将来,人工智能都不是适格的刑事责任主体。

二 人工智能:一个基本概念的框定

在理论研究中,经常存在这样一种现象:激辩双方的观点针尖对麦芒,水火不相容,但实际上指涉的对象根本就不是同一个事物或概念。“我们的意见之所以分歧,并不是由于有些人的理性多些,有些人的理性少些,而只是由于我们运用思想的路径不同,所考察的对象不是一回事。”〔5〕在人工智能刑事责任主体问题的探讨中,“人工智能”是一个关键词,因而明确其涵义是进行有效对话的前提。

人工智能,“本质上有别于自然智能,是一种由人工手段模仿的人造智能”。〔6〕人类的许多活动,如下棋、竞技、解算题、猜谜语、编写程序、驾驶汽车等,都需要“智能”。如果机器也能够执行这些任务,就具有某种“人工智能”。但这样理解,只是指出人工智能与人类智能的生成原因不同,而未能揭示其本质特征。而要揭示人工智能的本质特征,则必

〔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》,《法学》2018 年第 1 期,第 47 页。

〔2〕 Gabriel Hallevy, *When Robots Kill: Artificial Intelligence under Criminal Law*, Northeastern University Press, 2013, p. 21.

〔3〕 时方:《人工智能刑事主体地位之否定》,《法律科学》2018 年第 6 期,第 71 页。

〔4〕 〔瑞士〕萨比娜·格雷斯、〔德〕托马斯·魏根特:《智能代理与刑法》,赵阳译,载陈泽宪主编《刑事法前沿》(第十卷),社会科学文献出版社 2017 年版,第 222 页。

〔5〕 〔法〕笛卡尔著:《谈谈方法》,王太庆译,商务印书馆 2000 年版,第 3 页。

〔6〕 蔡自兴等著:《人工智能及其应用》,清华大学出版社 2016 年版,第 2 页。

须首先明确“智能”的定义。但关于智能的定义,理论上却存在严重的分歧。例如,心理学家斯腾伯格认为:“智能,是指个人从经验中学习、理性思考、记忆重要信息以及应付日常生活需求的认识能力。”〔7〕而《麻省理工学院认知科学百科全书》却将智能定义为“适应、影响(改变)和选择环境的能力”。〔8〕分歧的原因,主要在于论者在以下三个维度方面存在不同的理解:第一,智能是只能以精神状态存在还是可以独立地客观存在;第二,智能的主体是生物体还是非生物体,生物体是否包括动物、植物;第三,智能是否存在逻辑或算法复杂度的标准。〔9〕

智能定义的多样性,必然导致人工智能定义的多样性。有学者分析,人工智能可以从两个维度和四个途径来定义,共有八种定义。这两个维度分别是思维和行动,四个途径则是图灵测试、认知建模、思维法则和合理行动。根据智能成功与否的不同判断标准,这些定义可以分为两类:一类是根据人类逼真度来衡量,另一类是依靠一个称为合理性的理想的表现量来衡量。〔10〕另有学者统计,根据不同的理解,人工智能有十三种定义。〔11〕在这些定义中,人工智能既可能指一门科学,也可能指一种能力,还可能是指一种智能机器等。人工智能定义的多样性,也带来了人工智能分类的不统一性。最常见的分类,是将人工智能分为弱人工智能和强人工智能。前者是指专注于解决特定领域问题的人工智能,没有意识和意志,只能简单地执行交付的任务,而后者是指可以胜任人类所有活动的人工智能,能够感知自我和环境的存在,并对目标任务进行调整。简言之,前者是指“机器能够智能地行动(其行动看起来如同它们是有智能的)”,后者是指“能够智能行动的机器确实是在思考(不只是模仿思考)”。〔12〕前者重在程序可执行性,“人造物是否使用与人类相同的方式执行任务无关紧要,唯一的标准就是程序能够正确执行”;后者重在生物可仿真性,“当人造物展现智能行为时,它的表现基于人类所使用的相同方法。”〔13〕也有的在强人工智能的范围内还划分出一种超人工智能,即“在许多普遍的认知领域中,表现远远超越目前最聪明的人类头脑的智能”。〔14〕还有的将人工智能分为应用人工智能和通用人工智能,限制人工智能和完全人工智能,或者智能增强和人工智能,这些分类大体上相当于弱人工智能和强人工智能的分类。

上述简短的勾勒表明,给人工智能下一个精准的、一致认同的定义是相当困难的,笔者也无意作此努力。鉴于本文探讨的主题,笔者是在强人工智能的意义上来理解人工智能的,即将其理解为在学习记忆、思维决策、行动规划、情绪情感等方面具有与人类同等的感知、认知和意志能力的机器智能体或程序软件。应当注意的是,上文关于人工智能刑事

〔7〕 Robert J. Sternberg, *Search of the Human Mind*, Harcourt-Brace, 1994, pp. 395 – 496.

〔8〕 Robert A. Wilson and Frank C. Keil, *The MIT Encyclopedia of the Cognitive Sciences*, MIT Press, 2001, pp. 409 – 410.

〔9〕 杨学山著:《智能原理》,电子工业出版社 2018 年版,第 124 – 129 页。

〔10〕 Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010, pp. 1 – 5.

〔11〕 参见蔡自兴等著:《人工智能及其应用》,清华大学出版社 2016 年版,第 2 – 3 页。

〔12〕 Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010, p. 1020.

〔13〕 〔美〕史蒂芬·卢奇、丹尼·科佩克著:《人工智能》,林赐译,人民邮电出版社 2018 年版,第 11 页。

〔14〕 〔英〕尼克·波斯特洛姆著:《超级智能:路径、危险性与我们的战略》,张体伟、张玉青译,中信出版社 2015 年版,第 63 页。

责任主体的肯定论和否定论之争,虽然针对的对象都是人工智能,但二者所指涉的人工智能其实并不是同一个概念。肯定论者着眼于未来,其所指的人工智能是强人工智能;否定论者却立足于当下,其所指的人工智能是弱人工智能。就此而言,否定论"偷换"了肯定论的命题,因而是一种偏离靶标的反驳。真正的对话,应当直面"强人工智能是适格的刑事责任主体"这一命题,围绕这一命题进行充分的讨论。

具体而言,按照事先设计和编制的程序实施犯罪的人工智能(弱人工智能),是程序的研发者或使用者的工具,实现的是后者的意志,故应由后者承担相应的刑事责任。对此,肯定论和否定论的观点是一致的。形成聚讼的是,对于人工智能在设计和编制的程序范围之外实施的"犯罪行为",是否可以要求其独立承担相应的刑事责任(是否同时追究研发者或使用者的刑事责任暂且不论)?对此,肯定论的立场是完全可以,否定论的立场是,现在不可以,但不否定将来可以(取决于人工智能的发展程度)。事实上,这一问题本身隐含着一个前提性假设:人工智能发展到一定阶段后,将获得与人类完全相同(强人工智能)甚至更高(超人工智能)的认知能力和意志能力,即成为"人造人"。然而,值得认真反思的是,这一前提性假设是否成立?如果这一前提性假设本身是一个伪命题,即所谓的"强人工智能"永远不可能实现,那么关于人工智能的刑事责任主体资格的争论就类似于唐吉诃德大战风车,因为所谓的强敌纯粹是无端臆想的人物。而恰恰是在这一点上,否定论没有注意辨别,而是落入肯定论的圈套,毫不怀疑这一前提性假设成立的可能性。

三 奇点来临:一个想象的反乌托邦

肯定论的立论前提是强人工智能,确切地说,是强人工智能时代机器智能体"反控"人类所带来的风险。这一反控人类的时间点,被称为"奇点"。据说,届时人工智能体可以自我复制和自我提高,从而在数量和智力方面超过人类,并决定人类的所有事务。

"奇点"这一术语源自计算机之父冯·诺依曼和数学家乌拉姆的一次谈话,他们模糊地预测到,技术发展到一定程度将会达到某个重要的奇点,从而使人类生活模式发生革命性的变化。[15] 但他们并未提出技术将反控人类的观点,奇点理论的真正始作俑者是库兹韦尔。他将摩尔定律简单地包装后抛出一种"加速回归定律",大意是:技术正以指数级规模快速发展,在未来的某个时刻,人工智能的技术发展将接近于无限大,进而出现连续性发展过程中的突发性中断。这一神秘的时刻即为奇点。他还设定了奇点出现的时间表:强人工智能将在 2029 年出现,而超人工智能将于 2045 年出现,届时生物体的"碳基人"将受制于非生物体的"硅基人"。"我把奇点的日期设置为极具深刻性和分裂性的转变时间——2045 年。非生物智能在这一年将会十亿倍于今天所有人类的智慧。"[16] 由于他集发明家、演说家、未来学家于一身,加之一系列成功的公关活动,因而其鼓呼的奇点理论虽然极具震撼性和颠覆性,但却拥趸者众,无论是在政界、商界还是学界,无论是普罗大

〔15〕 See Stanislaw Ulam, Tribute to John von Neumann, 64 (3) *Bulletin of the American Mathematical Society* 5 (1958).

〔16〕〔美〕雷·库兹韦尔著:《奇点临近》,李庆诚等译,机械工业出版社 2011 年版,第 80 页。

众还是技术精英。

然而,也有不少学者对奇点理论提出强烈的质疑和批驳。美国哲学家休伯特·德雷福斯(Hubert Dreyfus)指出,人工智能的研究好比炼金术士炼金,在目标定位上就存在方向性错误。"把人类理智全部分解成离散的、确定的、与上下文环境无关元素的规则来支配运算,是可能的吗?逼近人工智能的这一目标究竟是不是可能的呢?两者的答案是一个:不可能。"[17]英国数学家罗杰·彭罗斯(Roger Penrose)认为,人类判断数学真理的过程是超越任何算法的。"任何特定的形式系统都具有临时和'人为'的品格,在数学的讨论中,这类系统的确起着非常有价值的作用,但是它只能为真理提供部分(或近似)的导引。"[18]英国哲学家玛格丽特·博登(Margaret Boden)认为,奇点信徒们关于奇点出现之后的预言违背常理,超人主义哲学近乎荒谬。"人类心智何其丰富,我们还需要与其工作方式相关的良好心理/计算理论。人类水平的强人工智能的前景看起来黯淡无光。"[19]美国人工智能专家皮埃罗·斯加鲁菲(Piero Scaruffi)将奇点理论譬喻为运用倒叙法讲故事的宗教。"奇点理论最坏的影响无疑是它逐渐成为既不研究历史和哲学,也不学习科学,甚至连计算机科学都不曾涉猎的高科技怪人的宗教信仰。"[20]

笔者认为,就技术层面而言,人工智能所拥有的知识总量达到甚至超越人类智能是有可能的,但认为人工智能具有人类的心智进而反控人类的观点,则无异于痴人说梦,甚至是有意混淆视听。[21] 人工智能绝对不可能超越人类,所谓的奇点来临的论断不是危言耸听,就是杞人忧天。事实上,这些人的预言相当随意,既没有严密的逻辑推理,也缺乏充分的事实根据。"因为是预言,专家们并不需要为背后的逻辑自洽负责,但这些随口说出的一个年份的预言会让敏感的公众忐忑不安。"[22]

四 认知思维:一个难以突破的技术瓶颈

从发展史来看,人工智能研究大体有以下三种技术路径:符号主义、连接主义和行为主义三大学派。[23] 符号主义学派认为,人类认知的基本单元是符号,认知过程就是对符号的逻辑运算过程。通过计算机来处理数理逻辑运算,可以模拟人类的抽象思维,实现机械化的人类认知。连接主义学派认为,人类的认知基本单元是大脑的神经元细胞,认知过程是神经元之间的连接和反馈。通过对大脑进行人工仿造,则可以实现人类智能的机器

〔17〕〔美〕休伯特·德雷福斯著:《计算机不能做什么:人工智能的极限》,宁春岩译,三联书店出版社 1986 年版,第 310 页。

〔18〕〔英〕罗杰·彭罗斯著:《皇帝的新脑》,许明贤、吴忠超译,湖南科学技术出版社 2018 年版,第 147 页。

〔19〕〔英〕玛格丽特·博登著:《AI:人工智能的本质与未来》,孙诗惠译,中国人民大学出版社 2017 年版,第 180 页。

〔20〕〔美〕皮埃罗·斯加鲁菲著:《智能的本质:人工智能与机器人领域的 64 个大问题》,人民邮电出版社 2017 年版,第 4 页。

〔21〕人工智能之父马文·明斯基就曾直言不讳地指出:"一门年轻的学科,一开始都需要一点'过度销售'。"尼克著:《人工智能简史》,人民邮电出版社 2017 年版,第 20 页。

〔22〕李开复、王咏刚著:《人工智能》,文化发展出版社 2017 年版,第 127 页。

〔23〕这三大学派的代表人物分别是麦卡锡、明斯基和麦卡洛克-皮茨。参见吕廷杰等编著:《信息技术简史》,电子工业出版社 2018 年版,第 242 页。

模拟。行为主义学派认为,人类智能是感知—运作的结果,经过了不同阶段的增强和进化过程。通过从复制简单动物的智能开始,沿着进化的阶梯逐步向上推进,最终可以复制人类的智能。

客观地说,上述三大学派在推动人工智能研发方面均取得了一定的突破,但这些突破均局限于弱人工智能的范畴,而在强人工智能的领域,实际的发展状况却是"路漫漫其修远兮",就连历史性的第一步都尚未迈出。这是因为,要从弱人工智能发展到与人类认知水平相当的强人工智能,这中间存在诸多难以逾越的技术阈值。

第一,上述三大学派在目标定位上均过于宏大,存在浓厚的理想主义色彩。符号主义学派的立论根基是物理符号假设,目标是发明一种通用问题求解器。但是,物理符号假设是一个经验假设,其表述并不是一个永真的重言式;这个判断排斥了认知构成中其他要素的重要性以及符号系统和非符号系统的外部环境之间的因果关系;这个路径预设了系统对于每类问题所有可能的求解方式的完备知识,而难以被修正或扩容。〔24〕 连接主义学派遵循的是一种仿生学路径,这种路径的典型模式是人脑仿真。然而,人脑是一个极其复杂且神秘的系统,对其进行扫描、建模,是一个不可能完成的任务。"人脑是一个异常复杂的组织,目前对人脑结构和活动机制的了解只是冰山一角,要建立一个与人类大脑相近的神经网络目前看来还是天方夜谭。"〔25〕 行为主义学派从生物进化的角度来完成智能的构建,但其模型是一种自下而上的系统,即事先储备相关知识,建立高层认知任务,然后把任务分解为子任务,从而实现系统的整体性。这种还原主义的方法论不仅导致研究模型不尽完善,而且难以构建复杂的适应系统,还明显忽视了意向性和主观性等人类行为的重要特质。〔26〕

第二,人工智能研发的每一次"重大突破",其实都局限在单一维度的弱人工智能领域。以人机大战为例,无论是 IBM 公司的"深蓝"险胜国际象棋冠军卡斯帕罗夫,还是 DeepMind 公司研发的阿尔法狗完胜世界围棋冠军李世石,都只不过是"海量数据 + 暴力算法"的结果,其实质只是一种统计学的应用,根本谈不上"智能"。具体而言,深蓝其实是一套用于国际象棋的硬件,大部分逻辑规则是以特定的象棋芯片电路实现,辅之以较少量负责调度与实现高阶功能的软件代码。其算法的核心则是暴力穷举:生成所有可能的下法,然后执行尽可能深的搜索,并不断对局面进行评估,尝试找出最佳下法。〔27〕 与深蓝一样,阿尔法狗也是一种大数据集的处理方法。"'阿尔法狗'主要是利用人类过去所有的棋谱数据,在机器的快速搜索和计算下,进行预测和判断。也就是说,机器是靠强大的记忆能力、快速的搜索能力和超级的计算能力战胜人类的,机器并没有多少智慧,只有'蛮力'"。〔28〕 二者所不同的,仅在计算能力方面存在差异:深蓝运行的是原初的机器学习模式;阿尔法狗运行的则是进阶的机器学习模式(深度学习、增强学习、对抗学习等)。

〔24〕 参见徐英瑾著:《心智、语言和机器》,人民出版社 2013 年版,第 42 - 43 页。
〔25〕 王天一著:《人工智能革命:历史、当下与未来》,北京时代文化书局 2017 年版,第 30 页。
〔26〕 参见魏斌等编著:《人工情感原理及其应用》,华中科技大学出版社 2017 年版,第 14 - 15 页。
〔27〕 参见王天一著:《人工智能革命:历史、当下与未来》,北京时代文化书局 2017 年版,第 66 - 69 页。
〔28〕 黄欣荣:《人工智能热潮的哲学反思》,《上海师范大学学报(哲学社会科学版)》2018 年第 4 期,第 40 页。

迄今为止,各种人机大战,犹如甲乙二人拼字比赛,甲全凭自己的记忆,乙背后则有个强大的团队帮忙查词典、找答案。最后,乙险胜。试问,这究竟是人类的胜利还是电子词典的胜利?! 总之,当前人工智能的发展仍然局限在弱人工智能的范畴,其认知水平也是单一的、低级的、浅层次的,与人类认知水平存在指数级差。“计算机只是在那些没有‘感知’的且单一维度的领域远远胜过人类,但是在有‘感知’能力下数以万计的综合性领域的学习上,实际上比人类差得不仅仅是几何级的问题,或许有着巨大的几乎无法逾越的鸿沟。”〔29〕

第三,人类认知方面的一些特殊的感知能力,如常识、框架、直觉等,对其模拟几乎是不可能的。迄今为止,所有机器智能的学习模式,无论是深度学习、迁移学习还是增强学习,本质上都是一种模仿,是对人类已有知识的学习。通常是建立一个数据库,对数据库储存的知识,机器经过反复训练,基本能够熟练掌握;但对数据库尚未储存的知识,机器则一无所知。“机器人没有自身积累的知识,其机器知识库的知识都是特定领域的,并且都是人类输入的。”〔30〕机器认知仅是对既有信息的提取、分析和归类;人类认知则不仅限于此,还包括视觉、听觉等感觉器官和注意、意识等高级认知功能之间高强度的交互。特别是,人类认知都是建立在常识的基础之上的,而机器认知恰恰缺乏这种常识。“人类解决问题的最有效的方法并非建立在大范围搜索的基础之上,而是基于如何使用大量的常识性知识来‘分割和克服’人们面对的问题。”〔31〕例如,即使没有学过欧氏几何,人们也知道两点之间直线距离最短,抄近路就要走直线不能走弯路,拿空纸杯和盛满水的纸杯,手的力度分配要有所不同。常识的获取是一个渐进漫长的过程,且与人类的成长过程不可分离。“人的常识涉及面十分广泛,是从胎儿期开始用十多年甚至更多的时间,从不停息的学习和周围所有人不厌其烦的教育、修正的结果。人对常识的学习所花的时间超过所有领域性、专业性知识的学习。”〔32〕常识源发于童年,机器智能没有童年,因此,要让机器获得这些常识,只能通过建模、建立各个不同的数据库,让机器分别学习。例如,“手抓积木”的动作,只会改变积木的位置,不会改变积木的形状、大小、颜色等。这是三岁小孩都知道的常识,但要让机器智能理解,则将涉及诸多复杂的背景性知识。“仅就人形机器人来说,问题就有手部及抓握算法、搬运过程整体平衡的保持、理解用户姿势及活动、可实现奔跑及跳跃或至少可以流畅运动的腿部设计、摔倒的处理等等”,〔33〕因而,“这种定义必定是非常冗长的,因为这会逼得你事先将事物的任何方面都罗列清楚,并将这些方面在相应的‘框架公理’中予以事先的排除。”〔34〕对于机器学习来说,解决框架问题,唯有进行更精细化的分类以及设置更多的变量。从理论上讲,对于某项任务,如果能够挖掘出其所有的

〔29〕 王骥著:《新未来简史:区块链、人工智能、大数据陷阱与数字化生活》,电子工业出版社 2018 年版,第 253 页。

〔30〕 吴汉东:《人工智能时代的制度安排与法律规制》,《法律科学》2017 年第 5 期,第 131 页。

〔31〕 〔美〕马文·明斯基著:《情感机器:人工智能与人类思维的未来》,王文革等译,浙江人民出版社 2016 年版,第 147 页。

〔32〕 杨学山著:《智能原理》,电子工业出版社 2018 年版,第 399 – 400 页。

〔33〕 〔法〕鲁道夫·格林著:《机器人是人类最好的朋友吗》,孙兆原、应远马译,上海科学技术文献出版社 2017 年版,第 81 页。

〔34〕 徐英瑾著:《心智、语言和机器》,人民出版社 2013 年版,第 8 页。

因量,且对每个因量又能想象到其所有的变量,则对该任务的执行而言,机器智能完全可以达到人类智能的水平。“只要有足够的工程师、时间和处理器,确实有可能创造一百万台机器,做到人类自然而然就会做的事情。”〔35〕但是,在实际操作层面,却存在一个“组合爆炸”问题。〔36〕“同一个问题,哪怕在只有两三个变量时很简单,当变量数目非常多时就会变得非常难,甚至有可能无法解决。所以,简单的人工智能应用在实验室环境中表现很好,可到了实际环境中却毫无用处。”〔37〕而且,现实的复杂性往往超出理论的预设。“人类行为过于复杂而无法通过任何简单的规则集合捕捉到,而计算机所能做的只不过是遵循规则集合,因而他们无法作出像人类一样的智能行为。”〔38〕

机器智能究竟能够在多大程度上习得与人类完全相同的认知能力?按照最初的“老式人工智能”学派的设想,机器智能不仅应具有认知能力,而且在认知程度上可以媲美甚至完胜人类。然而,人工智能的研发应用却表明,机器智能仍然只能在人为设计的有限领域内获得单项的认知能力。“计算机的知识都是这样加工编程而来,因而无法像人类大脑那样与常变常新的客观世界经常保持一致……纵然运算极快,计算机指令也无法对有机世界的永恒变化作出定性的反应。”〔39〕事实上,在当前的人工智能业界,务实的研发人员早已放弃实现强人工智能这一好高骛远的“梦想”,而是退而求其次,专注于在某一单项领域实现机器智能的突破。“大多数人工智能研究者认为弱人工智能假设是当然的,而并不关心强人工智能假设——只要他们的程序可行,他们并不在乎你将其称为‘模拟的智能’还是‘真正的智能’。”〔40〕总之,人类认知能力的魅力,不仅是对旧知的掌握,还有对新知的领悟,即能举一反三、触类旁通地创新学习。但从现有的设计来看,无论是哪种路径,人工智能都是对旧知的“复制”,而不是对新知的“探知”。并且,即使对于旧知的“复制”,人工智能也有相当的局限性。虽然从理论层面看,通用人工智能的实现是可以证成的,但要在技术层面解决,其间的难度或许比翻越珠穆朗玛峰还要大。而机器对新知的创新性学习,无论是在理论层面还是在技术层面,都根本不存在一条可走的路径。因为,人工智能并不具有真正的思维学习能力,所谓的“机器学习”“机器思维”只是拟人化的隐喻而已。

五 自由意志:一个无法模拟的精神实体

恩格斯曾经指出:“如果不谈所谓自由意志、人的责任、必然和自由的关系等问题,就

〔35〕〔美〕皮埃罗·斯加鲁菲著:《智能的本质:人工智能与机器人领域的 64 个大问题》,人民邮电出版社 2017 年版,第 132 页。

〔36〕1973 年,数学家詹姆斯·莱特希尔为英国科学院撰写的报告明确提到了这一问题。

〔37〕〔英〕卡鲁姆·蔡斯著:《人工智能革命:超级智能时代的人类命运》,张尧然译,机械工业出版社 2017 年版,第 11 - 12 页。

〔38〕Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Upper Saddle River, NJ: Prentice Hall, 2010, p. 1024.

〔39〕〔美〕刘易斯·芒福德著:《机器神话(下卷):权力五边形》,宋俊岭译,上海三联书店 2017 年版,第 272 页。

〔40〕Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010, p. 1020.

不能很好地讨论道德和法的问题。”[41]适格的刑事责任主体,不仅需要具备辨认能力,还应当具备控制能力。辨认能力取决于认知水平,控制能力则与意志自由密切相关。道义责任论认为,人的意志是自由的,在人能够选择犯罪也能够不选择犯罪而竟然选择了犯罪的意义上,就可以对人进行道义上的谴责和非难。[42] 随着科技不断的发展,也许未来的人工智能将在辨认能力方面能够达到甚至超过人类水平,但就控制能力而言,人工智能将永远处于“零”的状态,因为其既不能外在强行地输入“意识”,也不能内在自发地生成“意识”。

回顾人工智能的发展史,为了解决人工智能的意识和意志问题,各个学派曾尝试了多种解决方案。一是人脑仿真。根据机械唯物论,意识源自人脑这一物质实体。“人是一架机器;在整个宇宙里只存在着一个实体,只是它的形式有各种变化。”[43]如果一颗大脑能够产生意识,那么其复制品应该也能够产生意识。人工智能专家科赫认为:“如果你能建造一台计算机,其电路和大脑一模一样,这台计算机就能产生意识。”[44]发明家布鲁克斯也认为:“我个人相信我们都是机器,基于这一点,我认为理论上肯定可以用硅品和钢铁造出具有真正情感和意识的机器。”[45]既然意识是通过特定的生物机制产生的,则克隆一颗大脑就可以解决机器的意识难题。然而,人脑结构的高度复杂性、意识形成机理的复杂性以及人类对意识认知的有限性,表明这一路径根本不可能成功。二是机器学习。人类的智能源自童年,孩子在好奇心的驱动下,不断学习,不断获得新知。因此,如果能够开发关于好奇心的算法,复制人类的求知欲,则人工智能也可以获得与人类相同的意识和心智。然而,机器学习仍然是一种暴力算法,是主体在特定情境下作出的反应。即通过反复学习、修正后,大致按照普遍情形与分类来处理,或者通过观察大量事件中人类的案例,从中摸索出经验,通过无监督学习获得。这和孩童的好奇心存在天壤之别。三是脑机对接。它源自“钵中之脑”的设想:假设某人的大脑被从身体上截下并放入一个营养钵,以使之存活。神经末梢同一台超级计算机相连接,这台计算机使大脑的主人具有一切如常的感觉。人群、物体、天空……等等,似乎都存在,但实际上此人所体验到的一切都是从那台计算机传输到神经末梢的电子脉冲的结果。[46] 如果这一“思想实验”可以实现,则人类的意识、感知,就可以通过连线在人类的肉身和大脑之间穿梭传递,冷冰冰的机器将受命于人类的意识和命令。但是,脑机对接在实际应用层面面临着诸多挑战。目前的人与机器的交互方式仍然是单向的,即用户通过操控鼠标、键盘等设备来驱动机器,机器则根据既定的程序或数据库条件性地作出相应的反应。

---

〔41〕《马克思恩格斯选集》(第3卷),人民出版社2012年版,第490-491页。

〔42〕有学者主张用主体选择性这一概念来代替意志自由。即人是实践活动的主体,能够能动地改造和控制周围世界。刑事责任是人在应该而且能够选择符合法律规范的行为时却主体性地选择了违反法律规范的行为因而必须接受的谴责和惩罚。参见冯军著:《刑事责任论》,社会科学文献出版社2017年版,第103页。

〔43〕〔法〕拉·梅特里著:《人是机器》,顾寿观译,商务印书馆1996年版,第73页。

〔44〕转引自〔英〕卡鲁姆·蔡斯著:《人工智能革命:超级智能时代的人类命运》,张尧然译,机械工业出版社2017年版,第90页。

〔45〕Rodney Brooks, *Flesh and Machine: How Robots Will Change Us*, Rep. ed., Vintage Books, 2003, p. 180.

〔46〕参见〔美〕希拉里·普特南著:《理性、真理与历史》,童世骏、李光程译,上海译文出版社1997年版,第11页。

鉴于机器的意识问题的不可解决性,不少学者干脆放弃构建意识的努力,认为智能和意识可以彻底分离,强人工智能并不需要意识。“没有必要预判超级智能是否有意识,或者自我意识。从逻辑上来说,一种思维完全可以有能力做决定,有能力在学习的基础上,比人类更加高效地解决问题,同时对它所做的事情毫无概念。”〔47〕然而,人类智能的精妙之处,恰恰在于人类具有意识,即能够意识到自我和环境的存在,进行思考和判断,作出选择和决定等。对此,马克思早就指出:“有意识的生命活动把人同动物的生命活动区别开来。”〔48〕强人工智能的最初愿景,就是要建造一台“会思考、学习和创造的机器”,使其具有与人类相同甚至更高的智识、智力和智慧(人类级 AI),故而是不可能跳过“意识构造”这一步的。如果以人工智能本来就不同于人类智能为由而否定意识构造的必要性,则等于自我放弃实现最初愿景的努力。还有论者则玩起了文字游戏,认为意识只是个信仰问题,其有无取决于判断主体的主观确信。“当机器说出它们的感受和感知经验,而我们相信它们所说的是真实的时,它们就真正成了有意识的人。”〔49〕据此,机器是否真正具有意识并不重要,重要的是判断主体是否相信它们有意识。显然,这种鸵鸟策略掏空了意识的内涵,偷换了讨论的主题。

意志的核心,是主体的自主选择和决定能力。强人工智能的乐观派,一方面认为人工智能必将赶上甚至超越人类智能,另一方面又认为即使在超级人工智能时代,人类仍然是世界的主人——关键时刻可以关闭电源。这其实是一种骑墙摇摆的悖论:一方面,既然认为机器智能发展到一定程度,可以不受人类控制、自主地决定选择实施行为,则必然认为机器有自由意志;另一方面,既然认为人类始终掌握着控制的总开关,可以随时停止机器的运行,则必然又承认机器没有自由意志。笔者认为,机器智能在任何情形下都不可能具有意识和决定的能力。以无人驾驶汽车为例,尽管它装配各种先进的电子设备,可以精准地导航、纠正跑偏的方向、控制车速车距、稳当地泊车,但如果没有人类启动汽车,则它根本就不会转动。这一事象表明,无人驾驶汽车都是按照人类事先的设计和“授权”在行动,“服膺”于人类的指令,并没有自主决定的意志。因此,无论机器智能在程式化领域领先人类多远,也永远不能主宰人类,因为“我们永远都拥有一个法宝,能让它按我们的意愿行事”。〔50〕

意识之所以重要,是因为其能展现人类特质。“‘意识’和‘思想’的功能在于它们能使我们针对时空中遥远的东西而作出行动,即使那种东西当前并没有刺激我们的感官。”〔51〕机器智能在许多单项领域可以代替人类从事许多活动,甚至做得比人类更好,但由于不会思维,没有意识,不能自主决定,因而只能作出机械的反应。知识和技能可以通

〔47〕〔英〕卡鲁姆·蔡斯著:《人工智能革命:超级智能时代的人类命运》,张尧然译,机械工业出版社 2017 年版,第 111 页。

〔48〕〔德〕马克思著:《1844 年经济学哲学手稿》,人民出版社 2000 年版,第 162 页。

〔49〕〔美〕雷·库兹韦尔著:《人工智能的未来:揭示人类思维的奥秘》,盛杨燕译,浙江人民出版社 2016 年版,第 203 页。

〔50〕〔法〕鲁道夫·格林著:《机器人是人类最好的朋友吗》,孙兆原、应远马译,上海科学技术文献出版社 2017 年版,第 88 – 89 页。

〔51〕〔英〕伯特兰·罗素著:《心的分析》,贾可春译,商务印书馆 2010 年版,第 258 页。

过复杂的算法来习得,但是,意识绝对无法以同样的方式来获取。“意识精神不能像一台电脑那样运行是‘显而易见’的,即使真正涉及的精神活动中,有许多的确像电脑一样运行。”〔52〕当前人工智能的发展,都仅限于认知领域,在意识领域则根本没有触及,更遑论突破了。“从人工智能目前的发展方向看,无论它再怎么‘自动学习’‘自我改善’,都不会有‘征服’的意志,不会有‘利益’诉求和‘权利’意识。”〔53〕但问题的复杂性在于,意识是客观存在的主观映象,是对自身和周围的存在有所认识的一种心智状态。“清醒状态、心智与自我是意识最重要的铁三角。”〔54〕意识既是人脑的机能,更是人类社会交往的经验反映。它“并不仅仅是你在思考和阅读此页书籍时所浮现在脑海中的想法或意象,同时也指你每天在生活中可能经历的感知、感觉和情绪。”〔55〕一方面,人脑的结构、意识的机制等极为复杂,截至目前,科学界仍然知之甚少,“还没有人能够真实地看见实验情境下诞生意识,或甚至提出意识是如何诞生的理论”〔56〕;另一方面,即使哪一天科学界弄明白了意识的生成机理,也不可能在制造或升级机器、系统或程序等非生物体的过程中,将人类生物体的“意识”“情感”引入、留驻或整合进去。这不仅因为构成人工智能和人类智能的物质存在硅基和碳基的不同,更在于意识的形成过程与人的成长过程相伴而生,而人工智能不可能具有类人的“生长过程”。人工智能在本质上只是一个算法机器,无法创造出自己的符号,“如果没有真实的思维,即使是拥有密密麻麻、精巧细致、梳理整齐、带有反馈环路和稳态平衡器的交换电路网络,也无法成为一个有意识的实体。”〔57〕

意志自由,是“有意选择行为的自由,在于它不受感官冲动或刺激的决定。”〔58〕意识和意志,与人的道德情感密切相关。根据道义责任论,人类之所以可以成为犯罪主体和责任主体,正因为其具有自由意志,能够依据伦理参照系分辨是非善恶。对于基于自由意志而实施的行为,即“那种可以由纯粹理性决定的选择行为”,行为实施者是适格的刑事责任主体,国家具有对其惩罚的正当性;对于非基于自由意志而实施的行为,即“那种仅仅由感官冲动或刺激之类的爱好所决定的行为”,是非理性的兽性的选择,行为实施者不是适格的刑事责任主体,国家欠缺对其惩罚的正当性。机器智能也是一样,如果自由意志问题不解决,则其始终只是人类的工具,是不自由的,因而就不具备刑事责任主体的前提条件。“智能代理虽然能够‘自动地’完成某些任务,但它们即使具备学习能力,所遵循的最终仍然是由程序规定的选择,而没有对其行为给出自己的理解。”〔59〕机器智能不能认识到自己

〔52〕〔英〕罗杰·彭罗斯著:《皇帝的新脑》,许明贤、吴忠超译,湖南科学技术出版社2018年版,第565页。

〔53〕翟振明、彭晓芸:《“强人工智能”将如何改变世界——人工智能技术飞跃与应用伦理前瞻》,《学术前沿》2016年第4期(上),第27页。

〔54〕〔美〕安东尼奥·达马西奥著:《当自我来敲门:构建意识大脑》,李婷燕译,北京联合出版公司2018年版,第144页。

〔55〕〔美〕弗朗西斯·福山著:《我们的后人类未来:生物技术革命的后果》,黄立志译,广西师范大学出版社2017年版,第166-167页。

〔56〕〔美〕弗朗西斯·福山著:《我们的后人类未来:生物技术革命的后果》,黄立志译,广西师范大学出版社2017年版,第171页。

〔57〕〔美〕乔治·吉尔德著:《知识与权力》,蒋宗强译,中信出版社2016年版,第262页。

〔58〕〔德〕康德著:《法的形而上学原理——权利的科学》,沈叔平译,商务印书馆2002年版,第13页。

〔59〕〔瑞士〕萨比娜·格雷斯、〔德〕托马斯·魏根特:《智能代理与刑法》,赵阳译,载陈泽宪主编《刑事法前沿》(第十卷),社会科学文献出版社2017年版,第222页。

在做什么,也不知道自己行为的社会意义,缺乏自我意识,不能价值判断,故而对其行为进行刑法评价就没有根基。

六 情感动机:一个刑罚配置的基点

刑罚是一种必要的"恶"。这种"恶"之所以必要,在于其为预防犯罪所必需。边沁指出:"自然把人类置于两位主公——快乐和痛苦——的主宰之下。只有它们才指示我们应当干什么,决定我们将干什么。是非标准,因果关系,俱由其定夺。"[60] 因此,刑罚配置应当遵循功利原理,即"它按照看来势必增大或减小利益有关者之幸福的倾向,亦即促进或妨碍此种幸福的倾向,来赞成或非难任何一项行动。"[61] 费尔巴哈的心理强制说亦认为:"所有的违法行为在感性上都有其心理学上的原因,人的贪欲在一定程度上会因对行为的乐趣或者产生于行为的乐趣得到强化。这种内心的动机通过下列方式加以消除:让每个人知道,在其行为之后必然有一个恶在等待自己,且这种恶要大于源自于未满足的行为动机的恶。"[62] 无论是功利原理还是心理强制说,都以被惩罚者的情感动机作为基点。犯罪能够给行为人带来快乐,刑罚则对其赋加痛苦,抵消其快乐,从而抑制其犯罪的动机。

然而,人工智能却没有情感,不能体会到犯罪之乐和刑罚之苦,对其适用刑罚,难以实现刑罚的预防功能。为解决这一瓶颈问题,业界提出模拟人类情感的解决方案——打造情感机器人。情感机器人是具有情感计算功能的人工智能,情感计算这一设想最先由麻省理工学院媒体实验室皮卡德教授提出。她认为:"情感计算是一种与情感相关、源于情感或能够对情感施加影响的计算。"[63] 情感计算,通过赋予计算机像人类一样的观察、理解和表达各种情感特征的能力,建立和谐的人机交互环境,并使计算机具有更高级的、更全面的智能。明斯基亦指出:"情感、直觉和情绪并不是与智能(intelligence)不同的东西,而只是另一种人类特有的思维方式。情感是先于理智存在的,人工智能只有智力,没有情感,不是真正的智能。"[64] 根据这一理论,没有加入情感计算的人工智能,只能停留在单句指令、机械问答的程度,远不能与人类进行对等的、自然的交流;有了情感计算,人工智能能够通过语义、图像和语音精准识别人类情感和真实意图,从而提供一对一的专属个性化服务,使人类对其产生情感上的信任和依赖。然而,情感计算的应用离不开情感建模,这就决定了这一技术应用归根结底仍然是一种"暴力计算"。因此,"人造情感"与"人类情感"必然存在本质的区别,前者只能是一种"虚情假意",而后者才具有"真情实感"。即使在技术的场面,人工智能的模拟情感能力也相当有限。"在通过触觉、视觉和嗅觉等感官来理解世界并与世界进行交互的能力等方面,现在的人工智能还不如孩童。对于人类表

〔60〕〔英〕边沁著:《道德与立法原理导论》,时殷弘译,商务印书馆 2000 年版,第 57 页。
〔61〕〔英〕边沁著:《道德与立法原理导论》,时殷弘译,商务印书馆 2000 年版,第 58 页。
〔62〕〔德〕费尔巴哈著:《德国刑法教科书》,徐久生译,中国方正出版社 2010 年版,第 28 页。
〔63〕Rosalind W. Picard, *Affective Computing*, MIT Press, 1997, p. 3.
〔64〕〔美〕马文·明斯基著:《情感机器:人工智能与人类思维的未来》,王文革等译,浙江人民出版社 2016 年版,第 5 页。

情、语气、情感以及人类交往的微妙之处,人工智能系统只有最初级的理解能力。换言之,现在的人工智能'IQ'很强,但'EQ'很弱。"〔65〕既然如此,那么针对人工智能而设置的阶梯刑罚(包括删除数据、修改程序、永久销毁),则并不能让人工智能产生"惧怕"的情感,也就不能发挥应有的威慑功能。鉴于"智能代理不能真正感受到刑罚,现阶段也就无法对其科处刑罚。人类可以将捣毁机器人理解为对其错误行为的责难惩罚,然而这一层意思并不能传达给机器人"〔66〕,因此,惩罚机器人,无异于对牛鼓簧。

也有一些论者认为,即使人工智能没有情感动机,也不妨碍将其设定为承受责任和罪责的对象,前提是只要立法机关认为有其必要性。"在一个国家中,什么样的人或组织能够成为法律关系的主体,取决于该国的法律规定。"〔67〕例如,人类历史上就曾经出现过惩罚动物。既然动物、非人类生物可以成为刑事责任的适格主体,则没有理由否定人工智能不能成为刑事责任主体。"无论是罗马法实践,还是中世纪的老鼠审判,实际都深刻挑战了近代以降以自然人为鹄的的法律人格理论,也为探讨人工智能的法律身份留下了充分的想象空间。"〔68〕的确,在欧洲中世纪,动物审判相当流行。〔69〕对这些动物进行一场正式的刑事审判,其根本目的何在?比较法学者埃瓦尔德从历史的视角,对此进行了全面的考察,但也未能找到确定的答案,而只是指出了发现答案的切口和路径。〔70〕或许功能主义可以对此作出合理的解释。简言之,在中世纪人们认为审判动物能够实现惩罚的目的,因而导致这种审判相当流行。而在文艺复兴之后,整体人类概念—关系框架彻底重构,理念思维完全改变,从而导致这类审判的彻底消失。客观地说,这种转变是符合理性的。因为动物既不能将审判与其恶行关联起来,也不能理解审判的社会意义。将动物作为受罚主体既不能阻止其将来再次重犯类似的恶行,也不能阻止其他动物或者人类仿效实施,因此,在现代人看来,整个审判不仅是无效益的,更是一场滑稽的闹剧。所以,如果以动物曾经作为刑事责任主体为据来证明人工智能刑事责任主体的合理性,则无异于承认人类理性和文明的退步。

也有观点认为,法人"没有可谴责的灵魂,没有可受罚的身躯",没有意识情感,但却不影响其成为刑事责任主体。既然法人可以成为适格的刑事责任主体,则没有理由否定人工智能不能成为刑事责任主体。这一类比推理初看无懈可击、滴水不漏,但细辨却存在误读法人犯罪的惩罚机理这一前提性错误。法人的设立,表面上看是新增一种法律人格体,但真正关注的是幕后的自然人,确切地说,是要降低自然人在商事活动中的风险,"尤

〔65〕 沈向阳、〔美〕施博德著:《计算未来——人工智能及其社会角色》,北京大学出版社 2018 年版,第 8 页。

〔66〕 〔瑞士〕萨比娜·格雷斯、〔德〕托马斯·魏根特:《智能代理与刑法》,赵阳译,载陈泽宪主编《刑事法前沿》(第十卷),社会科学文献出版社 2017 年版,第 222 页。

〔67〕 张文显主编:《法理学》,高等教育出版社 2018 年版,第 155 页。

〔68〕 余成峰:《从老鼠审判到人工智能之法》,《读书》2017 年第 7 期,第 76 页。

〔69〕 法学家威廉·埃瓦尔德撰写的名篇《审判老鼠的意涵?》,对此即有描述。公元 1522 年,一群老鼠在欧坦教会法庭受到审判,它们因啮食和破坏该教区内的大麦作物而被指控犯有重罪。法学家沙萨内最终为这群可怜的老鼠做出了成功辩护,开启了他杰出的法律职业生涯。据埃瓦尔德统计,从公元 9 世纪到 19 世纪,西欧就有两百多件记录在案的动物审判,被放上被告席的动物包括:驴、甲虫、水蛭、公牛、毛虫、鸡、金龟子、奶牛、狗、海豚、黄鳝、田鼠、苍蝇、山羊、蝗虫等。

〔70〕 See William Ewald, Comparative Jurisprudence (I): What Was it Like to Try a Rat, 143 *U. Pa. L. Rev.* 1896 (1995).

其是航海经商,需要大量资金,利润大风险也大,个人根本无力经营而须采取合伙的方式。"[71]法人人格的突出后果是,"权利义务均由作为一个整体的团体承担,并将成员个人完全排除在外",[72]从而限制自然人的法律责任。[73] 同样,法人犯罪,表面上看追究的是法人的刑事责任,但法人由其股东组成,且法人责任的承担形式是罚金,因此,惩罚法人实质上是惩罚法人的股东,以促其改善公司治理、执行合规计划等。事实上,在法人犯罪中,虽然在法律意义上犯罪行为由法人这个"虚拟人"作出,行为的后果也由法人承担,但不可否认的是,股东会、董事会等法人机构仍然由自然人构成,其决定的作出离不开自然人的认识和意志。换言之,虽然法人犯罪中行为效果归属于法人,但是其行为的选择和执行仍然与自然人的认识和意志存在直接关联,后者的情感动机在法人犯罪生成中起着至关重要的作用,故而将法人拟制为犯罪主体,仍然可以有效发挥刑罚的预防犯罪功能。换言之,法人犯罪生成过程中隐含着自然人的认识和意志要素,而人工智能"犯罪"生成过程中却难觅自然人的主观因素,因而以刑法将法人设置为独立的责任主体为由而当然地推出人工智能也应是适格的责任主体,并不具有充分的说服力。

人之所以是适格的受罚主体,在于其能够认识到自己行为的伦理后果,并且权衡利弊后独立自主地决定采取行动。由于"机器人不可能有道德感,只有基于程序的反复和预先设计而总结出的规律,从而也就没有民事主体所必备的基于内心感知(良知)所做出的善恶评判和行为选择",[74]因而无法满足受罚的前提条件,也就无法实现通过对其"犯罪"行为定罪判刑等否定性评价而抑制或矫正其行为的效果。不过,也有论者提出,不是所有的犯罪者都要承受同样的后果,惩罚不仅要符合罪行,还要符合罪犯,而对合成智能最好的惩罚就是强制"失忆"——删除事件数据库。[75]然而,如果这样处理,则相当于将现行刑法中的犯罪、刑罚、罪犯等关键词全部改写,这将导致整个刑法体系的崩溃。为此,最好还是要坚持惩罚的基本理念,将受罚主体限定为具有意志情感的主体。由于"智能代理既不具备——可以让人类感知到的——感受刑罚的能力,也无法理解与惩罚相联系的伦理指责,因此针对它们'本人'动用刑罚是没有意义的"[76],它们不是适格的责任主体和受罚对象。

七 未尽的结语

人不仅具有思维属性,而且具有社会属性。"人是最名副其实的政治动物,不仅是

〔71〕 周枏著:《罗马法原论》(上册),商务印书馆 2014 年版,第 305 页。

〔72〕 〔意〕彼得罗·彭梵得著:《罗马法教科书》,黄风译,中国政法大学出版社 2018 年版,第 44 页。

〔73〕 在民法上,作为法人成员的股东,仅以其出资资产为限对法人的行为承担法律责任。而法人作为独立的法律人格体,其财产与其股东的财产完全分离,对以自己名义实施的行为,独立承担法律责任。

〔74〕 赵万一:《机器人的法律主体地位辨析——兼谈对机器人进行法律规制的基本要求》,《贵州民族大学学报(哲学社会科学版)》2018 年第 3 期,第 158 页。

〔75〕 参见〔美〕杰瑞·卡普兰著:《人工智能时代:人机共生下财富、工作与思维的大未来》,李盼译,浙江人民出版社 2016 年版,第 87 页。

〔76〕 〔瑞士〕萨比娜·格雷斯、〔德〕托马斯·魏根特:《智能代理与刑法》,赵阳译,载陈泽宪主编《刑事法前沿》(第十卷),社会科学文献出版社 2017 年版,第 238 页。

一种合群的动物,而且是只有在社会中才能独立的动物。"[77]人之所以为人,不仅在于其在自然界中具有独立的生物存在,更在于其在社会网络中与其他人相互交往。因此,"人的本质不是单个人所固有的抽象物,在其现实性上,它是一切社会关系的总和。"[78]人的本质有赖于在后天的社会生活中去形塑,每个个体参与社会生活的方式、需要、程度、经历、感受、目标等具有多样性和可变性,因而都是独一无二的,不可能被复制和模拟。因此,理解人类的行为,不能单纯用生物规律和自然规则来阐释,还应当将其纳入社会生活这个场域中给予人文的考察。反观人工智能,无论是对知识的获取、对行为的认知还是对情境的反映等,都是在特定的现实空间或虚拟空间里进行的,并非是参与社会活动的结果,因而不具有社会性,而只具有自然性、反射性。所以,人工智能永远不可能成为主宰地球的"新人类",担心人工智能控制或毁灭人类,显然是鳃鳃过虑、徒增烦扰。"最有可能变成现实的情形是全人类步入一个崭新的人机协作时代,在这个时代,以人工智能为驱动的机器将大幅提高人类的工作效率,但无论从哪个角度说,机器都只是人类的工具。"[79]

不过,人类一直在不断地发明创造工具,这些工具带来了深刻的社会变革。人工智能的出现,更是极大地影响着人类的社会生活。问题的复杂性在于,"机器能够渐进在它们所处环境的刺激下学习,并从它们自己的行为中获取知识和技能,因此不仅对于使用者,同样对于它们的设计者而言,机器人会逐渐地变得不可预测。"[80]奇点理论和强人工智能论营造的乌托邦多少有点分散人们的注意力,但其关于人工智能可能引发的技术风险的预警,却是值得认真对待的。"技术风险是人类和非人类行为的综合产物。任何一个人造物,无论是最不起眼的,还是最复杂的,如果设计低劣,处理不当,人类对它又管理不当的话,都会变得很危险。"[81]因此,警惕技术风险、倡导合理的科技伦理、设计全新的技术规范,是更为急迫的任务。一方面,在技术研发过程中,要结合新技术带来的法律关系的变化合理地分配法律责任。例如,关于自动驾驶车辆犯罪,亟需研究的不是自动驾驶车辆(人工智能体)本身是否应当承担刑事责任以及承担多重的刑事责任,而是研究如何在车辆的设计者、制造者、销售者、购买者、使用者等不同主体之间分配注意义务和法律责任。这些主体的注意义务与传统过失理论中由驾驶员承担直接、完全的结果避免义务完全不同。"这种差异,使得在具体确定自动驾驶车辆不同主体注意义务时不能照搬传统过失理论,而是需要重新确定注意义务的范畴和要求。"[82]另一方面,也要防止科学至上主义,"它不仅仅是将诸如量化的技巧错误地用在数字无法解答的问题上,还混淆了人类经历发生的物质和社会领域,并主张将自然科学的目标和方法应用到人类世界里。"[83]应当肯

〔77〕《马克思恩格斯选集》(第2卷),人民出版社1972年版,第87页。

〔78〕《马克思恩格斯选集》(第1卷),人民出版社1995年版,第56页。

〔79〕李开复、王咏刚著:《人工智能》,文化发展出版社2017年版,第160-161页。

〔80〕〔意〕乌戈·帕加罗著:《谁为机器人的行为负责?》,张卉林、王黎黎译,上海人民出版社2018年版,第48页。

〔81〕〔美〕希拉·贾萨诺夫著:《发明的伦理:技术与人类未来》,尚智丛等译,中国人民大学出版社2018年版,第27页。

〔82〕彭文华:《自动驾驶车辆犯罪的注意义务》,《政治与法律》2018年第5期,第91页。

〔83〕〔美〕尼尔·波兹著:《技术垄断:文明向技术投降》,蔡金栋、梁薇译,机械工业出版社2013年版,第150页。

定,在人工智能领域同样存在一定的“技术禁区”。犹如克隆技术一样,克隆羊可以,但克隆人则被严格禁止,因为维护人类的尊严,是技术发明必须坚守的伦理底线。技术绝非中立,而是始终具有意向性,对于人工智能,“必须以人类中心主义的关怀,来审慎对待它带来的社会风险,并予以积极的防御和规制。”〔84〕因此,必须对人工智能研发进行规范和引导,“需要对人工智能采取以人为本的态度,体现人类永恒的价值观,还需要坚定不移地秉承驾驭计算智能造福人类的主旨”,〔85〕以最大限度地增强人类的能力。

[本文为作者参与的 2018 年度浙江大学“人工智能与法学专项课题”的研究成果。]

[**Abstract**] The proposition that “artificial intelligence (AI) is a qualified subject of criminal responsibility” is based on the realization of strong AI in the future. But if intelligence is understood not only as programming algorithm, but also as mental awareness, then this proposition is untenable. It is fundamentally impossible to design an AI with the same mental consciousness as human beings. Firstly, in terms of cognition, AI has high cognitive ability only in certain single fields, but is faced with insurmountable technical bottlenecks in crossover fields. Secondly, in terms of consciousness, AI can neither form consciousness spontaneously within itself, nor import consciousness externally, so it is impossible for AI to have free will. Finally, in terms of emotion, AI cannot have the same emotional motivation as human beings, or experience the joy of crime and the suffering of punishment, so it is not a proper subject of punishment.

(责任编辑:郑 佳)

〔84〕 马长山:《人工智能的法律风险及其规制》,《法律科学》2018 年第 6 期,第 48 页。

〔85〕 沈向阳、〔美〕施博德著:《计算未来——人工智能及其社会角色》,北京大学出版社 2018 年版,第 8 页。