如何从刑法的维度对人工智能进行规制

唐亚南*

摘　要： 刑事责任能力是犯罪主体的核心要件，因此，人工智能机器人是否能成为刑事责任的主体，关键的标准之一是其是否具有辨认和控制自己的行为能力。对于弱人工智能主体而言，因它不具有辨认和控制自己的行为能力，只是一个工具而已，所以必须坚持传统犯罪主体理论，在现有的刑法框架内定罪处罚。对于强人工智能主体而言，从刑法内部构成的教义学层面分析，人工智能在一定程度上具备刑法上的认知控制能力及受刑能力，能够成为刑事责任主体。在刑法规制人工智能的过程中，即使有违于传统的刑法理论，面对有可能给人类带来灭顶之灾的强人工智能的规制，应该用更苛责的手段规制人工智能，以便警醒相关人员对智能机器人的编程等事项的“注意义务”，减少人类风险。中国社会在飞速发展，出现了大数据、云计算等，中国传统的刑法理论也应与时俱进，也就是说，在人工智能冲击下，刑罚体系可以进行部分重塑，世异则事异，事异则备变。

关键词： 人工智能　大数据　责任主体　刑法理论

人工智能是研究、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；该领域的研究包括机器人、图像识别、语音识别、专家系统等。2017 年 7 月 20 日，国务院印发的《关于新一代人工智能发展规划的通知》提出将人工智能放在国家的战略层面进行布局；因此人工智能这一人类的伟大发明对社会发展具有重要的战略作用。然而，正如有的学者所指出的，人工智能这一人类社会的伟大发明，同时也会给法律秩序带来巨大的风险。[1] 凡事预则立，不预则废。只有对人工智能领域的刑法问题具有前瞻性的分析并探索出行之有效的解决问题的途径，才能沉着应对人工智能领域出现的问题。笔者对人工智能的刑法应对提出粗浅的建议，以期抛转引玉。

一、人工智能的刑事责任主体和刑事责任能力

新事物的出现，以及社会观念的改变，往往会对现有的法律体系产生影响。以前，在人类的认知体系中，人工智能产品仅仅是人类发明创造出来的工具，当欧盟委员会法律事务委员会于 2016 年 5 月 31 日提交一项动议，要求欧盟委员会把正在不断增长的最先进的自动化机器“工人”的身份定位为“电子人（electronic persons）”，并赋予这些机器人依法享有著作权、劳动权等“特定的权利与义务”。时，人们逐渐意识到人工智能的智慧越来越靠近人类的智慧，中外学者对人工智能是否具有法律主体资格以及是

* 唐亚南，人民法院新闻传媒总社编辑、法学博士、博士后。

〔1〕 王志祥、张园国：《论人工智能时代刑事风险的刑法应对》，载《学术交流》2020 年第 11 期。

否应当承担刑事责任等问题展开探讨。

（一）刑事责任主体

在刑法领域，人工智能是否能成为刑事责任主体的讨论事关其责任模式的配置，就人工智能产品发展的现状看，在未来涉及刑事案件的可能性越来越大。从法人主体地位的获得，再到机器人主体地位的提倡，可以说是观念不断转变的过程。目前，理论界存在最大的争议是人工智能可不可以为刑事责任的主体？有专家认为，从刑法内部构成的教义学层面分析，人工智能实际具备刑法上的认知控制能力及受刑能力，能够成为责任主体。[2] 也有专家认为，人工智能不能成为刑事责任的主体，因为它不具备人自然人的特征，无法接受刑事处罚。笔者认为，两种说法都有道理；不过，仅仅因为不具有人的特征就否定人工智能机器人的主体资格，似乎不够周延，比如，传统的单位犯罪也不具有自然人的特征；法人在很多国家都是承认它的刑事责任主体地位的，包括我国。本体论探讨法律主体地位的时候，认为只有那些能够意识到自己的权利和义务并能够对自己的行为和自由进行反思的主体才能是法律上的主体；但实际上这样的理解显然存在偏颇，我们绝对不会否认一个植物人或者说一个精神病人的法律上的主体地位，所以从本体论的角度来设定法律主体显然是存在一定问题；因此所谓人格或者说法律主体地位并不仅仅是从本体论的角度来加以建构，事实上，是否承认一个人格体，或者是否承认人工智能是不是法律主体，更多要从社会角度加以考察，在理解人格体的时候，应该从社会角度加以理解。从社会角度看，人工智能应该具有刑事责任主体地位。从司法实践的角度看，只要一个人的举止不构成本能的反射活动，通常会承认这样的举止是行为，对于自然人的行为进行判断的时候，并没有完全遵照关于行为的理论，既然如此，对于人工智能机器人，承认它的举止是行为并不存在障碍。理论上争论比较多的是可不可以承认人工智能具有罪责，传统上对罪责的理解是对行为可能性的理解，行为人本来能够选择实施合法行为，但是选择了不法行为，从而违反了法律，这是传统上对于罪责的理解。但随着社会经济的迅速发展，人们的思想观念也在发生着深刻的变化，“大数据”、“云计算”这些看似虚无缥缈的东西，在现实中已经起到了很大的作用，也同时被大多数人所接受。对于罪责的理解，其实在理论上也有新的发展，比如，目前在德国对于罪责的理解有一种理论叫做基能的罪责理论，因此承认人工智能的罪责在理论上是说得通的。

虽然，人工智能犯罪属于犯罪中的新现象，但从犯罪客体看，它并不是新类型，因为它并没有在刑法分则的十种同类客体中创设新客体，而是涵括在这些同类客体之中，因此有专家提出在刑法中增设人工智能事故罪。[3] 一旦人工智能造成了损害或者实施了“犯罪”，从刑法目的的角度上应该对它进行预防，这是涉及到人工智能刑事责任主体地位与刑法预防的必要性之间的关系。刑法的预防是从一般立法和特殊立法这两个角度加以预防。从一般预防的角度上来看，惩罚人工智能的犯罪行为，对于预防或者是吓阻侵害的犯罪人有意义，如果仅仅从传统的角度来理解刑法的目的，就会认为惩罚人工智能不能达到刑法的目的，但事实上，惩罚人工智能不仅仅是符合刑法目的，而且人工智能也是有受刑能力的。未来对人工智能所施加的惩罚，比如说断电，或者从物理上消灭人工智能等，我们应该以人类中心主义作为我国的立法基础。刑法本来具有法律性、针对性和处罚性，法律对人类的一些权益进行剥夺，使人受到痛苦的惩罚，表现出一种惩罚性，而对于机器人删除数据，永久销毁等措施对智能机器人而言就是一种惩罚，当然可以成为刑罚种类。删除数据、修改程序和永久销毁，它不仅仅是一种法定措施，而且应该理解为是一种基础刑处罚。不可否认，在未来一定会出现强智能时代，这是一个大方向和大趋势，但不管智能时代的智能主体是通过利用缺乏认知能力的机器还是利用人类具有认知控制能力的新物种，我们应该接受它是一类责任主体，而不能仅仅认为它是一种工具。删除数据、修改程序以及永久销毁这种刑罚措施我们应

〔2〕 王耀彬：《类人型人工智能实体的刑事责任主体资格审示》，载《西安交通大学学报》2019 年第 4 期。

〔3〕 刘宪权：《人工智能时代的刑法风险与刑法应对》，载《紫光阁》2018 年第 5 期。

该认为是刑法上的行为人主体，行为人具备行为危害性。同时刑法是具有社会属性的性质，人工智能对于刑罚措施应该保障体系的完整性。犯罪事实本身的认定是通过什么样的犯罪情节和定罪量刑情节来抽取出准确的识别要素，然后再通过什么样的规则能够将这样的要素与争议焦点和裁判关联起来，本身就是需要论证和不断验证的过程。当然我们现在的应用很多时候是全方位的，比如，对行为数据之间的分析，实际上是一个通用领域，这个领域的影响可能会影响到我们生活当中的方方面面，再比如，在法律、大数据和人工智能深度应用中，我们会发现人工智能本身对于规范性和创造性是有一定的理解，并且能进行不断的深度学习和应用，必将会解决实践中一些重点、难点问题。

从人工智能时代整个发展的规律和趋势看，人工智能机器人实际上是从机器的因素逐渐减弱，人的因素逐步增加，机器人从代替人的手足，逐步替代人的大脑。作为刑事责任，就是一个人辨认和控制自己行为的能力，而辨认控制自己行为能力就是人的意识，如果这个机器人通过技术的进一步发展，他在意识和意志当中得到独立的发展，那么它就具有辨认和控制自己行为的能力。

事实上，这种发展在某种程度上是完全可能的，例如，在围棋领域中，阿尔法沟可以战胜中日韩的国手，围棋国手的下围棋的水平都是顶级的，阿尔法狗居然可以把他们战胜，而且研发阿尔法狗的研发者跟中日韩的围棋高手去下围棋不一定能战胜他们，也就是说他研发出来的阿尔法狗这个产品在下围棋的智能上已经超越了研发者本身，后来又出现了新阿尔法狗和阿尔法狗 zoro，最关键的区别是阿尔法狗是通过大量输入人类的棋谱，最后经过自己深度学习的能力战胜了国手。阿尔法狗 zoro 最大的特点就是不需要输入人类的棋谱，只需要告诉他下围棋的规则就把阿尔法狗战胜了。这说明在这个领域中，人工智能机器人的智能已经超越了人类。

人工智能一边给我们带来美好的未来，一边也给我们社会带来巨大的风险隐患[4]。人工智能时代将给我们的生活带来了巨大的变化，给我国的法律也必将带来较为深刻的影响。美国太空探索技术公司的 CEO 马斯克正在研究人脑和计算机对接的问题，俄罗斯已经在开发机器人杀手，而且已经实现了技术的应用。面对这样一种风险，需要未雨绸缪，一方面需要科技发展给人类带来福利，另一方面，也需要警惕技术的滥用所带来的风险，尤其是在人工智能这样一个快速发展的时代，把人工智能作为一个犯罪的主体，或者说把滥用人工智能的这些人作为犯罪加以规制，势在必行。那么，如何规制强人工智能所实施的犯罪行为？笔者认为，首先承认人工智能机器人的主体地位，同时对人工智能机器人相关的监管人员也应当追责；当人工智能机器人依据自己的意志实施犯罪行为的时，对机器人赋予监管责任的人员就要承担相应的刑事责任。具体来说就是为了对人工智能机器人进行更加有效的监管，可以考虑在人工智能机器人系统中设立毁灭系统，在人工智能机器人存在犯罪意识时，相关程序就会介入，强人工智能机器人的相关编程就会被摧毁，使其变成没有任何意识的机器，如果因为监管人没有充分履行监管义务而导致强人工智能机器人实施犯罪行为，不但机器人要承担相应的刑事责任，监管者也应当承担相应的刑事责任。机器人如何承担刑事责任要根据人工智能机器人的强弱来认定刑事责任的大小。人类发明和设计机器人的目的是为了更好地服务于人类的工作和生活，绝不能够容忍对人类命运构成毁灭性的打击，因此对人工智能机器人成为犯罪主体，一旦认定人工智能机器人构成犯罪，需要追究刑事责任，判处死刑，就意味着这款机器人永远不再使用。这昭示着人类在主宰一切，人是世界的核心。

（二）人工智能的刑事责任能力

人工智能是否具有承担刑事责任的能力，在理论界存在争议，有专家指出，财产权是承认损害赔偿责任的基础，人工智能有没有财产权，因此不能承担刑事责任。有学者认为，人工智能没有生命，因此不能承担刑事责任。有专家指出，虽然单位犯罪中的“单位”，只是一个组织，也没有生命，在刑法的规制中仍然要负刑事责任；但是单位犯罪需

〔4〕 潘庸鲁：《人工智能介入司法领域路径分析》，载《东方法学》2018 年第 3 期。

要单位意志，这种意志是单位的成员按照决策程序来做出的，单位背后的意志仍然是自然人的意志，也就是人类自己的意志，单位的大脑就是人脑，这跟人工智能还是不一样的。笔者认为，这种说法有一定的道理，的确，一个单位，一个组织确实没有生命，但是他所体现的所有的动态活动都是人的活动。事实上，人工智能的意志自由度可能比单位还要高，人工智能何尝不是人在制造，并体现人的意志，即便是强人工智能，也是出于制造者之手，也就是说“人工智能”中修饰限制“智能”的是“人工”，它所体现的是人的意志，确切的说是制造者的意志，“人工智能”只不过是以一种新的物种形式出现而已。因此“人工智能”的“财产责任”应归罪于制造者。

依据美国加利福尼亚大学约翰. 塞尔教授的观点，可以把人工智能分为“强人工智能”和“弱人工智能”。弱人工智能是指计算机在心灵研究中心的主要价值只为我们提供一个强有力的工具。强人工智能是指计算机不仅是我们心灵研究中心的工具，而且带有正确程序的计算机，其实就是一个心灵，在强人工智能中，由于编程的计算机具有任职状态，这些程序不仅是我们用来检验心理解释的一个工具，而且本身就是一种解释。[5] 由此可见，弱人工智能并不具有自己的独立意志，只是在编程的范围内实施相关的行为，是人类为实现自己的目的而制造的工具；而强人工智能则拥有自己独立的意志，具有认识能力和控制能力，能够在编程范围外实施行为。当前，弱人工智能是人工智能技术的主要表现形式，利用弱人工智能实施故意犯罪或涉及弱人工智能的过失犯罪行为也是人工智能犯罪主流形态。然而，随着人工智能技术的不断发展，人工智能技术由弱人工智能向强人工智能进行转变的技术也在不久将来变为可能。[6] 有专家对人工智能划分为三个时代，即普通机器人时代，弱人工智能机器人时代，强人工智能机器人时代。普通人工智能机器人时代是指人工智能只具有部分人脑的生理功能，大部分还是有人工来操作实现。弱人工智能机器人时代是指已经具备了部分深度学习的能力，而且在某些领域中深度学习能力已经超越了人类。强人工智能机器人时代是指它具有分析、判断、思考的能力，它的智能远超于人，人不能做到的它能做到。虽然强人工智能时代还没有到来，但势不可挡。强人工智能机器人时代和弱人工智能机器人时代的具体划分就是看是否超越了编程之外的有关领域。以大数据发展所支撑的技术，慢慢走向成熟，并且和人类对薄的过程中明显占有一定优势。超级计算机的诞生，加速了数据运算的速度，缩短了周期。人工神经网络研究的深度学习算法，搭建了多层神经网络框架，能够模拟人类处理信息的过程。

二、人工智能在刑事审判领域的应用

刑事司法人工智能的发展是以前期的刑事司法信息化为前提的，所以人工智能对刑事司法基本形态的塑造也应遵循信息化改造的基本形态，将刑事诉讼从线下搬到线上，而线上刑事诉讼不断地驱动着刑事司法人工智能在刑事司法实践中的应用，最高人民法院在加强智慧法院建设中提出，要努力实现法院“全业务网上办理、全流程依法公开、全方位智能服务”。事实上，刑事司法不仅仅是个案的裁判，亦是社会治理体系和治理能力的重要一环。通过刑事个案裁判助推刑事司法社会治理功能也是一个重要的手段。由于各国司法制度差异，有时会呈现出不同的表现形式：判例法国家的判例本身就蕴含着具有法律效力的规则体系，司法直接介入社会治理；非判例法国家的司法裁判的介入较为迂回。在大数据和人工智能背景下，拓展刑事司法的社会治理功能如虎添翼，国内已有法院通过人工智能技术积极参与刑事审判。运用刑事司法人工智能信息多维度的工具属性，从社会治理层面最大程度地提升了刑事司法效率。

（一）刑事司法重塑的必要限度

人工智能作为一种内生力量介入刑事司法，将实体意义的规则从现实走向虚拟网络，贯彻着

〔5〕［英］玛格丽特·A. 博登：《人工智能哲学》，刘西瑞、王汉琦译，上海译文出版社 2001 版，第 92 页。
〔6〕同前注〔1〕。

技术主义的属性，并体现出强大的效率导向；但这在一定程度上是以牺牲司法的价值为代价，因而必须将之控制在必要的限度之内。刑事司法意义上的程序与计算机意义上的程序大相径庭，刑事司法意义上的程序必须在法定的步骤中体现出刑事司法的程序价值，让参与其中的当事人和社会公众都能感知到刑事司法的专业性和权威性，从而强化刑事司法威信。如果司刑事法如同计算机程序化指令性的输入—输出，那就既不用事人的参与，也不用法官的介入，刑事司法已化为计算机的运行，这种刑事司法的电脑运营显然完全“阉割”了蕴含在司法程序中的内在价值。将刑事诉讼从线下搬到线上，最为显著的变化就是司法场景的变化，并进而导致以下几个方面的程序价值的减损：由法庭设施所营造的“司法剧场效应”受到削弱，法庭的威严感、神圣感不复存在；即有诉讼程序中的亲历性、对抗性、规范性、严密性等也因网络的非现场性而不可避免地受到了削弱，并进而产生一系列相关问题，因此司法程序价值理应成为司法人工智能应用的一条界限。

从本体上看，刑事司法是指运用国家权力对刑事诉讼案件进行审判的活动。刑事司法即国家的刑事审判活动，刑事司法权就是刑事审判权，通过个案裁判的方式维护法的价值的终局性的权力；然而随着大数据和人工智能技术的嵌入，不仅外在的刑事诉讼活动和程序发生了变化，而且刑事司法主体也在改变，进而影响刑事司法权的行驶。在当前法官普遍受制于技术门槛而不能很好地掌握相关技术的情况下，技术人员有可能通过算法的构建和引入，在事实上“分享”司法权，从而在一定意义上影响法官办案。而更复杂的是，这种算法在技术上存在着不可解释性的算法黑箱，在价值上又不可避免地带有技术人员的偏见，并以机器人的方式驱动司法之运转，此种苗头隐患已经在司法实践中开始显现。因此技术人员在司法自身逻辑框架内所开发的“要素式审判”等技术应用，也会因认知的局限而存在“简化实体审判”的问题。[7] 基于司法本体的限制，笔者认为司法改革对人工智能技术的强化会导致司法权本身被分化为司法权力和技术权力，同时人工智能技术人员将成为新一类司法辅助人员。例如2019 年 1 月 23 日 4 时，上海市第二中级人民法院公开开庭审理了由上海市检察院第二分院提起公诉的殷某抢劫一案。该案系全国法院首次运用“206 系统”辅助审理可能判处被告人 10 年以上有期徒刑、无期徒刑、死刑的重特大刑事案件。[8] 这是人工智能在刑事审判领域的深度应用，表现出了以下功能：第一，自动推送功能，电子证据在法庭上同步推送各方，运用“206 系统”的图文识别、智能语音识别、自动语义分析等人工智能技术，通过语音唤醒功能，实现了通过语音精准调取案件中的证人证言、物证、书证、现场勘验检查笔录等多种证据，并将其通过视频方式推送到法庭各处屏幕上，供审判人员、辩护人、被告人阅看，大大缩短了庭审示证时间，提高了庭审效率。第二是自动抓取功能，通过语音提示自动抓取并推送，在法庭调查阶段，“206 系统”自动抓取与公诉人、辩护人、审判长讯问、发问问题有关的证据材料，帮助法庭判断被告人供述、证人证言的可信度，以及是否存在疑点，实现了系统的单一证据校验功能。第三，自动识别功能，瑕疵证据自动提示进入人机深度互动，庭审中，合议庭运用“206 系统”的瑕疵证据提醒功能。第四，自动转化功能，书记员速记已被音字自动转换替代，庭审运用音字转换技术，将庭审与语音转换相结合，实现了对庭审的高效、准确记录，提高了庭审效率。第五，自动生成功能，实现大量卷宗无纸化扫描。打破了“信息壁垒”，实现了公检法司办理刑事案件网上运行、互联互通、数据共享。[9] 基于信息的多维度，刑事司法人工智能确实能够拓展司法的社会治理功能，从而赋予了刑事审判更广阔的社会意义；但需要认识到的是商业大数据具有极强的公共属性，数据公开，商业领域的数据垄断会以同样的原理

〔7〕 周玉华主编：《中国司法学》，法律出版社 2015 年版，第 16 页。

〔8〕 参见邱波：《人工智能“阿尔法狗”现身二中院，首次辅助审理抢劫杀人案》，载搜狐网，https://www.sohu.com/a/291275570_651791。

〔9〕 同前注〔8〕

和方式表现在司法领域，如一旦智慧法院建设中的平台设施和相应的数据要受制或依赖于网络技术公司，则当事人信息，甚至审判秘密的安全等也可能受到威胁。司法大数据是以司法公开为条件的，其本身并不涉及正当与否的问题，关键在于开发利用主体及其目的。

由于司法的特殊性，这就决定了对司法大数据的智能运用应有目的上的正当性，尤其是应当考虑国家安全。从现实情况来看，目前，我国还停留在技术应用层面，与此相关，更为重要的网络基础架构、操作系统等还存在着风险，并且这种风险已经超出了司法制度自身范畴，更多的是基于技术失控所带来的外溢风险，[10] 但司法从来不应当是点状思维，而是立足社会的系统思维，对此类风险的预估和评判也是刑事司法人工智能极限的另一方面。

（二）大数据层面的司法支撑和规制

从司法大数据的形成来看，目前的司法公开体量很大，在维度上也很丰富性，司法大数据已经初具规模；但让然需要加大司法公开的广度和管理力度。从公开深度上看，近年来关于强化裁判文书说理、司法文书质量进一步排查做的非常好，事实的认定和法律的适用分析的越来越好，越来越详尽。从管理力度上看，司法大数据最大范围地促成了在法院和相关部门之间、法院系统之间在管理体制机制上的障碍，真正实现数据的互联互通和司法人工智能的最大价值。[11] 不过，司法大数据的发展所带来的技术失控问题应得到重视，这主要是国家安全、个人隐私的保护和数据权力的规制等。目前我国的情况是重公开轻保护，对于公开裁判文书的获取应该采取更加严格的溯源性的程序控制，在促进司法大数据发展的同时，更好地维护意识形态的安全。而对于隐私的保护，除必要的技术处理外，如裁判文书的匿名化处理，更多地是与数据权力的正当行使相关。司法大数据的建设和共享应限于正当目的，不能无节制地扩大，否则会导致数据权力的失控。承办法官在调取相关数据时，要按照一定的程序获得，并要有相应的备案。这样会更方便追责。司法大数据的形成应主要由公、检、法合理完成，而不能交由企事业单位承担数据库建设、查询、开发之责，这样更能有效地保护相关涉案人员的隐私。

尽管理论界对于司法人工智能的算法等问题警惕性很高，但主要集中在商业研究领域，而司法领域的研究和介入还极其有限。且不论此种算法规制在技术、路径上如何可行，相关的配套机制的建设目前尚付之阙如。鉴于司法人工智能的规模化运作属性，域外已经对这方面设置了具体的考量规则，如当地时间 2019 年 4 月 8 日，欧盟委员会发布一份人工智能道德准则，该准则由 2018 年 12 月公布的人工智能道德准则草案演变而来，提出了实现可信赖人工智能的七个要素，要求不得使用公民个人资料做出伤害或歧视他们的行为。该七个要素是：确保人的能动性和监督性；人工智能的算法要足够安全、可靠和稳健；个人数据应受到正当控制，不得用来伤害或歧视他人；人工智能系统应具备透明性和可追溯性；要充分考虑到人类能力和技能要求的范围，要做到多样性、非歧视性和公平性；要有助于社会的可持续性发展，并且担负生态责任；应当有相应的问责机制，需要对它造成的后果承担责任。我国亦应参考上述人工智能的规制思路和具体要素，在法院系统或更高层面设置司法人工智能审查委员会，负责组织实施司法人工智能的开发许可、司法属性和伦理道德审查、应用测评、风险应对和后续的评估清退等。同时在具体的规制过程中，也要综合考量司法和技术的两方面要求和司法场景的具体差异，采用灵活多样的规制路径和方法。基于司法的稳定性和公共性、民主性和公开性，在总体倾向和基本原理层面，应选择成熟的、可视化的、可解释性的算法，努力击破算法黑箱，将算法转化公众、法官可以理解的自然语言，并接受伦理道德、司法政策之审查。[12] 事实上，这只是一种理想的事前规制，但从技术适配性角度考虑，这种规制很难现实。如深度学习中的神经网络

〔10〕 侯猛：《互联网技术对司法的影响——以杭州互联网法院为分析样本》，载《法律适用》2018 年第 1 期。

〔11〕 同前注〔8〕。

〔12〕 曾学原、王竹：《道路交通纠纷要素式审判探索——从四川高院的改革实践出发》，载《中国应用法学》2018 年第 2 期。

算法采用网状的非线性函数，在不相干的元素中建立一个假设的逻辑关系，然后通过海量的数据来检验这种假设的正确性，淘汰正确度低的建设，保留正确度高的建设，从而具备更强的学习能力，而一般人的线性逻辑很难理解神经网络算法等非线性逻辑。[13] 因此在没有办法兼顾司法人工智能的透明性和实用性的情况下，只能考虑其他的规制路径。对此，不妨通过事前数据输入质量的严格把控和事后的输出结果的校验，来淡化算法黑箱所带来的危害。

（三）刑事司法人工智能应用的总体方向

对刑事司法人工智能的基本态度是工具主义，并在刑事司法审判、刑事司法主体等各个层面强调司法的主导性。在应用场景上，要紧密结合刑事司法人工智能的基本理论；在刑事司法和技术的合作层面，要始终立足现实基线，强化刑事司法需求的技术导入和技术应用，并要有长远规划。在刑事司法审判方面，既要有以往经验的总结提炼，也要有超越历史的能动创新，同时还要彰显主体性。从刑事诉讼程序看，起诉、立案、送达、庭审、合议、宣判，还有刑事诉讼材料的送交签收各种笔录的记载签字等，这部分工作虽然是程式化的工作要求，似乎可以交给司法人工智能来处理，但事实上会出现一些误差和不准确性，对司法人工智能履行的上述行为必须施以严格的监督。因为该类工作不仅涉及到当事人的程序权益，而且也会影响到当事人的实体权益，这对刑事司法程序的自身价值是一种贬损。从理论上看，似乎存在着简单成型的标准化案件，在这种情况下，司法被类型化为“正义产品”的制造，诉辩各方诉讼材料的输入都会有确定的、可预期的裁判结果的输出，故要求法官做到标准化、流程化似乎是有可能的，理论上司法人工智能可以在一定程度上胜任审判任务。目前业界正在开发的“要素式审判”司法人工智能贯彻的就是这一思路；对要素式审判是以法律形式主义的预设为前提，事实上并不能解决价值判断和法律适用的问题，一旦超出预设的范围，司法人工智所得出的结论有可能和实际情况大相径庭。但从审判实践来看，司法人工智能的应用特别复杂，根本无法胜任司法审判。因而，即使在所谓“简单成型”的标准化案件中，法官也不能放松对司法人工智能产品的严密监督。

在司法实践中还存在不少引领社会价值观、促进规则精细化的“创造性案件”，此时裁判的做出更多地要依赖法官对社会经济条件和发展趋势的感知和积极地作为。事实上，司法人工智能最大的好处在于为法官提供更多更丰富的维度，以供其审视裁判的合法性及其内在的适当性。在一定意义上说，在标准化案件中，开发司法人工智能可以重点考虑裁判的具体规则，在“创造性”案件中，则需要重点关注认知维度。不过，无论在何种案件中，法官对司法人工智能的严格监督都必不可少，因为司法人工智能不能脱离自身条件而发展。[14] 因为现实中的司法大数据因各种原因并非准确无误，用以驱动司法大数据的算法也有可能存在算法独裁或算法歧视问题，并非总是正确的。

综上所述，司法人工智能的算法规制体系即算法层面的司法支撑，可以根据司法应用场景的具体差异，综合采用事前数据输入，严格把控输出结果并进行验证，用不同层面的各种手段对司法人工智能进行规制。

三、人工智能的刑事立法

随着人工智能技术的不断发展，人类社会将会遭遇人工智能发展过程中所涉及到的法律问题，面对不断出现的人工智能犯罪出现的新问题，而现行刑法并不能完全解决这些问题。为了防控人工智能的犯罪对人类社会造成的危害，对人工智能的立法势在必行。有学者认为，人工智能最好是用民法或行政法去调整，刑法是社会治理的最后一道防线，社会治理不能过度刑法化。笔者认为，人工智能作为科技界的重要产物，在未来的社会发展中，人工智将会给人类带来无法估量的风险，人工智能作为一种社会变革的先进技术，

〔13〕 华宇元典法律人工智能研究院编著：《让法律人读懂人工智能》，法律出版社 2018 年版，第 8 页。
〔14〕 同前注〔8〕。

理应受到刑法的保护，只有用刑法这种严苛的手段，才能保障人工智能健康有序的发展。

（一）刑事立法应遵循的原则

人工智能刑事立法的目的是保障人工智能有序发展，防控人工智能可能带来的风险，以应对人工智能对刑法的冲击，因此人工智能刑事立法应遵循以下原则：

第一，利益优先和风险管控原则。人工智能刑事立法一定要坚持以人为本，保障人类利益优先的地位。首先弱人工智能前期的旧法益类型，主要是人工智能与原有的法益类型出现了新的表现形式，法益本身没有发生质的改变，法益主要出现于新工业时代向弱人工智能时代转型期。其次弱人工智能的中后期，扩张型法益，转型期以及强人工智能时代，随着人工智能产品技术优势的逐渐增强，智能机器人的刑事责任主体就可以通过立法予以确立；三是作为新型法益予以确定增加型法益，是指人工智能发展过程中产生的独立保护的利益价值，人工智能研发、设计、应用活动具有独立保护的价值，并且在刑法中设立专章对破坏人工智能的研发设计活动予以规制。[15] 例如，侵犯公司企业管理秩序罪，这里面涉及两个罪名，违规纰漏不纰漏重要信息罪和隐秘故意销毁快递凭证，快递账户，财务会计报告罪；还可能会破坏金融管理秩序，再如，利用未公开的信息交易罪，编造并传播证券期货交易虚假罪；还有可能破坏知识产权保护秩序，因为现在的著作和商业秘密有的是以电子数据的形式存在于计算机系统种。目前国家是鼓励人工智能的研发，所以对于互联网行业的职业相关性的行为不能评价为刑法上的危害行为，应该以行为人所处阶段的一般人为标准来进行判断。根据不同的职业领域，评价时不仅要关注事实的性质，还要关注微观要素，例如，一些人在买房需要打印征信报告，个人的交易经历，贷款等，这是国家经营调控金融风险必须的行为，不能说是侵犯公民个人信息的行为，这种行为不具有刑事违法性。具体职业领域内拥有正当性和必要性的行为，必须与法律保持一致，才能最终被确定为具有社会相当性，法益的衡量要个人隐私权让位于公共利益。

人工智能刑事立法应该在一定程度汲取互联网的经验，发展的前中期预留给人工智能相当程度的成长、发展、试错的空间。中后期阶段，或者是在人工智能风险出露端倪的时候，这个时候可以结合实际，适当增加刑法规制的力度和扩大刑法规制的范围，以此来保护人们的利益，推动人类社会的健康发展。[16] 人工智能应用于信息、生命、材料、能源、环境等很多领域，有相应的风险，人工智能研发者、使用者对可能存在的风险应具有一定的注意义务，因违反注意义务所导致的重大事故，可归责于人工智能研发者、使用者，只有违反了前置性法律法规，便可推定其具有对结果规范性的可能性，危害发生的情况下，只要没有把证据推翻就可以成立过失犯罪。这些规则不仅适用于强人工智能还没有到来之前，一旦强人工智能时代到来，强人工智能虽然具备自主意识而可以自定规则，这种情况下人工智能设计者仍然应该承担责任，也就是说这些规则仍然适用。

第二，协调和效益原则。人工智能刑事立法应该与技术发展相协调，人工智能刑事立法应该与相关法律相协调，只有这样才能够实现人工智能刑事立法的科学性。例如，应该修改侵害计算机信息系统类犯罪，目前人工智能技术的实现仍然需要计算机信息系统作为媒介或者是基础。对于人工智能机器人而言，确定哪些风险需要特殊的防范措施，最重要的是看哪些因素会引起最大的潜在危险。有一些危险会导致人类的灾难；有一些危险只会产生轻微的伤害。如果一个风险的最大潜在损失的程度达到了人类无法承受的地步，那么，风险留存是不可行的，必须采取立法控制，使其被降低到一个可控制的程度，否则就必须将风险转移。如果一个风险既不能被降低到一个可控制的程度，而又无法转移的时候，那么必须用立法的方式将它规避，即不允许该款智能机器人上市和使用。不过，确定人工智能机器人的风险问题是非常复杂的问题，也是很有技术性的问题，每一个风险的留存都与对人的伤害

〔15〕 见2021年2月11日《人民法院报》理论版。

〔16〕 同前注〔15〕。

有关，而后者又取决于人的承受能力和在出现危险时的应对能力。

风险管理和其他财务管理一样，必须遵循成本效益原则，只有当风险管理方案的所得大于支出时，该风险管理才是成功的。成本效益原则就是要对风险管理活动中的所付出的代价与所得的收益进行分析比较，对管理行为的得失进行衡量，使成本与收益进行最优的结合，以求获得最多的利益。如果，在确定型风险决策中，各种损失发生的概率是可以知道的，而在不确定型的风险决策中是没有这些信息的，决策中信息越充分，决策的准确程度就越高。因此，对立法者而言，确定型风险立法相对容易，知道这个风险会引起损失的可能性是很小、中等或很大，将帮助立法者来决定如何立法才能排除风险。对于不特定的风险，在立法中较为困难，当然，这并不是说对不特定的风险的发生可以被忽略不去立法，恰恰相反，对不确定的潜在的风险如何立法，应采取谨慎的态度，必须把所有风险的可能性或概率作为一个重要的因素来考虑。

事实上，从司法的整个过程来看，几乎所有的场景都需要对算法做出解释。例如，法庭记录，有人认为只需要完整地记录各方当事人的陈述发言即可，一份好的庭审笔录的关键在于书记员的提炼、概括。对此，笔者不敢苟同，虽然现代语音识别、自动记录技术完全可以一字不漏地将当事人陈述发言记录下来，但这仍然需要我们对其背后的算法原理有充分的认识，有些司法场景还必须要求对当事人、社会公众有所回应，特别是核心的裁判形成环节和裁判文书的说理。在这些环节中，算法黑箱、算法的不可解释性是不被允许的。在运用司法人工智能获得新的信息维度时，法官有义务解释司法人工智能是如何获取的该信息，以及该信息对本案事实认定、法律适用将产生怎样的影响，并要努力将此方面的举证质证义务分配给各方当事人，以秉持司法中立性和当事人抗辩主义。也有学者提出，在某些核心环节，不妨通过司法人工智能的反向运用来避免算法黑箱、算法歧视等问题，针对司法实践中 206 系统“数据化、统一的证明标准”的努力存在的风险，不妨将该标准作为一个反向标准、否定性标准。形象地说，它可以作为英美法系排除合理怀疑的标准，而不是中国和大陆法系直接定罪的标准。从广义说，也是一种值得借鉴的规制思路，即通过司法人工智能的运用方式来规制算法。

综上所述，司法人工智能的算法规制体系，也即通过司法人工智能委员会的设置和相关审查准则的探索，确定算法透明性和无歧视性（或平等性）的原则，根据司法应用场景的具体差异，综合采用事前数据输入质量的严格把控和事后的输出结果校验，持续性的测评评估，和司法人工智能的运用方式等不同层面的规制手段。

（二）人工智能刑事立法的基本构想

第一，普通机器人相当于无刑事行为能力人，不应承担刑事责任。刑事责任能力是犯罪主体的核心要件，因此，人工智能机器人是否能成为刑事责任的主体，关键的标准之一是其是否具有辨认和控制自己的行为能力。刑事责任能力是辨认能力和控制能力的统一，其中前者是基础，后者是关键。普通机器人不具有辨认和控制能力，因此不应追究其刑事责任。

刑法修正案十一将刑法第十七条修改为：“已满十六周岁的人犯罪，应当负刑事责任。“已满十四周岁不满十六周岁的人，犯故意杀人、故意伤害致人重伤或者死亡、强奸、抢劫、贩卖毒品、放火、爆炸、投放危险物质罪的，应当负刑事责任。“已满十二周岁不满十四周岁的人，犯故意杀人、故意伤害罪，致人死亡或者以特别残忍手段致人重伤造成严重残疾，情节恶劣，经最高人民检察院核准追诉的，应当负刑事责任。”对依照前三款规定追究刑事责任的不满十八周岁的人，应当从轻或者减轻处罚。“因不满十六周岁不予刑事处罚的，责令其父母或者其他监护人加以管教；在必要的时候，依法进行专门矫治教育。”那么，人工智能中普通机器人相当于不满 12 周岁的人，因此不负刑事责任。理由是：普通机器人一般而言几乎全部由人来控制，因此，它完全体现人的意志，只要控制普通机器人的人没有犯罪的故意和过失，普通机器人对人类一般来说是没有危险的；可以比照刑法修正案十一的规定，不追究普通机器人的刑事责任；这意味着，其他人仍然可以使用该普通机器人。但对于操作该普通机器人的人按照刑法的规定，该定

什么罪就定什么罪，因为该普通机器人就相当于该人使用的一个工具，利用工具、动物等犯罪，当然应当负刑事责任。

第二，弱人工智能机器人相当于限制刑事行为能力人，应该有选择的承担刑事责任。弱人工智能机器人相当于已满12周岁，不满14周岁的人，应该有选择的负刑事责任。弱人工智能机器人只有造成特别严重的后果时，经指定的有关科研单位的有关专家认定后，应当承担刑事责任。即制造者和弱人工智能机器人都应当负刑事责任，但不一定负同等刑事责任，应视不同情况而定。制造者应负有保障该机器人在使用时安全的义务，这个所谓的安全是指弱人工智能机器人的编程等符合相关规定，人正当操作时不会发生危害后果。如果在人正当操作而由于编程等技术性问题造成被害者伤亡或重大的财产损失，那么制造者和弱人工智能机器人应当分别负刑事责任。制造者负刑事责任很好理解，那么弱人工智能机器人如何负刑事责任呢？让它负刑事责任的意义和目的又是什么呢？笔者认为，让弱人工智能机器人去模仿人承担刑事责任的目的是限制这款机器人在市面上的流通即在规定的时间内不能再出售这款机器人，目的是保障人们的安全，比如，张三在使用A款弱智能机器人时，在造成了多人死亡或重大财产损失，事故发生后，经有关科研单位的有关专家鉴定，张三在操作过程中并无不当，而是编程的参数出现了问题，由于它是一款弱人工智能机器人，相当于刑法修正案十一中规定的已满12周岁不满14周岁的人负刑事责任，假设对制造该款弱智能机器人的人和该款弱人工智能机器人同时判处5年有期徒刑，对于制造者的刑事责任很好理解，在这里不赘述，那么对于该款弱智能机器人判处的5年有期徒刑的刑事责任，意味着两层含义：第一，对该款机器人删除程序，在5年内该款机器人不能在市面上出售，也不允许使用；第二，制造者或其他合法持有这款弱智能机器人的人，可以在5年内对该款弱智能机器人进行改进、完善，也许制造者或改进者在3年之内就改进完毕，但仍然不能上市出售，只有在5年之后，经过指定的有关科研单位的认定，才能上市出售。它的意义在于，给制造者以充足的时间去完善和改进该款弱智能机器人。当然这些预设是要有立法支撑的，也就是说，对人工智能的立法迫在眉睫。

第三，强人工智能机器人相当于完全刑事行为能力人。强人工智能机器人相当于完全刑事行为能力人拥有自己独立的意志，具有认识能力和控制能力，能够在编程范围外实施行为，因此应当承担刑事责任。阿尔法狗zoro应当属于强人工智能机器人，它最大的一点就是不需要输入人类的棋谱，只需要告诉他下围棋的规则就把阿尔法狗战胜了。后来又出现新的阿尔法狗，新的阿尔法狗除了不需要输入棋谱之外，可以进一步地深度学习，然后可以接受各种各样下棋的相关规则，最后在所有的棋当中都能下。实际上在这一个领域中，人工智能机器人的智能已经超越了人类。这类人工智能机器人就属于强人工智能机器人，它相当于“已满十六周岁的人犯罪，应当负刑事责任。强人工智能机器人如果在正常操作使用中造成他人伤亡的可以根据不同的情况，判处不同的刑罚。例如，如果对该款强智能机器人判处死刑，这意味着对这款强智能机器人的彻底否定，没有改进的必要和价值，删除其数据，对智能机器人永久销毁，永远不得上市销售和使用。对于制造者可根据不同情况定罪量刑。再如，无人驾驶汽车因为汽车质量的问题，对人产生重大伤亡，那么汽车研发者和生产者就应当承担相关的刑事责任，如果是由于驾驶者违反交通规则而导致重大交通事故，驾驶者应承担交通肇事罪。非法利用驾驶汽车事故进行故意犯罪，主要是驾驶人故意或者过失操作车辆出现事故，也是根据自动驾驶程序的原理，人机协作的参与程度而来分别定罪。自动驾驶引起事故的责任考量应该根据驾驶人参与程度不同进行区分，应该允许驾驶人部分信赖自动驾驶系统，如果过度信赖，驾驶人应该承担过失责任。自动驾驶技术的程序原理主要是在于对于自动驾驶技术的应用，自动驾驶技术实际上就是系统自主执行部分或者是全部驾驶操纵，能够在驾驶人完全不参与驾驶的情况下进入安全的行驶状态。当下的现实状态仍然是人机结合，人工和智能的结合，尚未达到完全脱离人。如果是人和自动驾驶技术参与的程度不同，对于出现事故之后责任的划分也受到影响。

（责任编辑：李琦）