主题研讨

刑事合规视野下人工智能的刑法评价进路

于　冲

内容提要:人工智能犯罪研究的既有成果,大多忽视了作为人工智能核心问题的数据安全、算法规制问题,就“人工智能”谈“人工智能”,以至于针对人工智能的刑法回应性研究大多集中在不可知化、科幻化的“机器人规制”层面。人工智能犯罪的刑法规制,首先应当明确刑法评价的对象是人工智能体研发者的行为,而非人工智能体本身的“行为”,进而立足于传统刑法基础理论和刑法框架,确立人工智能犯罪防治的“共治”思维、预防性思维,推动刑法规制的重心由事后的结果性评价转向兼顾事前的危险性预防。对此,有必要引入刑事合规在人工智能中的评价机制,以刑法为手段,在明确人工智能算法可解释、人工智能决策数据透明的前提下,通过人工智能产业链条上的算法合规与数据合规,厘定算法过错、算法霸凌背后的人的过错,防范因数据瑕疵以及算法设计、算法部署、算法应用中人的罪过引发人工智能决策失误。具体要求,正视人工智能风险,实现刑事合规的犯罪预防机能,推动传统刑法事后制裁的评价模式转向事前预防。实现刑事合规的外部规制与自我管理的有机结合,在刑法规则具体化、情境化的基础上,实现刑法与人工智能企业规章制度的功能性协作。

关键词:人工智能　刑事合规计划　数据安全　算法规制　刑事责任

于冲,中国政法大学刑事司法学院副教授。

人工智能技术创新和发展产生的新风险与不确定性,给既有的犯罪体系、责任体系乃至刑罚体系带来了全新挑战。围绕人工智能发展可能带来或者已经出现的潜在威胁、“失控风险”以及现实危险,刑法学界从主体论、责任论、刑罚论的角度开展了密集研究。但是,既有成果要么偏离了刑法学轨迹,专注于算法技术研究;要么关注人工智能“未来刑法学”甚至是“科幻刑法学”,桎梏于人工智能的主体化与否,无法有效地为人工智能犯罪〔1〕的刑法评价与预防提供理论支撑。对此应当明确,对于人工智能犯罪的防治而

〔1〕 本文所称人工智能犯罪是指人工智能体引发的犯罪,不包括以人工智能为工具、为对象实施的犯罪。

言,仅靠刑法的单一手段已经无法有效应对。在人工智能风险日益增强的背景下,有必要以刑事合规为视野,推动人工智能刑事合规制度的确立,通过刑事合规计划实现刑法规则的细化,倒逼人工智能产业者加强内部风险控制和自我管理,降低企业犯罪风险、提供减免罪责甚至出罪化的同时,客观实现人工智能犯罪的预防。[2] 从横向上来看,人工智能刑事合规,应当跳出人工智能是否主体化问题,在可控的范围内,将决定人工智能运行的数据、算法作为刑法评价的重心。从纵向上来看,人工智能刑事合规应当将人工智能的研发、部署、使用,在不同的层级、阶段上进行刑法的分级规制,实现人工智能风险防控的全过程化。总体上,人工智能刑事合规,应当立足于网络时代的"共治"模式,实现人工智能犯罪从外部规制到内外共治的转变。

一 人工智能刑法规制研究的现有缺憾与偏差

人工智能犯罪评价与规制不断成为刑法学界的热点话题,围绕人工智能的本体化、人工智能刑事责任论等论题都有丰硕的成果。随着人工智能技术的不断发展,关于人工智能犯罪的问题将逐渐增多,也必然对现有刑法体系造成冲击。因此,如何为人工智能犯罪的刑法评价提供理论助力,进而为防治人工智能犯罪提供应对思路,成为刑法学界不可推却的时代课题。但应当注意的问题是,不能脱离基本刑法理论乃至基本法理常识,盲目开展想象式的未来式、科幻式研究,以及"泛人工智能化研究"。[3] 无论是人工智能算法问题,还是大数据问题,都应当立足于"工具论""客体论"的视角,[4] 在坚持人类本位的基础上,立足传统刑法基础理论、传统刑法框架,对人工智能犯罪的刑法评价提供现实化的应对模式和解决思路。

(一)人工智能刑法研究的观点争讼

刑法理论界关于人工智能的刑法评价思路,大体可以归纳为人工智能本体化与非本体化、人工智能刑事责任论、人工智能阶段规制论等观点。这些观点的共同性均在于强调人工智能犯罪的风险性和危害性,争议点多集中在人工智能是否主体化,以及人工智能致害后果的责任承担问题。[5] 人工智能主体化的观点认为,人工智能根据自我意思在人类既定的设计之外实施犯罪,表明其具有了独立的辨认能力与控制能力,应当赋予其刑法上的主体地位并承担相应的刑事责任。[6] 相似的观点将人工智能体拟定为介于自然人与

〔2〕 本文所称人工智能犯罪预防,是指基于人工智能工具论,强化人工智能研发者、部署者、应用者的保证人地位,通过加强前述主体在人工智能研发、部署、运行等阶段的风险识别与防控义务,进而预防人工智能体引发犯罪这一犯罪类型。

〔3〕 参见刘艳红:《人工智能法学研究中的反智化批判》,《东方法学》2019 年第 5 期,第 121 页。

〔4〕 参见高铭暄、王红:《互联网 + 人工智能全新时代的刑事风险与犯罪类型化分析》,《暨南学报(哲学社会科学版)》2018 年第 9 期,第 3 页。

〔5〕 除了涉及人工智能相关犯罪问题的争议之外,学界对人工智能的研究还集中在人工智能侵权责任问题、人工智能社会地位问题、人工智能算法的演变及回应问题。参见皮勇:《人工智能刑事法治的基本问题》,《比较法研究》2018 年第 5 期,第 156 页。

〔6〕 参见刘宪权、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 42 页;类似论述可见刘宪权:《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,第 136 页。

单位之间的特殊主体,根据人工智能的智能化程度判定其是否具备刑事责任能力。〔7〕 认为针对强人工智能阶段智能机器人独立实施的犯罪行为,应当对其进行独立的刑法评价,并建立适合该类主体的特殊刑罚体系。〔8〕 与之相对,人工智能主体化否定论的观点认为,无论人工智能发展到何种阶段,都不能具有刑法意义上的主体资格。例如,有学者指出,智能机器人的认知和决定能力是由人类赋予的,即使未来出现“心智”完整的智能机器人也不存在“自由意志”,其所谓的“类人化”行为都区别于人的行为,〔9〕人工智能主体化不符合刑法的目的、任务等问题。折中论的基本立场是,弱人工智能阶段距离未来可能产生具有自主意志、人格独立的强人工智能甚至是超强人工智能依然十分遥远,因此现阶段仅需讨论作为“工具”和“产品”的人工智能引发的刑事犯罪问题,对可能产生自主意志和人格独立的人工智能不予分析评价。〔10〕

整体上讲,学界关于人工智能刑法评价的思路,主要是围绕人工智能的主体化与否,以及在此基础上的刑法规制路径展开的。人工智能主体化观点,机械地以“拟人”的标准对人工智能进行思考,并呈现出“泛人工智能化”的特点,在以自然人特征为评判标准的传统刑法上,探究人工智能具有刑事责任能力的可能性,忽视了刑事责任能力本身即是自然人的“专属问题”,也就无法准确、可操作性地对人工智能主体化之后的评价思路给出令人信服的评价方案。

(二)人工智能刑法规制研究的错位与理论困境

通观学界关于人工智能刑法规制的研究,很大程度上偏离了以数据、算法为核心的人工智能本身的特有属性,尤其围绕人工智能主体化的研究更是偏离了传统犯罪论和刑罚论的基本轨道。

1. 人工智能刑法规制路径的错位与偏差

随着机器学习算法领域的飞跃性发展,自主能力和思维能力显著增强的智能机器人为传统法律制度带来了全新的挑战,人工智能不断被拟制为一种新兴法律主体。关于人工智能未来式、泡沫式的研究,大都忽视了“刑法是处罚人的法律”,〔11〕以及“人”作为刑事责任主体地位的不可撼动性。当前刑法学界集中讨论的人工智能主体化之争,很大程度上是对刑法责任论根基的违反,是对刑罚目的论和功能论的背反,更是对人类中心主义的失守。一方面,刑法上的责任能力不单单是从技术层面进行的,还要经过法的价值选择,即,人能够成为刑事责任主体的适格性,这种适格性是经过规范选择的“能不能成为

〔7〕 参见马治国、田小楚:《论人工智能体刑法适用之可能性》,《华中科技大学学报(社会科学版)》2018 年第 2 期,第 110 页。

〔8〕 参见吴波、俞小海:《人工智能时代刑事责任认定思路的挑战与更新》,《上海政法学院学报(法治论丛)》2018 年第 5 期,第 98 页。See Sabine Gless, Emily Siverman, Thomas Weigend, “If Robots Cause Harm, Who Is To Blame? Self-driving Cars And Criminal Liability”, *New Criminal Law Review: In International and Interdisciplinary Journal*, Vol. 19, no. 3, Summer 2016, pp. 412 – 436.

〔9〕 参见黄京平等:《人工智能与刑事法治的未来》,《人民检察》2018 年第 1 期,第 44 页。

〔10〕 参见高铭暄、王红:《互联网 + 人工智能全新时代的刑事风险与犯罪类型化分析》,《暨南学报(哲学社会科学版)》2018 年第 9 期,第 7 页。

〔11〕 [日]西原春夫著:《刑法的根基与哲学》,顾肖荣等译,中国法制出版社 2017 年版,第 2 页。

责难主体”的适格性。[12] 相对于人类的本位价值而言，其他一切非人物种所固有或衍生的价值，仅仅具有工具价值或手段价值的意义。[13] 人工智能本质上是一种工具，是为了追求更好的生活目标所创造出的科技产物。从服务人类的工具属性而言，人工智能并非是与人类并存的主体，仅具有客体属性。“机器人之父”阿西莫夫提出的机器人三定律其实已经揭示出人类在社会中的主导性，技术的发展永远要以人类的安全为终极目标。从社会构成的角度来看，若赋予人工智能以法律主体地位，则现行的社会秩序不得不演变成为新型的人机并存社会秩序，届时原有的由自然人组成的人类社会形态、社会关系、社会伦理、社会自治等一系列学说都将面临解构以及重构的巨大困境。[14] 另一方面，从刑罚论的角度看，关于人工智能具有接受刑罚的能力，因而可以成为刑事责任主体的观点，实质上颠倒了责任主体与接受刑罚能力的前后顺序。无论是借鉴刑法中的刑罚体系制裁人工智能，还是对人工智能施以诸如删除程序等特殊的制裁措施，处罚人工智能的“刑罚”本质上都不属于刑法的刑罚，而应将其视为人工智能的应用措施或技术措施，即属于一种对人工智能的处置性技术方法。可以说，现在人工智能是否主体化的理论旋涡与研究悖论，不仅会造成人工智能研究的“反智能化”和“泡沫化”，[15] 而且也无法切实有效地发挥基本理论对人工智能产业发展和相关司法实践的指导作用。

2. 人工智能刑法研究困境的症结

人工智能犯罪的刑法规制是一个系统化工程，当前研究成果主要集中在对人工智能本身的刑法规制层面，大都没有认识到人工智能背后的技术实质、犯罪实质。客观讲，人工智能数据处理技术水平的提升，将其解决问题的能力推进至接近甚至超过人类智能的水平，网络空间的主体呈现异化趋势。人的主体性特征在网络空间中变得愈发模糊，智能机器自主性决策的能力，使得人在世界中的绝对支配地位受到极大挑战，而法律体系的主体性条件也不再是唯一而无异议的。[16] 可以说，以数据、算法为基础的新型法权关系正在信息时代的大环境中孕育生长，冲击着传统的法权结构，但此时更要强调人作为社会主体和法律主体的主导性。此种背景下，造成人工智能犯罪刑法评价研究困境的根源主要体现为三个方面：(1)算法不可解释、不可控制甚至不可知论，造成人工智能研究的科幻化、泡沫化，这种不可知论实质上是一种放任算法的做法，不仅不利于人工智能的有效规制，甚至会颠覆人工智能行业发展，颠覆人类传统社会及其建立在人类社会之上的传统规则体系。例如，人工智能主体化观点，实质上反映了理论在人工智能面前的无奈和妥协，跳过人工智能背后的数据问题、算法问题的刑法评价，意图通过直接赋予人工智能主体资格解决犯罪的评价问题。(2)以传统人类社会的规则体系，以及对人类行为的评价标准来评价人工智能，忽视人工智能的迭代更新与技术实质，对人工智能的研发、运营等单位及其人员施加过度的监管义务与法律责任，同样会阻碍人工智能技术发展，还会产生违反

〔12〕 王韬著：《论刑法上的责任》，中国社会科学出版社2013年版，第112页。

〔13〕 参见舒年春：《走入真正的人类中心主义》，《广西大学学报(哲学社会科学版)》2002年第2期，第23页。

〔14〕 参见范忠信：《人工智能法理困惑的保守主义思考》，《探索与争鸣》2018年第9期，第79页。

〔15〕 参见刘艳红：《人工智能法学研究中的反智化批判》，《东方法学》2019年第5期，第124页。

〔16〕 参见陈璞：《论网络法权构建中的主体性原则》，《中国法学》2018年第3期，第77页。

传统刑法罪责主义的责任追究后果。客观讲,就人工智能引发的危害后果,基于传统刑法的责任主义原则,难以进行有效评价。尤其是随着人工智能决策的自主化,以评价人工智能危害后果为中心的刑法规制模式,已经无法适应人工智能的迭代式发展。[17] (3)单一的部门法甚至单一的法律手段,无力调整和评价人工智能,网络空间中的"共治"模式有待刑法层面的进一步发展。网络犯罪治理的经验表明,面对网络犯罪社会危害性的几何式倍增,以及网络空间的去中心化和扁平化、网监部门与司法机关的技术瓶颈、网络犯罪刑事责任认定的困难,例如,因果关系难以查明、罪过责任难以确立等,单纯依靠国家机关或者单纯依靠刑法,根本无法有效地防治网络犯罪。尤其对于本体化特色逐渐增强的人工智能犯罪,更是对传统的犯罪治理模式提出了挑战。

客观上,基于对大数据的挖掘预测与深度学习结合产生的自主性与分离性,人工智能时代的算法已经能够根据大数据自行进行深度学习进而具备独立的自主决策能力。此种人造神经网络算法(ANN)不同于监督学习状态下可受算法编写者控制的算法,算法编写者、使用者及其他相关利益者对基于深度学习算法的人工智能最终会作出何种行为,还难以预测和进行绝对的控制,人类也无法准确预知人工智能失去控制后导致的实害结果的影响范围和社会危害程度。因此,人工智能时代建立以风险防范为目的的法律制度已成必然趋势,必须从人工智能算法的外部行为与后果和算法内部的设计规则进行规制,建立与风险相配套的法律制度。人工智能发展所带来的技术风险、规范风险乃至人类本位风险的增强,不断表明针对人工智能犯罪的刑法回应,应当超越人工智能本身,通过人工智能背后的算法规制、数据保障等核心命题来明确技术过错与人的过错的本质关系,明确人工智能在人类范式控制下的刑法评价路径。

二 人工智能刑法规制困境的突破:刑事合规计划的引入

数据、算法作为人工智能发展进化的关键,人工智能的刑法规制应当由规制"机器人",转向对数据和算法研发的规制,转向对数据和算法过程本身的规制。有鉴于此,有必要以刑事合规为视野,从算法设计上对人工智能的识别能力、决策能力和行为方式预先进行合规性审查,[18] 确保人工智能决策系统公平地收集、存储和使用数据,从积极的一般预防的角度限缩、降低人工智能犯罪带来的危害后果。因此,通过人工智能合规计划的实施,使人工智能犯罪预防责任由人工智能产业者承担,通过人工智能刑事合规计划,倒逼人工智能产业者事前主动介入,增强人工智能犯罪的积极防控,以此带动人工智能产业者对人工智能风险防范与安全的自觉维护意识。

(一)人工智能刑事合规计划引入的必要性思考

人工智能犯罪所呈现出的严重态势以及其严重危害性已经逐渐被学界所认识,但是,

〔17〕 此种规制模式,主要限制为人工智能设计研发阶段具有主观罪过的研发设计行为,人工智能运行使用阶段具有主观罪过的部署和使用行为。

〔18〕 参见马长山:《人工智能的社会风险及其法律规制》,《法律科学》2018 年第 6 期,第 51 页。

传统刑法的单一治理模式已经滞后于人工智能犯罪的防治。刑事合规计划作为企业内部治理的手段之一,其核心目标在于通过加强企业内部治理规避刑事风险,通过刑事合规计划的制定和实施作为减免罪责,甚至出罪的根据,客观上实现预防企业犯罪的效果。详言之,专业性的旨在预防犯罪的合规计划是由一系列法律之外的措施组成的,而这些措施是由当事企业在预防违法犯罪中发展出来的。这些措施的范围包括技术上的自我保护,诱发犯罪之体制的消除,以及以企业内部制裁制度进行的预防。〔19〕从这个层面上讲,刑事合规计划同刑法预防犯罪的功能具有异曲同工之处。〔20〕

1. 人工智能犯罪的迭代异化:刑事合规是对人工智能刑事风险的回应

随着传统犯罪的网络异化、网络犯罪的传统化,人工智能带来的犯罪异化对于传统刑法理论、刑法规则更具挑战性。人工智能犯罪中,由于人的行为与人工智能决策之间的分离,使得刑事责任的确定和承担产生了模糊地带。整体上讲,人工智能犯罪的异化主要体现在:(1)危害行为的异化。危害行为作为犯罪论中的基石性概念,成为判定犯罪是否存在的基础和前提。但是,随着人工智能深度学习、自主决策能力的增强,"工具化"的人工智能逐渐向"本体化"的人工智能演变,使得"危害行为"的实施由人转变为人工智能体,造成了"危害行为"与人的分离。(2)因果关系的异化。人工智能引发的危害结果与行为之间的因果关系,呈现出极为复杂的态势,即使判定行为人违反了相应的注意义务或者作为义务,造成了危害结果,但这种注意义务或者作为义务的违反,是否与危害结果之间具有因果关系依旧难以判定。(3)主观罪过的异化。同因果关系的认定一样,对于人工智能犯罪责任的追究,行为人的罪过也是基本前提。但是,由于"危害行为"与人的分离,使得人的过错认定较之网络犯罪更为复杂,更多地需要基于对注意义务或者作为义务的违反加以判断。(4)刑事责任的异化。人工智能犯罪同传统犯罪、其他网络犯罪不同,危害后果往往具有不可控性,且危害后果具有超越人类意思的算法自觉性,刑法传统的事后评价和谦抑性评价,既无法客观上有效评价人工智能犯罪的倍增性、不可控性的危害结果,又无法主观上解决人工智能的罪过认定问题。〔21〕因此,对于人工智能犯罪的刑法评价,应当积极呼应人工智能及相关犯罪的客观规律,增强人工智能犯罪的预防与风险防范。此种背景下,刑事合规计划作为企业防范犯罪风险、加强内部管理的内控制度,能够有效解决人工智能犯罪异化所带来的刑法挑战。诚如有学者所言:"国家制定的规范有时并不符合公司的具体情况,而与这些国家规范相比,公司的自治可以是一个有效得多的方法。对于控制公司犯罪而言,在一个自制框架内,效率的额外提高是可能发生的。"〔22〕

---

〔19〕 参见[德]乌尔里希·齐白著:《全球风险社会与信息社会中的刑法:二十一世纪刑法模式的转换》,周遵友、江溯等译,中国法制出版社2012年版,第267页。

〔20〕 刑事合规计划客观起到的犯罪预防功能,也是信息时代人工智能风险增强的背景下,应被重点强调的功能。

〔21〕 除非认定人工智能的主体地位,才能基于人工智能的罪过,对人工智能判处刑罚,例如,报废、消除数据或者存储记忆,但此种"未来刑法"的思路罔顾人类中心主义,将报废、删除数据作为刑罚也违背了基本的法律认知。See Gabriel Hallevy, *When Robots Kill: Artificial Intelligence under Criminal Law*, Northeastern University Press, 2013, pp. 144 – 159.

〔22〕 [德]乌尔里希·齐白著:《全球风险社会与信息社会中的刑法:二十一世纪刑法模式的转换》,周遵友、江溯等译,中国法制出版社2012年版,第264页。

一方面,人工智能风险的不断倍增及人工智能犯罪的异化,要求刑法前置化评价以应对犯罪风险,人工智能刑事合规在这个层面上同风险刑法具有共鸣之处,即强调犯罪预防与及时干预。换言之,人工智能刑事合规通过刑法以及其他法律规则的细化,从犯罪的外部规制到自我管理,通过刑事合规实现事前的积极预防,在人工智能设计研发、部署、应用等阶段提前介入,加强风险防范。另一方面,人工智能犯罪的异化对传统刑法的挑战也在于,受人工智能深度学习、自主决策的影响,传统刑法的危害行为、危害后果、因果关系、主观罪过、刑事责任等均受到了挑战,发生犯罪之后再动用刑法评价存在重大的理论障碍。详言之,算法的设计、部署、运行,在有人的过错时,均可以评价,但随着人工智能利用阶段化的发展,越接近于应用端,应用的期间越长,人类的可控性则越差,所谓的过错也更加难以评断。人工智能刑事合规计划作为人工智能研发、部署、运行领域相关产业者的内部制度,结合其特殊的业务范围、运作流程、产品特征所设计的具有针对性的犯罪风险识别、判定与防范机制,以及其事实上是否有效履行了合规计划,可以作为判定罪责的重要依据。因此,通过人工智能刑事合规计划的引入,明确人工智能产业者的合规义务,明晰相关主体的义务范围与责任边界,客观上也为进而确定相关主体的主观罪过、因果关系等提供了根据。但是,刑事合规计划的前提在于计划的可实施性,鉴于人工智能具有很大程度的不可控性,对于超出算法透明、可解释性要求,超出数据安全管控要求等不可控的范围,便不再属于刑事合规计划所规范和评价的对象。

2. 单一的刑法规制模式滞后:刑事合规实现刑法与企业制度间的功能性协作

人工智能背景下,网络犯罪防治面临着“前门拒虎、后门进狼”的严峻挑战,即,传统网络犯罪尚未完全解决,人工智能犯罪又接踵而来。在传统刑法“救火式”被动解决网络犯罪时尚且疲于应对,对于算法异化后的人工智能犯罪更是治理效果有限,网络犯罪尤其是人工智能犯罪的刑法规制思路,单纯依靠刑法单一的制裁手段已经明显不足。[23] 同时,人工智能较之传统互联网具有更强大的技术支撑,人工智能犯罪的防治更加依赖于专业的知识水平、技术水平,尤其是人工智能算法的掌握、数据抓取的筛查与管理。传统的政府监管手段与犯罪防治手段,在人工智能犯罪面前更加显得捉襟见肘。因此,有必要强化人工智能犯罪预防理念,在人工智能犯罪的防治中引入刑事合规计划。刑事合规通过加强企业内部管理降低犯罪风险,其实质便在于将企业纳入到了犯罪预防体系之中。与之不同,“刑事司法的传统观念排外地仅仅关注刑法和刑事司法本身,这显得太过狭窄因而难以捕捉有意义的行为”。[24] 因此,人工智能刑事合规是在刑法与刑事司法的框架之外,发挥人工智能产业者在犯罪预防中的积极作用,通过刑法与企业内部刑事合规制度的结合,赋予人工智能产业者合理有效的作为义务,倒逼人工智能产业者积极履行作为义务,防范人工智能犯罪及其发展中的潜在风险。因此,通过人工智能刑事合规,可以有效实现相关法律法规的具体化、可操作化,将法律责任、法定义务具体化地、情境化地细化、

〔23〕 网络法有别于其他部门法的特有属性便在于,改变了传统的单一法律调整模式,将法律评价与内部规则相统一、国家治理与企业治理相统一。

〔24〕 See Sally S. Simpson, Making Sense of White-Collar Crime: Theory and Research, *Ohio State Journal of Criminal Law*, vol. 8, 2011, p. 481.

落实到人工智能研发、部署、应用及相关管理的日常工作中。实质上讲,人工智能刑事合规便是对现行单一刑罚惩罚模式的一种有益补充。[25]

(二)人工智能刑事合规计划引入的正当性基础

人工智能发展背后同时关涉国家、社会、公民以及互联网行业利益。人工智能技术性强,尤其是算法的更新迭代发展使得国家监管更加困难,这也是互联网空间权力扁平化和去中心化的进一步体现。无论是基于强化第三人责任以控制违法犯罪行为的功能主义论,[26]还是基于犯罪评价的社会危害性论,都将人工智能产业者的网络安全管理义务推向人工智能风险防治的前台。[27] 人工智能刑事合规计划通过人工智能产业者内部运行机制与安全防控制度,以具有可能性、必要性与可期待性的方式实现人工智能犯罪的系统化防治。

1. 网络"共治"的模式优化:人工智能外部规制与内控管理的统一

实践中,对于公司和立法者而言,公司良治的新理念在世界范围内备受青睐。合规计划(Compliance Programs)、风险治理(Risk Management)、价值管理(Value Management)、公司治理(Corporate Governance)以及商业伦理(Bussiness Ethics)、诚信守则(Integrity Codes)、行为守则(Codes of Conduct)、公司社会责任(Corporate Social Responsibility)等都是最常用的概念。[28] 事实上,由于整体商业环境对企业承担社会责任要求变高,各国对企业责任的策略性和主动性的要求界限也在不断提高。[29] 人工智能刑事合规作为单位内部对违法、犯罪进行预防、发现及治理的内控制度,其存在的核心价值便在于,督促企业积极主动承担人工智能风险的防范处置义务,进而实现人工智能犯罪的预防。例如,美国于2017年7月通过的《自动驾驶法案》(*Self Drive Act*)[30]通过"纵向行政职权配置方式,要求各州履行无人驾驶汽车的监管职权及其安全责任,并要求交通行政部门制定无人驾驶汽车的安全标准和公众评价标准等事项"。[31] 整体上讲,在当前网络犯罪"共治"模式的基本共识下,刑事合规计划通过量刑减轻、免除甚至是正当化事由的认定,引导、倒逼企业积极履行基于法律规定的、业务要求的作为义务,实现企业自身在预防违法犯罪行为上的自觉性、内在性,实现犯罪防治的"共治"合力。因此,刑事合规计划作为防治企业犯罪的模式创新,对于人工智能犯罪的防治具有紧密的契合性。对此,需要明确,要求人工智能产业者承担相关设计研发、部署、运用过程中的风险识别、监控、防治等义务,并非是简单地转嫁政府管理责任,而是基于网络时代背景下网络管理的去中心化和扁平化,人工智

〔25〕 参见李本灿:《企业犯罪预防中合规计划制度的借鉴》,《中国法学》2015年第5期,第185页。

〔26〕 See Jonathan Zittrain, A History of Online Gatekeeping, *19 Harvard Journal of Law and Technology* 253 (2006).

〔27〕 参见于冲:《"二分法"视野下网络服务提供者不作为的刑事责任划界》,《当代法学》2019年第5期,第23页。

〔28〕 参见[德]乌尔里希·齐白著:《全球风险社会与信息社会中的刑法:二十一世纪刑法模式的转换》,周遵友、江溯等译,中国法制出版社2012年版,第237页。

〔29〕 See Jooyoung Park, Nohora Díaz-Posada, Santiago Mejía-Dugand, Challenges in implementing the extended producer responsibility in an emerging economy: The end of life tire management in Colombia, *Journal of Cleaner Production* 189, 2018, pp. 754 - 762.

〔30〕 参见美国:《自动驾驶法案》(*Self Drive Act*), https://www.congress.gov/bill/115th - congress/house - bill/3388/text, 最近访问时间[2019 - 05 - 30]。

〔31〕 参见张玉洁:《论无人驾驶汽车的行政法规制》,《行政法学研究》2018年第1期,第72页。

能产业者既具有强大的管理能力,更具有维护人工智能犯罪防治的保证人地位,这种保证人地位是基于其业务属性、经营范围,在其业务属性内履行合理限度的安全保障义务具有技术可能性和业务要求性,即网络服务提供者对于信息网络安全管理应当承担起与其经营活动相一致的注意义务,这种合规义务是一种强制性义务,〔32〕是与其经营现状、服务类型等相对应的安全防控义务,不仅不会阻碍人工智能行业技术发展,而是规范人工智能更加有序健康发展,使人工智能技术发展永远以人类安全为终极目标的同时,实现人工智能犯罪的有效预防。

2. 人工智能致害责任认定的最优路径:基于责任、预防与刑罚的关系

责任和预防作为刑罚论中的基石性概念,对于刑法功能的发挥具有关键作用。简言之,责任是科处刑罚的前提,预防是科处刑罚的目的,如何实现二者在犯罪评价中的融合,关涉刑法基础理论的构建。被具体化的刑事合规概念通过刑法的具体化,外部规制与内控管理的统一,能够有效地将人工智能犯罪防治中的责任、预防与刑罚有机地统一起来。从人工智能责任追究的演变历程来看,以民事责任为例,人工智能侵权责任经历了严格责任到过错责任,再到间接责任的演变。人工智能初期沿用传统的开发商严格责任,在维护社会公共利益的同时,却给企业增加了过重的责任。随着产品化的算法、产品化的人工智能的发展,算法设计的过错责任成为人工智能责任追究的主要根据,但过错责任对人工智能产业者却存在过度宽容之嫌。因此,传统的严格责任、过错责任模式已经不能全面、有效地评价人工智能的"侵害后果"。有鉴于此,人工智能刑事合规的核心目标或者功能,在于通过企业内控制度体系和具体化的法律规则,督促人工智能产业者在研发、部署、应用人工智能的过程中,采取可能的、必要的以及可期待的措施预防犯罪,使其免受刑事追诉和刑罚处罚,〔33〕通过积极的刑法奖励引导实现人工智能犯罪的防治。〔34〕例如,针对无人驾驶汽车、智能机器人等可能造成人类人身伤害和财产损失的智能产品,通过刑事合规让人工智能设计者、生产者乃至应用者,针对人工智能的设计、部署、运行,构建科学有效并可付诸实施的合规计划,引导他们选择更具可预见性、更可控的算法。〔35〕依照此种逻辑,在关于人工智能犯罪的责任认定中,刑事合规对于相关主体的作为可能性、结果避免可能性的认定具有重要价值。详言之,人工智能产业者在犯罪发生时如果明知不法行为存在而不履行安全管理义务,在具有结果避免可能性、作为相当性的情况下,则可能构成不作为犯罪。反之,如果人工智能产业者履行了相应的作为义务,制定了系统有效的刑事合规计划并付诸实施,便不再具备刑事苛难的基础。从这个层面来讲,刑事合规计划的引

〔32〕 See Doug Lichtman, Eric Posner, Holding Internet Service Providers Accountable, *Supreme Court Economic Review*, vol. 14, 2006.

〔33〕 Dennis Bock, Strafrechtliche Aspekte der Compliance-Diskussion- § 130 QWiG als zentrale Norm der Criminal Compliance, *ZIS*, 2009, p. 73.

〔34〕 作为消费者的人工智能用户,在对人工智能犯罪没有故意或者过失的情况下,一般应认定为意外事件,不承担刑事责任。人工智能犯罪的防治主体,应主要聚焦于人工智能产业者。

〔35〕 See Trevor N. White, "Seth D. Baum, Liability for Present and Robotics Technology", in Patrick Lin, Ryan Jenkins, and Keith Abney (eds.), *Robotics Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Oxford University Press, 2017, pp. 66 – 79.

入并非仅仅止步于解决人工智能犯罪的责任认定问题,无论是对人工智能犯罪的刑法评价,还是对人工智能犯罪的预防,都具有远远超出刑事合规制度原初仅用以减免责任的价值。[36] 尤其是基于刑事合规在犯罪预防的客观效果上,通过刑法手段倒逼相关主体履行安全管理义务实现犯罪预防已经具有了相关立法探索。例如,拒不履行信息网络安全管理义务罪的增设,很大程度上对于网络服务提供者刑事合规计划的实施具有推动作用,可以倒逼网络服务提供者通过完善内控机制与刑事合规计划履行网络安全管理义务。

(三)人工智能刑事合规计划引入的可行性探究

从整个互联网产业来看,刑事合规正不断受到重视并逐步展开。[37] 互联网产业逐步成熟的刑事合规,为人工智能刑事合规的推行提供了借鉴模式和实践基础。同时也要看到,人工智能产业发展较之传统互联网产业,具有更大的不确定性,责任更难以认定,引入刑事合规的必要性更为突出,因此,其在制度设计、可操作等层面的可行性也就成为无法回避且有必要探究的问题。

1. 人工智能算法的可解释与人工智能风险的可识别

人工智能犯罪责任认定上,算法的可解释性成为关键性问题。2019 年 4 月,欧盟委员会颁布的《可信赖人工智能道德准则》(*Ethics Guidelines For Trustworthy AI*)指出:"可靠的人工智能在应提供关于人工智能系统影响和塑造组织决策过程的程度、系统的设计选择以及开发合理性的解释。"[38] 由此可见,《可信赖人工智能道德准则》将人工智能的算法和数据的可追溯性与可解释性,作为评价人工智能是否可信赖的依据之一。同时,人工智能发展的基础也在于相关算法的可解释性,算法可解释性应当成为人工智能研发、投入应用的技术前提和法定义务。2018 年谷歌大脑团队研发出一项名为"可解释性的基础构件"的技术成果,[39] 意图对人工神经网络算法进行可视化操作,使算法的工作状态回到"人类尺度",即,使算法可以被非专业人士识别和理解。因此,从技术层面来讲,打破人工智能的不可知论和技术恫吓,正视人工智能算法的可解释性,是人工智能刑事合规的基础和前提。从刑事合规层面来讲,要求人工智能算法的可解释性和透明化,也成为人工智能产业者的刑事合规义务。例如,2019 年 5 月,经济合作与发展组织(OECD)通过的《关于人工智能设计国际标准的建议》(*Recommendation of the Council on Artificial Intelligence*)中提出,"为确保人类可以理解并质疑人工智能的决策结果,人工智能系统应具备透明度并能为其决策做出负责的解释"。[40]

通过刑事合规计划加强人工智能产业者内部管理制度和风控措施,对于深藏在内部

---

〔36〕 目前国内关于刑事合规制度的研究,几乎都在强调刑事合规在加强犯罪预防中的特殊功能,尤其对产生各种犯罪异化和刑法挑战的人工智能犯罪,强调其刑事合规计划的犯罪预防功能,更具必要性。

〔37〕 参见比特律:《网络信息安全合规刻不容缓——工信部责令 7 家互联网企业进行合规管理整改》,https://www.chainnews.com/articles/811499015487.htm,最近访问时间[2019-06-21]。

〔38〕 Ethics Guidelines For Trustworthy AI, https://ec.europa.eu/futurium/en/ai-alliance-consultation,最近访问时间[2019-06-13]。

〔39〕 参见马长山:《人工智能的社会风险及其法律规制》,《法律科学》2018 年第 6 期。

〔40〕 Organisation for Economic Co-operation and Development: Recommendation of the Council on Artificial-Intelligence, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449,最近访问时间[2019-06-13]。

管理制度、平台运营中的网络安全风险,最大限度地及时发现,提前解决人工智能更加不可控、危害性更大的网络犯罪隐患。人工智能刑事合规体系中关于贯彻风险管理的义务,使得对人工智能犯罪防治涉及的风险识别、防范等进行调查和评价成为可能。[41] 在刑事合规制度体系下,人工智能研发设计者应当承担起透明、公开、程序合法、说明理由等义务,使人工智能算法等技术化的监控和决策手段不再是无法被问责的"黑箱"。同时,刑事合规计划的实施,也能督促人工智能相关产业者"通过技术正当程序,加强自主决策系统中的透明性、可责性以及被写进代码中的规则的准确性。例如,谷歌大脑团队公布的'可解释性的基础构件'的研究成果,将算法比喻成人工神经网络的核磁共振成像,这种开源化处理使得其他技术人员能够在此基础上编写适用于不同算法和场景的解释性算法"。[42]

2. 人工智能刑事合规实现刑事规则的具体化、情境化

与核技术、基因技术应用一样,人工智能带来巨大便利的同时,也伴随重大风险,而且这一风险不仅在于人工智能是否能进化出自主意识和超级智能,更在于人工智能是否会基于其不可知论、不可控论成为少数人控制社会的工具,两种风险同时需要刑法的有效规制。防控人工智能风险,应当对人工智能的研发、产品提供和应用进行全过程的安全控制,刑法应发挥其不可或缺的作用。[43] 通过人工智能刑事合规体系,实现相关法律法规的具体化、可操作化,将法律责任、法定义务具体化地、情境化地细化、落实到人工智能研发、部署应用及管理的日常工作中,[44] 根据人工智能发展的不同阶段,不同的应用类型、应用领域、服务内容,以及安全风险等具体情况制定系统的合规方案,将外化的法律规则转化为内在的公司章程,不仅有助于法律法规的有效实施,更有助于履行人工智能产业者在智能化时代维护网络空间安全、预防人工智能犯罪的企业社会责任。因此,人工智能刑事合规的关键点,便在于基于人工智能风险管理的刑法义务的确定以及相关主体内部风险管理在刑法上的要求,通过风险管理、内部标准的制定与实施,在具体措施和方法上将刑法上的义务具体化、可操作化,弥补刑法中缺乏明确性规定的不足。

三 人工智能刑事合规计划的基本立场:以数据安全与算法规制为中心

算法、数据和算力作为人工智能发展的三要素,缺一不可。[45] 人工智能的本质在于数据和算法处理,"类人化"的物理形体并非构成人工智能的必备要素,即使是人形机器

〔41〕 Dennis Bock, Strafrechtliche Aspekte der Compliance-Diskussion- § 130 QWiG als zentrale Norm der Criminal Compliance, *ZIS*, 2009.

〔42〕 李婕:《人工智能中的算法与法治公正》,《人民法院报》2018 年 5 月 23 日第 2 版。

〔43〕 参见皮勇:《人工智能刑事法治的基本问题》,《比较法研究》2018 年第 5 期,第 158 页。

〔44〕 参见石磊:《刑事合规:最优企业犯罪预防方法》,《检察日报》2019 年 1 月 26 日第 3 版。

〔45〕 参见赛迪顾问股份有限公司:《2018 人工智能核心产业发展白皮书》,《中国计算机报》2018 年 11 月 26 日第 8 版。

人,也只不过是算法主导下的一个硬件系统。[46] 因此,现有人工智能存在的失控问题主要与人工智能体中所应用的数据和算法有关,人工智能刑事合规制度建设的关键,就在于规制在人工智能体中的数据和算法。整体上讲,当前关于人工智能的法律规制模式,主要体现为以专业化审查、算法规制为核心的美国模式,以及以个人数据权利与数据安全的保护,以数据源规制为核心的欧盟模式。[47] 我国人工智能刑事合规的引入,应当立足于人工智能的风险防范,通过人工智能产业链上的算法合规与数据合规,实现人工智能风险防范与犯罪的防治。

(一)人工智能刑事合规计划的源头治理:数据安全保障

人工智能作为根据数据输入、决策参数自动提供结果的决策过程,[48] 数据成为人工智能运作的基础原料,并在人工智能运作下得到进一步分析、产生新数据。随着大数据分析方法逐渐应用于人工智能决策指导,数据瑕疵带来的风险也将会日益增加。[49] 例如,工程机械、医疗、自动驾驶等与人类性命攸关的系统,如果发生因原始数据瑕疵而导致输出瑕疵结果的情况,则会导致严重的危害后果。[50] 为了避免这类结果发生,人工智能决策所依托的数据监管则显得愈发必要。例如,由于人工智能加密技术的缺失,采用默认密码等技术措施的不足,使得网络安全存在风险、隐私保护存在威胁。[51] 再如,2018 年初,剑桥分析公司在未经允许的情况下收集了 5000 万 Facebook 用户的个人信息,严重侵犯了用户个人信息权。[52] 在这种情况下,必须对人工智能产业链各个环节上的不同主体进行制约和监督,确保数据的收集和使用合法有序进行。[53]

因此,人工智能刑事合规的基础任务,就是对人工智能决策源头数据安全的风险识别、动态评估、监测预警、举报调查、奖惩机制、事后救济,以及相关架构系统风控管理的规制。一方面,人工智能原始训练数据的刑事合规,应当成为人工智能刑事合规的基础工作。人工智能深度学习乃至人工智能算法的自我进化,都很大程度上依赖于原始训练数据。如果原始训练数据本身存在故意或者疏忽的瑕疵,那么经由算法运算后的结果数据、新增数据也会存在瑕疵,人工智能的进化则将会进一步放大瑕疵及其带来的不良后果。一般认为,数据具有天然的过时性、滞后性,[54] 数据采集者、数据抓取算法设计过程中故

---

〔46〕 参见胡凌:《人工智能的法律想象》,《文化纵横》2017 年第 2 期,第 113 页。

〔47〕 See Drew Simshaw, Nicolas Terry, Kris Hauser, M. L. Cummings, Regulating Healthcare Robots: Maximizing Opportunities While Minimizing Risks, *Richmond Journal of Law & Technology*, 2016, 22 (3): 1 - 38.

〔48〕 参见沈亮亮:《算法在市场竞争中的应用与法律难题——从大数据杀熟谈起》,《太原学院学报(社会科学版)》2019 年第 3 期,第 27 页。

〔49〕 参见彭兰:《假象、算法囚徒与权利让渡:数据与算法时代的新风险》,《西北师大学报(社会科学版)》2018 年第 5 期,第 25 页。

〔50〕 参见李智勇著:《终极复制:人工智能将如何推动社会巨变》,机械工业出版社 2016 年版,第 90 页。

〔51〕 参见腾讯研究院:《人工智能安全的法律治理:围绕系统安全的检视》,http://www.sohu.com/a/207616322_455313,最近访问时间[2019 - 06 - 08]。

〔52〕 Mae Anderson, Facebook privacy scandal explained, https://www.ctvnews.ca/sci - tech/facebook - privacy - scandal - explained - 1.3874533,最近访问时间[2018 - 04 - 06]。

〔53〕 参见雷悦:《人工智能发展中的法律问题探析》,《北京邮电大学学报(社会科学版)》2018 年第 1 期,第 21 页。

〔54〕 参见王珊、萨师煊著:《数据库系统概率》(第 5 版),高等教育出版社 2014 年版,第 15 页。

意或者疏忽的价值偏见的影响,都可能会进一步导致人工智能数据失真和片面的输入结果,[55]由此造成数据瑕疵、引发人工智能不良决策,甚至违法犯罪行为。因此,人工智能犯罪的基础本源在于数据采集、数据抓取算法过程中人的过错。基于此,人工智能刑事合规的首要任务,应当是在数据的收集和使用环节中,通过引导和规制尽可能避免包含价值性歧视的数据被纳入到人工智能机器学习的数据库中,识别和挑战数据应用中的歧视和偏见,实现人工智能的"数据透明"。[56] 因此,人工智能数据合规,就是要通过合规体系将数据的收集、使用和处理过程透明化,进而成为人工智能因数据问题决策失误时的归责、问责基础。具体言之,人工智能刑事合规的关键,即在于对人工智能赖以存在和发展的数据合规,要求制定人工智能数据收集、使用过程的监管与安全制度,将"必要原则""公开原则"和"有限原则"贯穿到刑事合规体系的制定、实施、监督等过程中。事实上,我国 2019 年发布的《数据安全管理办法(征求意见稿)》也为刑事合规的具体制度设计提供了规范依据,其中第 17 条规定:"网络运营者以经营为目的收集重要数据或个人敏感信息的,应当明确数据安全责任人。"

(二)人工智能刑事合规计划的关键核心:算法规制与风险管理

人工智能本质上是以大数据为基础,以深度学习算法为核心的自动化分析决策智能系统,即人工智能从某种意义上来说可视为一种能够自主学习、判断和决策的算法,[57]算法设计对人工智能的"行为"及其自主决策具有重要影响。人工智能时代,人类将选择权和决策权让渡于算法,算法取得了广大领域的决策权、控制权,甚至引发对公权力和管理部门的权力挑战。算法的高度技术性,使其运行机制和决策依据往往掌握在少数专业人士手中。[58] 故而,我们认为,应当打破算法的神秘化、不可知性的"技术恫吓",打破算法规制与传统法律规制之间的技术鸿沟和法律隔阂,将人工智能算法合规作为刑事合规的重中之重。因此,无论是网站推荐、检索服务,亦或是智能机器人,人工智能刑事合规的重心都应当立足于人工智能算法研发者、部署者、使用者的行为。

首先需要明确的问题是,算法并非是绝对中立的,即使技术的中立也不等于价值中立。人工智能研发者、应用者很大程度上决定了人工智能影响社会的具体模型、路径,实质上,不当的人工智能算法设计、算法应用都会对既有的社会秩序、法律规则形成巨大挑战甚至颠覆。诚如有研究者指出,人工智能的发展并不能以技术中立为由来回避它的商业偏好、善恶价值和社会风险。[59] 人工智能算法所产生的偏见、偏差乃至违法犯罪"行为",实质上都可以归结为技术过错、技术霸凌。人工智能决策过程看起来并非人为、也非人控制的决策行为,但算法内嵌程序在算法设计阶段不可避免地会带入设计者的价值取向、伦理道德素质、指标标准、结果导读等因素。因此,机器学习算法所产生的偏见和偏

〔55〕 参见周涛:《数据的偏见》,《金融博览》2017 年第 5 期,第 22 页。

〔56〕 参见许可:《人工智能的算法黑箱与数据正义》,https://blog.csdn.net/UFv59to8/article/details/79947730,最近访问时间[2019-05-28]。

〔57〕 参见郑戈:《算法的法律与法律的算法》,《中国法律评论》2018 年第 2 期,第 77 页。

〔58〕 参见赛迪智库、周游:《我国亟待建立人工智能算法审查机制》,《中国计算机报》2018 年 5 月 14 日第 12 版。

〔59〕 参见马长山:《人工智能的社会风险及其法律规制》,《法律科学》2018 年第 6 期,第 49 页。

差问题,本质而言属于人的过错。例如,谷歌公司开发的 Google Photos 智能机器人误将两名黑人标注为“大猩猩”并截图发至 Twitter 上引发的“人种歧视”。〔60〕 微软公司智能聊天机器人上线不到 24 小时,由于发布涉及种族主义、性别歧视和纳粹主义的言论被紧急叫停。〔61〕 与个人歧视的针对性、个别性不同,算法歧视体现为明显的系统化、规模化,对于整个社会的有序发展均会造成严重影响。因此,人工智能刑事合规的重心,应当也必须落脚到人工智能背后的人的行为上去。解决人工智能刑法规制的核心问题在于,正视人工智能的技术实质及其犯罪规律,通过刑事合规的引入,在人类可控、传统规则可控的范围内,解决人工智能犯罪问题。事实上,人工智能犯罪的刑法防治,通过对人工智能背后的算法合规,是在人类规则可控范围内最大限度防范人工智能风险的有效路径。通过刑事合规制度的架构,将刑法等法律法规所赋予的作为义务实现内部的具体化和可操作化,在人工智能研发、部署、应用的各环节、各流程,构建可量化、可评估的风险预防、风险监测和风险消除机制,通过算法设计、算法部署、算法应用的合规实现人工智能犯罪的防治,实现人工智能算法过错背后人的过错责任认定。同时,除了人工智能算法研发设计中隐含的歧视与不公,有些算法甚至还存在利益集团的操控。〔62〕 因此,通过刑事合规促进算法公开、算法透明,促进“黑箱社会”转向“可被了解的社会”,〔63〕更加有力地对算法暴政进行规制。故而,算法合规的重点,应当着重增强算法透明度以及可解释性,避免算法开发者及相关应用者以技术中立的原则作为开脱罪责的理由。

四　人工智能刑事合规计划的制度设计与路径架构

在刑事合规体系下,为了实现人工智能合规制度的应有价值和功能,应当明确刑事合规的出罪功能与量刑奖励功能,通过量刑激励甚至是出罪激励模式,推动人工智能企业加强内控与合规,实现人工智能犯罪的积极预防。通过在人工智能设计研发、部署、运行以及事后保障、救济阶段的刑事合规,规范人工智能的合法生产、部署和运行,架构相应的程序规则、职责规则及技术规则,用来识别、评估和消除人工智能犯罪风险。〔64〕

(一)人工智能刑事合规计划的目标定位

随着人工智能由“工具化”向“本体化”转变,人工智能自我决策能力不断增强,对于人工智能犯罪的致害后果往往难以归责于人的行为,预防性刑法对于人工智能的规制愈加具有必要性。换言之,人工智能犯罪的本体化特征意味着一旦发生危害后果将难以有效控制,并且难以评价。对于人工智能的犯罪,刑法的事后评价机制应当转化为事前的积

〔60〕 参见腾讯科技:《谷歌照片应用误把黑人标记成“大猩猩”》,http://tech.qq.com/a/20150703/004879.htm,最近访问时间[2018-12-23]。

〔61〕 参见何熠:《微软智能机器人竟然是“种族主义者”》,http://www.donews.com/idonews/article/8293.shtm,最近访问时间[2018-12-28]。

〔62〕 参见姜野:《算法的规训与规训的算法:人工智能时代算法的法律规制》,《河北法学》2018 年第 12 期,第 146 页。

〔63〕 [美]弗兰克·帕斯奎尔著:《黑箱社会——控制金钱和信息的数据法则》,赵亚男译,中信出版社 2015 年版,第 25 页。

〔64〕 Vgl. nur BGH NStZ 1997, 545 (546).

极预防机制。刑事合规本质上体现为为预防违法犯罪行为而设置的程序规则、责任规则及技术规则,主要价值在于使公司遵守刑事实体法。[65] 因此,人工智能刑事合规计划的核心目标在于,通过刑法手段促进人工智能的“法律规划”或者“法律驯化”,降低其社会危害性和技术风险性,通过积极的一般预防模式将人工智能的技术风险转变为刑法的可评价、可规制对象。根据积极的犯罪一般预防理念,刑罚预防犯罪最合理且唯一的目的应当是在犯罪之前培养守法文化和法律信仰,通过基于守法的自觉和对法秩序的积极维护,主动、积极地预防犯罪的发生。[66] 预防功能是刑事合规最为主要的功能之一,人工智能刑事合规的关键也在于根据人工智能产业者业务流程、服务类型与内部管理制度,为预防人工智能犯罪对关涉刑法的不当举止进行预防、调查与制裁。[67] 同时不容回避的是,关于刑事合规的基本制度功能,一致意见显然占据主导地位,即避免“刑事责任”。[68] 刑事合规作为企业防范法律风险的内控制度,对于人工智能犯罪预防作用的实现,应当赋予其积极的量刑功能,甚至是出罪价值,使得人工智能刑事合规同刑法紧密相连,使其成为人工智能犯罪预防体系中的重要一环。具体言之,人工智能刑事合规计划的实质在于,通过合规计划的实施避免因为人工智能研发、部署、应用过程中可能的违法犯罪行为,以及降低相应不法行为的刑事可罚性,[69] 以此实现对相关产业者预防人工智能犯罪的激励和推动。因此,刑事合规风险的分配,在预防人工智能犯罪的同时,实质上也是最大限度地兼顾技术发展与人类利益保护的平衡。在风险配置上,人工智能的设计者、编程者、部署者、使用者在合理边界内承担人工智能“失控”的风险。

(二)人工智能刑事合规计划的内容

明确人工智能刑事合规计划的目标定位之后,关键的问题则在于刑事合规计划的内容设置,尤其是防止人工智能产业者犯罪的合规计划。根据域外成熟立法的范例,人工智能刑事合规计划应当主要包括:(1)人工智能产业者相关的特定风险分析,以及基于风险预防而制定的企业内部执行的制度与程序。对那些影响公众基本权利、涉及重大社会公共利益的人工智能算法,通过建立完整的企业文化与网络安全责任意识、内部制度,明确预防人工智能犯罪的防控机制,从研发内容、内容判断标准、推荐标准、干预手段等关键环节,加强内部监管。[70] (2)具有促进制定、实施和保障合规计划得以顺利实施的平台高管参与,具有专业而独立的合规人员。确立最高领导层的责任,即关于防治人工智能犯罪的既定目标、价值和程序方面的责任;规定中层领导者的责任,即负责组建相应的专业部门(比如合规部门)以及向员工进行解释和培训。(3)建立旨在揭露和查明人工智能犯罪及

〔65〕 B. Fateh-Moghadam, Criminal Compliance ernst genommen-zur Garantenstellung des Compliance Beauftragten usw., in Steinberg/Valerius/Popp (Hrsg), Das Wirtschaftsstrafrecht des StGB, 2011, S. 25 (26 ff.).

〔66〕 参见张明楷著:《刑法学》(第五版),法律出版社 2016 年版,第 24 - 26 页。

〔67〕 D. Bock, Criminal Compliance, *Nomos*, 2011, S. 19 ff. (22).

〔68〕 D. Bock, Compliance und Aufsichtspflichten in Unternehmen, in: Kuhlen u. a. (Hrsg.), Compliance und Strafrecht, 2013, S. 57.

〔69〕 T. Rotsch, *Compliance*, in: Achenbach/Ransiek (Hrsg), Handbuch Wirtschaftsstrafrecht, 3. Aufl. 2012, Rn. 6.

〔70〕 See Stefano Manacorda, Francesco Centonze and Gabrio Forti (eds.), *Preventing Corporate Corruption*, Springer, 2014, p. 334.

其风险的信息系统，尤其是对内部人员和事务进行的控制、报告义务，接收匿名举报的“内部告发制度”，以及自我申报制度等。（4）实施层面上，开展有效的内部合规计划培训，对合规计划的实施进行有效的动态监管和审查，对合规计划定期进行评估和完善；为合规计划要素设置外部的控制人员和控制方式，建立用以防范、制裁技术滥用行为的内部措施。（5）建立激励机制，设置完善的奖励、惩戒、考评机制。[71] 对此，需要进一步明确的是，刑事合规计划的制度基础在于为确定单位刑事责任，尤其是责任减免和出罪，提供规范依据。因此，人工智能刑事合规计划重点解决的是人工智能产业者在研发、生产、部署过程中的责任问题，在罪过上应当对故意不履行合规计划、过失未能履行合规计划进行有差别的区分判定，并结合预见可能性等问题进行综合判断，以此确定是否给予相关主体减轻、免除处罚或者予以出罪。同时，对于直接责任人员、主管人员故意或者过失不履行合规计划引发危害结果的，且单位没有过错的，在现有刑法框架内追究相关人员的刑事责任。

除了刑事合规计划的制度架构之外，基于人工智能产业者风险管理义务的确立，以及对于相关企业内部风险管理刑法层面的要求，人工智能刑事合规计划还需要满足两个条件：一是在方法上将刑法的义务具体化；二是组织上成为刑法上欠缺合规体系的组成部分。即自治、共治的统一。[72] 具体言之，人工智能刑事合规计划功能目标以及制度架构实现的关键，在于完整而有效的合规体系以及风险防治体系。刑事合规体系的重要性体现为对法律风险的防范，因此，人工智能刑事合规的风险识别、风险评估、风险消除至关重要。在风险识别上，对人工智能风险管理尤其是人工智能研发、部署、应用过程中可能的风险应当给予足够重视，应当体系化、持续性地对潜在的安全风险、实然的损害后果进行识别、梳理和确认。在此基础上，对于所发现的风险在设计方法、组织程序上进行风险评估和风险消除，对可能的合规风险进行量化，以此确定应当采取的防治措施，进而根据风险发生的可能性和可能造成损害后果的严重程度，确定具体的防范方法和实施方案。[73]

（三）人工智能刑事合规计划的路径架构

人工智能具有主观性、价值性、可修改性、可解释性等特性，这些特性在不同阶段的突出程度有所不同，也决定了刑事合规计划在人工智能的不同应用阶段具有差异性。在人工智能研发阶段，通过刑事合规计划的制度架构、有效实施，以及对算法研发者的引导和规制，实现对算法的间接规制。尽管人工智能（尤其是超人工智能）自主决策主要是基于数据自动化处理得出的算法结果，但实质上仍然体现为人为编制的运算法则，其中的回报函数很大程度上体现了算法设计研发者的价值取向和设计意图。[74] 因此，在人工智能算法设计研发过程中，需要通过刑事合规计划实施，督促人工智能算法治理的法律归化。[75]

---

〔71〕 参见[德]乌尔里希·齐白著：《全球风险社会与信息社会中的刑法：二十一世纪刑法模式的转换》，周遵友、江溯等译，中国法制出版社2012年版，第246页。

〔72〕 参见李本灿等编译：《合规与刑法：全球视野的考察》，中国政法大学出版社2018年版，第53页。

〔73〕 Kromschröder/Lück, DB 1998, 1573 (1574).

〔74〕 参见姜野：《算法的规训与规训的算法：人工智能时代算法的法律规制》，《河北法学》2018年第12期，第148页。

〔75〕 参见何明升：《中国网络治理的定位及现实路径》，《中国社会科学》2016年第7期，第115页。

通过刑事合规计划的制定,在风险识别上,要求人工智能研发者、具有监督人工智能行为义务的相关人员,以"对计算机行为之技术能力的常规化预期"为标准,[76]能够预见智能机器的行为可能性。对此,有必要借鉴美国关于人工智能算法的监管模式,建立可供评估的企业内部的算法委员会,[77]赋予人工智能研发阶段的算法伦理审查义务、算法的测试义务,以及应用过程中的配套技术服务、止损义务。在此之外,在合规计划的设置中,构建算法伦理标准,避免算法歧视、算法黑箱,防止人工智能研发阶段的算法操纵所产生的人工智能应用阶段的损害后果。制定算法禁止性标准,在选定算法及数据库时,尽量剔除任何可能产生歧视的因素,避免造成人工智能算法运行的结果歧视。[78] 在此基础上,对基于深度学习在输出结果上存在异化可能性的人工智能部署、应用,设置针对性的防范措施和监控手段。由此可见,人工智能研发阶段的算法责任更多地是一种风险预防责任,不在于后果的责难,而在于风险防范的督促。例如,针对无人驾驶汽车以及智能机器人等,可能造成人身伤害和财产损失的智能产品,刑事合规计划中可以明确相应的风险识别、风险评估的防范措施与责任人员,从而引导他们选择更有可预见性、更可控的算法。[79] 事实上,2017 年 7 月,国务院发布的《新一代人工智能发展规划》就要求,"开展与人工智能应用相关的民事与刑事责任确认、隐私和产权保护、信息安全利用等法律问题研究,建立追溯和问责制度,明确人工智能法律主体以及相关权利、义务和责任等"。[80] 某种程度上讲,这一外部规制的明确化,也为人工智能刑事合规计划的具体架构提供了规则性指引。

在人工智能部署、应用阶段,算法的可修改性成为对其进行风险管理与合规"自治"的关键——算法投入运行过程中,赋予设计研发人员在人工智能使用、运行过程中的监管义务,适时对人工智能使用当中包含偏见、瑕疵、漏洞的代码进行修改,从规则与技术的双重进路防范人工智能技术异化引发的危害后果。因此,人工智能部署、应用阶段的风险识别、风险评估以及风险消除义务,对于防止人工智能"非因果性和不确定性"带来的安全风险,[81]更具有关键意义。通过刑事合规计划在刑事合规目标定位、合规内容、制度架构上,严格人工智能部署、应用者的监管义务、证据保存义务,确保人工智能使用过程中的备案、可追溯。有鉴于此,各国针对人工智能领域中的安全故障问题,出台了相应的法律法规,要求安装数据记录的相关设置,以实现证据保存。例如,英国《自动驾驶汽车测试实践准则》[82]明确规定了自动驾驶汽车应当配备专门的数据记录设备等。[83] 德国则在《道

〔76〕 See Gunther Teubner, Rights of Non Humans? Electronic Agents and Animals as New Actors in Politics and Law, *Journal of Law and Society*, Vol. 33 (2006), p. 509.

〔77〕 See T. Macaulay, Pioneering computer scientist calls for National Algorithm Safety Board, https://www.techworld.com/data/pioneering-computer-scientist-calls-for-national-algorithms-safety-board-3659664/,最近访问时间[2019-03-20]。

〔78〕 参见董妍:《人工智能的行政法规制》,《人民法治》2018 年第 11 期,第 10 页。

〔79〕 See Joseph Savirimuthu, Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence, *International Journal of Law and Information Technology*, Volume 26, Issue 4, 2018, pp. 337-346.

〔80〕 参见国务院《新一代人工智能发展规划的通知》,国发〔2017〕35 号,2017 年 07 月 20 日发布。

〔81〕 参见张成岗:《人工智能时代:技术发展、风险挑战与秩序重构》,《南京社会科学》2018 年第 5 期,第 48 页。

〔82〕 Department for Transport in UK, *The Pathway to Driverless Cars: A Code of Practice for Testing*, 2015.

〔83〕 参见唐钧:《人工智能的风险善治研究》,《中国行政管理》2019 年第 4 期,第 49 页。

路交通行为法修正案》中提出自动驾驶汽车要安装黑匣子用于数据记录和证据保存。[84] 总体上讲,在人工智能刑事合规计划的制度架构与执行过程中,风险识别、风险评估与风险消除尤为重要。因此,人工智能刑事合规计划体系中,贯彻风险管理的义务使得对人工智能风险及其防范进行调查和评价变为可能,[85] 进而实现人工智能犯罪的自治、共治的有效统一,实现人工智能风险背景下相关犯罪的积极预防。

[本文为作者主持的2017年度国家社会科学基金项目“网络共同犯罪基本原理及其对传统共犯理论的突破研究”(17CFX023)的研究成果。]

[**Abstract**] Most of the existing researches on AI crime ignore data security and algorithm regulation, which are the core issues of AI, and talk about AI in isolation, so that most of the criminal law responsiveness researches on AI focus on the level of the unknowable and sci-fi “robot regulation”. In the criminal law regulation of AI, it should first of all be made clear that the object of criminal regulation is the behavior of AI developers, rather than the “behavior” of AI. Then, we should, on the basis of the basic traditional theory and the framework of criminal law, establish the “co-governance” thinking and preventive thinking in the prevention and control of AI crimes, and shift the focus of criminal law regulation from evaluation of results after the event to the prevention of risks in advance. Therefore, it is necessary to introduce the evaluation mechanism of criminal compliance in AI, and on the condition that the algorithm can be interpreted and the AI decision-making data is transparent and by means of criminal law, use algorithmic compliance and data compliance on the AI industry chain to determine the fault of the algorithm and the fault of the person behind the algorithm bullying, and prevent AI decision-making errors caused by data flaws and human faults in algorithm design, algorithm deployment and algorithm application. More specifically, we should face up to the risks of AI, realize the crime prevention function of criminal compliance, and promote the transition from traditional criminal law evaluation model of after-the-event sanction to that of prevention, organically combine the external regulation of criminal compliance and self-management and, on the basis of the concretization and situationalization of criminal law rules, realize the functional coordination between criminal law and the regulatory frame work of AI enterprises.

(责任编辑:王雪梅)

〔84〕 Federal Ministry of Transport and Digital Infrastructure, *The German Road Traffic Act Amendment Regulating the Use of Motor Vehicles with Highly or Fully Automated Driving Function*, July 17, 2017.

〔85〕 Heine, Die strafrechtliche Verantwortlichkeit von Unternehmen, *Nomos*, 1995, S. 44, 129.