【法学研究】

人工智能机器人的刑事可罚性*

[德]拉塞·夸尔克 著 王德政 译

摘 要:为了在人工智能的运用中不形成刑法上的规范漏洞,必须在教义学上讨论智能代理的可罚性。无论是出于法律政策还是从刑罚目的理论层面考虑,电子人地位的采用都是有意义和必要的。对实现一般预防目的而言,仅确定可罚性是不够的,还必须向社会大众明确说明已实施的不法行为之可罚性,通过尽可能适当的量刑,对罪责非难进行量化,具体包括进行公益劳动、对机器人的身体进行干预、执行算法所违反规范的内容而进行重新编程、关闭机器人。人工智能的可罚性由此成为可能。

关键词:人工智能;机器人;电子人;算法;可罚性

中图分类号:D924.1 **文献标识码**:A **文章编号**:1003-0751(2020)10-0048-05

一、数字革命与法

在亚历克斯·普罗亚斯的科幻小说《智能机器人》中,有这样一幕:类人机器人桑尼对那位人类主角说自己"做梦了"。在这一幕中,人与机器之间的界限消失了。不明确的是,桑尼是一个机器人还是一个类人的生物?艾萨克·阿西莫夫著作全集中的短篇小说卷是普罗亚斯小说的渊源,从该著作全集可以看出,最晚从第二次工业革命末,人类就开始思考人工智能及其影响。很多文献和电影作品都涉及智能机器及其与人类的互动,这些作者和导演所展示出来的幻想一般都是反乌托邦式的。从弗里茨·朗所著的被列入电影艺术类世界记忆项目的《大都会》,到斯坦利·库布里克的永恒经典《2001:太空漫游》,再到沃卓斯基姐妹的《黑客帝国》,在这些作品中,机器人和超级计算机大多被描绘成对主人公的威胁。

幸运的是,这种风格化其实与自然科学的小说相关。迄今为止,机器和算法始终是人类有益的帮助者,我们多年来在日常生活中是无法忽略其存在的。数字语音助手、自动驾驶汽车、医用或护理机器人只是少数例子,却说明当代科技是无处不在的。在美国,算法甚至被用来支持判例的形成。比如,几年前罪犯埃里克·卢米斯被判处6年有期徒刑,因为算法"Compas"证明他有很高的再犯风险。①

即使上述无论如何都存在疑问的做法导致产生了诸如机器人法官这种值得向往的幻想,所有这些例子也都令人印象深刻地说明,技术在当今所能带来的一切。这一点可能通过瞬间处理巨量的各种来源的数据而做到,这在专业术语上被称为"大数据"。在此期间,算法所面临的任务是处理复杂情况并进行学习。②

(一)人工智能变得自主化

智能系统的学习能力超越了单纯的"试验与错误"之范畴。借助于神经网络和模仿生物性大脑结构的电路,系统能同时在多个层面处理信息。③通过多种关于机器学习的拟设,如强度学习、监督或无监督的深度学习等,神经网络能得到训练,甚至能作出决定和基于学习而富于创造性。④系统的能力由此产生,其自主地实施行为,无须人类的帮助就能(可

收稿日期:2020-06-10

*基金项目:教育部人文社会科学研究青年项目"刑法目的解释研究"(18YJC820057)。

著者简介:[德]拉塞·夸尔克(Lasse Quarck),男,基尔大学法学院教授,法学博士(德国基尔 24118)。

译者简介:王德政,男,成都大学法学院讲师,法学博士(成都 610106)。

能是无意识的)对很多情况和难题独立而适当地作出反应。[5]这种灵活性和适应能力蕴含巨大的潜力。然而,人工智能未来对我们社会的具体影响,当前还无法预见。面对几十年来和近几年来技术的飞速进步,我们在不太遥远的将来,会遇到类似于本文开头提到的极为先进的桑尼机器人,似乎不是不可能的。

无论如何,显而易见的是,自主化或部分自主化的人工智能通过分担人类的特定任务或使之更容易完成,会改善我们的生活条件。[6]智能代理造成的法益侵害可能由此无法避免。这尤其适用于以下情况:智能系统不但不断变得更加自主化,不再始终处于人类的监管之下,而且不断进入公共区域。智能代理可能实施的越来越多的独立行为合乎逻辑地导致法益侵害风险的升高。

(二)刑罚的目的是产生人工智能可罚性需要的出发点

那么,一种人工智能,比如以智能机器人的形式,侵害了一个保护法益,谁对此承担责任?这就成为问题所在。由于人工智能的学习能力和自主性的提升,在不远的将来,这种情况是可以想象到并且很可能发生的,这种情况下的刑法结果基于各种原因不能再归责于幕后的人,即编程者、制造者和使用者。该结果是可以想象到的,因为欠缺预见可能性[7],或者是因为,人工智能的普遍运用和不可能完全排除其导致的损害,人工智能的运用具备社会相当性,由此不再违反注意义务。[8]

自主性的前提恰恰在于,不依赖于人类从外部施加的影响。上述法律后果对智能代理而言是内在的。系统所实施的无法预见的行为是自主性合乎逻辑导致的结果,在这一点上,人类与人工智能相同。[9]因此,法学的任务必然是,为新科技的运用与该运用导致的结果设定法学上的框架条件。这导致两点:其一,在测试环境之外去运用新科技,编程的法学标准也能得到考虑。其二,法律问题的澄清对运用人工智能的社会认同之形成有决定性的贡献。[10]

社会大众形成了这样一个印象:数字化的进步对法学提出挑战,法治国原则对此没有适当的回应,规范秩序的信任受到了撼动。[11]对保护法益被侵害缺乏反应是不可容忍的。因为对规范秩序的信任而言,需要的是全面的保护,而非违反计划的规范漏洞。刑法及其适用者无法对法益受侵害作出相应的反应,这等于使值得保护的法益变相失效。[12]没有人能对一个刑法上的结果负责,这使得法的基础被撼动。[13]

刑法的目的不仅是对既有的不法情况作出反应,还在于通过这种反应防止将来出现此类不法情况。[14]这一方面通过作用于行为人而产生特殊预防的效果,另一方面通过展示给社会(刑法适用的对象)——违反法律将导致法治国意义上的回应,从而产生一般预防的效果。[15]只有当判例正式将一个行为认定为违法行为并予以处罚时,社会大众才能全面遵守法律规范。如果人们形成这样一个印象——可以侵害法益而无须承担后果,刑法的目的就无法实现。

为了在人工智能的运用中不以责任分散[16]的方式形成刑法上的规范漏洞,必须讨论智能代理的可罚性。更确切地说,讨论这一问题并不取决于,是否人们不仅认为人工智能很可能在外表上立即存在,还认为人工智能的认识能力很可能与人类一致。这要求法学在我们的规范之价值秩序的意义上,参与这一数字化的变革。因此,只有当这一点在法益侵害中表现出来之后才去寻找有关人工智能问题的解决方案,从刑罚的预防目的考虑,是令人无法容忍的。对于这类问题,必须预先在教义学上进行解释。

二、人工智能的可罚性对刑法教义学的挑战

关于刑法对智能代理的实际适用性,可以提出三个重要的批评点。这些批评点说明,我们以自然人为出发点的刑法教义学,对于自主运行的人工智能之运用,还没做好准备。[17]这些批评点即:首先,人工智能缺乏一种在刑法意义上实施行为的能力;其次,人工智能缺乏罪责能力;最后,智能代理不是刑罚的合适对象。

(一)有关人工智能的行为概念

对行为能力这一问题而言,重要的是,人们基于既定的概念理解,是多么强烈地想要维护有关人工智能的行为概念。实现这一愿望可能首先取决于,人们在多大程度上规范性地赋予这个概念以内涵。由于人们以至少是潜在的规范理解(此即行为能力的前提)来看待人工智能的能力,所以该能力至少在当前要被否定。[18]智能代理还没有这种能力去义务性地识别未知的感觉冲动以及实施相应的行为。只有当人们仅要求一个举止是在意志的支配下实施

时,行为的认定才成为问题。这种意义上的意志(不同于单纯的条件反射)是否必须从因果论的角度去理解,或者说,行为的成立是否要具备目的指向性即目的性,成为问题所在。[19]

在此,要讨论两种情况:智能代理在多大程度上有意识地去培养规范理解,或者在多大程度上有能力形成指向目的之意志。技术这个术语说明,我们以人文主义为特征、以自然人为出发点的刑法教义学,只是部分适合于解决有关人工智能的法律问题。刑法是人类为自己创造的。直到几十年前,人工智能机器还没有存在于立法者的想象和法学中。因此,人类的概念范畴一开始就欠缺一种直接的可转用性。然而,如果可罚性与自然人实施的不法行为无关,而与系统或算法实现的不法情况有关,那么,具备人类意义上的意志这一要求,可能就得放弃。[20]不同于德国法秩序的其他法秩序放弃了上述要求,其已认识到企业的可罚性。据此,企业的可罚行为并不以人类的代理行为为基础,而是以其内部结构——企业内部的组织和交流为基础。[21]人类的机关和企业的代理人实施的行为仅是其上一级交流过程的付诸实践,因此,每个人自己都没有实施行为,只有企业实施了行为。[22]

当前,关于智能代理的讨论还有:就企业的行为能力而言,其自我动力重新被提及,以使之能适应相应的算法程序。[23]在检验可罚性的前提时,行为这一构成要件要素与构成要件结果之间具备因果关系和客观归责性。这样一个后来导致结果发生的法所不容许的风险,可能由于企业的交流结构内的错误而被创设,或者由于算法内的错误而被创设。该行为概念超越了自然人不法的情况,关于智能代理行为能力的观点因赞同该行为概念而被驳倒。

(二)智能代理的罪责能力

智能代理必须由于行为或算法错误而导致不法情况出现,从而受到直接和间接的非难。其一定是具备罪责能力的。这种意义上的罪责能力的意思是,行为人能够作出决定不去实施不法行为而实施合法行为,却有意地不利用这种可能性。[24]罪责能力的基本前提是意志自由。因为只有具备自由意志的人才处于这样一个位置——在认识和理解法律规范的情况下使其行为遵守或不遵守该规范。

1.意志自由是罪责能力的必要前提吗?

这种意义上的意志自由当然是无法证明其存在的。决定论的支持者在某种程度上甚至否定(至少是怀疑)意志自由在神经科学上存在。[25]因此,只有自由作出决定的经历是真实的,而导致作出决定的过程完全是注定的(被决定了的)。[26]该决定是由遗传易感性、教育和人的社会化导致的结果,是当时的情绪情况、其他情况和诸多其他因素导致的结果。这些因素在细节上并不能完全被人们理解。下述相反的说法也是正确的:就像当前人类的意志自由无法被积极地证明一样,肯定不能认为,这样一种意志自由是存在的。

意志自由的这种现实存在应当是可罚性的必要前提,这在上述论断的背景下显得不合情理。这样一个无法证明且由此不具备法的安定性之要素,何以决定是否判处刑罚?为了解决这个问题,对行为人的罪责问题而言,需要一个现实可行的解决方案。仅在以下基础上才存在自由意志:自身有与人类相同的经历。如果我自己有自由意志,其他所有人也必定有自由意志。作为可罚性前提的罪责由此对负责性进行了分配。之所以进行这种分配,是因为已实施的不法行为在我们的社会体系内引发了冲突。通过这种分配应当能解决此冲突。[27]解决此冲突的社会需要是存在的。这是因为,基于上述原因,如果没有相应的处罚,不法情况就可能继续存在。

对罪责的这样一种功能性的理解实现了对智能代理之负责性的分配。如果基于我们的社会体制,认为算法具有意志自由,则与人类的不当行为一样,人工智能对法的违反易衍生出一种解决社会冲突的需要。对人类和机器而言,所作出的不法决定是否基于既定的生物或算法过程,或者基于在法律上有瑕疵的自由意志的形成,都是不重要的。

2.对智能机器人地位的思考

只有当智能代理被视为法律意义上的人时,这样一种基于社会冲突的责任分配才会肯定成功。[28]只有当社会大众这一对象不再被视为物,而是被视为法律主体,同时受到法律上的对待时,刑罚才能产生一般预防的效果。

人工智能具备人类的身份,即人格化。这在20世纪60年代就发生了。计算机科学家约瑟夫·维森鲍姆开发的程序伊莉莎,通过对特定的、人类在电子对话中给出的关键词和预设的词组作出反应,模拟了一个心理疗法上的治疗。[29]尽管伊莉莎的能力受到很大限制,但这场对话导致一些参与者认为程

序能理解它自身的情况。人工智能程序举止的人格化由此也被称为伊莉莎效应。[30]该效应如果能在一个连图灵测试都无法通过的程序中发生,对极为先进的、外表更像人类(如有可能)的人工智能而言,其发生更是可想而知。这种社会存在被视为某种事物,其拥有一种超越物的地位。这是极有可能的。此外,这样一种电子人地位能够在法律上得到承认。对无论如何都要采用电子人地位进行讨论而言,存在政策上的抵触。例如,2017 年欧盟议会要求关心民法问题。[31]

电子人并非教义学上的空想,这说明人的概念在法律上存在变异性。基于刑事责任能力,自然人与法人之间、成年人与未成年人之间的区别或者不同清楚地表明,有一种评价始终存在于不同的人的概念中:相关权利主体有何种社会、法律和道德地位?法秩序想要从中得出何种结论?[32]人们已观察到这种极其重要且将来愈加重要的地位、数字科技的无处不在性以及相关挑战和法律问题。我的结论如下:无论是出于法律政策还是从刑罚目的理论层面考虑,电子人地位的采用都是有意义和必要的。

(三)智能代理的可罚性

最后也要反对以下论断:电子人完全无法受到处罚。这里存在的疑问首先是,刑罚有何种特殊特性?其次是,对智能代理而言,刑罚能否适用?在这里,刑罚有三个特征:恶害性、社会伦理上的无价值评价性、否定性。

我们先来考察一下恶害施加。可以通过刑罚对行为人产生一种效果,行为人将该效果感知为不利的。[33]该效果看上去是无害的,事实上并非如此,因为宪法上的保障(如行动自由和个性的发挥)无论如何都受到了限制。在此,要举一个教科书上的案例——无家可归者案。该案中的无家可归者打算在冬天来临之前让自己受到关押,以便在有暖气的监狱里度过这个寒冷的季节。[34]这种恶害(即感受到的恶害)如果是罚金刑或者罚款,是否要进行金额上的计算,当然可能让相关者不太感兴趣。[35]这同样适用于自由刑和(作为保安处分中警察拘留的)保安监禁。在所有的情况下,同样的宪法保障都受到了限制。恶害施加并非刑罚的特殊特性。[36]

与恶害施加一样,社会伦理上的无价值评价性也不是刑罚的特殊特性。[37]行为人违反刑法、惩戒法和秩序违反法导致国家对其进行非难——行为人实施了不法举止。对违规停车这一举止开具罚单所表达出来的是——该举止与整体社会中以法秩序为形式的价值共识相矛盾。[38]此外,无法确定的是,社会伦理应在何种程度上充当刑罚的特征。它在概念上的可选择性是很差的。[39]最终产生了宪法上的如下考虑:通过这种价值判断去明确表明刑罚的特征。对刑罚而言,行为人降低了道德伦理的价值是较为重要的。如果有人想要这样做,国家实施的欺凌就会利用人性尊严保障的个人效力要求,降临到行为人身上。[40]。

由此,仅剩下否定性成为刑罚的重要特征,它直接连接刑罚与行为人的罪责。[41]个人的可非难性对于国家判处多余的刑罚而言,恰恰不是其前提。该可非难性表明了对刑法上的重要的举止更高程度的否定,也表明了该举止导致的刑罚。就像刚才所说的,如果人工智能在负有责任(源自我们的社会现实)的意义上是有罪责能力的,则刑罚也能发挥对人工智能的否定性。

然而,对实现一般预防目的而言,仅确定可罚性是不够的。必须向作为刑罚对象的社会大众明确说明已实施的不法行为之可罚性,因此,需要通过尽可能适当的量刑,对罪责非难进行量化。在此,可以考虑进行公益劳动、对机器人的身体进行干预,或者作为最后手段去关闭机器人。[42]此外,还可以通过执行算法所违反规范的内容而进行重新编程。[43]只要几个智能代理互相连接并由此全面执行该内容,就会产生最大限度的特殊预防效果,如果可能的话,还会产生最大限度的一般预防效果。人工智能的可罚性也将成为可能。

三、结论

第四次工业革命方兴未艾。相关学科要积极参与这场革命以及相互交融,否则,这些学科将落在后面或完全止步不前。基于此种原因,法秩序必须对随之而来的疑问和难题准备好答案。人工智能可罚性的采用由此是长期的,并且是必然的。反对该可罚性的观点可能遭到教义学上有说服力的反对,通过这一方式,在人工智能领域,一种扩展性的、超越当今以自然人为出发点的刑法教义学的观点,会在刑法的概念性上得到认可。该观点涉及行为能力,(尤其是)罪责能力,以及与罪责能力相关的刑罚论。

即使科技的进步在规模和速度方面或许现在有点令人惊恐,这可能是无法阻挡的,也完全不应该去阻挡这种进步。如果及时识别法律和技术上的风险并予以最小化或完全消除,则科技进步的优点无论如何都能超越这些风险。因此,“技术阻碍者”至少不会来自法学领域。在这方面,我想以德国皇帝威廉二世的一句名言作为结尾。他在 20 世纪初说:“我相信马。汽车是一个转瞬即逝的现象。”

(译者注:本文译自德语论文“Zur Strafbarkeit von intelligenten Robotern”,译者已获得著者的中文翻译和发表授权,本文内容系全球首发。)

注释

①Smith. New York Times v. 22.6.2016, abrufbar unter https://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html? module = inline (4.2.2020); FAZ v. 11.6.2019, abrufbar unter https://www.faz.net/aktuell/rhein-main/algorithmen-werden-in-amerika-bei-gerichtsprozessen-genutzt-16230589.html (4.2.2020).②Vgl. Bräutigam/Klindt. NJW 2015, 1137 (1138 f.).③Styczynski/Rudion/Naumann. Einführung in Expertensyste-me, 2017, S. 132 ff. ④ Zu den verschiedenen Prozessen beim Maschinenlernen Görz/Nebel. Künstliche Intelligenz, 2003, S. 108 ff.; Gopnik. SdW kompakt v. 1.10.2018, S. 15 ff.; Wolfangel. SdW kom-pakt v. 17.20.2016.⑤Hilgendorf. ZStW 130 (2018), 674 (675).⑥Vgl. Kirchschläger. AJP/PJA 2017, 240 (241 f.).⑦Dazu Gless/Weigend. ZStW 126 (2014), 561 (581); vgl. auch Markwalder/Simmler, AJP/PJA 2017, 171 (177).⑧Gless/Weigend. ZStW 126 (2014), 561 (583 f.); krit. Gless/ Janal, JR 2016, 561 (566).⑨Borges. NJW 2018, 977 (978).⑩Sternberg-Lieben. in: Hilgendorf (Hrsg.), Robotik im Kon-text von Recht und Moral, Robotik und Recht, Bd. 3, 2013, S. 119.⑪Brüning. Das Verhältnis des Strafrechts zum Disziplinar-recht, 2017, S. 186 f.⑫Vgl. Meier. Strafrechtliche Sanktionen, 4. Aufl. 2015, S. 36.⑬Brüning. in: Gesk/Jing (Hrsg.), Digitalisierung und Straf-recht in Deutschland und China, noch unveröffentlichtes Manuskript des Vortrags vom 22.11.2018 an der Universität Osnabrück.⑭⑮㉝㉞Meier (Fn. 12), S. 18 ff., 21 f., 16, 16.⑯Beck. JR 2009, 225 (227 f.).⑰②Brüning (Fn. 13); Seher. in: Gless/Seelmann (Hrsg.), Intel-ligente Agenten und das Recht, Robotik und Recht Bd. 9, 2016, S. 46 f.⑱Seher (Fn. 17), S. 48 ff.⑲Roxin. Strafrecht, Allgemeiner Teil, Bd. 1, 4. Aufl. 2006, § 8 Rn. 10 ff.; Heinrich. Strafrecht Allgemeiner Teil, 5. Aufl. 2016, Rn. 96 ff.⑳Brüning (Fn. 13).㉑Ortmann. NZWiSt 2017, 241 f.㉒Dannecker/Dannecker. NZWiSt 2016, 162 (164).㉓Vgl. Teubner. AcP 218 (2018), 155 (157 ff., 165 f.).㉔Roxin (Fn. 19), § 19 Rn. 1 ff.; BGHSt 2, 194 (200); 18, 87 (94).㉕Vgl. Marlie. ZJS 2008, 41 (44 in Fn. 47-49).㉖Vgl. Marlie. ZJS 2008, 41 (44 in Fn. 52).㉗ Markwalder/Simmler. AJP/PJA 2017, 171 (180); Gless/Weigend. ZStW (126), 2014, 561 (574 f.); vgl. auch Roxin (Fn. 19), § 16 Rn. 39 ff.㉘Seher (Fn. 17), S. 58; Zur Frage, ob eine Sta-tusdebatte überhaupt sinnvoll ist, Beck, in: Hilgendorf/Günther (Hrsg.), Robotik und Gesetzgebung, Robotik und Recht, Bd. 2, 2013, S. 239.㉙Siehe Österreichische Akademie der Wissenschaften v. 1.12. 2017, abrufbar unter https://www.oeaw.ac.at/detail/news/gefangen-im-eliza-effekt/ (4.2.2020).㉚ Vgl. dazu Gless/Weigend. ZStW 126 (2014), 561 (565); Herberger. NJW 2018, 2825 (2826).㉛ Entschließung des Europäischen Parlaments von 16. Feb-ruar 2017 mit Empfehlungen an die Kommission zu zivil-rechtlichen Regelungen im Bereich Robotik (2015/2103 [INL]), abrufbar unter http://www.europarl.europa.eu/sides/getDoc.do? pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//DE (4.2.2020).㉜Teubner. AcP 218 (2018), 155 (168 f.); Vgl. zum Perso-nenbegriff im Zoll- und Mehrwertsteuerrecht Scheller/ Zaczek, UR 2015, 937.㉟Vgl. Roxin. in: Hassemer/Kempf/Dörr/Moccia (Hrsg.), In dubio pro libertate, Festschrift für Klaus Volk zum 65. Ge-burtstag, 2009, S. 602.㊱A.A. Meier (Fn. 12), S. 16.㊲㊳Brüning (Fn. 11), S. 547 ff., 549.㊴Roxin (Fn. 35), S. 603.㊵Brüning (Fn. 11), S. 543 f. m.w.N.㊶Brüning (Fn. 11), S. 547 f. ㊷ Gless. GA 2017, 324 (328).㊸ Gless/Weigend. ZStW 126 (2014), 561 (589).

责任编辑:邓 林

The Criminal Punishability of Artificial Intelligent Robots

[*Germany*] *Lasse Quarck the author* *Wang Dezheng the translator*

Abstract: In order not to form the normative loopholes in criminal law in the application of artificial intelligence, we must discuss the punishability of intelligent agent in doctrines. Whether it is from the perspective of legal policy or from the perspective of the theory of the purpose of punishment, the adoption of the status of electronic person is meaningful and necessary. In order to achieve the general purpose of prevention, it is not enough to determine the punishability only. It is also necessary to clearly explain to the public the punishability of the illegal acts that have been committed. Through appropriate sentencing as far as possible, we can quantify the blame for the crime, including public welfare labor, intervening in the body of the robot, reprogramming the content of the algorithm that violates the norms and turning off the robot. Therefore, it is possible to punish artificial intelligence..

Key words: artificial intelligence; robot; electronic person; algorithm; punishability