

人工智能系统的刑事责任认定^{*}

罗曼·德雷姆利加 娜塔莉亚·普里塞金娜 著^{**}

刘 源 译^{***}

薛前强 校^{****}

摘要:本文以刑法的立法规范和基本原则为切入点,研究如何将现行刑法体系中的刑事责任认定运用到以人工智能作为犯罪工具的场景中。本文立足于俄罗斯刑法中故意犯罪和过失犯罪的立法与司法现状,对两种犯罪行为进行比较和分析。其一是使用人工智能系统实施故意犯罪的罪责判定问题。有研究认为,在此情形下立法者和司法者应重新考虑罪责认定方式,因为人工智能系统在某种意义上可能具备自我认知与意志,而传统故意犯罪的责任认定需要犯罪人的主观犯罪故意,故法院不能再依赖传统故意犯罪的犯罪要件进行判断。其二是开发者是否需要承担过失犯罪的刑事责任的问题。人工智能开发者可能影响其创建的人工智能系统并使之具备伤害能力,假设人工智能研发者对人工智能产品所造成的危害不承担任何形式的刑事责任,则很容易导致负面的社会影响。但是,如果要求开发者必须考虑研发产品所有的潜在危险,也同样会产生不利后果。应当在上述两种极端情况之间寻求一个平衡点。总而言之,本文认为目前刑法体系框架并不适用于人工智能犯罪领域。

关键词:人工智能 犯罪意图 故意犯罪 过失犯罪

一、导言

人工智能的发明起初是为了方便公众的日常生活,但是对于一些“特殊分

^{*} 本文原题为“The Concept of Culpability in Criminal Law and AI Systems”,原刊于 *Journal of Politics and Law* 13(3),经授权翻译发表。本文为中央民族大学2022年研究生课程建设和教育教学改革项目“新时代法学类研究生教育教学多维改革研究”(项目批准号:GRSJYGG013)的阶段性成果。

^{**} 罗曼·德雷姆利加(Roman Dremluga),俄罗斯海参崴远东联邦大学法学院;娜塔莉亚·普里塞金娜(Natalia Prisekina),俄罗斯海参崴远东联邦大学法学院。

^{***} 刘源,中央民族大学法学院。

^{****} 薛前强,中央民族大学法学院。

子”而言,犯罪即是他们的“日常生活”。古往今来,犯罪分子一直试图利用技术手段提高犯罪的效益和收益,例如海盗利用快艇躲避海军军舰并追击受害者。据统计,银行劫匪最喜欢的犯罪工具是汽车,同样,无人机难逃毒贩的魔爪,互联网也成了网络黑客的帮凶。大多数情况下,每当新技术研发成功,其或多或少都会被用于非法目的。更有甚者,新技术在合法领域中的应用平平无奇,在犯罪领域反倒“颇受好评”。

此类例证屡见不鲜。首先,在所有州都被禁止的非法儿童色情制品大部分都是借助互联网进行传播的。其次,起初基于优秀理念研发的技术,现今却常被用作犯罪的“暗网”(dark net),其非法领域的收益远远超过了合法收益。如今,“暗网”已成为非法毒品交易以及盗窃信用卡分赃的主要据点(Pergolizzi, Lequang & Taylor et al. 2017; Duxbury & Haynie, 2017)。设想,尽管开发者研发此类技术是基于一个良好的出发点,但仍有一些技术被自然地运用于实现非法目的。互联网是促进科学发展,促进情感沟通,消除种族、宗教和文化障碍的进步工具,但现实中互联网衍生出的社会问题不在少数,其中之一就是以互联网为媒介实施犯罪,而“暗网”则以工具的身份得以幸存。

互联网及暗网是颇具代表性的例证,理想性的互联网技术却分化出了暗网的应用,这表明技术发展的理想和实际之间存在偏差。然而,在相关法律法规尚未制定时,互联网早已在全世界投入应用。显然,大多数问题在技术开发和引进的初始阶段是可以避免发生的。只有在少数例外情况下,法律规则才会随着技术的发展而变化。因此,法律应与时俱进并且做好准备迎接未来的挑战。

诸多文献聚焦于人工智能开发和应用中的伦理问题或法律问题。例如,部分文章建议将道德或法律中的规则落实到人工智能行为原则之中,以防止此类系统造成伤害,或对伤害进行补救。从法理维度上看,这似乎符合逻辑,但正如帕特里克·林(Patrick Lin)在谈到自主机器时所言:“考虑将机器人造成伤害的风险降到最低的一个自然方法是让它们遵守我们的法律或道德准则。当然,这说起来容易做起来难,因为法律可能是模糊的、情境性的,至少在可预见的未来,机器人还不够成熟,它们无法理解这些。”(Lin, 2012:9)正是因为法律和伦理均有着如此复杂的规范体系,高度依赖于适用语境,所以绝无可能设计出一种完全不会被用作犯罪工具的机器。

本文探讨了人工智能应用对刑法中“罪责”概念的影响。当一个工具拥有

了自己的意愿和决策自由,这意味着什么?一个甚至对于研发者来说都是黑匣子的人工智能的自治系统,是否应成为影响被告罪责的认定因素?

二、刑法中的犯罪意图

在刑法理论中,犯罪故意是指犯罪主体的一种犯罪心理状态。只有将不法行为归咎于某一特定的人,才可能使其承担刑事责任(Fletcher,1998)。“犯罪意图”(或“罪责”)一词意味着,如果此人处于可归责的精神状态,他/她可能需要对其犯罪行为负责。在刑法体系中,犯罪故意通常可以有二种、三种或四种罪责模式(种类),一般取决于具有管辖权的国家的刑法体系。将犯罪归于一个人的做法在许多法律制度中很常见(Fletcher,1998)。例如,在美国法律学会制定的《示范刑法》(*Model Penal Code*)中存在着四种传统的罪责形式:蓄意、明知、轻率和过失(Alexander, Ferzan & Morse, 2009: 23)。在英国刑法中则有三种模式:故意(分为直接故意和间接故意)、轻率和过失(Mohamed & Marchuk, 2013)。俄罗斯的刑法典规定的犯罪意图包括:故意、间接故意、极端轻率和过失。最后,中国的刑法规定了两种罪责模式——故意和过失,但在刑法的实际适用中采用了三种形式的故意(Mohamed & Marchuk, 2013)。因此,故意、过失和其他形式的罪责是犯罪的方式,通常被描述为犯罪的主观要件。然而,究竟什么类型的罪责才是犯罪意义上的罪责?其中对于犯罪意图的判断极为重要,因为责任性质、刑罚性质和法律后果都取决于此。

除少数情况外,所有法域都将“犯罪意图解构为认知和意志的双重因素”,是“这些构成因素间的相互作用和强弱程度决定了犯罪主体的罪责程度”(Mohamed & Marchuk, 2013: 28)。而人工智能被定义为一种间歇性具备认知能力(认识、理解和思考)和自主意志的机器或软件。一些学者的研究为该观点提供了理论支撑,他们将自治机器描述为反应性的(对环境的变化做出及时反应)、自我控制的(对自己的行为进行控制,而不是由任何其他代理直接控制)、目标性的(不只对环境做出反应)和连续性的(一个连续运行的过程)(Brolek & Jakubiec, 2017)。因此,在理论上,这意味着当人工智能系统成为犯罪工具时,犯罪主体的罪责程度可能会受到人工智能的认知水平和意志特征的影响。

三、故意犯罪

人们通常认为故意犯罪比过失犯罪更严重(Fletcher, 1998)。因此,从犯罪心理因素来看,故意犯罪是最危险的犯罪类型。故意犯罪(目的)是指犯罪主体预见并期望其行为产生消极后果,或者在其行为造成相应消极后果的情况下,犯罪人知道该后果的存在或者希望该后果发生。例如,除极少数的例外情况,如果用枪杀人,大部分人都清楚,扣动手枪扳机并将其对准受害者的结果是可以预测的。如果一个人否认他/她已预测到该结果并以其不愿意致被害人死亡的理由抗辩,那么法院可适用“可预见性测试”(virtual certainty)来进行判断(Wallerstein, 2014)(该术语来源于英国刑法,其他司法管辖区可适用不同的术语)。这意味着某些行为的可能后果对任何一个正常人来说都是显而易见的。

然而,如果人工智能被用作犯罪工具,罪责问题的认定就变得更加困难。关键点在于,人工智能系统在其运行环境中所做出的决策和其学习的能力。从某种意义上说,这些特征使人工智能构成了实施犯罪的独立行为体。然而,行为的不确定性、意志的自由性及自身复杂的认知为犯罪主体的犯罪意图认定带来了困难。设想,当有人向已经能够理解人类语言或文本,并能够学习和控制自己行为的智能系统发出指令时,会发生什么情况?想象一下:在执行任务的过程中,这个系统不是因为操作员的意志,而是由于产生的“副作用”杀了人。合理的推断是,指挥系统的操作员可能不必承担相应责任,因为他/她无法预料这一后果,毕竟他/她既没有实际按下扳机,也没有瞄准。

基于上述推断,在此种情形下,可以排除所有军事机器人指挥官对战争罪和危害人类罪的任何形式的刑事责任。例如,灭绝种族罪是最严重的战争罪,是一种“意图全部或部分摧毁一个民族、族裔、种族或宗教团体”的严重暴力行为(Rome Statute of the International Criminal Court, 1998: Article 6)。更严重的是,一些人工智能系统已经显示出种族歧视和其他歧视的倾向(Caliskan, Bryson & Narayanan, 2017)。可想而知,当一个军官给装备了人工智能的机器人下达消灭所有敌人的命令时,这台承载种族歧视系统的机器人可能只会消灭具有特定种族特征的人。就客观后果而言,这可以被视为严重罪行,但军官或其他人不会为此承担刑事责任,因为灭绝种族罪只规制“故意”的行为(Rome Statute of the

International Criminal Court, 1998: Article 6)。若指挥者因过失实施犯罪,则无法对其进行制裁。这个生动的例子揭示了,由于智能系统存在自主认知和意志,因此对自然人犯罪意图的进一步判定可能会更加复杂,甚至可能会出现不存在犯罪责任人的情况。

直接故意意味着罪犯要直接地认识到他/她的行为很有可能产生消极后果,并积极实施这些行为。故有必要特别估算,主体有多大的可能性认识到其实施的有目的(直接故意)的犯罪行为会产生消极后果。然而,人工智能系统通常根据开发者或程序员的目的进行编程或学习。通常情况下,用户可能只能依赖研发人员提供的一些信息来预测人工智能系统会采取的行动。因此,用户对系统的理解是基于研发人员提供的信息,并不能完全理解系统如何对他/她输入指令做出反应。

罪犯运用人工智能工具是很难预测其行为后果的。例如,毒品交易可以通过部署自主机器来实现(King, Aggarwal & Taddeo et al, 2020)。这些机器可以“发送和回复电子邮件或使用即时消息每天与数十万甚至数百万人进行一对一的对话,提供色情文字或毒品”(Kerr & Bornfreund, 2005)。各种后果可能来自人工智能销售,并且可能会涉及数十亿人的毒品消费。在人工智能的这种“辅助”下,实施犯罪的自然人主体就不用对其行为预期或希望产生的结果做出很明显的表态。

同样,具有自主学习行为能力的人工智能系统不仅对用户来说不可预测,从研发者角度看也同样是不可预测的。微软开发的 Tay-bot 就是这种不可预测性的典型例子。该聊天机器人配备了人工智能系统并具备自主学习能力,可以通过分析其他用户的聊天记录来学习如何参与会话,并且还能从用户的回复中学习一些短语。人工智能聊天机器人的这一特性引发了一场大丑闻,因为 Tay-bot 在对话中使用了一些在与其他用户的对话中学习到的纳粹短语。毫无疑问,微软并没有故意开发出在互联网上侮辱公众的人工智能。由于使用侮辱性言辞的用户可能也并非故意使聊天机器人学习侮辱性的短语,因而最终结果是——可能没有人就此受到谴责。

因此,执法者和立法者应重新考虑在使用人工智能系统实施故意犯罪的情况下,如何确定罪责的路径。由于人工智能有自主的认知和意志,法院在故意犯罪中不能依赖传统的罪责概念,即仅仅通过犯罪人的行为来判断是否为故意。

除此之外,另一个令人关切的问题是自动致命武器的使用,因为国际刑法没有规定在战争罪、危害人类罪和蓄意灭绝种族罪中考虑犯罪意图这一因素,仅通过行为便可认定该罪责。国际社会应就人工智能应用的刑事责任商讨新的规制规则,探讨指挥官、研发人员、机器研究专家和其他相关人员的责任程度。

四、过失犯罪

可以说最有可能影响人工智能系统实施危害行为的主体是研发者(程序员)。那么,当人工智能造成犯罪、引发伤害时,追究开发者的刑事责任是否合理?若程序员开发的自主程序或机器人只用于故意犯罪或(和)造成实质伤害,则构成故意犯罪;然而,如果开发者并无故意犯罪的意图,那么开发者则可能被认定为过失犯罪。这意味着一个人“应当意识到存在实质且不可预测的风险,或其行为将产生重大且不合理的风险。风险的性质和程度必须使行为人考虑到其行为的性质和目的以及他所知道的情况而未能察觉到此种风险,这就严重偏离了一个理性人在做出相关行为时所能注意到的‘谨慎标准’(Model Penal Code)”。此外,在大多数司法管辖区,过失一般指:应当预见自己的行为可能出现危害社会的结果,因为疏忽大意而没有预见,或者已经预见而轻信能够避免,以致发生这种结果。

美国法学会《示范刑法》中对过失犯罪的定义引发了一个问题:在人工智能的开发和技术设计方面,谨慎的标准是什么?如果该领域发展迅速,这是否意味着无人就此承担责任?或者,是否只要开发、设计者没有考虑其产品所造成的所有潜在损害,就要为此负责?显然,以上两种观点都过于片面。但是,如果产品开发者不负刑事责任,就将引发消极的社会后果。当前,市场上充斥着很多含有人工智能的危险产品。尽管产品开发者仍将对其产品给受害人造成的损害承担民事责任,但如果收益超过赔偿金额,那么开发者会继续选择生产此类危险产品。当研发人工智能系统的主体不得不考虑产品所造成的所有潜在危害时,也会产生消极的社会影响。鉴于存在不可预知的法律风险,这种要求将会导致合法参与人工智能研发的人员越来越少,这种对研发人员归责的司法态度势必阻碍人工智能领域的发展。

在实践中,由于缺乏对于人工智能研发者责任的追究,因而在研究这一问题

时有必要对制造商的共同责任进行审查。事实上,当法官判定研发者是否存在过失时,他/她必须确定研发者采取了哪些必要措施来降低损害消费者的风险。如需研发者承担责任,“可以在法庭上证明,公司对其产品使用过程中产生的缺陷、风险和潜在危害存在疏忽,那么即可追究公司因其产品给受害人造成损害所应承担的刑事和民事责任”(Asaro,2012:170)。

基于法律实践,追究研发者的过失责任具有一定的局限性,这表现在研发者大多利用行业标准(Asaro,2012)或行业规范逃避过失犯罪的指控。例如,行业规范要求自动驾驶汽车的研发者必须遵守道路交通安全要求,如果违反强制性法律中的安全要求,便可能构成刑事犯罪,警察便有权进行干预管控(Vellinga,2017)。所以,行业规范反而使开发商能够更加简单地逃避过失指控,即如果意图逃避刑事责任的追究,仅需生产符合标准的产品即可。

研发者盲目遵循行业标准还可能引发社会规范的冲突。设想,在马路中央,一个孩子跳到由人工智能操控的汽车前,汽车操纵程序的一个选择是转向另一旁的小道,以避免撞到孩子。但是,这种操作会违反安全规范要求,即汽车必须在行车道上行驶的规定,在此情形下反而会引发其他风险。因此,如果政策制定者将行业标准作为判处研发者是否应承担刑事责任的依据,那么在研发者违反相关规定并造成负面后果的情况下,将会被判处玩忽职守罪。但是从公众视角而言,这辆车如果选择撞死或重伤了孩子才是更加难以接受的结果。然而,研发者却能因此避免承担刑事责任。一般来讲,标准的制定是为了减少潜在的危害,但在危急情况下,智能系统可以搜寻最优解来解决问题。

自动驾驶领域中的决策问题受到了社会的广泛关注,通常被称为“电车难题”(Wu,2020):假设汽车研发者不严格遵守法律及行业标准,但能够在危急时刻做出正确决策,使损失降到最低,那么研发者可能会挽救很多人的生命,并因此得到社会的广泛认可。有问卷调查模拟了道路上自动驾驶汽车处于危险状态时所采取决策的情境,结果显示,更多的人希望自动驾驶汽车选择救人,而不是死板地遵守法律。但人们在上述问卷中给出了对危险情况的多种不同反应,这表明,即使对于品行较好的人而言,在存在额外干扰因素(年龄、社会价值观、性别等)和受害者的特殊情况下,也很难选择应该做出什么样的决定。而这个困境仍然存在:很难判断一个研发者究竟会为自己的产品选择哪种倾向。

研发者避免承担过失犯罪责任的另一个方式是告知消费者产品的潜在风

险,即如果开发者在产品使用过程中就可预见的风险分别向消费者发出通知,他/她将不必承担责任。各国立法对告知消费者潜在风险的义务有着不同的规定,但内容大同小异:研发者必须告知消费者产品可能造成的严重伤害或死亡风险。例如,2005年英国颁布的《一般产品安全条例》宣称:“在其活动范围内,生产者应向消费者提供相关信息,使他们能够(a)评估产品在正常或合理可预见的的使用期内固有的风险,在没有充分警告的情况下,这些风险不应是立即显现的,并且(b)对这些风险采取预防措施。”(The General Product Safety Regulations, 2005: Article 7)。这一要求的难点不仅在于预测风险,而且在于需要提供适当的信息。人工智能系统作为一种软件,具有非常复杂的环境响应系统,并且在软件自主更新的情形下这种系统有着更为积极的响应。因此,很难用每个人都能理解的方式来描述该系统,而且将通知覆盖到每一个用户同样存在难度。

然而,能够自主更新的系统可能会带来额外的挑战。事实上,研发者能够预测在系统自主更新期间将出现的一些未知风险。但 Tay-bot 的例子表明,对于能够自主更新的软件来讲,并不是所有的风险都可以被预测到。此外,有些人可能会有意干预系统更新的过程,从而引发消极后果,就如何分配程序员和第三方的责任,目前还没有答案。有公正的评论认为:“关于故意的构成要件,在某些情况下,犯罪意图可能是缺失的,造成犯罪意图缺失的潜在原因是,尽管自主代理可以自动实施犯罪行为,但自主代理程序的复杂性使得设计人员、开发人员或部署人员(人类代理)既不知道也无法预测自主代理犯罪的作为行为或不作为行为。”(King, Aggarwal & Taddeo et al., 2020)

正如芬威克(M. Fenwick)所强调的:“没有过错的惩罚,或者与过错程度不成比例的惩罚,会被认为是不适当和不公正的。”(Fenwick, 2009: 113)如果对犯罪分子贸然进行惩罚或者惩罚的力度超过了根据其参与程度所应承担的水平,就会引起社会的广泛关注,这意味着法院将面临新的挑战。如果违规者没有意识到应用人工所产生的所有风险,却和存在主观故意者受到同样的惩罚,那显然是不公平的。所有者、操纵者、设计者对自己拥有的人工智能的所有行为都承担严格责任是不公平的。但如果所有者、操纵者、设计者仅仅因为应用人工智能而逃避了本应承担的刑事责任,那也是不公平的。因此,有必要在两种极端情形之间寻求一个平衡点。

五、结论

本文的目的是强调将有关的人工智能以及人工智能在刑法中的罪责概念纳入关键立法范畴。本文的分析结果并不意味着我们需要停止发展这项颇有前景的技术,而是通过简短概述表明,人工智能的出现导致刑法罪责问题变得更加复杂而非简单。当然,本文中提出的问题多于答案,笔者希望未来能够引起相关研究更加广泛的关注。

此外,对任何立法者来说,一个明确的建议是重新界定传统刑法中的罪责概念。并且,在以人工智能作为犯罪工具的情况下,需要重新制定新的标准评估个人的罪责程度。这种新的罪责范式不可能具有普遍性,因为新的刑事责任认定方式必须符合一个国家的传统法律框架。不过,尽管法律传统千差万别,但成熟的概念仍应具有一些普遍性的特点。在本文提出的问题中,不同立法面临着相同的挑战,即,毋庸置疑,一个造成严重后果的人必须承担相应程度的责任,但一个无辜的人也可能会被认定为应负责任。此外,还必须根据人工智能技术的特性设计新的罪责评估规则。例如,新的法律认定必须考虑到人工智能在某种意义上是否拥有自己的认知和意志。

传统刑法理论无法克服的困难在于,人工智能系统可以被视为一把有着自己认知和意志的“枪”,这是一种有着复杂程式的、可以自主更新及研发的“枪”,它可以根据输入的信号和经验“射击”不同的目标。通常,它的行为模式复杂到研发人员也无法预测。如果犯罪分子将人工智能作为犯罪工具,则无法明确认定犯罪分子对其行为负面后果的认知程度、过失程度等因素。

参考文献

- Alexander, L., K. Ferzan & S. Morse 2009, *Crime and Culpability: A Theory of Criminal Law*, Cambridge: Cambridge University Press.
- Asaro, P. 2012, “A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics.” in *Robot Ethics: The Ethical and Social Implications of Robotics*, London: MIT Press.
- BroĹek, B. & M. Jakubiec 2017, “On the Legal Responsibility of Autonomous Machines.” *Artificial Intelligence and Law* 25(3).

- Caliskan, A. , J. Bryson & A. Narayanan 2017, “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.” *Science* 356(6334).
- Duxbury, S. & D. Haynie 2017, “The Network Structure of Opioid Distribution on A Darknet Cryptomarket.” *Journal of Quantitative Criminology* 34(4).
- Fenwick, M. 2009, *Corporate Wrongdoing and the Limits of the Criminal Law in Facing the Limits of the Law*, Berlin: Springer.
- Fletcher, G. 1998, *Basic Concepts of Criminal Law*, New York: Oxford University Press.
- Kerr, I. & M. Bornfreund 2005, “Buddy Bots: How Turing’ s Fast Friends Are Under – Mining Consumer Privacy.” *Presence: Teleoperators and Virtual Environments* 14(6).
- King, T. , N. Aggarwal & M. Taddeo et al. 2020, “Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions.” *Science and Engineering Ethics* 26.
- Lin, P. 2012, “Introduction to Robot Ethics.” in *Robot Ethics: The Ethical and Social Implications of Robotics*, London: MIT Press.
- Mohamed, E. & I. Marchuk 2013, “A Comparative Study of the Principles Governing Criminal Responsibility in the Major Legal Systems of the World (England, United States, Germany, France, Denmark, Russia, China, and Islamic Legal Tradition).” *Criminal Law Forum* 24(1).
- Pergolizzi, J. , J. LeQuang & R. Taylor et al. 2017, “The ‘ Darknet’ : The New Street for Street Drugs.” *Journal of Clinical Pharmacy and Therapeutics* 42(6).
- Rome Statute of the International Criminal Court 1998, Text of the Rome Statute circulated as document A/CONF. 183/9 of 17 July 1998 and corrected by process-verbaux of 10 November 1998, 12 July 1999, 30 November 1999, 8 May 2000, 17 January 2001 and 16 January 2002. The Statute entered into force on 1 July 2002. Retrieved from https://www.icc-cpi.int/NR/rdonlyres/EA9AEFF7-5752-4F84-BE94-0A655EB30E16/0/Rome_Statute_English.pdf.
- The General Product Safety Regulations 2005, The United Kingdom, Retrieved from <http://www.legislation.gov.uk/uksi/2005/1803/regulation/7/made>.
- Vellinga, N. 2017, “From the Testing to the Deployment of Self-Driving Cars: Legal Challenges to Policymakers on the Road Ahead.” *Computer Law and Security Review* 33(6).
- Wallerstein, S. 2014, “Oblique Intent in English and Jewish Law.” *Oxford Journal of Law and Religion* 3(2).
- Wu, S. 2020, “Autonomous Vehicles, Trolley Problems, and the Law.” *Ethics and Information Technology* 22(1).