人工智能犯罪责任主体的划分

杨钧雅 杨飞霞

兰州理工大学，甘肃兰州，730050；

摘要：对于人工智能能否构成犯罪，国外主要有“心理要素说”“智能代理说”“法人类比说”“ 法定实体说”“当然主体说”等观点，国内主要有否定说与肯定说两种观点。希望通过本文研究，了解有关人工智能犯罪的前沿研究成果，在人工智能是否能够成为刑事责任主体方面的探讨上提出理论与实践见解，为切实解决人工智能犯罪问题提供一定帮助，以推动人工智能技术与法律责任的平衡发展，为社会的和谐稳定贡献力量。

关键词：人工智能犯罪；责任主体；刑事责任

DOI:10.69979/3041-0673.25.03.009

引言

在科技浪潮中，人工智能技术正以令人惊叹的速度迅猛发展，当下，无人驾驶汽车这类典型的人工智能产品，已不再仅仅是技术革新的象征，与之相关的涉嫌犯罪案例陆续浮出水面。面对这一全新法律挑战，传统法律体系暴露出明显短板。在过往立法框架中，对于人工智能产品自主行为导致的违法犯罪情形，从责任主体界定、归责原则到惩处机制，均存在大片空白地带。鉴于此，法学界在近年内敏锐捕捉到这一前沿问题，积极组织研讨交流，展开了多维度的初步探讨。人工智能发展带来的新情况和新问题，势必会对传统法律规定形成冲击，仅通过立法并不能彻底解决人工智能产品涉嫌犯罪的问题，应在人工智能发展不同阶段，通过法律解释和立法等不同规制方式应对[1]。而在其中最重要的一点就是研究人工智能是否能成为独立的犯罪主体，并承担相应的法律责任。研究人工智能犯罪不仅是现实的需要，解决当今社会人工智能犯罪问题，弥补立法的空白，更重要的是合理预测未来人工智能的发展，高瞻远瞩，未雨绸缪。总而言之，研究人工智能犯罪问题有其必要性和合理性。

国内外学者大都从刑法角度对人工智能犯罪问题进行规制，且重点集中在规制、犯罪风险或集中某一有关社会问题进行研究。本文主要采用文献研究方法（通过研究已有的法律文献、案例、法规等法学资料，来获取、整理、分析和评价相关的法学知识和信息）与实证分析方法（通过实证研究法律现象，探寻其中的规律性和因果关系，以提高对法律问题的认识和理解），以求对人工智能犯罪的主体问题进行研究。

1 从民法角度进行责任主体划分

触犯民法的主体是民事法律关系中享受权利，承担义务的当事人和参与者，包括自然人、法人和其他组织，即民法规制的是以人为主体的违法行为。在民法范围中，人工智能问题能依靠现有的法律法规得到比较妥善的规制，因为违法犯罪主体是人，不涉及人工智能自主意识的问题。

2 从刑法角度进行责任主体划分

2.1 否定人工智能刑事责任的主体地位

（1）人工智能不具备主观意识与自由意志，不具备人格性。人工智能体不能理解其行为的实质意义及违反法律的消极后果，无法具备犯罪故意和犯罪过失，其行为本质上不可归于人工智能体本身。人工智能没有对法律规范应有的敬重和遵守的意识，无法主动依照法律规范来行动以躲避相应的处罚。且人工智能无法做出恰当的价值判断，其认识能力和控制能力是算法所赋予的，在人工智能眼中只有数据，它的选择只是基于算法和数据的反馈，这种以程序为核心的“智能”自始至终都与人类的自由意志存在本质差异，与自然人和单位都存在无法跨越的差距，因此人工智能不能成为犯罪主体。

（2）如果承认人工智能成为刑事责任主体，将对传统刑法体系产生冲击，且人工智能不符合我国刑法对犯罪主体的规定。我国刑法规定的犯罪主体有两类，分别是自然人和单位。人工智能的行为是基于程序和算法，其既不归属于自然人也不归属于单位，在本质上与现有刑事责任主体存在巨大差别。若将人工智能拟制为新的刑事责任主体，就要对现行刑法的理论进行修改，将会造成现有刑法理论陷入困境，人工智能的主体地位可以拟制而来，但是其主观心态却不能因此而变得可知，致

使人工智能体刑事责任的认定被迫片面化为只注重行为客观方面的客观责任，还会带来共犯认定和未完成形态界定困难等问题。

（3）对人工智能体施加刑罚不具有刑法意义。刑法处罚的意义在于通过对犯罪行为的惩罚，给与犯罪分子警示作用维护社会秩序，预防和减少犯罪。人工智能在传统刑罚种类的适用上呈现出明显的局限性。我国的刑罚体系主要由主刑和附加刑共同构成，其中又能进一步精准地划分成权利刑、财产刑、自由刑和生命刑等不同类别。这四类刑罚都不能对人工智能犯罪进行规制：其一，人工智能不具备生命体征，这就决定了在对其实施危害行为进行惩处时，自由刑和生命刑这两类刑罚根本无法适用；其二，在法律还没有将政治权利和财产权赋予智能机器人的情况下，从逻辑和法理上讲，权利刑和财产刑自然也没有被施加于它的可能性。这就使得在处理人工智能犯罪这一全新的法律难题时，传统刑罚体系陷入了颇为棘手的困境，亟需我们从全新的角度去探索和构建适合人工智能的刑罚规范和处置方式，以确保法律在人工智能时代能够保持其应有的公正性和有效性。

2.2 人工智能犯罪责任主体的划分

2.2.1 从设计研发者角度

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。但在法律实践中，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或行业规范逃避过失犯罪的指控。若设计者明确设计人工智能来进行犯罪，则要追究设计者的故意犯罪责任。若由于设计缺陷使得人工智能产生了严重的损害后果，且该设计缺陷并非是处在现有技术条件下，属于完全无法被发现的那种隐蔽漏洞之中，在这种情况下，就应判定为是设计者的过失责任。从人工智能的设计者的立场出发，他们需要尽一切可能去排查和消除设计缺陷，只有这样，才能切实保障人工智能在实际运行过程中能够始终保持安全、可靠的状态，避免因为设计缺陷而引发一系列本可避免的危害和损失。但基于人工智能体的工具性，以及主客观相一致原则，同时也为了不阻碍人工智能相关科技的发展[3]，不能将设计者所有设计具有犯罪功能的人工智能体的研发行为都认定为犯罪，应当基于设计者的设计意图予以具体分析、分别认定。

若设计者在研发人工智能体时，主观上以实施犯罪行为，或为向他人提供犯罪工具为明确目的，且在研发筹备阶段，便积极与生产商沟通定制特殊规格、性能以契合犯罪用途的技术细节，同时在研发过程中及完成后，频繁与犯罪团伙等进行关于出售、使用该人工智能体的深度联络，那么其设计时的主观心态显然缺乏正当性。在此情形下，依据刑法中主客观相统一的归责原则，设计者应当对其设计的人工智能体所造成的各类社会危害承担刑事责任。反之，若设计者进行研发纯粹是为了科学研究，旨在推动人工智能技术在合法、有益领域的边界拓展，或是为预防诸如特定行业安全事故、公共卫生风险等相应损害，且在研发成果出现之前的规划阶段，以及成果诞生后的后续考量中，均未产生将其应用于不当领域的意图。同时，在技术实现过程中，通过设置多层加密防护系统、严格限制访问权限、部署实时监控预警机制等妥善的安全防护措施，将人工智能体谨慎管控于符合安全标准的环境内，那么即便有人通过窃取或其他非法手段获取该人工智能体并实施犯罪行为，追究设计者的刑事责任亦属不当。因为设计者已在其能力范围内尽到了合理且充分的注意与防范义务，不应为他人的违法犯罪行为承担无端责任。

若设计者在整个研发过程中，完全未采取任何措施防止人工智能体被用于犯罪目的，任由其外泄风险存在，那么设计者极有可能构成不作为的间接故意犯罪或过失犯罪。这是由于设计者明知人工智能体若落入不法分子之手可能引发严重危害后果，却消极地不履行应尽的防范义务。举一案例，当地时间 2018 年 3 月 18 日晚上 10 点左右，美国亚利桑那州一名推着自行车横穿道路的女子被优步自动驾驶汽车撞伤，被送往医院后不治身亡。当时优步测试车辆处于自动驾驶模式，根据美国国家运输安全委员会调查结果显示，事故发生时此自动驾驶系统存在超速和识别错误等问题。当时车辆时速为 62km/h，而道路限速 50km/h。而在撞到行人前的 5~6 秒时，车辆的自动驾驶解决方案已经检测到行人，但从未准确地将她归类为“人”。几秒钟之内，自动驾驶系统把它识别为未知物体，然后识别为车辆，归类为“其他”，却一直没有识别为“人”并自动刹车，最终导致事故发生。在这起案件中，生产研发者对于自动驾驶系统具有支配力，危险也是由于人工智能算法错误造成，设计者可能因为其过失行为承担刑事责任。

2.2.2 从生产者角度

生产者如果明知是用于犯罪或具有重大缺陷的人工智能，仍然生产出来，就需要承担故意犯罪的刑事责任[4]。在人工智能的应用场景中，无论是使用方还是所有权方，一旦蓄意利用人工智能系统中存在的漏洞，或是将其作为达成非法目的的工具，这种主观上具有明确恶意和犯罪意图的行为，无疑构成了故意犯罪。另一方面，若使用方或所有权方仅仅是因为疏忽大意，未能切实履行自身所肩负的监管义务，进而导致损害结果的发生，那么基于其在监管上的过失，应当承担与之相应的刑事责任。

2.2.3 从使用者角度

使用者利用人工智能进行犯罪，毫无疑问追究其法律责任。举国外案例，加拿大一对夫妇因接到“儿子”来电被诈骗，爱子心切的他们为帮助“儿子”迅速筹措资金汇款，直到儿子班杰明当天晚上打电话过来，他们才惊觉被骗，而电话中熟悉的“儿子”的声音竟是诈骗集团根据班杰明在网上发布的视频通过 AI 仿真而成。这起案件中人工智能犯罪的发生是基于使用者的行为，由使用者承担全部刑事责任。

2.2.4 共同犯罪角度

有学者认为，人工智能犯罪刑事责任的追究和承担，应当与行为人对法益侵害发生的支配、控制力相联系，即支配力的有无决定着应否追究行为人的刑事责任，支配力的大小决定着行为人所应承担刑事责任的轻重。基于刑事责任与法益侵害支配力紧密关联的原则，在面对内部结构较为复杂的人工智能犯罪情形时，务必要将法益侵害支配力这一核心要点作为开展责任追究工作的重中之重。在涉及人工智能犯罪的特定情境里，仅仅由于产品操作不当而引发犯罪，其根源既可能是使用者的操作不当，也有可能是算法在决策过程中出现了失误。而算法作为弱人工智能的核心部分，其编译、训练以及学习等各个环节，均完全处于研发生产者的掌控范围之内。因此，当面临涉及人工智能的犯罪情形时，责任认定需区别对待：当犯罪是因使用者的行为而产生，那么使用者理应对此负起全部的刑事责任；要是犯罪是由算法的决策失误所导致，研发生产者就应当承担全部的刑事责任；而当犯罪是使用者的行为与算法的决策失误共同作用的结果时，使用者和研发生产者需共同承担相应的刑事责任。

3 结语

在现实实践与理论研究的双重维度下，人工智能体蕴含的潜在犯罪风险，如今已成为人工智能发展进程中难以逾越的重大障碍。在现阶段，人工智能技术的运行模式决定了其仍需人的深度操控，无论是基础算法的编写、数据的标注与训练，还是实际应用场景中的指令下达与参数调整，均离不开人的参与，因此犯罪的责任主体无疑是人。但随着科技以指数级速度迅猛进步，从理论与技术发展趋势来看，人工智能在未来极有可能实现高度自主的独立行动。届时，其决策过程不再单纯依赖预设程序与人的实时干预，而是基于自身强大的学习与推理能力，在复杂环境中自主做出关键行动决策。当这种情况出现时，它便具备了独立的主体地位，应承担相应的法律责任。法律作为社会秩序的维护者与保障者，除了聚焦于解决当下的现实问题，更要秉持前瞻性与预防性思维，对未来可能发生的事作出合理的预测。通过开展法律前瞻性研究，模拟不同人工智能发展路径下可能衍生的法律风险场景，提前制定针对性的法律规范与应对预案，构建起严密的风险防范体系，从源头上遏制风险萌芽，防止风险的扩大化，确保人工智能技术在法治轨道上稳健前行，实现技术发展与社会安全稳定的良性互动。

参考文献

[1]杨仕兵,张闯玉.涉人工智能产品犯罪的刑法规制研究[J].东北农业大学学报：社会科学版,2019.
[2]张建军,毕旭君.论强人工智能体的刑事责任主体资格[J].贵州社会科学,2023(4):94-99.
[3]张磊,梁田.涉弱人工智能犯罪刑事责任问题研究[J].警学研究,2021,000(004):P.71-83.
[4]王充,董璞玉.人工智能时代刑事责任主体之再审视[J].广西社会科学,2020.

作者简介：杨钧雅（2004 年 8 月），女，汉族，浙江龙游，本科，兰州理工大学。
杨飞霞（2002 年 6 月），女，汉族，甘肃陇南，本科，兰州理工大学。