人工智能对传统刑法因果关系理论的挑战与重构

张瑞杰

山东科技大学　山东青岛

【摘要】人工智能技术的快速应用使刑法因果关系判断面临全新困境，传统理论以“人的主观罪过支配下的实行行为”为核心构建的因果关系与结果归属体系，在智能算法自主决策、多主体介入的场景中出现适用障碍。本文剖析人工智能对传统理论中实行行为定型性、因果流程认定、结果归属规则的三重挑战，依托我国刑法立法与司法实践，提出以重构路径，明确人工智能场景下因果关系判断的层级规则与责任归属边界，为司法实践提供理论参考。

【关键词】人工智能；刑法因果关系；危险现实化；结果归属；实行行为

【收稿日期】2026 年 2 月 18 日　　【出刊日期】2026 年 3 月 31 日　【DOI】10.12208/j.ssr.20260116

The challenge and reconstruction of traditional criminal law causal relationship theory by Artificial Intelligence

Ruijie Zhang

Shandong University of Science and Technology, Qingdao, Shandong

【**Abstract**】The rapid application of artificial intelligence technology has presented new challenges to the judgment of criminal law causal relationship. The causal relationship and result attribution system constructed by traditional theories based on "the act performed under the subjective fault of the person" has encountered application obstacles in scenarios where intelligent algorithms make autonomous decisions and multiple entities are involved. This paper analyzes the threefold challenges of artificial intelligence to the traditional theory regarding the stereotypical nature of the act, the determination of causal process, and the rules of result attribution. Relying on China's criminal law legislation and judicial practice, it proposes a reconstruction approach to clarify the hierarchical rules and responsibility attribution boundaries for causal relationship judgment in artificial intelligence scenarios, providing theoretical references for judicial practice.

【**Keywords**】Artificial intelligence; Criminal law causal relationship; Danger realization; Result attribution; Act of execution

引言：刑法因果关系是认定犯罪成立与刑事责任承担的核心要件，我国传统刑法理论将其界定为“危害行为与危害结果之间引起与被引起的客观联系”，并围绕必然因果关系与偶然因果关系展开长期争论。张明楷教授在《刑法学》（第七版）中提出的因果关系二分理论，将因果关系判断区分为事实层面的条件关系与规范层面的危险现实化判断，成为当前我国刑法理论与司法实践的重要参考。随着人工智能技术在自动驾驶、智能算法决策、工业自动化等领域的普及，智能系统的自主决策行为打破了“人作为唯一行为主体”的传统刑法框架，算法介入下的因果流程呈现出多环节、隐蔽性、不确定性特征，传统理论中实行行为的认定标准、因果关系的判断逻辑、结果归属的规范规则均面临前所未有的挑战。

1 传统刑法因果关系理论的现实困境

1.1 实行行为定型性消解，传统行为主体范畴受限

传统刑法因果关系的判断以“人的实行行为”为起点，张明楷教授强调实行行为必须具有“类型化的法益侵害危险”，且该行为由人的主观罪过支配，这一特征成为区分刑法因果关系与事实因果关系的核心标准。我国《刑法》第十四条、第十五条明确规定，故意与过失犯罪均以“人的行为”为前提，自然力或机器的单纯运作因缺乏主观罪过，无法成为刑法意义上的实行行为。但人工智能场景下，智能算法具备一定的自主决策能力，其行为并非完全由人直接控制，而是通过算法模型自主分析、判断并作出反应，形成了“人-算法-危害

结果”的新型因果链条。例如，自动驾驶车辆的算法系统在遇到突发状况时，会自主选择制动、避让或继续行驶，该决策行为并非由驾驶者直接操控，却可能直接导致交通事故的发生。

1.2 因果流程高度复杂，事实层面条件关系判断受阻

传统理论中事实因果关系的判断以条件说为基础，即“若无实行行为，则无该具体侵害结果”，该规则在单一主体、单一行为的场景中具有较强的操作性。但人工智能场景下的因果流程呈现出多主体、多环节、多因素交织的特征，算法的自主决策、数据输入的偏差、算法模型的漏洞、多个主体的不同行为均可能成为危害结果发生的原因，形成复杂的因果网络。根据张明楷教授对“重叠的因果关系”“二重的因果关系”的界定，多个相互独立的行为若共同导致危害结果发生，需分别肯定各自行为与结果的因果关系，但人工智能场景下的因果关系更为复杂，部分因素并非“人的行为”，而是算法的自主运作或技术漏洞。例如，智能算法因数据输入错误与模型漏洞共同导致金融诈骗结果发生，数据提供者、算法设计者、算法使用者的行为与危害结果之间均存在一定的条件关系，且算法的自主运作成为重要的介入因素，此时依据条件说难以准确筛选出与危害结果具有事实因果关系的行为，甚至可能出现归责泛化的问题。

1.3 介入因素形态异化，规范层面危险现实化判断失灵

结果归属的核心是判断“行为制造的法益侵害危险是否现实化为实害结果”，张明楷教授提出，在介入因素场合，需综合考量实行行为的危险性、介入因素的异常性、介入因素的作用大小及是否属于行为人管辖范围。传统理论中的介入因素多为被害人行为、第三者行为或行为人自身行为，均具有可预测性和可评价性，但人工智能场景下的介入因素呈现出异化特征，智能算法的自主决策成为新型介入因素，其具有不可预测性、技术性和独立性，难以依据传统规则判断其异常性与作用大小。例如，驾驶者在使用自动驾驶功能时，算法系统突然出现自主决策偏差，导致车辆撞击行人，此时算法的自主决策属于介入因素，但其是否具有“异常性”，能否中断驾驶者或算法设计者的行为与危害结果之间的危险现实化进程，传统理论均无法作出合理判断。同时，根据传统理论，若介入因素属于“第三者专属责任领域”，则否定前行为的结果归属，但算法的自主决策既非传统意义上的“第三者行为”，也非行为人可控制的范围，导致责任领域的划分标准失效。

2 人工智能场景下刑法因果关系的规则构建

2.1 坚守二分理论框架，明确双层判断逻辑

人工智能场景下的刑法因果关系重构，需以张明楷教授提出的“事实因果与规范归属二分”理论为基础，坚守构成要件符合性与实行行为概念的核心地位，构建“事实层面的条件关系判断-规范层面的算法危险现实化判断”的双层逻辑体系，这一框架与我国刑法主客观相统一的基本原则相契合。在事实层面，仍以条件说为基本判断标准，即“若无相关主体的行为，则无算法介入下的危害结果”，但需对条件说的适用范围进行拓展，将“与算法运作相关的人的行为”均纳入事实因果关系的判断范围，包括算法设计者的开发行为、开发者的优化行为、使用者的操作行为、数据提供者的输入行为等。在规范层面，摒弃传统的“必然/偶然因果关系”判断标准，以“算法危险现实化”为核心构建结果归属规则，判断各主体的行为是否制造了“算法运作的法益侵害危险”，且该危险是否通过算法的运作或决策现实化为实害结果，实现事实判断与规范判断的分层区分，防止归责问题前置化。

2.2 重塑实行行为标准，界定新型行为主体范畴

结合我国刑法对“危害行为”的界定，重塑人工智能场景下的实行行为认定标准，将实行行为界定为“主体通过对算法的设计、开发、使用、数据输入等行为，制造了类型化的法益侵害危险，且该行为对算法的自主决策具有支配或影响作用”，明确实行行为的核心是“对算法的支配力或影响力”，而非传统的“身体动静”。根据该标准，算法设计者、开发者、使用者等主体的行为若对算法的运作具有支配或影响作用，且制造了法益侵害危险，则构成刑法意义上的实行行为；单纯的算法自主决策因缺乏主观罪过，无法成为实行行为，算法本身也不能成为刑法意义上的行为主体，刑事责任的承担主体仍限定为“人”，这一界定符合我国《刑法》第三条罪刑法定原则的要求。同时，根据实行行为的定型性程度，对不同主体的行为进行类型化区分：算法设计者的开发行为、开发者的漏洞修复行为具有较高的定型性，若其行为存在故意或过失，且制造了算法运作的危险，则直接认定为实行行为；算法使用者的操作行为需结合其对算法的控制能力进行判断，若使用者对算法具有实际控制能力，其未按规定操作的行为则构成实行行为；数据提供者的输入行为若存在故意提供虚假数据的情形，且该数据成为算法决策的重要依据，则构成实行行为。

2.3 构建算法介入规则，完善危险现实化判断体系

针对算法自主决策这一新型介入因素，借鉴张明楷教授提出的介入因素判断规则，构建“算法介入下的危险现实化判断体系”，综合考量三个核心要素：一是实行行为的危险性程度，若主体的行为制造了高度的法益侵害危险，即使算法介入成为危害结果发生的直接原因，也应肯定危险的现实化；二是算法介入因素的异常性，若算法的自主决策是因主体的行为所引发或具有通常性，如算法设计者未修复已知漏洞导致算法决策偏差，则该介入因素不具有异常性，不中断危险的现实化进程；若算法的自主决策是因不可预测的技术故障所导致，且与主体的行为无关，则该介入因素具有异常性，否定前行为的危险现实化；三是算法介入因素的作用大小，若主体的行为制造的危险已经处于现实化的进程中，算法的介入仅起到加速作用，则肯定前行为的结果归属；若算法的介入成为危害结果发生的决定性原因，且与主体的行为无关，则否定前行为的结果归属。同时，明确算法介入因素的“管辖范围”判断标准，若主体对算法的运作具有监督、控制或修复义务，如算法开发者对算法漏洞具有修复义务，其未履行义务导致算法介入引发危害结果，则该介入因素属于主体的管辖范围，应肯定结果归属；若算法的运作属于他人专属管辖范围，如网络服务提供者对算法的日常维护义务，则否定前行为的结果归属。

3 结语

人工智能技术的发展打破了传统刑法因果关系理论的适用框架，但其并未否定刑法因果关系的核心价值，而是对理论的适应性提出了更高要求。刑法因果关系理论的重构，并非对传统理论的全盘否定，而是结合人工智能场景的特征进行的理论拓展与完善。通过坚守“事实因果与规范归属二分”的判断逻辑，重塑实行行为认定标准，构建算法介入下的危险现实化判断体系，细化结果归属规则，能够有效解决人工智能场景下因果关系认定的困境，实现对法益的有效保护与罪刑相适应原则的落实。

参考文献

[1] 余鑫扬. 刑法因果关系判断的理论反思：基于“复合条件说导向的二阶模式”[J].中南法律评论,2025,(00):150-165.

[2] 陈璐. 过失竞合理论下人工智能致损的刑事责任分配[J].郑州大学学报(哲学社会科学版),2025,58(06):73-79+171.

[3] 刘艳红. 刑法因果关系理论的横断切面与危险的现实说之确立[J].政治与法律,2024,(08):64-82.

[4] 石经海,赵春雨.刑法治理的现代化与本土化讲演录[M].中国社会科学出版社:202311:567.

[5] 宋振武. 作为冗余理论的刑法学因果关系学说[J].烟台大学学报(哲学社会科学版),2023,36(05):106-113.

[6] 蒋太珂. 因果力比较规则的刑法理论构造[J].法学研究,2023,45(01):108-124.

[7] 李阜蒙,储陈城. 刑法因果关系的立场选择和理论纠偏[J].哈尔滨师范大学社会科学学报,2023,14(01):75-79.

[8] 钱叶六. 刑法因果关系理论的重要发展与立场选择[J].中国刑事法杂志,2022,(04):95-111.

[9] 车浩,于改之.因果关系的理论与实务[M].北京大学出版社:202206:134.