경북대학교 법학연구원『법학논고』
제94집 (2026.07) 429~459면.
Kyungpook Natl. Univ. Law Journal
Vol.94 (July 2026) pp.429~459.
http://doi.org/10.17248/knulaw..94.202607.429

AI기본법상 AI사업자 의무와 형사법상 조직책임*

- 책임구조 · 의무층위 · 피해자구제를 중심으로 -

안 정 빈**

〈국문초록〉

인공지능기본법은 인공지능 자체를 형사책임의 주체로 승인하지 않고, AI를 개발 · 제공 · 이용하는 인공지능사업자(이하 'AI사업자')에게 투명성 · 고지 · 위험관리 · 안전성 확보 등 행정법상 의무를 부과하는 방식을 선택하였다. 그러나 이러한 선택은 형사법과 무관한 단순한 행정규제에 그치지 않는다. 고영향 인공지능의 확인 · 관리, 생성형 인공지능 산출물 표시, 딥페이크 등 현실과 구별하기 어려운 AI 생성물의 표시 · 고지, 안전성 · 신뢰성 확보, 설명 방안의 수립 · 시행 및 관련 문서의 작성 · 보관과 같은 의무는 향후 AI 관련 사고에서 주의의무, 감독의무, 보증인지위 및 법인의 조직책임을 판단하는 외부적 준거규범으로 기능할 수 있다. 본고는 AI기본법의 책임구조에 관한 논의와 AI 형사책임론의 문제의식을 하나의 축으로 결합하여, AI 자체를 처벌할 수 있는가라는 추상적 질문에서 벗어나 AI 시스템을 개발 · 배포 · 운영하는 조직의 위험관리 실패를 어떻게 형사법적으로 평가할 것인가를 검토한다.

이를 위해 첫째, AI기본법의 입법적 선택을 형사책임 · 행정제재 · 민사적 피해구제라는 세 층위에서 정리하고, 둘째, AI사업자 의무를 형식이행형 의무(formal-compliance obligation)와 위험관리형 주의의무(risk-based duty of care)로 구분하여 책임귀속의 실천적 기준을 제시하며, 셋째, 법인 형사책임론에서 발전한 동일시이론과 조직모델이 AI사업자 조직책임에 갖는 참조 가능성과 현행법상 한계를 분석한다. 다만 조직모델은 법인 자체의 일반적 형사책임을 창설하지 않으며, 법인의 처벌에는 별도의 명시적 근거가 필요하다. 결론적으로 AI 형사법의 과제는 AI를 사람처럼 처벌하는 데 있지 않다. 핵심은 AI 시스템을 사회에 배포하고 운영하는 조직이 위험을 예견하고 통제할 수 있었는지, 그 통제의무를 위반하여 법익침해가 현실화되었는지, 그리고 피해자가 정보비대칭 속에서도 실질적으로 구제받을 수 있는지를 묻는 데 있다.

주제어 : 인공지능기본법, 과실범, 보증인지위, 조직책임, 의무의 강약설

• 투고일 : 2026.06.20. / 심사일 : 2026.07.21. / 게재확정일 : 2026.07.25.

---

* 이 글은 필자가 2025년 11월 8일 대만 타이베이에서 개최된 The 8th International Conference on AI and Law 2025 학술대회 제1세션에서 "Korea's AI Basic Act and Its Impact on Industry"라는 제목으로 발표한 내용을 바탕으로, AI사업자 의무의 형사법적 의미와 조직책임 문제를 확장 · 재구성한 것이다.

** 경남대학교 법학과 부교수.

Ⅰ. 서론

1. 문제의 소재

인공지능에 관한 형사법 논의는 오랫동안 “AI에게 형사책임을 물을 수 있는가”라는 질문을 중심으로 전개되어 왔다. 그러나 AI의 형법상 주체성에 관해서는 현 단계에서 행위능력·책임능력 및 형벌수용 가능성을 부정하거나 유보하는 견해와, 장래 독자적 법인격 또는 형법주체성의 부여 가능성을 검토하는 견해가 병존한다. 본고는 그 장래의 가능성 자체보다 현행법상 AI를 개발·제공·이용하는 사람과 조직의 통제책임에 초점을 둔다.[1)]

「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 ‘AI기본법’)은 2025년 1월 21일 공포되어 2026년 1월 22일 시행되었고, 일부개정을 거쳐 현재는 2026년 7월 21일 시행 법률 제21311호가 적용된다. 이 법은 AI 자체를 처벌 대상으로 삼지 않고, AI를 개발·제공·이용하는 인공지능사업자(이하 ‘AI사업자’)에게 투명성, 고지, 위험관리, 안전성 확보 등 행정법상 의무를 부과하는 방식을 택한다. 이 선택은 AI를 의제적 행위자로 구성하기보다, AI 시스템을 개발·제공·이용하는 조직과 사람에게 규제의 초점을 맞춘다는 점에서 현실적이다.[2)]

한편 한국 AI기본법은 유럽연합 인공지능법에 이어 세계 두 번째로 제정된 포괄적 인공지능 법제로 소개되어 왔다.[3)] 필자도 2025년 11월 대만 국제학술대회에서 같은 취지로 그 의의를 설명한 바 있고,[4)] 이 세션에서는 한국·미

---

1) 현 단계의 AI에 형법상 책임능력과 형벌수용 가능성을 인정하기 어렵다는 견해로 임석순, 「형법상 인공지능의 책임귀속」, 「형사정책연구」 제27권 제4호, 한국형사정책연구원, 2016, 73면, 77-78면 참조. 전통적 형법이론에 대한 도전과 장래의 형법주체성 문제를 제기하는 논의로 김성돈, 「전통적 형법이론에 대한 인공지능 기술의 도전」, 「형사법연구」 제30권 제2호, 한국형사법학회, 2018, 81-116면; 독자적 법인격 부여의 가능성과 필요성을 적극 검토하는 견해로 강지현, 「4차 산업혁명시대, 법학의 역할 – 독자적 법인격 주체로서 인공지능 –」, 「한국부패학회보」 제27권 제1호, 한국부패학회, 2022, 59-71면 참조. 필자의 선행연구로 안정빈, 「인공지능 로봇에 관한 형사책임 – 인공지능형법에 의한 형사처벌가능성을 중심으로 –」, 「외법논집」 제46권 제4호, 한국외국어대학교 법학연구소, 2022, 63-86면 참조. 위 문헌들의 결론이 동일한 것은 아니다.

2) 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 [시행 2026. 7. 21.] [법률 제21311호, 2026. 1. 20., 일부개정]. 이하 특별한 언급이 없는 한 현행법을 기준으로 한다.

3) 한국 AI기본법을 유럽연합에 이은 두 번째 포괄적 AI 법제로 평가한 견해로 정찬모, 「인공지능기본법 주요 내용의 검토와 향후 과제」, 「법학연구」 제28집 제1호, 인하대학교 법학연구소, 2025, 73면 참조.

4) 위 학술대회의 제1세션은 鄧振中·施立成이 공동좌장을 맡고, Nathaniel Persily, Florence

국 · 프랑스 · 일본 · 싱가포르 및 대만 여섯 나라의 시각에서 각국 AI 법제와 거버넌스가 비교 · 검토되었다.[5]

대만의 현직 입법자와 법무부장(관)까지 참석한 이 국제 컨퍼런스가 개최된 직후에 대만의 「人工智慧基本法」이 2025년 12월 23일 입법원에서 통과되고 2026년 1월 14일 공포 · 시행되어, 한국 AI기본법의 최초 시행일인 2026년 1월 22일보다 8일 앞서 시행되었다.[6] 대만법은 한국법보다 늦게 제정되었으나 시행유예기간을 두지 않아 시행일이 앞서게 된 것이다.

그럼에도 불구하고, 더 중요한 차이는 규범의 밀도라고 생각한다. 대만 「人工智慧基本法」은 전문 20개조의 원칙법으로서 민간사업자에 대한 구체적 운영의무를 상세히 규정하기보다 관계 행정기관의 후속 부문별 규율에 상당 부분을 맡기고 있는 반면, 한국 AI기본법은 인공지능사업자에게 투명성 · 안전성 · 고영향 인공지능 관련 책무를 직접 부과한다.[7] 본고는 이러한 규범밀도의 차이가 한국 AI기본법상 사업자 의무를 형사상 주의의무와 감독의무를 구체화하는 외부적 판단자료로 활용할 가능성을 높인다고 본다.

그러나 AI기본법이 AI사업자의 의무위반을 범죄로 구성하지 않았다고 해서 형사법적 의미가 없는 것은 아니다. 형사법은 구성요건 문언만으로 주의의무와 보증인지위를 언제나 완결적으로 정하지 않는다. 예컨대 산업안전과 제품안전 영역에서는 법령상 안전조치 · 검사의무가 업무상 주의의무의 내용과 결합하여

G′sell, Gen Goto(後藤元), Weitseng Chen(陳維曾), 張麗卿 교수 및 필자가 발표하였다. 이 학술대회는 財團法人人工智慧法律國際研究基金會가 주최하고 實踐大學 · 東海大學이 공동주최하였다. 공식 일정은 「2025 第八屆人工智慧與法律國際學術研討會 － AI時代下的産業鏈重構與法治政策」 프로그램, https://law.thu.edu.tw/upload/law/news_upload/790_議程_20251108第八屆國際學術研討會1031.pdf (최종접속 2026. 7. 26.) 참조.

5) 項家麟, 「第八屆人工智慧與法律國際學術研討會 六國專家剖析AI法治與産業布局」, 「經濟日報」, 2025. 11. 11., https://money.udn.com/money/story/11799/9131073 (최종접속 2026. 7. 26.) 등; 法務部司法官學院犯罪防治研究中心, 「探索AI應用的刑事風險與法律議題, 本中心攜手産官學界共同舉辦「2025第八屆人工智慧與法律國際學術研討會」」, 2025. 11. 17., https://www.cprc.moj.gov.tw/1563/36715/1627/45130/post (최종접속 2026. 7. 26.) 참조. 위 보도는 참가자들이 AI 규제의 필요성 자체에는 인식을 같이하였으나 규제의 방식과 강도는 국가마다 상이하였다고 전한다. 이는 후술하는 규범밀도의 문제와 연결된다.

6) 대만 「人工智慧基本法」은 2025년 12월 23일 입법원에서 3독 통과되고, 2026년 1월 14일 총통 명령(總統令, 華總一義字第11500001671號)으로 공포되어 제20조에 따라 같은 날 시행되었다. 立法院議事暨公報資訊網, 「人工智慧基本法草案」 의안심의정보, https://ppg.ly.gov.tw/ppg/bills/202110165050000/details (최종접속 2026. 7. 26.); 中華民國總統府, 「制定人工智慧基本法」, 「總統府公報」 제7837호, 2026. 1. 14., https://www.president.gov.tw/Page/294/50131 (최종접속 2026. 7. 26.) 참조.

7) 본문의 규범밀도 비교는 한국 AI기본법 제31조부터 제36조까지와 대만 「人工智慧基本法」 전문 20개조의 조문구조를 대조한 본고에서의 평가이다.

평가되어 왔다.[8] 본고는 AI기본법상 사업자 의무도 그 자체가 형벌구성요건은 아니지만, 구체적 사건에서 주의의무와 보증인지위를 판단하는 보조적 외부규범으로 작동할 수 있다고 본다.[9]

따라서 본고에서의 질문은 AI를 형사책임의 독립 주체로 인정할 것인가가 아니다. 본고는 AI기본법상 사업자 의무가 AI 관련 사고와 범죄에서 형사법상 책임귀속 구조를 어떻게 변화시키는가를 묻는다. 특히 AI사업자 의무가 과실범의 주의의무, 부작위범의 보증인지위, 법인의 조직책임, 인간-AI 협업범죄의 책임귀속, 피해자 구제의 구조에서 어떠한 의미를 갖는지 검토한다.

이 점에서 현행 AI기본법 체계에서 중요한 것은 AI사업자의 의무위반을 곧바로 독자적 형사범죄로 구성할 것인가의 문제만이 아니다. 오히려 현행법상 그 의무위반이 주로 행정법적 규제의 대상에 머무른다고 하더라도, 그러한 의무위반이 AI 시스템 이용 과정에서 발생한 범죄에 대하여 AI사업자의 과실범 성립 여부, 부작위범의 보증인지위, 고의 방조책임, 조직책임 판단에 어떠한 규범적 의미를 갖는지가 핵심 쟁점이 된다.

2. 선행연구와 본고의 방향

설명의 편의를 위하여 본고는 종래의 AI 형사책임 논의를 세 방향으로 재분류한다. 첫째, AI 자체의 행위주체성 또는 책임능력 인정 가능성을 논하는 접근이다. 둘째, AI가 야기한 결과를 개발자, 제조자, 이용자 등 자연인에게 귀속시키는 접근이다. 셋째, 형사책임보다는 행정규제나 민사책임, 보험제도, 제품

8) 대법원은 산업재해 사망 사안에서 중대재해처벌법위반(산업재해치사)죄, 산업안전보건법위반죄 및 업무상과실치사죄를 상상적 경합 관계로 판단하였다(대법원 2023. 12. 28. 선고 2023도12316 판결). 직접적 판시사항은 세 죄의 죄수관계이지만, 법령상 안전조치의무와 업무상 주의의무가 동일한 행위사실을 매개로 결합할 수 있음을 보여준다. 이에 관하여 강현석, 「중대재해처벌법위반죄와 산업안전보건법위반죄의 죄수관계에 관한 연구 － 대법원 2023. 12. 28. 선고 2023도12316 판결을 중심으로 －」, 「법이론실무연구」 제12권 제2호, 한국법이론실무학회, 2024, 11-38면 참조. 가습기살균제 사건에서는 주원료에 관한 안전성검사 미실시가 업무상 주의의무 위반으로 문제되었으나, 대법원은 서로 다른 원료제품의 제조·유통 관계자들 사이에 공동의 주의의무 위반에 관한 인식과 의사연락이 인정되지 않는다는 이유로 과실범의 공동정범 성립을 부정하고 원심을 파기환송하였다(대법원 2024. 12. 26. 선고 2024도1856 판결). 두 판결은 행정법상 의무위반이 형사상 과실로 자동 전환된다는 뜻이 아니라고 보고 있다고 판단된다.

9) 이 문장은 판례가 AI기본법에 관하여 직접 판시한 내용을 옮긴 것이 아니라, 앞의 판례와 AI기본법상 의무구조를 토대로 한 본고에서의 해석이다. AI기본법은 고영향 인공지능 해당 여부의 사전검토, 투명성·표시·고지, 위험관리, 사람에 의한 관리·감독, 문서 작성·보관 등 위험관리의 법정 기준을 제시한다.

책임을 통하여 피해자 구제와 위험관리를 도모하려는 접근이다.[10]

본고는 위 세 방향 중 둘째와 셋째의 접점에 선다. AI 자체의 형사책임을 부정하거나 유보하는 결론은 이미 상당 부분 공유되고 있다. 문제는 그 다음이다. AI를 처벌하지 않는다면, 형사법은 AI 시스템을 둘러싼 조직적 위험창출을 어떻게 다루어야 하는가. AI사업자가 법령상 위험관리의무를 부담한다면, 그 의무 위반은 과실범에서 어떤 의미를 갖는가. 법인 내부의 의사결정이 알고리즘 설계, 데이터 선택, 모델 배포, 사후 모니터링으로 분산되어 있을 때, 기존의 동일시이론이나 자연인 중심 책임귀속 모델은 여전히 충분한가?

본고에서는 종래 분리되어 논의되어 온 두 흐름을 하나의 축으로 결합하고자 한다. 하나는 AI기본법의 입법구조와 사업자 의무의 층위를 분석하는 흐름이고, 다른 하나는 AI 형사책임론과 법인 형사책임론에서 책임주체 및 귀속구조를 검토하는 흐름이다. 전자는 규제구조와 의무의 성질 분석에 강점이 있고, 후자는 고의·과실·보증인지위·조직책임의 이론에 강점이 있다. 본고는 두 논의를 단순히 병렬시키지 않고, AI사업자 의무가 형사법상 책임귀속의 판단기준으로 작동하는 과정을 중심축으로 재구성한다.[11]

이러한 재구성의 장점은 세 가지이다. 첫째, AI기본법을 단순한 산업진흥법 또는 행정규제법으로만 보지 않고 형사법상 주의의무의 외부규범으로 읽을 수 있다. 둘째, AI 형사책임론을 AI의 책임능력이라는 추상적 문제에서 AI사업자의 위험관리 실패라는 실천적 문제로 이동시킬 수 있다. 셋째, 피해자 구제를 형사처벌의 부수효과가 아니라 정보비대칭·기록보존·설명가능성의 문제와 연결할 수 있다.

---

10) 이 삼분법은 특정 선행문헌의 분류를 그대로 전재한 것이 아니라 설명의 편의를 위하여 본고에서 재구성한 것이다. AI 자체의 주체성에 관한 논의로 강지현, 앞의 논문, 59-71면; 자연인에 대한 책임귀속에 관한 논의로 임석순, 앞의 논문, 73면, 77-78면; 위험관리·규제 및 피해구제에 관한 논의로 정찬모, 앞의 논문, 83-87면, 97면 참조.

11) AI기본법의 입법구조와 사업자 의무에 관하여 박상철, 「인공지능 기본법의 시행 전 개정 필요성 – 규제 조항의 체계·축조상 문제점을 중심으로 –」, 「정보법학」 제28권 제3호, 한국정보법학회, 2024, 3-50면; 정찬모, 앞의 논문, 83-87면, 97면. AI 관련 형사책임과 귀속구조에 관하여 김성돈, 앞의 논문, 81-116면; 임석순, 앞의 논문, 73면, 77-78면; 신혜진, 「인공지능(AI) 시대의 신종 범죄에 대한 형사법적 대응 방향 – 형사책임의 주체 및 범위를 중심으로 –」, 「형사법의 신동향」 제87호, 대검찰청, 2025, 253-255면, 267면 참조. 두 흐름을 조직의 위험관리 실패라는 축으로 결합하는 것은 본고의 연구설계이다.

Ⅱ. AI기본법의 입법적 선택과 AI사업자 의무

1. 법이 선택한 방향: 형법이 아닌 행정법

AI기본법의 가장 중요한 특징은 형사처벌 중심의 규율을 선택하지 않았다는 점이다. 이 법은 AI 자체를 형벌의 대상으로 삼지 않고, AI를 개발·제공·이용하는 사업자에게 사전적·절차적 의무를 부과한다. 이는 AI 시스템이 야기할 수 있는 위험을 사후 처벌보다 사전 통제와 관리의 영역에서 다루려는 입법적 선택으로 이해할 수 있다. 물론 이 법에 벌칙조항이 전혀 없는 것은 아니다. 제42조는 제7조 제9항을 위반하여 직무상 알게 된 비밀을 타인에게 누설하거나 직무상 목적 외의 용도로 사용한 자를 3년 이하의 징역 또는 3천만 원 이하의 벌금에 처하도록 규정한다. 그러나 이는 국가인공지능전략위원회의 심의·의결 과정에 관여한 자의 비밀유지의무 위반을 처벌하는 조항일 뿐이고, AI사업자의 투명성·안전성·위험관리 의무 위반 그 자체를 범죄로 구성한 규정은 존재하지 않는다. 본고에서 'AI기본법이 형사처벌을 선택하지 않았다'고 할 때의 의미는 이러한 한도에서의 것이다.[12]

이 선택은 책임주의의 관점에서 일단 타당하다. 현 단계의 AI를 형법상 책임 있는 인격적 주체로 보기 어렵고, 자유형·응보·교화 등 형벌의 전통적 기능을 AI에 그대로 적용하기도 어렵다. 현재의 AI는 이러한 의미에서 형벌책임의 주체로 인정되기 어렵다.[13]

그러나 사업자 의무위반을 범죄화하지 않았다는 사실만으로 법이 책임 문제를 해결한 것은 아니다. AI기본법은 AI사업자의 위험관리의무 위반에 대한 형사처벌을 중심적 규율수단으로 삼지 않았을 뿐, AI 시스템의 위험을 누가 예견하고 통제해야 하는가라는 문제를 제거하지 않는다. 오히려 법이 AI사업자를 의무주체로 특정했다는 점은, 향후 AI 사고에서 사업자의 예견가능성·회피가능성·통제가능성을 판단할 때 중요한 규범적 출발점이 된다.

따라서 AI기본법이 사업자 의무위반을 범죄화하지 않은 것은 형사법의 후퇴를 의미하지 않는다. 그것은 AI사업자 의무위반을 곧바로 범죄화하지 않으면

12) 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제7조 제9항 및 제42조 참조. 현행법상 위원회 명칭은 '국가인공지능전략위원회'이다.

13) 현 단계 AI의 책임귀속을 AI 자체가 아니라 배후 인간에게 묻는 접근과 AI의 책임능력 인정 한계에 관하여 임석순, 앞의 논문, 73면, 77-78면; 전통적 형벌주체성에 대한 인공지능 기술의 도전에 관하여 김성돈, 앞의 논문, 81-116면 참조.

서도, 그 의무위반이 구체적 법익침해와 결합한 경우 형사상 주의의무 위반, 감독의무 위반, 방조책임 또는 조직책임 판단에서 어떠한 의미를 갖는지를 묻게 만드는 구조이다.

2. 책임 구조의 세 층위

본고는 AI기본법을 둘러싼 책임구조를 세 층위로 나누어 본다. 첫째는 행정규제상 책임이다. 이는 법령상 의무위반에 대하여 행정청이 부과하는 제재를 말한다. 과태료 대상은 사전고지의무 위반, 국내대리인 미지정, 중지명령 · 시정명령 불이행의 세 경우에 한정된다(제31조 제1항, 제36조 제1항, 제40조 제3항 및 제43조 제1항). 표시의무나 안전성 확보 의무 위반은 시정명령을 거치지 않으면 그 자체로 직접 제재되지 않는다. 둘째는 형사법상 책임이다. 행정의무 위반 자체가 곧바로 범죄가 되는 것이 아니라, 그 위반이 사망 · 상해 · 재산상 손해 · 명예훼손 · 성적 자기결정권 침해 등 개별 법익침해와 결합할 때 문제된다. 셋째는 민사적 피해구제이다. 이는 손해배상, 결함책임, 설명의무 위반, 정보제공의무 위반, 보험 또는 기금에 의한 손해분산을 포함한다.

이 세 층위는 서로 대체관계가 아니라 보완관계에 있다. 행정제재가 있다고 해서 형사책임이 당연히 배제되는 것은 아니고, 형사책임이 논의된다고 해서 피해자 구제가 자동으로 이루어지는 것도 아니다. 특히 과태료는 국가에 귀속되는 금전제재이지 피해자에게 직접 배상되는 돈이 아니다. 따라서 AI기본법 이후의 책임논의는 “처벌할 것인가 말 것인가”라는 이분법을 넘어서 행정규제 · 형사책임 · 민사구제를 어떻게 연결할 것인가를 물어야 한다.[14]

이러한 삼분 구조를 전제로 할 때, AI기본법은 사업자 의무에 관한 한 형벌법규가 아니지만 형사법적으로 무의미하지 않다. 법령상 의무는 형사법원이 사후적으로 주의의무의 내용을 판단할 때 참고할 수 있는 최소한의 행위기준을 제공한다. 반대로 형사법은 AI기본법상 의무 위반이 중대한 법익침해로 현실화된 경우, 단순한 행정위반을 넘어 책임비난 가능성이 있는지 심사한다.

---

14) 행정규제 · 형사책임 · 민사구제의 세 층위 구분은 본고의 분석틀이다. AI기본법의 행정규제 중심 구조, 투명성 · 고영향 인공지능 의무 및 권리구제 보완과제에 관하여 정찬모, 앞의 논문, 83-87면, 97면 참조. 과태료의 정확한 범위는 AI기본법 제43조 제1항에 따른다.

3. 혼합적 규율구조와 준연성규범

AI기본법은 강한 명령적 규범도 아니고 단순한 권고적 규범도 아니다. 산업 진흥과 신뢰 확보를 동시에 내세우면서 법률상 구속적 의무 · 행정제재와 노력의무 · 자율규제 · 가이드라인을 혼합한다. 본고는 이러한 혼합적 규율구조를 편의상 '준연성규범(semi-soft norm)'이라 부른다.

이러한 혼합적 구조는 장점과 한계를 동시에 갖는다. 장점은 빠르게 변화하는 AI 기술에 대하여 과도한 형사처벌을 앞세우지 않고, 사업자의 자율적 위험관리와 산업 발전을 병행할 수 있다는 점이다. 한계는 의무의 강도와 책임배분이 불명확할 경우, 실제 사고가 발생했을 때 피해자는 누구에게 어떤 근거로 책임을 물어야 하는지 알기 어렵다는 점이다.[15]

따라서 AI기본법의 해석론은 단순히 조문을 나열하는 데 그쳐서는 안 된다. 어느 의무가 외형적 이행 여부를 중심으로 판단되는지, 어느 의무가 위험의 예견 · 통제 · 회피 가능성에 따라 실질적 주의의무로 구체화되는지를 구분해야 한다. 이 구분이 본고가 말하는 형식이행형 의무와 위험관리형 주의의무의 출발점이다.[16]

4. AI기본법상 AI사업자 의무의 구조

(1) AI사업자의 유형과 책임 분산

AI기본법은 AI사업자를 하나의 단일한 주체로만 보지 않고, 개발 · 제공 · 이용의 단계에 따라 다양한 지위를 전제한다. 개발사업자는 모델의 설계, 학습데이터 선택, 알고리즘 구조, 사전 검증, 배포 기준에 관여한다. 이용사업자는 실제 서비스 환경에서 AI를 적용하고, 이용자에게 고지하며, 사람의 감독 절차를 운영한다. 플랫폼 또는 서비스 제공자는 모델과 이용자를 연결하면서 접근제

15) 정찬모 교수는 AI기본법이 완결된 규율이라기보다 토대와 얼개를 마련한 법이고, 구체적 규제수준은 시행령 · 가이드라인을 통하여 보충되어야 한다고 평가한다. 정찬모, 앞의 논문, 71면, 97면. 다만 '준연성규범'은 그 논문에서 사용한 용어가 아니라, 법률상 구속적 의무와 노력의무 · 자율규제 · 가이드라인이 혼합된 구조를 설명하기 위한 본고의 분석용어이다.

16) 본고의 '형식이행형 의무(formal-compliance obligation)'와 '위험관리형 주의의무(risk-based duty of care)' 구분은 일반적으로 확립된 법적 분류를 재현한 것이 아니라, AI기본법상 사업자 의무를 행정제재의 기준에 그치지 않고 형사상 주의의무 · 보증인지위 · 피해자 구제의 연결규범으로 읽기 위한 분석틀이다.

한, 이용조건, 오남용 방지, 사후 모니터링을 수행한다.

다만 이하에서 말하는 개발사업자 · 제공사업자 · 이용사업자는 AI기본법상 법정 개념을 엄밀히 재현한 것이 아니라, AI 시스템의 개발 · 제공 · 운영 단계별 책임분배를 설명하기 위한 기능적 구분이다.

이러한 다층 구조는 책임귀속을 어렵게 만든다. AI 사고가 발생했을 때 개발자가 학습데이터 편향을 방치했는지, 이용사업자가 부적절한 환경에 모델을 적용했는지, 플랫폼이 범죄적 이용을 반복적으로 방치했는지, 최종 이용자가 경고를 무시하고 범죄에 활용했는지를 구분해야 한다. 하나의 결과가 여러 단계의 작은 의무위반이 결합되어 발생할 수 있기 때문이다.

법인 조직책임론의 일반적 문제를 AI 개발 · 운영 구조에 적용하면, 특정 자연인의 고의 · 과실만으로 책임을 구성하기 어려운 상황이 더욱 뚜렷하게 나타난다. AI 시스템의 개발 · 운영은 데이터 수집팀, 모델 설계팀, 제품화 부서, 준법감시 부서, 영업부서, 외주업체, 클라우드 인프라 제공자 등이 결합하는 분산적 의사결정에 의존하기 때문이다. 이러한 경우 특정 자연인 한 명의 심리상태만으로 전체 위험을 설명하기 어렵다.[17]

(2) 고영향 AI와 제32조상 안전성 확보 의무 대상 AI의 구별

AI기본법상 고영향 인공지능은 사람의 생명 · 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있는 인공지능시스템으로서 법 제2조 제4호 각 목에서 정한 영역에 활용되는 것을 말한다. 의료 · 교통 · 채용 · 금융 · 교육 · 공공서비스 등이 주요 적용영역으로 논의되며, 형사사법과 관련해서는 법이 범죄 수사나 체포 업무를 위한 생체인식정보의 분석 · 활용을 명시하고 있다. 이러한 영역에서는 단순한 오류도 개인의 삶에 중대한 결과를 초래할 수 있다.

한편 AI기본법은 '고성능 AI'를 별도의 법정 개념으로 정의하지 않는다. 다만 제32조는 시행령상 세 요건, 즉 학습에 사용된 누적 연산량, 최첨단 인공지능기술의 적용, 사람의 생명 · 신체의 안전 및 기본권에 대한 광범위하고 중대한 위험 가능성을 모두 충족하는 인공지능시스템에 대하여 인공지능 수명주기

17) 분산된 의사결정구조에서 특정 대표자의 심리상태만으로 법인의 불법과 책임을 포착하기 어렵다는 조직모델의 문제의식에 관하여 안정빈, 「법인의 형사책임 – 동일시이론과 조직모델이론의 구별을 중심으로 –」, 「일감법학」 제44호, 건국대학교 법학연구소, 2019, 207-235면 참조. 데이터 수집팀 · 모델 설계팀 · 클라우드 인프라 제공자 등 AI 생애주기의 구체적 배열은 위 일반론을 AI 개발 · 운영 구조에 적용한 본고의 분석이다.

전반의 위험을 식별·평가·완화하고 안전사고를 모니터링·대응하기 위한 위험관리체계 구축 등의 안전성 확보 의무를 부과한다.[18]

형사법상 중요한 것은 명칭이 아니라 위험의 구체성이다. AI가 고영향 인공지능에 해당하거나 제32조상 안전성 확보 의무의 대상이 되었다는 사실은 사업자가 더 높은 수준의 위험관리와 사후검증을 예상할 수 있었다는 간접사실이 된다. 그러나 그러한 분류만으로 곧바로 형사책임이 인정되는 것은 아니다. 결과발생과 의무위반 사이의 규범적 관련성, 예견가능성, 회피가능성, 통제가능성이 별도로 검토되어야 한다.

(3) 투명성·고지·위험관리·영향평가·기록의무

개별 의무의 검토에 앞서 그 조문상 위치를 정리해 둘 필요가 있다. AI기본법은 제31조에서 투명성 확보 의무를, 제32조에서 안전성 확보 의무를, 제33조에서 고영향 인공지능 해당 여부의 사전검토와 확인을, 제34조에서 고영향 인공지능과 관련한 사업자의 책무를, 제35조에서 고영향 인공지능에 대한 영향평가를, 제36조에서 국외 사업자의 국내대리인 지정을 각각 규정한다. 이 가운데 형사법상 주의의무·감독의무의 판단에 가장 직접적으로 연결되는 것은 제34조 제1항이다. 같은 항은 ① 위험관리방안의 수립·운영, ② 기술적으로 가능한 범위에서 인공지능이 도출한 최종결과와 그 도출에 활용된 주요 기준 및 학습용데이터의 개요 등에 관한 설명 방안의 수립·시행, ③ 이용자 보호방안의 수립·운영, ④ 고영향 인공지능에 대한 사람의 관리·감독, ⑤ 안전성·신뢰성 확보를 위한 조치의 내용을 확인할 수 있는 문서의 작성과 보관, ⑥ 그 밖에 고영향 인공지능의 안전성·신뢰성 확보를 위하여 국가인공지능전략위원회에서 심의·의결된 사항을 요구한다. 특히 제4호는 법률이 명시적으로 '사람에 의한 관리·감독'을 요구한다는 점에서, 후술하는 감독의무 위반과 보증인지위 판단의 실정법적 출발점이 된다.

AI기본법 제31조의 투명성 확보 의무는 고영향 인공지능 또는 생성형 인공지능에 기반한 제품·서비스라는 사실을 사전에 고지하고, 생성형 인공지능의 결과물 및 현실과 구별하기 어려운 생성물에는 그 생성 사실을 표시·고지하

18) 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령」 [시행 2026. 7. 21.] [대통령령 제36506호, 2026. 7. 20., 일부개정] 제24조 제1항은 법 제32조의 안전성 확보의무 대상에 관하여 ① 학습에 사용된 누적 연산량이 10의 26승 부동소수점연산 이상일 것, ② 현재 인공지능기술 중 최첨단 인공지능기술을 적용하여 구성·운영될 것, ③ 인공지능시스템의 위험도가 사람의 생명·신체의 안전 및 기본권에 광범위하고 중대한 영향을 미칠 우려가 있을 것이라는 요건을 함께 요구한다.

도록 한다. 이러한 고지·표시는 이용자가 AI의 개입을 인식하고 자기결정을 형성하며, 피해자가 사후 권리행사의 출발점을 확보하기 위한 절차적 전제가 된다.

위험관리 의무(제32조 제1항 제1호, 제34조 제1항 제1호)는 사전적 통제의 핵심이다. 고영향 AI가 의료진단, 교통통제, 채용평가, 금융심사, 교육배치 등에서 사용되는 경우, 사업자는 모델의 정확도와 편향성, 오류 가능성, 사용환경, 인간 감독절차, 오류 보고체계를 점검해야 한다. 이러한 절차가 존재하지 않거나 문서상으로만 존재했다면, 사고 발생 후 조직책임 판단에서 불리한 사정으로 평가될 수 있다.

제34조 제1항 제2호의 설명 방안 수립·시행과 제5호의 문서 작성·보관은 피해자 구제와 밀접하게 연결된다. 피해자는 AI가 자신에게 불리한 결정을 내렸다는 사실은 알 수 있어도, 그 판단이 어떤 데이터와 주요 기준, 어떤 위험관리 실패에서 비롯되었는지 알기 어렵다. 사업자가 관련 기록을 보유하지 않거나 사후 검증이 불가능한 방식으로 관리한다면 피해자는 과실과 인과관계를 주장·입증하기 어렵다.[19]

다만 현행 AI기본법이 피해자에게 일반적 설명청구권이나 손해배상상 입증책임의 전환을 직접 부여한다고 보기는 어렵다. 설명 방안과 문서 작성·보관 의무의 형사법적 의미는 우선 주의의무 이행 여부와 조직적 위험관리 실패를 판단하는 간접자료로 이해되어야 하며, 피해자 구제와의 직접 연결은 별도의 입법적 과제로 남는다.

제35조의 영향평가는 고영향 인공지능을 이용한 제품 또는 서비스를 제공하는 사업자가 사전에 사람의 기본권에 미치는 영향을 평가하기 위하여 노력하도록 하는 의무이다. 현행법은 영향평가를 실시하는 경우 제품 또는 서비스의 성격을 고려하여 인공지능취약계층의 특성이 반영될 수 있도록 요구한다. 따라서 영향평가의 불실시만으로 곧바로 형사상 주의의무 위반을 인정할 수는 없다. 다만 고위험 이용환경과 구체적 위험정보가 존재하였는데도 영향평가를 전혀 검토하지 않은 사정은 예견가능성과 조직적 위험관리의 적정성을 판단하는 간접자료가 될 수 있다.

---

19) AI기본법 제34조 제1항 제2호와 제5호는 설명 방안의 수립·시행 및 안전성·신뢰성 확보조치에 관한 문서의 작성·보관을 요구한다. 이는 현행법상 일반적 피해자 설명청구권이나 손해배상상 입증책임 전환을 곧바로 부여하는 규정은 아니며, 피해자 입증곤란의 완화, 감독기관의 사후 검증 및 형사법상 주의의무 판단을 연결하는 절차적 장치로 이해할 수 있다.

다음 표는 AI사업자 의무를 형사법상 책임판단의 관점에서 재정리한 것이다.

<표 1> AI사업자 의무와 형사법상 책임판단의 연결구조

의무 유형	주된 규율 대상	형사법상 의미	책임 판단에서의 기능
투명성 · 고지 의무 (제31조)	고영향 AI · 생성형 AI 이용 제품 · 서비스 및 실제와 구분하기 어려운 AI 생성물	AI 개입 · 생성 사실의 고지와 이용자 자기결정 보장	주의의무 위반 및 예견가능성 판단의 간접사실
위험관리 의무 (제32조 · 제34조)	제32조상 안전성 확보 대상 AI 및 고영향 AI의 개발 · 제공 · 운영	위험창출자에게 요구되는 사전 통제의무	과실범, 감독책임, 조직책임 판단의 중심 자료
설명 방안 · 문서 작성 · 보관 (제34조 제1항 제2호 · 제5호)	AI 판단의 주요 기준 · 학습용데이터 개요 · 안전성 조치의 사후 검증	원인규명과 피해자 입증곤란 완화를 위한 기초자료	인과관계 · 주의의무 판단에서 정보비대칭 조정 기능
영향평가 노력의무 (제35조)	고영향 AI 이용 제품 · 서비스	기본권 영향의 사전 검토	불실시 자체가 곧 과실은 아니나 예견가능성 · 위험관리 적정성의 간접자료
국내대리인 지정의무 (제36조)	국내에 주소 · 영업소가 없는 일정한 국외 사업자	책임 회피 방지와 규제관할 확보	국외 사업자에 대한 규제관할 · 집행의 확보(조직 내 책임귀속과는 간접적 관련)

(4) 과태료 체계의 한계

AI기본법 제43조 제1항은 3천만 원 이하의 과태료 부과대상을 사전고지의무 위반, 국내대리인 미지정, 중지명령 · 시정명령 불이행의 세 경우로 한정한다.[20] 과학기술정보통신부는 법 시행 초기 최소 1년 이상의 계도기간을 운영하고, 그 기간에는 사실조사와 과태료 부과를 원칙적으로 유예하되 인명사고나 인권침해 등 중대한 사회적 문제가 발생한 경우에는 예외를 두는 집행방침을 밝혔다.[21] 계도기간은 법률상 의무의 효력을 정지하는 규정이 아니라 행정상 집행방침이다. 과태료 체계가 존재하더라도 사망 · 상해 · 재산상 손해 등 구체적 법익침해가 발생하면 개별 형사구성요건의 적용 가능성은 별도로 검토되어야 한다. 본고는 이처럼 제한된 행정제재 구조가 중대한 법익침해 국면의 형사법적

20) 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제43조 제1항.

21) 과학기술정보통신부, 「'인공지능기본법' 22일 시행…생성형 AI 결과물 '워터마크' 표시」, 2026. 1. 21., 대한민국 정책브리핑, https://www.korea.kr/news/policyNewsView.do?newsId=148958380 (최종접속 2026. 7. 26.).

평가를 대체하지 못한다고 본다.

본고는 AI기본법상 의무의 이행 여부가 형사상 주의의무의 존부를 자동으로 결정하지는 않지만, 구체적 사건에서 예견가능성 · 회피가능성 · 통제가능성과 조직적 위험관리의 적정성을 판단하는 간접자료가 될 수 있다고 본다. 따라서 사업자가 기본법상 의무를 이행하였다는 사정은 상당한 주의의무 이행을 뒷받침할 수 있고, 최소한의 위험관리 · 고지 · 기록의무조차 이행하지 않은 사정은 주의의무 또는 감독의무 위반을 판단하는 불리한 간접사실이 될 수 있다.

이 점에서 AI기본법과 형법의 관계는 대체관계가 아니라 보완관계이다. AI기본법은 사전적 위험관리의 틀을 제공하고, 형법은 그 위험관리의 실패가 중대한 법익침해로 현실화된 경우 책임귀속의 최종 판단을 담당한다.

5. 형식이행형 의무와 위험관리형 주의의무

(1) 구별의 필요성

AI기본법상 사업자 의무를 모두 같은 성질의 의무로 취급하면 책임 판단이 거칠어진다. 어떤 의무는 이행 여부가 비교적 형식적으로 판단된다. 예컨대 특정 산출물에 AI 생성물 표시를 했는지, 일정한 고지를 했는지, 서류를 제출했는지는 비교적 명확하게 확인할 수 있다. 이러한 의무를 본고에서는 형식이행형 의무(formal-compliance obligation)라 부르기로 한다.

반면 어떤 의무는 단순히 “했는지 안 했는지”로 판단하기 어렵다. 위험평가를 충분히 했는지, 검증절차가 합리적이었는지, 인간 감독체계가 실효적으로 작동했는지, 사후 모니터링이 충분했는지는 구체적 위험, 기술수준, 사업자의 지배가능성, 피해자의 의존성에 따라 달라진다. 이러한 의무를 본고에서는 위험관리형 주의의무(risk-based duty of care)라 부른다. 이 명칭은 본고의 분석용어이고, 그 실질은 위험도에 따라 주의의무의 내용을 구체화하여야 한다는 문제의식과 연결된다.

이 구별은 단순한 개념 분류가 아니다. 형식이행형 의무의 위반은 행정제재 판단에 특히 적합하고, 형사법에서는 예견가능성 또는 고지의무 위반을 뒷받침하는 간접사실로 기능한다. 위험관리형 주의의무의 위반은 형사상 주의의무, 부작위범의 보증인지위, 법인 조직책임 판단에서 더욱 직접적인 의미를 갖는다.

(2) 형식이행형 의무

형식이행형 의무는 법령이 일정한 행위를 외형적으로 요구하고, 그 이행 여

부가 비교적 명확하게 판정되는 의무이다. 표시의무, 고지의무, 자료제출의무, 국내대리인 지정의무, 일정한 절차의 개시의무 등이 여기에 속한다. 이 의무들은 행정법상 집행가능성이 높고, 사업자에게 예측가능한 준법기준을 제공한다.

그러나 형식이행형 의무가 형사법과 무관한 것은 아니다. 예컨대 AI 생성물 또는 딥페이크가 선거홍보물, 성적 허위영상, 유명인 사칭 광고 등에 이용되고도 제31조상 표시·고지가 이루어지지 않았다면, 그 위반은 이용자와 피해자의 착오 및 자기결정 침해를 판단하는 자료가 될 수 있다.[22] 다만 표시·고지의무 위반만으로 형사책임이 성립하는 것은 아니며, 표시가 있었다면 구체적 결과를 방지할 수 있었는지와 위반된 의무의 보호목적을 별도로 심사해야 한다.

(3) 위험관리형 주의의무

위험관리형 주의의무는 법령상 문언만으로 구체적 이행수준이 완전히 정해지지 않고, 위험의 성격과 사업자의 통제가능성에 따라 내용이 구체화되는 의무이다. 위험관리체계 구축, 사전 검증, 편향성 점검, 인간 감독, 사후 모니터링, 오류 보고와 수정, 기록보존 등이 대표적이다. 이 목록은 대체로 제32조 제1항 각 호와 제34조 제1항 각 호가 요구하는 조치에 대응한다. 법문이 요구하는 조치의 '내용'은 특정되어 있으나 그 '수준'은 특정되어 있지 않다는 점에 위험관리형 주의의무의 특질이 있다.

위험관리형 주의의무의 핵심은 "합리적으로 기대 가능한 조치"이다. 형사법은 완전한 무오류 AI를 요구할 수 없다. 본고의 관점에서 사업자가 고영향 AI를 배포하면서 합리적으로 요구되는 사전검증을 하지 않았거나, 반복적 오류를 구체적으로 인식하고도 수정하지 않았거나, 특정 범죄유형의 이용을 인식하면서 가능한 차단조치를 취하지 않았다면 주의의무 위반 여부가 문제될 수 있다.[23]

---

22) AI기본법 제31조는 고영향 인공지능·생성형 인공지능 이용 사실의 사전고지, 생성형 인공지능 결과물의 표시 및 실제와 구분하기 어려운 가상 결과물의 고지·표시를 규정한다. 딥페이크의 문제유형을 선거홍보물·음란물 등 공익침해, 퍼블리시티권·저작권 등 사익침해, 소비자 기만으로 나누어 설명한 문헌으로 박준우, 「인공지능(AI)·딥페이크 생성물의 퍼블리시티권 침해 연구」, 「산업재산권」 제79호, 한국지식재산학회, 2024, 321면; 딥페이크 성범죄·사칭 사기·선거 허위정보의 사례에 관하여 신혜진, 앞의 논문, 239-244면 참조. 표시·고지 위반을 형사책임의 간접사실로 평가하는 것은 본고의 해석이다.

23) 본문의 세 가지 조치는 AI기본법상 위험관리의무를 과실범론에 적용한 본고의 구체적 예시이다. AI 공급자의 위험 식별·완화, 품질·보안관리, 출시 전 테스트·검증, 출시 후 로그관리와 지속적 모니터링에 관하여 신혜진, 앞의 논문, 253-255면; 범죄목적 공급자의 공동정범·방조와 예측 가능한 결과에 대한 과실책임에 관하여 같은 논문, 267면 참조.

위험관리형 주의의무는 결과책임으로 흐르지 않게 하기 위해 엄격하게 다루어야 한다. 사고가 발생했다는 이유만으로 곧바로 의무위반이 인정되어서는 안 된다. 의무위반이 인정되기 위해서는 구체적 위험의 예견가능성, 조치의 회피가능성, 사업자의 통제가능성, 의무가 방지하려던 위험과 실제 결과 사이의 규범적 관련성이 함께 확인되어야 한다.

(4) 의무층위와 책임분배

AI 시스템에는 여러 의무주체가 동시에 관여한다. 개발사업자는 모델 구조와 학습데이터에 관한 1차적 정보를 보유한다. 이용사업자는 실제 적용환경과 사용자 고지, 인간 감독절차를 지배한다. 플랫폼은 접근권한, 이용조건, 오남용 탐지, 계정제재, API 사용 제한을 통제한다. 최종 이용자는 AI 산출물을 범죄 또는 위법행위에 활용할 수 있다.

따라서 책임분배는 단순히 “개발자 책임” 또는 “이용자 책임”으로 나눌 수 없다. 위험이 모델 설계 단계에서 비롯되었는지, 적용환경의 부적절성에서 비롯되었는지, 사후 모니터링 실패에서 비롯되었는지, 최종 이용자의 범죄적 의도에서 비롯되었는지에 따라 의무의 중심이 달라진다.

이 지점에서 ‘의무의 강약’[24]이라는 사고가 유용하다. 같은 AI사업자라도 위험에 대한 정보, 지배, 통제, 이익, 피해자 의존성을 더 많이 가진 주체에게 더 강한 의무가 인정된다. 반대로 범용기술을 일반적으로 제공했을 뿐 구체적 범죄이용 가능성을 알기 어려웠던 주체에게 형사책임을 쉽게 인정해서는 안 된다.

의무의 강약이라는 관점은 AI기본법의 차등화된 의무구조와 접점을 갖는다. 이 법은 모든 AI사업자에게 동일한 의무를 부과하지 않는다. 고영향 인공지능을 제공하는 사업자에게는 제34조의 가중된 책무가, 누적 연산량 · 최첨단 기술의 적용 · 광범위하고 중대한 위험 가능성이라는 시행령상 요건을 모두 충족하는 인공지능시스템의 사업자에게는 제32조의 안전성 확보 의무가 각각 추가된다. 즉 위험에 대한 정보와 지배를 더 많이 가진 주체일수록 더 강한 의무를 지도록 설계되어 있다. 이는 부작위 경합 상황에서 의무의 강약에 따라 정범성

24) 의무의 성질 · 강도에 따라 보증인지위와 부작위범의 책임구조가 달라질 수 있다는 문제의식에 관하여 안정빈, 「보증인지위와 적극적 · 소극적 의무 개념 – 의무범 이론과의 연계성을 중심으로 –」, 「비교법연구」 제22권 제3호, 동국대학교 비교법문화연구원, 2022, 407-459면; 의무의 강약에 따른 정범 · 공범 구별에 관하여 안정빈, 「의무범에서의 정범과 공범을 구별 짓는 기제로서의 ‘의무의 강약설’ 및 사회적 · 규범적 판단에 대한 연구 – 부작위 경합 상황에서의 의무의 충돌 및 의무부과의 개념을 중심으로 –」, 「중앙법학」 제24집 제4호, 중앙법학회, 2022, 특히 126-128면 참조.

과 공범성이 분배된다는 종래의 논의와 같은 방향을 향한다.[25] 형사법적으로도 제34조 제1항 제4호에 따라 사람에 의한 관리·감독의무를 부담하는 고영향 AI 사업자와, 그러한 의무를 부담하지 않는 범용 모델 제공자를 동일한 강도의 보증인으로 취급할 수는 없다. 의무의 강약은 여기서 보증인지위의 인정 여부만이 아니라 그 지위의 밀도, 나아가 정범과 공범의 분배를 결정하는 기제로 작동한다.

Ⅲ. AI사업자 의무의 형사법적 의미와 책임귀속

1. 행정법상 의무와 과실범의 주의의무

과실범에서 주의의무는 구성요건적 결과를 예견하고 회피할 수 있었음에도 요구되는 주의를 다하지 않은 경우에 인정된다. 문제는 AI 시스템처럼 기술적으로 복잡하고 불투명한 영역에서 요구되는 주의의무의 내용을 어떻게 구체화할 것인가이다. 순수한 사후적 관점에서 "결과가 발생했으므로 주의의무 위반이 있다"고 판단하면 결과책임에 가까워질 위험이 있다. 반대로 사업자의 기술적 재량을 과도하게 존중하면 피해자는 사실상 구제받기 어렵다.

이때 AI기본법상 의무는 주의의무의 내용을 구체화하는 기준으로 기능한다. 법령이 특정한 설명, 고지, 위험관리, 기록, 사후 모니터링을 요구하였다면, 이는 해당 분야의 평균적·객관적 주의수준을 보여주는 자료가 된다. 물론 행정법상 의무 위반이 곧바로 형사상 과실을 의미하는 것은 아니다. 형사책임을 인정하기 위해서는 위반된 의무가 보호하려는 위험이 실제 결과로 실현되었는지, 그 결과가 예견가능하고 회피가능했는지, 의무 위반과 결과 사이에 규범적 관련성이 있는지를 별도로 검토해야 한다.

그러나 법령상 의무 위반이 형사책임과 무관하다고 볼 수도 없다. AI사업자가 고영향 AI의 위험성을 알고 있었거나 알 수 있었음에도 위험평가, 검증, 고지, 인간의 감독체계 구축을 하지 않았다면, 그 위반은 과실범에서 주의의무 위반을 인정하는 강한 간접사실이 된다.

---

25) 안정빈, 「부작위 행위자들 사이에서 정범과 공범의 구별」, 「형사법연구」 제31권 제2호, 한국형사법학회, 2019, 3-38면 참조.

2. 예견가능성과 회피가능성

AI 관련 사고에서 예견가능성은 단순히 “AI가 오류를 낼 수 있다”는 일반적 가능성으로 충분하지 않다. 특정 유형의 오류, 특정 이용환경, 특정 피해 결과가 사업자에게 어느 정도 구체적으로 인식 가능했는지가 문제된다. 예컨대 의료영상 판독 AI가 특정 집단의 데이터 부족으로 오진 가능성이 높다는 사실을 사업자가 알 수 있었다면 예견가능성은 강화된다.

회피가능성 역시 기술적 가능성과 조직적 가능성을 함께 보아야 한다. 완전한 무오류 AI를 요구할 수는 없지만, 사전 검증, 인간 검토 절차, 위험등급 분류, 이용자 고지, 사후 모니터링, 오류 보고체계 등 합리적으로 기대할 수 있는 조치를 취할 수 있었다면 회피가능성은 인정될 수 있다. 이때 AI기본법상 사업자 의무는 무엇이 “합리적으로 기대 가능한 조치”인지 판단하는 기준을 제공한다.

마지막으로 규범의 보호목적도 중요하다. 투명성 의무가 보호하려는 것은 단순히 형식적 고지가 아니라 이용자와 피해자가 AI의 개입과 생성 사실을 인식할 수 있도록 하는 것이다. 위험관리의무가 보호하려는 것은 단순한 내부문서 작성이 아니라 고영향 AI로 인한 중대한 법익침해를 예방하는 것이다. 따라서 해당 의무가 보호하려는 위험이 실제로 발생한 결과와 연결되는 경우, 그 의무 위반은 형사법상 책임귀속의 중요한 근거가 될 수 있다.

3. 위험창출과 보증인지위

AI사업자의 책임을 논할 때 핵심은 그가 단순한 도구 제공자인가, 아니면 위험원을 창출·지배·관리하는 자인가이다. 모든 소프트웨어 제공자에게 광범위한 형사책임을 인정할 수는 없다. 그러나 고영향 AI처럼 생명·신체·재산·사회적 기회에 중대한 영향을 미치는 시스템을 설계하고 배포한 사업자는 일정한 범위에서 위험원을 창출하고 관리하는 지위에 선다.

이러한 지위는 부작위범론의 관점에서도 중요하다. 사업자가 자신의 선행행위로 위험원을 창출하였고, 그 위험이 이용 과정에서 현실화될 가능성을 예견할 수 있었으며, 기술적·조직적으로 이를 통제할 수 있었음에도 필요한 조치를 취하지 않았다면, 일정한 경우 보증인지위가 문제될 수 있다. 물론 보증인지위를 무제한 인정해서는 안 된다. 보증인지위는 위험의 구체성, 통제가능성,

법령상 의무의 명확성, 사업자의 역할, 피해자의 의존성 등을 종합하여 신중하게 판단해야 한다.

AI기본법상 의무는 바로 이 판단에서 중요한 의미를 갖는다. 법률이 특정 사업자에게 위험관리와 고지의무를 부과하였다면, 이는 그 사업자가 해당 위험과 무관한 제3자가 아니라 일정한 보호·감독 책임을 부담하는 지위에 있음을 보여준다. 따라서 AI기본법은 보증인지위를 직접 창설하는 형법규범은 아니지만, 형사법상 보증인지위 인정 여부를 판단하는 외부적 근거로 활용될 수 있다.

4. 법인 형사책임론과 AI사업자 조직책임

(1) 법인 형사책임론의 참조 가능성과 한계

AI 형사책임론은 법인 형사책임론과 자주 비교된다. 양자는 모두 자연인이 아닌 존재를 둘러싼 책임귀속 문제라는 점에서 표면적으로 유사하지만, 법인책임론과 AI의 형법상 주체성론은 서로 다른 이론적 문제이다. 법인책임론은 인간의 조직행위를 법인에 어떠한 요건으로 귀속할 것인지가 중심이다.[26] AI 주체성론은 AI 자체에 행위능력·책임능력과 형벌수용 가능성을 인정할 수 있는지가 중심이다.[27]

그러나 두 논의는 결정적 차이를 가진다. 법인은 인간들의 조직이다. 법인의 의사결정은 이사, 대표자, 임직원, 부서, 내부통제 시스템의 결합을 통해 형성된다. 따라서 법인 형사책임은 궁극적으로 인간 책임의 집합적·조직적 귀속 문제이다. 반면 AI는 인간이 설계하고 운용하지만, 특정 결과가 언제나 특정 인간의 의사결정으로 환원되는 것은 아니다. 학습 이후의 모델은 데이터, 알고리즘, 피드백, 배포환경, 이용자의 입력이 결합된 결과를 산출한다.

본고의 관점에서는 AI 관련 책임론을 법인 형사책임론의 단순한 확장으로 볼 수 없다. 법인 형사책임론은 “조직 안의 인간 의사”를 어떠한 요건으로 법인에게 귀속할 것인가를 묻는다. AI사업자의 책임은 “인간 의사로 완전히 환원되지 않는 시스템 위험”을 조직이 어떻게 예견하고 통제했어야 하는가를 추가로 묻는다. 이러한 차이 때문에 AI 관련 책임론은 법인 동일시이론의 기계

---

26) 법인책임론에서 동일시이론과 조직모델의 구별에 관하여 안정빈, 앞의 「법인의 형사책임」, 207-235면 참조.

27) 현 단계 AI의 책임능력과 인간에 대한 책임귀속 문제에 관하여 임석순, 앞의 논문, 73면, 77-78면; AI에 독자적 법인격을 부여할 가능성을 검토하는 견해로 강지현, 앞의 논문, 59-71면 참조.

적 유추만으로 해결될 수 없다.

(2) 동일시이론의 한계와 조직모델

법인 형사책임에서 동일시이론은 대표자 또는 고위관리자의 행위와 의사를 법인의 행위와 의사로 동일시하는 방식이다. 이 이론은 책임귀속의 명확성을 제공하지만, 현대 기업의 분산적 의사결정 구조를 충분히 포착하지 못한다. 특히 AI 시스템의 개발과 운영이 앞서 본 바와 같이(Ⅱ. 4. (1)) 다수 부서와 외부 사업자의 결합으로 이루어지는 경우, 특정 자연인 한 명의 고의나 과실로 전체 위험을 설명하기 어려운 경우가 많다.

이 점에서 AI사업자 책임은 동일시이론보다 조직모델에 가깝다. 문제는 누가 최종 결정을 했는가만이 아니라, 조직이 위험을 식별하고 통제할 수 있는 절차와 구조를 갖추었는가이다. 위험평가 체계가 있었는지, 고영향 AI 여부를 심사했는지, 데이터 편향과 오류 가능성을 검증했는지, 배포 후 모니터링과 리콜 또는 수정 절차가 있었는지, 이용자에게 한계와 위험을 고지했는지가 핵심 판단요소가 된다.

AI기본법상 사업자 의무는 바로 이러한 조직모델적 책임 판단과 친화적이다. 법은 특정 개인의 심리상태보다 사업자의 절차, 체계, 고지, 관리, 기록을 문제 삼는다. 형사법이 이를 분석에 수용한다면, AI 관련 책임귀속은 대표자의 고의·과실만을 찾는 방식에 머무르지 않고 조직 전체의 위험관리 실패를 평가하는 방향으로 보완될 수 있다.

다만 조직책임 모델은 법인의 결과책임을 인정하기 위한 이론이 아니다. 오히려 특정 자연인의 과실을 발견하기 어렵다는 이유만으로 무리하게 책임을 의제하거나, 반대로 조직 전체의 구조적 위험관리 실패를 전혀 평가하지 못하는 양극단을 피하기 위한 귀속통제 장치로 이해되어야 한다.

그러나 조직모델이 현행법상 법인 자체의 일반적 형사책임을 곧바로 창설하는 것은 아니다. 현행법상 자연인의 범죄행위가 법인에게 당연히 귀속되는 것은 아니고, 개별 법률의 양벌규정 등 명시적 처벌근거가 있는 경우에 한하여 법인에게 벌금형을 부과할 수 있다.[28] 따라서 본고의 조직책임 모델은 우선

28) 현행법상 법인의 처벌은 개별 법률의 양벌규정 등 명시적 근거가 있는 경우에 한하여 가능하고, 양벌규정이 적용되는 경우에도 법인이 해당 업무와 관련하여 상당한 주의 또는 관리감독의무를 게을리하였는지가 별도로 심사되어야 한다. 대법원 2011. 7. 14. 선고 2009도5516 판결 참조. AI기본법은 AI사업자의 일반적 조직과실을 처벌하는 포괄적 양벌규정을 두고 있지 않다.

조직 내부의 자연인에게 귀속되는 주의의무·감독의무 위반을 규명하고, 별도의 법인 처벌근거가 존재하는 경우에는 법인의 상당한 주의·감독의무 위반 여부를 심사하기 위한 분석틀로 이해되어야 한다. AI기본법상 사업자 의무는 그 자체로 법인에 대한 새로운 형사처벌의 근거를 창설하지 않는다.

(3) 조직책임의 판단요소

AI사업자 조직책임을 인정하기 위해서는 적어도 다음 요소가 검토되어야 한다. 첫째, 위험의 중대성과 구체성이다. AI 시스템이 생명·신체·재산·사회적 기회에 중대한 영향을 미치는 영역에서 사용될수록 사업자에게 요구되는 주의수준은 높아진다. 둘째, 통제가능성이다. 사업자가 모델 설계, 데이터 선택, 배포환경, 이용조건, 업데이트, 사후 모니터링을 실질적으로 통제할 수 있었는지가 중요하다.

셋째, 법령상 의무의 명확성이다. AI기본법이나 하위법령, 가이드라인이 특정 조치를 요구하고 있었다면, 사업자는 이를 기준으로 위험관리체계를 설계했어야 한다. 넷째, 내부통제의 실효성이다. 문서상 위험관리 절차가 존재하더라도 실제로 작동하지 않았다면 책임을 회피하기 어렵다. 다섯째, 결과와 의무위반 사이의 관련성이다. 의무위반이 추상적으로 존재한다는 사정만으로 형사책임을 인정해서는 안 되며, 그 의무가 방지하려던 위험이 현실화되었는지를 보아야 한다.

이러한 요소들은 AI사업자 책임을 결과책임으로 흐르지 않게 하는 장치이기도 하다. AI 사고가 발생했다는 이유만으로 사업자에게 형사책임을 묻는 것은 허용될 수 없다. 그러나 사업자가 고위험 시스템을 시장에 내놓으면서 위험평가와 사후관리 체계를 갖추지 않았고, 그 위험이 실제 법익침해로 현실화되었다면, 조직책임을 논의할 수 있다.

5. 인간-AI 협업범죄와 책임귀속

(1) AI는 공범인가, 도구인가

생성형 AI의 확산 이후 형사실무에서 가장 현실적인 문제는 AI가 단독으로 범죄를 저지르는 경우가 아니라, 인간이 AI를 이용하여 범죄를 수행하는 경우이다. 딥페이크 영상 제작, 보이스피싱 문안 생성, 피싱 메일 자동화, 허위 투자보고서 작성, 악성코드 생성 보조 등이 대표적이다. 이 경우 AI를 공범으로 볼 것인가, 아니면 범죄도구로 볼 것인가가 문제된다.[29)]

현행 형법의 관점에서 AI를 교사범이나 방조범으로 구성하기는 어렵다. 공범은 정범의 실행행위를 촉진하거나 결의를 유발하는 인간 행위자의 고의와 책임을 전제로 한다. AI는 범행결의를 갖는 존재가 아니며, 형법상 책임비난의 대상이 되기 어렵다. 따라서 일반적으로 AI는 공범이 아니라 도구로 보아야 한다.

다만 AI가 단순한 도구라는 결론이 모든 문제를 해결하지는 않는다. AI의 개입은 범죄의 규모, 속도, 정교함, 은닉성을 현저히 증대시킨다. 사람이 직접 작성한 사기문서와 AI가 대량으로 생성한 맞춤형 피싱문서는 사회적 위험성이 다르다. 따라서 AI를 공범으로 보지는 않더라도, AI 활용의 조직성·자동화·대량성은 구성요건 해석과 양형에서 별도로 고려될 필요가 있다.

(2) 제공자의 방조책임과 중립행위

AI 제공자의 형사책임은 더 어렵다. 범죄자가 범용 AI 서비스를 이용하여 범죄문안을 생성한 경우, AI 제공자를 방조범으로 볼 수 있는가.

일반적 거래·서비스 제공과 범죄방조의 경계는 중립적 방조행위론에서 논의되어 왔다.[30]

검색엔진, 워드프로세서 서비스가 범죄에 이용되었다는 이유만으로 제공자를 방조범으로 처벌할 수 없는 것처럼, 범용 AI 제공자를 곧바로 방조범으로 볼 수는 없다.

이 경우 고의 방조책임의 성립 여부는 AI 제공자가 범죄적 이용 가능성을 추상적으로 예견할 수 있었는지가 아니라, 특정 범죄유형의 반복적 이용, 차단 가능성, 경고·모니터링 조치의 부작위, 서비스 설계·홍보 방식 등을 통하여 범죄 실행을 실질적으로 용이하게 하였다고 평가할 수 있는지에 달려 있다.

그러나 중립행위라는 이유만으로 모든 책임을 배제할 수도 없다. 제공자가

29) 인간이 AI를 범죄수단으로 이용하는 유형, 딥페이크 성범죄·사칭 사기·개인정보 침해 등의 사례 및 공급자·사용자의 책임문제에 관하여 신혜진, 앞의 논문, 239-244면, 267면 참조. 본문의 구체적 예시 가운데 피싱 메일 자동화·허위 투자보고서·악성코드 생성 보조는 위 문제유형을 설명하기 위한 본고의 예시이다.

30) 중립적 방조행위의 개념과 불법귀속 구조에 관하여 이용식, 「일상적·중립적 행위와 방조 — 방조행위의 개념적 의미내용과 방조의 불법귀속의 이해구조 —」, 「법학」 제48권 제1호, 서울대학교 법학연구소, 2007, 302-337면; 미국법상 논의에 관하여 김종구, 「미국 형법상 중립적 행위에 의한 방조의 가벌성」, 「형사법연구」 제24권 제4호, 한국형사법학회, 2012, 77-108면; 가벌성 판단구조에 관하여 장승일, 「방조범의 불법구조와 중립적 방조행위의 가벌성판단」, 「법학논문집」 제40권 제1호, 중앙대학교 법학연구원, 2016, 205-235면 참조. 범용 AI 서비스에 대한 적용은 본고의 검토대상이며, 방조의 고의와 부작위방조의 보증인지위는 별도의 성립요건이다.

특정 범죄유형에 반복적으로 이용된다는 사실을 구체적으로 인식하고도 아무런 차단·경고·모니터링 조치를 취하지 않았거나, 범죄적 이용을 사실상 조장하는 방식으로 서비스를 설계·홍보하였다면 사정은 달라질 수 있다. 이때 핵심은 추상적 위험 인식이 아니라 구체적 범죄이용 가능성에 대한 인식과 통제 가능성이다.

AI기본법상 투명성·위험관리 의무는 여기서도 의미를 갖는다. 법령이 특정 위험에 대한 관리체계를 요구하고 있었음에도 사업자가 이를 방치하였다면, 중립행위의 범위를 넘어서는지 판단하는 자료가 될 수 있다. 다만 형사책임은 최후수단이어야 하므로, 제공자의 방조책임은 엄격하게 제한되어야 한다.

AI기본법상 위험관리의무 위반은 방조의 고의를 대체하지 않는다. 제공자가 구체적 정범범죄를 인식하고 그 실행을 실질적으로 촉진한다는 방조의 고의가 별도로 확인되어야 한다. 특히 부작위에 의한 방조책임을 인정하려면 구체적 범죄이용에 대한 인식과 더불어 정범의 범죄실행을 방지하여야 할 보증인지위가 인정되어야 한다. 반면 단순한 위험관리의 과실은 고의 방조가 아니라 해당 법익침해에 관한 과실범 처벌규정의 적용 문제로 다루어야 한다.

(3) 딥페이크와 생성형 AI 산출물

딥페이크는 AI 활용 범죄의 대표적 사례이다. 딥페이크는 성적 허위영상, 유명인 사칭 투자사기, 보이스피싱, 선거 관련 허위정보 등에 활용되어 피해자의 인격권·명예·성적 자기결정권·재산권 및 민주적 의사형성에 중대한 침해를 야기할 수 있다.[31]

생성형 AI 산출물의 문제는 딥페이크에 한정되지 않는다. AI 창작물은 저작권·퍼블리시티권·부정경쟁 등 민사·지식재산법 영역과 명예훼손·개인정보·사기·업무방해 등 형사법 영역을 교차한다. 형사법은 모든 침해를 형벌로 흡수해서는 안 되지만, AI 산출물이 기망·협박·명예훼손·성적 침해·재산범죄의 수단으로 활용되는 경우에는 기존 구성요건의 해석을 통해 대응해야 한다.[32]

---

31) 딥페이크 성적 허위영상, 사칭 투자사기·보이스피싱 및 선거 관련 허위정보의 사례에 관하여 신혜진, 앞의 논문, 239-244면; 딥페이크의 공익침해·사익침해·소비자 기만 유형과 서비스 제공자의 인식 문제에 관하여 박준우, 앞의 논문, 321면, 339-341면 참조.

32) AI 창작물의 생성·이용과 저작권 침해, 이용자 및 플랫폼의 책임문제에 관하여 차상육, 「인공지능 창작물 관련 저작권 침해 쟁점」, 「경영법률」 제32집 제4호, 한국경영법률학회, 2022, 45-90면; 퍼블리시티권 쟁점에 관하여 박준우, 앞의 논문, 315-354면; AI 사용 범죄의 형사책임에 관하여 신혜진, 앞의 논문, 239-244면, 267면 참조. 각 문헌의 검토대상은

법 제31조의 표시·고지의무는 이러한 책임판단에서 하나의 자료가 된다. 본고는 그 이행 여부가 이용자의 AI 생성 사실 인식, 피해자의 생성경위 확인 가능성, 제공자의 반복적 악용 인식 및 회피가능성을 판단하는 간접자료가 될 수 있다고 본다. 다만 표시·고지의무 위반이 곧바로 고의나 과실을 증명하는 것은 아니다.[33]

(4) 양형에서의 AI 활용

본고는 AI를 사용하였다는 사실 자체를 독립된 가중사유로 보아서는 안 된다고 본다. 다만 AI 사용으로 범행의 자동화·대량화·정교화·은닉성 및 피해회복의 곤란성이 현저히 증대된 경우에는 기존 양형요소인 범행수법, 계획성, 피해규모, 조직성 등을 평가하는 과정에서 이를 고려할 수 있다.

예컨대 딥페이크 성범죄에서 AI를 이용한 영상 합성은 피해자의 인격권과 성적 자기결정권을 중대하게 침해하고, 복제·확산 가능성 때문에 피해회복이 매우 어렵다. 보이스피싱에서 AI 음성합성이나 자동화된 대화시스템은 피해자의 착오를 강화하고 범행의 조직성을 높인다. 이러한 사정은 기존 양형기준의 피해규모, 범행수법, 계획성, 조직성, 피해회복 가능성 항목과 연결하여 평가될 수 있다.

결국 인간-AI 협업범죄의 형사법적 과제는 AI를 공범으로 의제하는 데 있지 않다. 인간 행위자의 책임을 중심으로 하되, AI 활용이 범죄의 위험성과 법익침해의 정도를 어떻게 변화시켰는지를 세밀하게 반영하는 데 있다.

6. 피해자 구제와 위험책임 모델

(1) 과태료와 피해자 구제의 분리

AI기본법상 과태료는 국가가 부과하는 행정상 제재이다. 과태료가 부과되더라도 그 금전은 피해자에게 직접 귀속되지 않는다. 따라서 과태료 체계만으로는 AI 사고 피해자의 치료비, 일실수입, 정신적 손해, 기회상실, 명예훼손, 개인정보 침해를 회복하기 어렵다. 이 점에서 AI기본법은 피해자 구제법이라기

서로 다르다.

33) AI기본법 제31조의 표시·고지의무가 구체적 형사사건에서 이용자의 착오, 제공자의 위험인식 및 회피가능성을 판단하는 간접자료가 될 수 있다는 점은 본고의 해석이다. 딥페이크 문제유형과 서비스 제공자의 실제 인식이 책임판단에서 문제된 사례에 관하여 박준우, 앞의 논문, 321면, 339-341면; 딥페이크 성범죄·사칭 사기의 구체적 양상에 관하여 신혜진, 앞의 논문, 239-244면 참조.

보다 위험관리법에 가깝다.

이 한계는 AI기본법이 산업진흥과 신뢰 기반 조성에 중점을 두면서도, AI 시스템으로부터 영향을 받는 개인의 권리구제 장치를 충분히 마련하고 있는지에 관한 비판적 검토와도 연결된다. 따라서 피해자 구제 논의는 행정제재의 강도뿐 아니라 정보접근, 기록보존, 설명가능성, 손해배상, 보험·기금 모델을 함께 보아야 한다.

형사법 역시 피해자 구제를 직접 담당하는 제도는 아니다. 그러나 형사법상 주의의무와 조직책임 판단은 민사책임과 피해자 구제에 간접적으로 큰 영향을 미친다. 형사절차에서 사업자의 위험관리 실패가 밝혀지면 민사상 과실, 결함, 인과관계 판단에도 영향을 줄 수 있다. 반대로 형사법이 사업자 의무와 책임구조를 지나치게 좁게 해석하면 피해자는 정보비대칭 속에서 손해배상청구를 입증하기 어렵다.

따라서 AI기본법 이후의 과제는 과태료의 상향 여부에만 있지 않다. 더 중요한 것은 AI사업자의 설명 방안, 문서 작성·보관 및 위험관리 조치를 피해자 구제와 연결하는 제도적 통로를 마련하는 것이다. 사업자가 AI 시스템의 의사결정 과정, 학습데이터의 한계, 업데이트 이력, 위험평가 결과를 보존하지 않았다면, 그 불이익을 피해자에게 전가해서는 안 된다.

(2) 위험사회와 책임구조의 변화

위험사회론은 전통적 책임모델과 현대적 귀속모델의 차이를 통하여 책임구조의 변화를 설명한다. 전통적 모델이 과거의 특정 행위와 결과의 귀속을 중심으로 한다면, 현대적 위험은 기술적 복잡성·장기적 누적성·예측곤란성·다수주체의 결합 때문에 장래 위험의 예방과 관리라는 기능을 강조한다. 그러나 이러한 전환은 책임주체를 무한히 확장하거나 과거의 유책행위에 관한 귀속요건을 약화시키는 근거가 될 수 없다.[34]

이러한 측면은 AI사업자 책임에도 나타난다. AI 사고는 특정 행위자 한 명의 잘못으로만 설명되지 않을 수 있다. 데이터 수집, 모델 학습, 제품화, 서비스 적용, 사후 모니터링, 이용자 입력이 결합하여 결과가 발생하기 때문이다. 따라서

34) 김영환 교수는 전통적 책임모델과 현대적 귀속모델의 변화를 설명하면서도, 책임주체의 과도한 확장과 책임의 기능화가 유책자와 비유책자를 구별하는 책임개념의 한계·경감기능을 약화할 위험을 지적한다. 김영환, 「위험사회에서의 책임구조: 자연 재해에 대한 법적 담론」, 「홍익법학」 제14권 제1호, 홍익대학교 법학연구소, 2013, 825-833면, 837면. 본고는 이러한 비판을 전제로 위험관리의 필요성을 책임주의의 한계 안에서만 수용한다.

“누가 위험을 알고 통제할 수 있었는가”를 묻는 것은 필요하지만, 그것이 “누가 마지막 버튼을 눌렀는가”라는 개별 귀속의 질문을 대체해서는 안 된다.

따라서 본고는 위험사회론을 형사책임의 예방책임화 근거로 사용하지 않는다. AI기본법상 행정의무는 고의·과실, 개인적 책임비난 가능성, 죄형법정주의와 책임주의의 한계 안에서 주의의무를 구체화하는 외부적 자료로만 기능하여야 한다.

(3) 제조물책임·환경책임·원자력책임의 시사점

AI 위험은 전통적인 형사책임 모델만으로 포착하기 어렵다. 형사책임은 고의·과실과 책임비난 가능성을 요구하지만, AI 사고에서는 원인규명과 인과관계 입증이 어렵고 책임이 여러 주체에게 분산된다. 이 때문에 제조물책임, 환경책임, 원자력책임 모델은 일정한 시사점을 제공한다.

제조물책임 모델은 결함 있는 제품으로 인한 손해에 대하여 제조업자에게 책임을 묻는 구조이다. AI가 동산에 내장된 형태로 제공되는 경우에는 제조물책임법의 적용이 문제될 수 있지만, 순수한 소프트웨어나 서비스로 제공되는 경우에는 현행법상 ‘제조 또는 가공된 동산’에 해당하는지부터 검토해야 한다. 또한 AI는 출시 이후 학습·업데이트·운영을 통하여 변화할 수 있으므로, 출시 당시의 결함과 사후 운영·모니터링 실패를 구별할 필요가 있다.

「환경오염피해 배상책임 및 구제에 관한 법률」은 일정 시설 사업자의 무과실책임, 인과관계 추정 및 환경책임보험을 결합하고, 「원자력 손해배상법」은 원자력사업자에게 무과실책임을 집중시키고 손해배상조치를 요구한다. 이러한 제도는 대규모 위험과 입증곤란을 분산시키는 민사적 모델이라는 점에서 참고가 된다.[35] 다만 현행 제조물 책임법상 AI 시스템이 곧바로 제조물에 해당하는지, 고영향 AI에 의무보험이나 보상기금을 도입할지는 별도의 해석·입법 문제이다. 본고는 이를 장기적 입법론으로 제안한다.

35) 「제조물 책임법」 제2조 제1호·제2호 및 제3조는 제조 또는 가공된 동산의 결함으로 인한 제조업자의 손해배상책임을 규정한다. 「환경오염피해 배상책임 및 구제에 관한 법률」 제6조, 제9조 및 제17조는 일정 시설 사업자의 무과실책임, 인과관계 추정 및 환경책임보험을 규정한다. 「원자력 손해배상법」 제3조 및 제5조부터 제7조까지는 원자력사업자의 무과실책임과 손해배상조치·책임보험을 규정한다. 순수 소프트웨어형 AI가 현행 제조물책임법상 제조물에 해당하는지와 고영향 AI에 보험·기금 모델을 도입할지는 별도의 해석·입법 문제이며, 본문은 이를 장기적 입법론으로 제안한다.

(4) 설명·기록의무와 입증곤란의 완화

AI 피해자에게 가장 큰 어려움은 앞서 지적한 정보비대칭이다(Ⅱ. 4. (3)). 피해자는 자신에게 불리한 판단이 내려졌다는 결과를 확인할 수 있을 뿐, 그 판단이 형성된 과정에는 접근할 수단을 갖지 못한다. 여기서는 그 정보비대칭이 형사·민사·행정 각 영역에서 어떻게 완화될 수 있는지를 살핀다.

따라서 제34조 제1항 제2호의 설명 방안과 제5호의 문서 작성·보관은 형사법과 민사법을 연결하는 절차적 장치가 될 수 있다. 형사법상으로는 사업자가 요구되는 설명·기록 체계를 마련하고 유지하였는지가 주의의무 이행 여부의 판단자료가 된다. 민사법상으로는 피해자가 인과관계를 주장·입증하기 위한 기초자료가 되고, 행정법상으로는 감독기관의 사후 검증을 가능하게 한다.

이러한 점에서 AI기본법상 투명성·설명·기록 관련 의무는 현행법상 일반적 설명청구권을 곧바로 부여하지는 않지만, 이용자와 피해자가 AI의 개입을 인식하고 사후적으로 판단과정을 검증할 수 있도록 하는 절차적 토대가 될 수 있다. 피해자의 직접적 정보접근권, 자료제출명령, 입증부담 조정은 별도의 입법적 설계를 필요로 한다. 이와 같이 한계를 분명히 할 때 AI기본법은 피해자 구제와 형사책임론을 연결하는 매개규범으로 기능할 수 있다.

Ⅳ. 결론

AI기본법은 AI 자체의 형사책임 문제를 해결한 법이 아니다. 이 법은 AI를 처벌하지 않고, AI를 개발·제공·이용하는 인공지능사업자에게 사전적·절차적 의무를 부과하는 방식을 선택하였다. 그러나 바로 그 점에서 AI기본법은 형사법적으로 중요하다. AI기본법상 사업자 의무는 향후 AI 관련 사고와 범죄에서 주의의무, 감독의무, 보증인지위, 조직책임을 판단하는 외부적 준거규범으로 기능할 수 있기 때문이다.

본고의 결론은 다섯 가지이다. 첫째, AI기본법상 투명성·고지·위험관리의무는 단순한 행정규제에 그치지 않고 형사법상 주의의무의 내용을 구체화하는 자료가 된다. 둘째, AI사업자 의무는 형식이행형 의무(formal-compliance obligation)와 위험관리형 주의의무(risk-based duty of care)로 구분하여 해석할 필요가 있다. 셋째, AI사업자 책임은 전통적인 자연인 중심 과실책임이나 대표자 동일시이론만으로 충분히 설명되기 어렵고, 조직 전체의 위험관리 실패를 평가하는

조직책임 모델로 보완될 필요가 있다. 다만 이 모델은 법인 자체의 일반적 형사책임을 창설하지 않으며, 법인 처벌에는 별도의 명시적 근거가 필요하다.

넷째, 인간-AI 협업범죄에서는 AI를 공범으로 의제하기보다 인간 행위자의 책임을 중심으로 하되, AI 활용이 범죄의 규모와 위험성을 어떻게 변화시켰는지를 구성요건 해석과 기존 양형요소의 평가에 반영해야 한다. 다섯째, AI기본법상 과태료 체계는 피해자 구제를 직접 수행하지 못하므로, 설명 방안과 문서 작성·보관, 손해배상, 의무보험, 보상기금 등 위험책임 모델과 결합되어야 한다.

결국 AI 형사법의 과제는 AI를 사람처럼 처벌하는 데 있지 않다. 핵심은 AI 시스템을 사회에 배포하고 운영하는 조직이 위험을 예견하고 통제할 수 있었는지, 그 통제의무를 위반하여 법익침해가 현실화되었는지, 그리고 피해자가 정보비대칭 속에서도 실질적으로 구제받을 수 있는지를 묻는 데 있다. AI기본법은 그 질문에 대한 완성된 답은 아니지만, 형사법이 참조해야 할 새로운 책임구조의 출발점이다.

[참고문헌]

강지현, 「4차 산업혁명시대, 법학의 역할 – 독자적 법인격 주체로서 인공지능 –」, 「한국부패학회보」 제27권 제1호, 한국부패학회, 2022.

강현석, 「중대재해처벌법위반죄와 산업안전보건법위반죄의 죄수관계에 관한 연구 – 대법원 2023. 12. 28. 선고 2023도12316 판결을 중심으로 –」, 「법이론실무연구」 제12권 제2호, 한국법이론실무학회, 2024.

김성돈, 「전통적 형법이론에 대한 인공지능 기술의 도전」, 「형사법연구」 제30권 제2호, 한국형사법학회, 2018.

김영환, 「위험사회에서의 책임구조: 자연 재해에 대한 법적 담론」, 「홍익법학」 제14권 제1호, 홍익대학교 법학연구소, 2013.

김종구, 「미국 형법상 중립적 행위에 의한 방조의 가벌성」, 「형사법연구」 제24권 제4호, 한국형사법학회, 2012.

박상철, 「인공지능 기본법의 시행 전 개정 필요성 – 규제 조항의 체계·축조상 문제점을 중심으로 –」, 「정보법학」 제28권 제3호, 한국정보법학회, 2024.

박준우, 「인공지능(AI)·딥페이크 생성물의 퍼블리시티권 침해 연구」, 「산업재산권」 제79호, 한국지식재산학회, 2024.

신혜진, 「인공지능(AI) 시대의 신종 범죄에 대한 형사법적 대응 방향 – 형사책임의 주체 및 범위를 중심으로 –」, 「형사법의 신동향」 제87호, 대검찰청, 2025.

안정빈, 「법인의 형사책임 – 동일시이론과 조직모델이론의 구별을 중심으로 –」, 「일감법학」 제44호, 건국대학교 법학연구소, 2019.

_____, 「보증인지위와 적극적·소극적 의무 개념 – 의무범 이론과의 연계성을 중심으로 –」, 「비교법연구」 제22권 제3호, 동국대학교 비교법문화연구원, 2022.

_____, 「부작위 행위자들 사이에서 정범과 공범의 구별」, 「형사법연구」 제31권 제2호, 한국형사법학회, 2019.

_____, 「의무범에서의 정범과 공범을 구별 짓는 기제로서의 '의무의 강약설' 및 사회적·규범적 판단에 대한 연구 – 부작위 경합 상황에서의 의무의 충돌 및 의무부과의 개념을 중심으로 –」, 「중앙법학」 제24집 제4호, 중앙법학회, 2022.

_____, 「인공지능 로봇에 관한 형사책임 – 인공지능형법에 의한 형사처벌가능성을 중심으로 –」, 「외법논집」 제46권 제4호, 한국외국어대학교 법학연구소, 2022.

이용식, 「일상적·중립적 행위와 방조 － 방조행위의 개념적 의미내용과 방조의 불법귀속의 이해구조 －」, 「법학」 제48권 제1호, 서울대학교 법학연구소, 2007.

임석순, 「형법상 인공지능의 책임귀속」, 「형사정책연구」 제27권 제4호, 한국형사정책연구원, 2016.

장승일, 「방조범의 불법구조와 중립적 방조행위의 가벌성판단」, 「법학논문집」 제40권 제1호, 중앙대학교 법학연구원, 2016.

정찬모, 「인공지능기본법 주요 내용의 검토와 향후 과제」, 「법학연구」 제28집 제1호, 인하대학교 법학연구소, 2025.

차상육, 「인공지능 창작물 관련 저작권 침해 쟁점」, 「경영법률」 제32집 제4호, 한국경영법률학회, 2022.

[Abstract]

Duties of AI Business Operators under the AI Basic Act and Organizational Responsibility under Criminal Law*

- Focusing on Responsibility Structure, Layers of Duties, and Victim Redress -

AHN JEONGBIN**

The Korean AI Basic Act does not recognize artificial intelligence itself as a subject of criminal responsibility. Instead, it adopts an administrative-law framework centered on business operators that develop, provide, or deploy AI systems. This legislative choice, however, is not devoid of criminal-law significance. Duties concerning the identification and management of high-impact AI, notice and labeling of generative AI outputs and realistic synthetic content, safety and trustworthiness, the establishment and implementation of explanatory procedures, and the preparation and retention of relevant documentation may serve as external normative benchmarks for negligence, omission liability, and organizational risk-management failures. This article integrates the responsibility structure of the AI Basic Act with the criminal-law theory of attribution. It distinguishes formal-compliance obligations, whose performance can be determined relatively clearly, from risk-based duties of care, whose content depends on the foreseeability, controllability, and avoidability of risk. These duties do not automatically establish criminal liability. Rather, they provide criteria for determining whether a risk was foreseeable, whether reasonable preventive measures were available, and whether an organization failed to manage AI-related risks. The organizational approach advanced here does not itself establish general criminal liability of corporations under current

---

* This article is based on the author's presentation, "Korea's AI Basic Act and Its Impact on Industry," delivered in Session I, "AI Basic Act, AI Governance and Global Industrial Policy Trends," at The 8th International Conference on AI and Law 2025, held in Taipei on November 8, 2025. It substantially expands and reconstructs that presentation by focusing on the criminal-law significance of AI business operators' duties and organizational responsibility.

** Associate Professor and Chair, Department of Law, Kyungnam University.

Korean law; corporate punishment still requires an express statutory basis, typically a statutory dual-punishment provision. The article further examines the limits of administrative sanctions, the requirements for aiding and abetting by providers, and the need to connect criminal-law attribution with victim redress. Ultimately, the task of criminal law in the age of the AI Basic Act is not to expand punishment, but to allocate responsibility accurately among developers, deployers, platforms, users, and corporate organizations while respecting the principles of culpability and legality.

Keywords : AI Basic Act, negligence, guarantor status, organizational responsibility, intensity of duty theory