「法과 政策」第24輯 第1號, 2018. 3. 30.
濟州大學校 法과政策硏究院

인공지능의 형사책임에 관한 소고*

A Study about Criminal Liability of Artificial Intelligence

황 만 성**
Hwang, Man-Seong

목 차

국문초록

인공지능이 이미 우리 생활 속에 다양하게 기능하고 있지만, 인공지능으로 인하여 발생한 법익침해의 결과에 대하여 인공지능에게 직접 형사책임을 묻는 것이 가능할 것인지가 문제된다. 이를 위해 전통적 형법이론의 측면에서 검토하는 것과 아울러 인공지능과 인간의 책임의 관계를 살펴보는 것이 필요하다고 할 것이다.

법학일반에 있어서 뿐만 아니라 형법학에 있어서도 행위책임을 묻기 위하여는 그 전제로 행위능력이 있을 것을 필요로 한다. 인공지능의 행위능력을 인정할 수 없다면 인공지능의 형사책임을 논하는 것도 무의미하므로, 인공지능의 행

논문접수일 : 2018.01.30.
심사완료일 : 2018.02.21.
게재확정일 : 2018.02.21.
* 이 논문은 2017학년도 원광대학교의 교비지원에 의해 수행됨
** 법학박사 · 원광대학교 법학전문대학원 부교수

위능력과 책임능력을 고찰하는 것은 논쟁의 핵심을 이룬다고 할 것이다. 이에 관하여 부정론과 긍정론이 있으나, 인공지능의 활동으로 인한 피해의 규모가 크고 반복적으로 이루어질 수 있다는 위험을 고려하면 사회적으로 용납할 수 없는 침해행위를 한 인공지능을 자연인과 다르다 해서 형사처벌로부터 무조건 자유로운 행위주체라고 하는 것은 적정하다고 할 수 없으며 인공지능의 반사회적 활동으로부터 사회를 방위해야 할 필요가 있고, 책임의 근거를 반사회적 위험성으로 이해하면 인공지능에게도 사회적 책임을 물을 수 있다고 할 것이다.

한편, 인공지능 로봇에 대해 책임비난이 가능하다고 하더라도 이들은 형사처벌의 의미를 이해할 수 없기 때문에 처벌이 무의미하다고 한다. 즉, 인공지능은 형사처벌의 의미를 이해할 수 없고 따라서 자신의 죄책과 자신에게 가해 진 처벌의 연관성을 깨닫지 못하며, 종래의 형벌이 가지고 있는 일반예방효과나 특별예방효과를 인공지능에게서는 기대할 수 없기 때문에 형벌능력을 인정하기는 어렵다고 할 것이다.

주제어 : 인공지능, 인공지능로봇, 형사책임, 범죄능력, 형벌능력

Ⅰ. 서 론

인공지능이 발달하여 일정한 수준에 이르게 되면 인간에게 해를 끼치는 것을 넘어서서 인류의 생존까지 위협할 것이라는 두려움은 수십년전부터 공상과학소설이나 영화에서 간접적으로 표출되고 있었다. 터미네이터에서의 '스카이넷', 아이 로봇의 '비키'를 비롯하여 매트릭스, 엑스맨 데이즈 오브 퓨처패스트 등 많은 영화에서 인공지능을 가진 컴퓨터에 의한 위협을 다루고 있다.

실제로도 최근에 세계적인 물리학자 스티븐 호킹 박사가 인공지능(AI)에 대해 또다시 경고하였는데, 그는 2017년 11월 6일 포르투갈 리스본에서 열린 한 기술 컨퍼런스에 참석하여, "AI는 인류 문명사의 최악의 사건이 될 수 있다"고 주장했다. 호킹 박사는 "이론적으로 컴퓨터는 인간의 지능을 모방함은 물론, 그것을 뛰어넘을 수도 있다. 효과적인 인공지능을 만드는 것은 우리 문명사에서

가장 큰 사건이 될 수도 있고, 최악의 사건이 될 수도 있다"면서 "인류가 과연 AI로부터 큰 도움을 받을지, 아니면 그것에 의해 옆으로 밀려나거나 파괴될지 아직 알 수 없다"고 전했다. 호킹 박사는 AI에 따른 재앙을 방지하기 위해 AI 제작사들이 모범 사례와 효율적인 관리 체계를 채택해야 한다고 촉구하며 특히 유럽에서 추진되고 있는 새로운 법안의 필요성을 강조하기도 했다.[1)]

인공지능은 외부환경을 인식할 수 있는 능력, 상황판단능력, 자율적 행위결정 능력 및 상호작용능력 등을 포함하고[2)] 스스로 학습할 수 있는 능력까지 확대되고 있다. 1950년대 초반 과학자들이 인공적인 두뇌의 가능성을 거론하기 시작한 이후, 인공지능이 학문적으로 연구되었고, 1980년대에 이르러서 반도체 기술의 발전으로 컴퓨터가 소형화, 고성능화되면서 인공지능의 하드웨어적 조건이 갖추어졌다.[3)] 이를 바탕으로 인식기술 및 센서의 발달, 인공신경망, 자연어 처리 등 소프트웨어 기술이 발전하면서 인공지능은 비약적으로 발달하였다.

어떤 과학자들은 2024년이면 언어를 번역하는 데 로봇이 인간을 앞지를 것이고, 2026년이면 인공지능이 인간보다 더 훌륭한 에세이를 쓸 수 있게 되고, 2051년에는 인간이 수행하는 모든 업무가 자동화될 것이라고 예측하기도 하였다.[4)]

인공지능이 이미 우리 생활 속에 다양하게 기능하고 있지만, 인공지능으로 인하여 발생한 법익침해의 결과에 있어서 인공지능의 형사책임문제를 논하는 것 자체에 대해 냉소적 시각을 지닌 입장도 있으나, 독일에서는 이미 10여년 전부터 소위 '로봇 형법(Strafrecht für Roboter)'에 대한 논의가 시작되었고[5)], 우리나라에서는 최근에야 이에 관련된 논의가 활발히 이루어지고 있다.[6)]

---

1) http://www.fnnews.com/news/201711071527028838

2) 윤지영 외, 법과학을 적용한 형사사법의 선진화 방안(Ⅵ), 한국형사정책연구원, 2015, 21면.

3) 이인영, "인공지능로봇에 관한 형사책임과 책임주의", 홍익법학 제18권 제2호, 2017, 33면.

4) http://www.koreatimes.co.kr/www/tech/2017/06/133_231554.html

5) 김영환, "로봇 형법(Strafrecht für Roboter)?," 법철학연구 제19권 제3호, 2016, 148면.

6) 이에 관한 대표적 논의로는, 김영환, "로봇 형법(Strafrecht für Roboter)?," 법철학연구 제19권 제3호, 2016; 안성조, "인공지능 로봇의 형사책임", 법철학연구 제20권 제2호, 2017; 장연화·백경희, "왓슨의 진단조력에 대한 현행법상 형사책임에 관한 소고," 형사법의 신동향 제55호, 2017; 이인영, "인공지능 로봇에 관한 형사책임과 책임주의," 홍익법학 제18권 제2호, 2017; 이원상, "4차 산업혁명에 있어 형법의 도전과제," 조선대 법학논총 제24권 제1호,2017; 송승현, "트랜스휴먼 및 포스트휴먼 그리고 안드로이드(로봇)에 대한 형법상 범죄주체의 인정여부," 홍익법학 제17권 제3호, 2016; 이주희, "인공지능과 법-지능형 로봇 및 운영자의 형사책임에 관한

관련된 논문 등에서는 그 대상으로서 '인공지능 로봇', '자율주행 자동차', '로봇' 등 다양한 유형을 다루고 있으나, 이러한 유형들의 본질적 요소는 인공지능 그 자체이므로 이 글에서는 '인공지능'을 중심으로 논의하고자 한다.[7)]

이 글에서는 인공지능의 일정한 작용 결과에 대하여 인공지능에게 직접 형사책임을 묻는 것이 가능할 것인지를 전통적 형법이론의 측면에서 검토하는 것과 아울러 인공지능과 인간의 책임의 관계에 대하여도 다루기로 한다.

인공지능의 형사책임을 논함에 있어서는, 우선 인공지능의 작동을 행위론의 관점에서 '행위'로 이해할 수 있을지, 더 나아가 범죄능력을 가진 주체성을 인정할 수 있을지를 살펴보고, 인공지능에 대한 형벌이라고 할 수 있는 여러 가지 강제조치의 가능성과 그에 대한 인공지능의 감수능력이라는 측면에서의 형벌능력의 여부를 검토하기로 한다. 또한 구체적인 상황에서의 인공지능의 형사책임과 관련하여, 인공지능, 그 설계자, 이용자 등 누구에게 형사책임을 인정할 수 있을 것인가의 논의가 필요해 보이며, 이와 관련하여 형법이론상 과실책임의 제한원리를 원용할 수 있을지, 구체적인 상황에서의 형사책임을 누구에게 인정할 수 있을지에 관하여 살펴보기로 한다.

Ⅱ. 인공지능의 행위와 범죄주체

1. 행위론 관점에서의 인공지능의 활동

전통적 형법이론에 의하면 범죄를 구성하는 첫 번째 요소는 『행위』이다. 또

고찰," 한국사회과학연구 제38권 제1호, 2016; 임석순, "형법상 인공지능의 책임귀속," 형사정책연구 제27권 제4호, 2016; 정정원, "인공지능(AI)의 발달에 따른 형법적 논의," 과학기술과 법 제7권 제1호, 2016; 윤지영·윤정숙·임석순·김대식·김영환·오영근, 법과학을 적용한 형사사법의 선진화방안(VI), 연구총서 15-B-16, 한국형사정책연구원, 2015 등

7) 인공지능의 실체성에 관하여 의문을 제기할 수도 있을 것인데, 컴퓨터 메모리나 하드디스크의 일정 영역을 차지하고 있기는 하지만, 네트워크 등을 통하여 이전이 가능하고 인터넷이 연결되어 있는 상황에서 스스로 네트워크상에서 옮겨다니는 것이 가능할 것이기 때문에 실체를 확인하거나 확보할 수 없는 인공지능 자체(프로그램, 실체가 없는 존재)에 대한 형사책임 여부를 논하는 것은 이 논문에서 주로 다루는 내용과는 별도의 논의가 필요해 보이므로 이 글에서는 다루지 않기로 한다.

한 형법적 심사의 대상이 되는 것은 '인간'의 행위에 한정되는 것이라고 보는 것이 오랜 형법의 기초적인 관념이었다.[8] 구체적으로 행위를 정의함에 있어서는 인과적 행위론, 목적적 행위론, 인격적 행위론, 사회적 행위론 등의 논의가 있다. 인공지능의 형법적 주체성을 논함에 있어서 제일 먼저 넘어야 할 관문이 바로 인공지능에 대하여 행위주체성을 인정할 수 있을지의 문제라고 할 수 있다. 즉, 인공지능의 실행이나 작동을 형법적으로 의미 있는 『행위』라고 볼 수 있는지가 검토되어야 할 것이다.

인과적 행위론에서는 행위를 "의사에 의한 신체 활동", 또는 "유의적 거동에 의한 외부세계의 변화"라고 한다.[9] 학습을 통해 다양한 상황에 대처할 수 있는 인공지능의 반응은 일련의 프로세스를 거친 "의지적" 판단이라고 볼 수 있다는 점에서는 인공지능은 형법상 의미 있는 행위주체가 될 수도 있다는 견해가 있다.[10]

한편, 목적적 행위론에서는 인간행위의 본질적 요소는 '목적활동성의 작용'이며, 인간행위는 언제나 일정한 '목적성'을 가진다고 주장한다. 이에 따르면 행위자의 목적지향적 의사의 표출이 형법적으로 의미있는 행위라고 이해한다.[11] 학습능력이 있는 인공지능은 상황에 따라 자신의 임무를 가장 잘 수행하기 위하여 독자적으로 결정하기 때문에 목적적 행위론에서도 인공지능의 행위주체성을 인정할 여지가 있다는 것이다.[12]

한편, 사회적 행위론에서는 '사회적 의미성' 또는 '사회적 중요성'의 기준이 평가적인 의미에서 전면에 등장한다. '객관적으로 예견가능한 사회적 결과를 지향하는 객관적으로 지배가능한 일체의 행태'라고 이해하는 마이호프(W. Maihofer)의 사회적 행위개념은 가장 순수하게 사회적 의미성만을 가지고 행위를 파악하기 때문에[13] 인공지능의 사회적 행위성을 인정함에 다소 유연한 입장일 것으로

8) 중세에는 인간이 아닌 동물에 대하여 형사책임을 묻는 재판이 실제로 있기도 하였다. 1474년 스위스 바젤에서는 닭에게 사형이 선고된 바 있다. Gleß/Weigend, Intelligent Agenten und das Strafrecht, ZStW 126(2014), 566. 임석순, 앞의 글, 78면.

9) 김일수, 한국형법 I [총론 상] 개정판, 1996, 253쪽; 배종대, 형법총론 제12판, 2016, 37/1; 이재상, 형법총론 제7판, 2011, 6/7 참조.

10) Gleß/Weigend, 앞의 글, 571.

11) 김일수, 앞의 책, 254쪽; 배종대, 앞의 책, 37/15; 이재상, 앞의 책, 6/9 참조.

12) Gleß/Weigend, 앞의 글, 572.

13) 배종대, 앞의 책, 139면.

이해된다.

사회적 행위론은 다른 행위론과는 달리 '사회적 의미성'이라는 행위 외부의 제3자적 평가관점에서 행위를 바라보기 때문에 인공지능과 인간과의 근본적인 차이를 염두에 두지 않고 그 사회적 의미만을 기준으로 판단한다는 점에서 인공지능의 행위성을 인정하기가 수월해 보인다.

2. 인공지능의 범죄능력과 범죄주체성

법학일반에 있어서 뿐만 아니라 형법학에 있어서도 행위책임을 묻기 위하여는 그 전제로 행위능력이 있을 것을 필요로 한다. 인공지능의 행위능력을 인정할 수 없다면 인공지능의 형사책임을 논하는 것도 무의미하므로, 인공지능의 행위능력과 책임능력을 고찰하는 것은 논쟁의 핵심을 이룬다고 할 것이다. 인공지능의 행위능력에 관하여는 종래 형법이론에 바탕을 두고 부정해야 한다는 입장과 인공지능에 대하여도 일정한 영역에서는 인간과 마찬가지로 행위능력을 인정할 수 있다는 입장이 대비된다.

가. 인공지능의 범죄능력 부정론

인공지능의 형사책임을 부정하는 이유로 가장 먼저 제시되는 것은 인공지능은 '인간성(personhood)'의 요건을 충족시키지 못한다는 점이다. 철학적 논의에 의하면 인간성의 의미는 자기성찰능력(self-reflection capacity) 또는 자의식(self-consciousness)이 있어야 하며 이러한 조건을 갖춘 인격체만이 법의 수범자가 될 자격이 있다는 것이다. 인공지능이 스스로 학습하고 결정을 내릴 수 있지만, 인공지능은 자의식이 없고, 자신의 의지에 의해 행동을 선택할 수 있는 자유의지도 없으며, 따라서 자신의 행동에 대한 책임을 질 수 없으며, 자기성찰능력의 결여로 인하여 하나의 인격체로서의 행위주체가 될 수 없다는 것이다. 요컨대 인공지능은 자유의지에서 비롯되는 양심이 없기 때문에 자신의 행위의 선악을 판단할 수 없고, 그 자신이 범한 해악에 대해 '인격적으로(personally)' 책임을 질 수 없다고 한다.[14)]

다른 한편으로, 인공지능은 인지능력이 있고 자율적으로 행동할 수 있지만, 도덕적 판단능력이 없기 때문에 법적 책임과 의무의 주체가 될 수 없고, 설령 인공지능 로봇에게 법인의 형사책임을 인정한 것과 유사하게 '전자적 인격(elektronische Person)'을 부여하여 '자연적 차원'이 아닌 '규범적 차원'에서 행위능력과 책임능력을 인정할 수 있다 하더라도 만일 이를 '형사처벌'하게 될 경우 '인간의 존엄성 원리'로부터 도출되는 책임원칙의 핵심내용인 '비난가능성'이라는 요소를 빼어 버리는 것이 되어 형벌의 진지함과 도덕적 요소가 사라져 '형벌의 존엄성'이 훼손될 수 있다는 것이다.[15]

인공지능에게 형법적 책임을 지우는 것은 그것이 "도덕적 인공지능"이 되지 않는 한 적어도 오늘날의 형법관 내지 법철학에 의하면 요원하다고 볼 수밖에 없다고 보는 견해[16]도 있다. 형법규범에서 "~한 자"를 구성요건실현의 주체로 규정하고 있는 것은 형벌규범의 수범자이자 국가형벌권의 대상을 오로지 종(種)으로서 인간으로 한정한 것이기 때문에 인공지능을 형벌규범의 수범자로 받아들일 수 없다는 것이다.[17] 이러한 입장에서는 "책임비난의 내적 근거는 인간이 자유롭고, 자신의 책임에 따른 것이며, 도덕적인 자기결정을 의도하였으며, 따라서 스스로 법에 따르고 불법에 저항하도록 결정하고, 법적 당위규범에 따라 자신의 행동을 조정하고 법적으로 금지된 것을 회피할 능력이 있다는 점에 있다"고 한 독일연방법원의 판례[18]는 여전히 유효하다고 한다.[19]

또한 책임능력은 자유의사를 가진 자연인을 전제로 하여 규정되어 있고, 책임은 이러한 자유의사를 전제로 하여 '다르게 행위할 수 있었거나' 즉 '적법하게 행위할 수 있다'는 점이 인정될 경우 그에 대한 비난가능성이 인정되므로, 현재 기술수준의 인공지능은 스스로를 과거와 미래를 가진 개체로서 이해할 수

14) 안성조, 앞의 글, 81면, Sabine Gless, Emily Silverman, & Thomas Weigend, "If Robots Cause Harm, Who Is To Blame? Self-Driving Cars and Criminal Liability," 19 New Crim. L. Rev. 412 (2016), 416면 이하 참조.

15) 김영환, "로봇 형법(Strafrecht für Roboter)?," 법철학연구 제19권 제3호, 2016, 151-160면.

16) Gleß/Weigend, 앞의 글, 589.

17) Ziemann, Zur Diskussion über ein Strafrecht für Maschinen, 2013, S. 184-185; Joerden, Strafrechtliche Perspektiven der Robotik, 2013, S. 203.

18) BGHSt 2, 194, 200 = BGH NJW 1952, 593, 594.

19) 임석순, "형법상 인공지능의 책임귀속," 형사정책연구 제27권 제4호, 2016, 76-77면.

없고, 권리·의무의 귀속주체임을 인식하지 못한다는 점에서 자유의지를 가진 행위자로 볼 수 없다는 견해도 같은 입장이라고 할 수 있다.[20]

이러한 입장에서는 현재 기술수준의 소위 '약'인공지능에 대해서는 직접적인 형사책임을 부과할 수 없으며, 그 배후에 있는 자연인 행위자(human behind the machine)에 대해서 형사책임을 부과할 수 있는 방법을 모색해야 한다는 것이다.[21]

나. 인공지능의 범죄능력 긍정론

인공지능에게 범죄능력이 있다고 보는 입장의 핵심적인 논지는 인공지능이 작동되는 체계가 인간의 사고-행동체계와 유사하다는 것이다.

인간성을 지닌 존재의 필수적 자격인 '지적 존재(intelligent entity)'로서의 요건은 크게 다섯 가지인데, 다른 지적 존재와의 '의사소통(communication)', 자기 자신에 대한 '내적 인식(internal knowledge)', '외부 환경을 인식하고 학습하고 정보를 이용할 수 있는 능력', '목표 지향적 행위(goal driven behaviour)' 및 '원래의 조치가 실패했을 때 다른 대안조치를 취할 수 있는 창조적(creative)' 능력 등이 그것이다.[22]

이러한 인간 존재의 요건이 범죄능력이나 형사책임을 묻는데 모두 필요한 것은 아니며, 범죄의 주관적 요소로 필요한 인식(knowledge), 의도(intention) 및 과실(negligence) 등이 있으면 범죄능력을 인정할 수 있다고 본다.

대부분의 인공지능에게 있어서 인식은 가장 흔한 기능으로, 수용된 사실적 자료를 중앙처리장치(central processing units)에서 분석하는 과정은 바로 인간의 이해과정을 닮아 있기 때문에 인공지능도 인식능력을 갖추고 있다고 볼 수 있으며[23], 인공지능은 특정한 목표를 성취하기 위해 프로그램되어 목적달성을 위하기 적절히 반응하도록 되어있다는 점에서 범죄의 주관적 요소인 '의도'도

20) Sabine Gless 외, 앞의 글, 417면.
21) 안성조, 앞의 글, 82면.
22) 안성조, 앞의 글, 87면.
23) 안성조, 앞의 글, 88면.

갖추고 있다는 것이다.[24] 이와 같이 인공지능에게 범죄의 주관적 성립요소인 인식과 의도를 인정할 수 있고, 범죄의 소극적(negative) 요소인 정당방위의 법리도 같은 이유로 적용될 수 있다고 한다.[25]

'약한 인공지능(weak AI)'과 '강한 인공지능(strong AI)'을 구분[26]하여 논의하는 다른 입장에서는 미래의 강인공지능에게 관하여 형사책임을 더욱 쉽게 인정하게 될 것이라고 한다. 강인공지능은 인간의 이성적 사고과정과 거의 다를 바 없는 '마음을 지닌' 자율적 행위주체로서의 능력을 구비할 것이고, 도덕적 추론능력까지도 갖춘 '인공적 도덕 행위자(AMA: artificial moral agent)'의 수준에 도달 할 것으로 전망되므로[27] 인공지능 로봇을 형법상 행위주체로 인정하는 데 문제될 것이 없다고 한다.[28]

또한, 인공지능은 인간의 삶에 매우 중요한 부분을 구성하며 단순하고 간단한 도구로부터 점차 많은 생활영역의 정교한 영역까지 침투해 오고 있는 것과 아울러 인간에 대한 다양한 법익침해 상황을 초래할 수 있는데, 인공지능을 형법의 적용범위 밖에 둔다면 처벌의 공백이 생기고[29] 또 다른 사회적 위협의 대상이 된다는 것이다.[30] 실용적인 측면에서 인공지능에 대한 형사적 규율의 필요성이 엄연히 존재한다는 견해도 넓은 의미에서 긍정론의 한 논거로 원용될 수 있을 것이다.

오늘날 민사법 영역에서는 행위능력의 전제로서 권리능력의 인정에 있어서도 불변의 원칙이 고수되는 것은 아니다. 태아에게 제한된 권리능력을 인정하는 것이나, 권리능력을 갖는 주체의 범위가 확장되고 있는 것에 비추어 볼 때, 형사

24) *Id.*

25) Gabriel Hallevy, op.cit., at 186-188, 192-193.

26) 강한 인공지능(strong AI)과 약한 인공지능(weak AI)란 구분은 본래 철학자 존 설(John Searle)에서 유래된 것이다. 그는 '강한 AI'란 '정확한 입력과 출력을 갖추고 적절하게 프로그램된 컴퓨터는 인간이 마음을 가지는 것과 완전히 같은 의미로 마음을 가진 것'이고, '약한 AI'는 '반드시 마음을 지닐 필요는 없고 한정된 지능에 의해서 지적 문제를 해결할 수 있는 인공지능'이라고 구별한다. 자세한 내용은 마쓰오 유타카/박기원 역, 인공지능과 딥러닝(동아 엠앤비, 2016), 참조.

27) 송승현, 앞의 논문, 491면-512면.

28) 송승현, 앞의 논문, 518면 이하.

29) 처벌의 공백 문제에 대한 논의로는 김영환, 앞의 논문, 162-163면 참조.

30) 안성조, 앞의 글, 86면.

책임의 주체도 확장될 여지가 있다 할 것이다.[31] 이러한 점에서 '행위'나 '책임' 등의 법률적 기본개념도 원칙적으로 인공지능에게 적용할 수 있다는 주장도 제기되고 있다.[32]

최근 인간의 뇌와 인공지능에 대한 연구가 활발해지면서 그에 관한 새로운 지식과 함께 책임개념의 변화가능성도 있다고 할 수 있는데, 인간의 필수적 속성과 능력에 대한 어떠한 형상이 형법상 유죄판단 가능성의 근거가 되는지를 명확하게 확정하는 것 자체가 애초부터 불가능하기 때문에 권리, 유책성(Verantwortlichkeit), 책임(Schuld) 등과 같은 개념을 근본에서부터 되묻는 것도 가능할 것이다.[33] 권리주체는 더 이상 자신 스스로 정한 행위와 의사표명으로 특정할 수 있는 사람인 개인에 국한될 수 없으며, 특히 정보화 시대에 사람은 ID와 같은 인위적 속성으로도 정체성을 규명할 수 있을 것이라는 견해도 조심스럽게 제기되고 있는데, 이러한 입장에 대하여는 정보기술 시스템이 한편으로는 인격발현의 새로운 영역을, 다른 한편으로는 새로운 사회적 행위를 형성하고 있기 때문에 일정한 조건 하에서 비신체적이고 살아있지 않으며 정보기술적인 인공물에 인격을 부여하는 것으로 이해할 수 있을 것이다.[34] 인공지능의 알고리즘이 기능적으로 인간의 의사결정구조와 일치한다면 인공지능에 형법적 책임비난을 가하는 것도 가능할 것이라는 견해도 있다.[35]

다. 소결

인공지능의 형사책임을 인정할 것인가에 관한 논의는 법인의 형사책임을 가부를 논하는 것과 몇가지 측면에서 교차하는 부분이 있다고 생각된다. 법인의 형사책임을 논함에 있어 법인에게 귀속되는 행위를 구체적으로 한 것은 대표자 등의 자연인이라는 점에서 인공지능에서의 논의와는 근본적인 차이점이 있지만,

31) 김영환 "위험사회에서의 책임구조: 자연 재해에 대한 법적 담론", 홍익법학 제14권 제1호, 2013, 825면.
32) Hilgendorf, 앞의 글, S. 122, 125.
33) 임석순, "형법상 인공지능의 책임귀속", 형사정책연구, 제27권 제4호(2016), 76면.
34) Gruber, 앞의 글, S. 157.
35) Gleß/Weigend, 앞의 글, 576.

법인 고유의 형사책임을 인정하여야 하는가의 문제는 법인에게 자기정체성이 인정되는가에 의하여서가 아니라, 입법자가 효과적인 법익보호를 위하여 법인 자체를 형사처벌의 대상으로 볼 것인지 여부를 결정해야 할 성격이 핵심이라는 점[36]에서는 인공지능의 형사책임과 접점이 있다고 할 것이다. 사회적으로 용납할 수 없는 침해행위를 한 법인을 자연인과 다르다 해서 형사처벌로부터 무조건 자유로운 행위주체라고 하는 것은 법인의 활동으로 인한 피해의 규모가 크고 반복적으로 이루어지고 있다는 현실에 비추어보면 적정하다고 할 수 없으며 법인의 반사회적 활동으로부터 사회를 방위해야 할 필요가 있고, 책임의 근거를 반사회적 위험성으로 이해하면 법인에게도 사회적 책임을 물을 수 있다는 견해[37]에 동조하며 이러한 논거는 인공지능의 범죄능력을 인정함에 있어서도 원용할 수 있을 것으로 생각된다.

한편, 부정하는 입장에서 부가하고 있는 단서는 '현재의 기술수준에서의' '약한' 수준의 인공지능에 대하여는 범죄능력을 인정할 수 없다는 것이다. 이러한 입장은 강 인공지능의 단계에서는 의사결정과 행동 이행의 과정이 인간의 뇌 기능과 유사하다는 면에서 범죄능력을 인정할 수 있다는 점을 애써 덮고 있는 것으로 여겨진다.

현재의 인공지능의 수준은 급속한 발달과정을 거치고 있으며, '딥 러닝'의 단계를 지나 더 고차원적인 사고 알고리즘을 시도하고 있다는 점에서[38] 약 인공지능과 강 인공지능의 구별을 모호하게 만드는 것과 동시에 인공지능의 책임문제를 기술수준을 핑계로 회피할 수 없는 수준에 이른 것으로 생각된다. 다만

36) 김호기, 법익적대적 법인문화, 위험관리 실패와 법인의 형사책임, 형사정책연구 제22권 제4호, 2011, 16면.

37) 이인영, 앞의 글, 43면.

38) 자의식을 지니고 있는지 여부를 확인하는 가장 기본적인 증명방법은 찰스 다윈이 최초로 시도한 것으로 알려진 '거울테스트(MSR; mirror self-recognition test)'라고 할 수 있는데, 거의 대부분의 동 물은 거울 앞에 서면 거울에 비친 영상을 다른 동물로 인식한다고 한다. 그런데 2012년 예일대에서 제작한 로봇 '니코(Nico)'는 이 테스트를 통과했다고 한다.(니코 연구사례에 대한 상세한 내용은 웬델 월러치·콜린 알렌/노태복 역, 왜 로봇의 도덕인가(메디치, 2014), 281-282면 참조). 한편, 버클리 대학의 저명한 이론 물리학자인 미치오 가쿠는 인공지능이 향후 인간과 비슷한 감정을 지닐 수 있고 따라서 인간의 감정을 읽을 수 있도록 프로그램될 수 있고 고통을 느낄 수 있도록 프로그램될 것이라고 전망한다. 미치오 가쿠/박병철 역, 마음의 미래(김영사, 2014), 334면 이하. 참조

장차 자아를 인식한 강한 인공지능이 출현면, 인공지능에게 인간도덕에 따른 규칙체계를 이식할 수 있는지 또는 해야 하는지, 이식한다면 어느 수준으로 할 수 있는지 또는 해야 하는지 하는 새로운 문제가 야기될 것이고[39] 이에 관한 논의는 별도로 필요할 것으로 생각된다.

Ⅲ. 형벌관점에서의 인공지능의 지위

1. 인공지능의 형벌능력

독일형법학계에서 이른바 로봇형법론을 다루는 에릭 힐겐도르프(Eric Hilgendorf) 교수[40]는 필요에 의해 "법인" 개념이 만들어진 것과 마찬가지로 독립적 책임주체로서 이른바 "e-인격(e-Person)" 개념을 도입할 것을 주장하고 있다.[41] 이러한 착안은 법인에 대한 형사처벌이 존재하지 않는 독일에서도 반향을 불러일으키고 있다.

그러나 법인과 인공지능은 행위의 발현양태가 다르며, 법인은 재산권의 주체가 될 수 있는 반면에 인공지능에게는 재산권을 인정하기 어렵기 때문에 법인의 형벌(적응)능력과 같은 수준에서 논의하기는 어렵다고 할 것이다. 법인의 행위는 그 구성원인 자연인인 인간의 의사결정에 의해 이루어지는 것이므로 인공지능의 행위(실행)와는 구별된다고 할 것이다. 뿐만 아니라 현행법상 법인에 대한 처벌은 대부분 법인이 소유한 재산에 대한 제재인데, 현행 재산법상 인공지능은 법인과 달리 재산을 소유할 수 없다. 인공지능은 자산을 소유하고 있지도 않고, 소유하고 있다는 인식도 없기 때문에 벌금형의 의미를 지닐 수 없다고 할 것이다. 인공지능에게는 물리적인 손상이나 파괴를 가하더라도 신체의 완전성을 유지하며 삶을 유지하려는 의지가 없기 때문에 사형이나 자유형이 의미가

39) Gleß/Weigend, 앞의 글, 577.

40) 힐겐도르프교수는 2010년 독일 뷔르츠부르크 대학교에 로봇법 연구소를 설립하여 인공지능 로봇과 관련된 다양한 법적 문제들을 연구하고 있다(윤지영 외, 앞의 보고서, 154-155쪽 참조).

41) Hilgendorf, Recht, Maschinen und die Idee des Posthumanen, Telepolis, 25.05.2014.; Fitzi, Roboter als "legale Personen" mit begrenzter Haftung, 2013, S. 391-394.

없으며, 사형이나 자유형과 같은 형사처벌은 인공지능의 해체나 파괴, 프로그램 재설치, 가동중지 등으로 대체될 수 있을 것으로 보이지만, 이를 굳이 "형벌"이라고 해야 하는지에 대하여는 의문이 든다.

또한 인공지능 로봇에 대해 책임비난이 가능하다고 하더라도 이들은 형사처벌의 의미를 이해할 수 없기 때문에 처벌이 무의미하다고 한다. 즉, 인공지능은 형사처벌의 의미를 이해할 수 없고 따라서 자신의 죄책과 자신에게 가해 진 처벌의 연관성을 깨닫지 못하며,[42] 종래의 형벌이 가지고 있는 일반예방효과나 특별예방효과를 인공지능에게서는 기대할 수 없기 때문에 형벌능력을 인정하기는 어렵다고 할 것이다.

2. 강한 인공지능 로봇에 대한 양벌규정의 수용 여부

특정 존재가 우리 사회에 중대한 위협이 된다면 입법자는 결단에 의해 그 특정한 존재에게 사회적 책임의 이행을 요구할 수 있을 것이며, 가까운 미래에 인공지능 로봇이 그 자신의 축적된 경험 또는 지식에 기초를 두고 외부의 자극에 의해 범의를 유발하여 스스로 범행계획과 진행경과를 통제해 나가는 단계로 발전한다면 이때 인공지능 로봇에 대하여도 강제조치를 수반한 사회적 책임을 물을 수 있을 것이다.

처벌이 요구되는 주체는 범죄행위를 계획하고, 실현하여 우리의 삶을 근본적으로 침해할 수 있는 활동주체이면서 책임을 질 수 있는 자이면 되며, 미리부터 자연인에게만 한정해야 한다는 가치규범이 존재하는 것은 아니다.[43] 인간의 노동이 미래에 거의 대부분 자율적인 인공지능 로봇으로 대체되고 로봇의 활동으로 인한 사고발생이 충분히 예상되고 피해가 폭넓게 일어난다면, 그러한 반복적인 위반행위의 발생을 방지하여야 한다는 점에서 단지 인간의 행위 즉 자연인이 아니라는 근거만으로 처벌의 대상에서 배제될 수는 없다. 인공지능이 초래

42) Sabine Gless, Emily Silverman, & Thomas Weigend, op.cit., at 416-425. 유사한 맥락에서 현 단계에서는 인공지능의 형사책임 인정이 어렵다는 견해로 보이는 이주희, 앞의 논문, 136면; 정정원, 앞의 논문, 202면.

43) 허일태, 위험사회에 있어서 형법의 임무, 비교형사법연구 제5권 제2호, 2003, 19-20면.

할 수 있는 위험의 성격, 사회적 위험으로서의 분배, 책임의 귀속주체 등에 관한 새로운 법정책이 필요할 수 있을 것이다.[44]

인공지능과 관련된 책임의 문제도 단순히 누가 책임의 주체인가 또는 면책할 수 있는가의 이분법적 접근이 아니라 개발자와 사용자의 다양한 형태의 지위와 그에 따른 책임도 다양화될 수 있다는 것을 고려하여야 한다고 한다. 또한 인공지능에게도 또 다른 형태의 사회적 책임과 그에 따른 양벌규정을 적용될 수 있을 것이다.[45]

만약 인공지능 로봇에게 양벌규정에 의한 형사책임을 인정할 수 있다면 그에 부합하는 적정한 형벌이 무엇인지, 책임주의의 원칙에 부합하는지 여부의 논의가 필요할 것이다. 인공지능 로봇의 재프로그래밍 또는 폐기 등이 인간에게 부과되는 사형과 유사한 형으로 대체될 수 있다는 주장[46]이 있지만, 재프로그래밍이나 폐기가 인공지능에게 실효적인 수단인지, 판결을 실제 선고할 수 있는지 그리고 어떻게 형벌을 실행할 수 있는지 등에 관한 논의가 선행되어야 할 것이다.

Ⅳ. 인공지능 실행 결과에 대한 책임귀속

1. 인공지능 실행의 위험과 사회적 용인 여부

우리 사회가 인공지능의 사용으로 인해 어느 정도 피해가 발생할 수 있다는 것을 받아들인다면, 결과적으로 침해가 발생하였다 할지라도 인공지능의 생산자나 운영자, 설계자를 형사처벌해서는 안 된다는 의견이 있다.[47] 사회가 혁신적 기술을 이용함으로 인해 일정한 이익을 보고 있음에도 그로부터 불가피하게 발생할 수 있는 위험을 온전히 생산자나 운영자에게 전가하는 것은 공평하지 않다는 것이다.[48]

---

44) 이인영, 앞의 글, 46-47면
45) 이인영, 47면.
46) Gabriel Hallevy, at. 30-31
47) Kuhlen, 앞의 글, Rn. 27 ff.

이러한 논지에서 인공지능을 이용하는 것에 대한 사회적 공감대가 형성된다면, 주의의무를 경감함으로써 인공지능의 적극적 개발 및 활용 가능성을 열어둘 수 있다는 것이다. 다만 그 주의의무의 경감을 어느 정도로 정할 것인지, 또는 활용가능성 극대화와 요구되는 주의의무 정도 사이의 적절한 접점을 어디로 결정할 것인지는 등의 문제는 남아있다.

현재 논의될 수 있는 적절한 접점으로는 우선 판매자에게 포괄적 경고의무 및 안내의무를 부과하고, 생산자에게는 사후 경과를 지속적으로 관찰하게 하고 피해가 발생한 경우 피해보고를 접수하여 신속하고 적절한 피드백을 이행할 의무를 부과하는 것을 고려해 볼 수 있다.[49] 이러한 방식의 주의의무의 조정은 설계·제작·판매 단계에서는 주의의무를 감경하는 한편, 사후 관리단계에서는 주의의무를 강화하는 형태라고 할 것이다.

구체적으로 생산자나 운영자, 설계자가 법적으로 의무화된 조치를 전혀 취하지 않으면 의무불이행에 대한 형사처벌할 수 있으며, 지배할 수 없는 인공지능을 설계·제작·판매함으로써 예측할 수 없는 위험을 야기하거나 증대시켰음에도 사후적으로 인지된 위험으로부터 야기될 수 있는 결과를 예방하거나 발생한 결과를 제거할 의무를 이행하지 않은 경우에는 형사처벌한다는 것이다.[50] 이러한 대응은 형사법적 제조물책임의 범주로 이해하는 입장에서 쉽게 찾을 수 있다.[51]

다른 한편, 인공지능은 주변환경으로부터 습득한 정보를 스스로 분석하고 그 결과에 따라 자신에게 부여된 임무를 최적의 방법으로 수행하려고 할 것이다. 사용자가 복잡한 인공지능의 반응을 관찰한다 할지라도, 인공지능이 어떠한 방식으로 데이터를 인식하고, 어떻게 해석하며, 어떻게 반응할지를 전부 예측할 수는 없다. 사용자뿐만 아니라 생산자나 관리자도 인공지능이 활용되는 상황에서 발생할 수 있는 모든 상황을 예측할 수는 없으며, 예측할 수 없는 위험을 회피

48) 김호기, "개발위험의 항변과 형법적 제조물책임", 형사정책연구 제27권 제1호, 2016, 192쪽; Gleß/Weigend, 앞의 글, 583.

49) 임석순, 앞의 글, 81면.

50) 임석순, 앞의 글, 81-82면.

51) 형법상 제조물책임의 법리에 대한 연구로는 전지연, "형법상 제 조물책임에서 주의의무위반에 대한 비교법적 고찰," 연세대학교 법학연구 제18권 제4호, 2008; 김호기, "개발위험의 항변과 형법적 제조물책임", 형사정책연구 제27권 제1호, 2016 등이 있다.

하도록 사전에 완벽하게 프로그래밍하는 것도 불가능하다. 또한 인공지능 이 독자적으로 학습을 하고 이에 따라 행동한다면 인공지능의 판단과 행동을 예측하기는 더욱 어려워진다.

인공지능이 실행된 결과 발생한 불법과 관련해서는 상반된 논의가 가능할 것이다. 인공지능의 사용자에게 책임을 물을 수 있는가라는 문제에 관하여, 어떤 입장은 인공지능이 행한 불법적이거나 유해한 행동은 인공지능의 독자적인 정보처리를 근거로 한 것으로 사용자가 예측하거나 회피할 수 없는 것이기 때문에 사용자에게는 그 책임을 귀속시킬 수 없다고 한다.[52]

다른 입장은 인공지능의 사용자는 인공지능의 소유자이기 때문에 인공지능의 활동으로 인해 발생한 "모든 것"에 책임을 져야 한다는 것이다. 즉 예측가능성 및 회피가능성과 무관하게 인공지능의 행동으로 인해 발생한 모든 피해에 (무과실)책임을 져야 한다는 것이다.[53] 사용자가 완벽하게 지배할 수 없는 인공지능을 출시한 자는 인공지능의 행동에 대한 예측불가능성을 근거로 인공지능의 오작용(Fehlreaktion)의 책임을 부인해서는 안 된다는 주장[54]도 이러한 입장이라고 할 수 있다..

2. 예측곤란성과 과실책임 - 자율주행 자동차의 인공지능에 대한 형사책임

2016년 5월, 미국 플로리다 주에서는 테슬라의 자율주행 기능인 오토파일럿(autopilot)의 운행중 오토파일럿 시스템이 밝은 하늘 배경의 흰 트럭을 감지하는데 실패하여 트럭과의 충돌사고로 운전자가 사망한 사고가 발생하였다. 인공지능의 활용이 가까운 시일내에 가장 현실적으로 이루어질 것으로 기대되는 것이 바로 자율주행 자동차의 운행이고 이에 관련한 형사책임의 문제는 현재진행형이라고 할 수 있다.

전통적 형법이론에서도 자동차운전자의 형사책임은 과실론에서 중요한 부분

52) 임석순, 앞의 글, 79면.
53) 임석순, 앞의 글, 80면.
54) 윤지영 외, 앞의 책, 162면.

을 차지하고 있고, 허용된 위험의 법리나 그 구체적 내용으로서 신뢰의 원칙이 일반적으로 받아들여지고 있다. 그러나 종래의 형법상 논의만으로는 자율주행자동차를 운행한 인공지능에게 형사책임을 인정할 것인지, 인공지능 프로그램을 설계한 개발자에게 책임을 물을 것인지, 아니면 언제든지 수동전환을 할 수 있었던 인간 운전자에게 책임을 물을 것인지 답을 찾기 쉽지 않다.[55]

이는 실제 자율주행자동차를 프로그래밍하는 개발자나 제작자는 자신이 설계, 제작하는 자율주행프로그램의 실행 상황이나 결과에 대하여 어느 정도까지 예측할 것을 요구할 것인지와 관련되어 과실의 주의의무의 기준을 정하는 중요한 의미를 가진다고 할 것이다. 실제 자동차 사고는 많은 비정상적인 상황이 중첩적으로 겹쳐서 일어나는 경우가 많고 그러한 상황을 모두 사전에 예측하고 이에 대비한 프로그램을 미리 만드는 것은 불가능하다고 할 것이다. 이러한 점에서 모든 자율주행 자동차의 사고로 인한 결과를 인공지능 프로그램의 설계자에 귀속시키는 것은 정당하지 못하다고 할 것이다. 많은 전문가들이 자율주행자동차를 운행하기 위한 지도정보와 위치정보 등이 제대로 반영되지 않거나, 전송과정에서 문제가 발생하거나 소프트웨어의 결함 또는 충돌로 인하여 사고가 발생할 수 있다고 경고하고 있는 것도 염두에 두어야 할 것이다.[56]

다른 한편, 인공지능 프로그램 설계자의 책임과 관련하여 의무의 충돌 상황에서 인공지능의 기초적 행동기준이나 선택기준을 설계하는 행위에 대하여 형사책임을 물을 수 있는가라는 점도 논의되어야 할 것이다. 예를 들어, 자율주행자동차가 스스로 위험판단을 하고 위험을 회피하는 수단을 선택하도록 프로그래밍하는 중에, 무단횡단하는 한 사람을 피하기 위하여 선택할 수 있는 방법들

55) 일반적이지는 않겠지만 인공지능의 해킹을 통한 사고발생도 가능할 것이다. 자율주행자동차의 해킹 가능성은 이전부터 이론적으로 지적되었지만, 실제 실험을 통해 해킹이 가능한 것으로 입증되었다(윤성현, 자율주행자동차시대 국민의 생명 신체의 안전보호를 위한 공법적 검토, 헌법학연구 제22권 제3호, 2016.9. 263-264면. 참조). 해킹을 한 자가 자율주행자동차를 외부에서 고의적으로 해킹을 해서 자동차 운행을 장악하여 테러나 복수 등에 이용할 수 있는 것이다. 이 경우에는 해킹하여 새로운 명령을 내린 자가 자율주행자동차를 도구로 이용하여 자신의 범죄를 실현하였으므로 결과에 대해 형사책임을 진다고 할 것이다.

56) 자동차 제조업체 볼보(Volvo)는 모든 인공지능 을 채용한 자율주행 차량의 사고는 모두 회사가 책임을 지겠다는 방침을 발표했는데 여기에는 인간이 운전하는 차량과 달리 대형사고로 이어지는 경우는 극히 드물 것이라는 확률과 예측에 바탕하고 있다고 이해하기도 한다. 최은창, 인공지능시대의 법적 윤리적 쟁점, 20면. 참조

이 더 큰 인명피해를 가져올 수 있는 경우(예를 들면, 좌측으로 피하면 유치원생 3명을 칠 수 있고, 우측으로 피하면 노인 3명을 칠 수 있는 상황)에 어떤 판단을 하도록 프로그래밍할 것인가하는 것이다. 그러한 프로그래밍의 결과에 따른 형사책임을 지울 수 있는가? 객관적 귀속이론으로 설명이 충분한지에 관한 검토가 필요한 부분이라고 하겠다.

주변 상황의 인식이나 분석, 나아가 구체적인 조치(결정)과정에서 복합적인 오류로 사고가 발생한 경우 그 인과관계의 판단이나 입증이 쉽지 않을 것이다. 인공지능의 소프트웨어 자체의 결함에 의하여 일어난 것인지, 인식에 오류가 있는지 등을 가리기 쉽지 않을 뿐 아니라 결과발생의 주된 원인이 어디에 있는지 그리고 예견가능한 결과인지 등을 판단함에 있어 종래의 과실범의 기준을 그대로 인공지능 시스템에 적용하기가 어렵다.[57] 미국 캘리포니아, 미시간, 워싱턴 DC, 버지니아주 등은 자율운행 차량의 주행을 허용하는 법을 마련했지만 인공지능의 오작동에 따른 교통사고에서 어떻게 과실을 인정하고, 예견가능성을 인정할 것인지 그리고 손해와의 인과관계를 인정할 것인지에 대해서는 법원의 사법적 판단으로 해결하도록 하였다.[58]

3. 사람과 인공지능 사이의 책임귀속

최근 산업체와 가정에서 사용자(user)의 의도 내지 이익을 위하여 인간의 노동을 대신하는 인공지능 로봇의 사용이 증가하고 있으며 향후 15년 내에는 사용자의 명령에 따라 작동되는 것으로 디자인된 인간과 유사한 감성과 지능을 가진 로봇을 산업체는 물론 가정에서도 구매할 것이라고 예상된다.[59] 사용자가 다른 가전제품처럼 가정용 로봇을 구매하고, 그 로봇을 사용자의 명령에 따라 가사를 돕도록 하여 편리하게 생활할 수 있을 것이다. 그런데 사용자가 자신의 명령에 따라 가사도우미 역할을 하던 인공지능 로봇에게 다른 사람을 상해하도록 명령

57) 양종모, 인공지능의 위험의 특성과 법적 규제방안, 홍익법학 제17권 제4호, 2016, 554면.

58) 이중기/오병두, 자율주행자동차와 로봇윤리: 그 법적 시사점, 홍익법학 제17권 제2호, 2016, 18면.

59) ARTIFICIAL INTELLIGENCE AND LIFE IN 2030, One Hundred Year Study on Artificial Intelligence (AI100), Stanford University, REPORT OF THE 2015 STUDY PANE, 2006, p.15. (https://ai100.stanford.edu.)

하여 상해를 입혔다면 그 로봇이 인공지능을 가지고 있다고 하여도, 로봇은 단지 명령을 받은 대로 그대로 이행한 것이고, 사용자가 로봇을 도구(instrument)로 이용한 것에 지나지 않는다.

이 경우에는, 로봇이 비록 뛰어난 인공지능을 가지고 있다고 하더라도 사용자의 명령을 단순히 이행하는 수준으로 작동하고 단지 명령을 이행하는 도구로서의 성격을 가지고 있기 때문에, 인공지능 로봇의 작동은 전적으로 이용자의 실행행위로 보아야 할 것이다.

한편, 인공지능이 분석하고 판단하여 결정한 작업명령에 따라 행동한 사람이 범죄적 결과를 발생시킨 경우, 형사책임을 누구에게 물을 것인지 어떤 근거로 처벌할 것인지 문제가 될 것이다.[60] 이 경우 인공지능의 명령에 따라 행동한 사람은 인공지능이 합리적으로 판단하여 명령하였을 것이라는 신뢰를 가졌다고 볼 수 있으므로 신뢰의 원칙이 적용될 여지가 크다고 할 것이다. 따라서 인공지능의 명령에 따라 행동한 단순 노동자에게 그 책임을 묻기는 어려워 보인다.

그렇다면, 그러한 명령을 내린 인공지능에게 형사책임을 물을 것이지가 문제되는데, 앞서 살펴본 바와 같이 인공지능에게 형법적 주체성을 인정 여부에 관한 논의를 비롯하여, 즉 적용할 수 있는 형벌의 종류와 내용에 관한 논의로 돌아가게 된다.

다른 한편, 인공지능의 작동에 의해 범죄적 결과가 발생했다 할지라도 인공지능의 형사책임을 부정한다면 인공지능의 배후에 있는 사람, 즉 인공지능의 사용자나 관리자, 생산자, 또는 설계자에게 책임을 물을 수밖에 없을 것이다.

설계자나 생산자가 예측하지 못했던 오류나 하자로 인해 법익침해적 결과가 발생한 경우에는 형사책임 귀속문제가 발생하게 되는데, 인공지능 자체의 책임문제와 별개로 이에 관련된 사람들의 형사책임에 관하여는 이에 관하여는 앞서 살펴본 바와 같이 설계자, 생산자, 이용자 사이의 적정한 주의의무의 경감 내지 가중이 별도로 논의되어야 할 것으로 생각된다. 설계자, 생산자, 이용자 각각의 예측가능성과 주의의무에 따라 누구에게 책임이 귀속될 것인가가 판단되어야

60) Joerden, Strafrechtliche Perspektiven der Robotik, in: Eric Hilgendorf/Jan-Philipp Günther (Hrsg.), Robotik und Gesetzgebung. Beiträge der Tagung vom 7. bis 9. Mai 2012 in Bielefeld, 2013., S. 206.

하고 구체적으로는 사용자나 관리자, 생산자, 또는 설계자가 각각 불법결과를 예측하고 예방했어야 하는 지위와 권한, 능력을 갖고 있었는지를 입증하는 문제로 귀결될 것이다.

Ⅴ. 결론

이미 우리 생활속에 인공지능은 많은 역할을 하고 있고 그 편의를 누리고 있지만, 호킹박사가 지적한 것처럼 그 미래가 어떨지에 대하여는 누구도 장담할 수 없을 것이다. 먼 미래가 아니라도 자율주행자동차 사고와 같이 그 책임을 판단하는 논의는 시작되었다고 할 것이다.

인공지능에 대하여 형사책임을 묻기 위한 몇가지 전제 내지 요건을 검토한 결과, 사회적 행위론의 입장에서 행위의 본질적 요소인 '사회적 의미성'이라는 행위 외부의 제3자적 평가관점에서 행위를 이해하여 인공지능과 인간과의 근본적인 차이를 염두에 두지 않고 그 사회적 의미만을 기준으로 판단할 수 있다는 인공지능의 행위성을 인정할 수 있다.

그리고 현재의 인공지능의 수준은 급속한 발달과정을 거치고 있으며, '딥 러닝'의 단계를 지나 더 고차원적인 사고 알고리즘을 시도하고 있다는 점, 사회적으로 용납할 수 없는 침해행위를 한 인공지능을 형사처벌로부터 자유로운 행위주체라고 하는 것은 인공지능의 실행으로 인한 피해의 규모가 크고 반복적으로 이루어지고 있다는 현실에 비추어보면 적정하다고 할 수 없으며, 인공지능의 반사회적 활동으로부터 사회를 방위해야 할 필요가 있고, 책임의 근거를 반사회적 위험성으로 이해하면 인공지능에게도 그 책임을 물을 수 있다고 보아야 한다.

반면, 법익침해를 일으킨 인공지능에 대하여 가해질 조치들에 관하여 이를 형벌로 이해할 수 있을 지는 의문이다. 인공지능은 자산을 소유하고 있지도 않고, 소유하고 있다는 인식도 없기 때문에 벌금형의 의미를 지닐 수 없으며 인공지능에게는 물리적인 손상이나 파괴를 가하더라도 신체의 완전성을 유지하며 삶을 유지하려는 의지나 필요[61]가 없기 때문에 사형이나 자유형이 의미가 없으며, 사형이나 자유형과 같은 형사처벌은 인공지능의 해체나 파괴, 프로그램 재

설치, 가동중지 등으로 대체될 수 있을 것으로 보이지만, 이를 굳이 형벌이라고 해야 하는지에 대하여는 의문이며, 인공지능에 대하여 가능한 조치의 종류와 내용에 대하여는 별도의 논의가 필요하다고 보인다.

참고문헌

김성규, “법인처벌의 법리와 규정형식”, 『법조』 578권, 2004. 11.

김영두, “인공지능과 자유의지, 인공지능시대의 법적 과제”, 『연세대학교 제60회 학술대회 발표집』, 2017. 2.

김영환, “로봇 형법(Strafrecht für Roboter)?,” 『법철학연구』 제19권 제3호, 2016.

김호기, “법익적대적 법인문화, 위험관리 실패와 법인의 형사책임”, 『형사정책연구』 제22권 제4호, 2011.

류화진, “지능형 로봇의 범죄주체성과 형사책임”, 『과학기술과 법』 제7권 제2호, 2016.

박기석, “양벌규정의 문제점과 법인 범죄의 새로운 구성”, 『형사정책』 제10호, 1988.

변종필, “형벌조항에 대한 위헌심사와 책임주의”, 『헌법실무연구』 제11권, 2010.

송승현, “트랜스휴먼 및 포스트휴먼 그리고 안드로이드(로봇)에 대한 형법상 범죄주체의 인정여부,” 『홍익법학』 제17권 제3호, 2016.

안성조, “인공지능 로봇의 형사책임”, 『법철학연구』 제20권 제2호, 한국법철학회, 2017.

양종모, “인공지능의 위험의 특성과 법적 규제방안”, 『홍익법학』 제17권 제4호, 2016.

양천수, “법인의 범죄능력 : 법 이론과 형법정책의 측면에서”, 『형사정책연구』 18권 2호, 2007. 6.

윤성현, “자율주행자동차시대 국민의 생명 신체의 안전보호를 위한 공법적 검토”,

61) 인공지능은 저장, 백업 및 복원이 가능할 것이기 때문에, 복구(부활) 가능성이라는 측면에서 그 형벌의 효과는 크게 감소할 것이라고 생각된다.

『헌법학연구』 제22권 제3호, 2016.
이원상, “4차 산업혁명에 있어 형법의 도전과제,”『조선대 법학논총』 제24권 제1호, 2017.
이인영, “인공지능 로봇에 관한 형사책임과 책임주의,”『홍익법학』 제18권 제2호, 2017.
이주희, “인공지능과 법－지능형 로봇 및 운영자의 형사책임에 관한 고찰,”『한국사회과학연구』 제38권 제1호, 2016.
이중기/오병두, “자율주행자동차와 로봇윤리: 그 법적 시사점”,『홍익법학』 제17권 제2호, 2016.
임석순, “형법상 인공지능의 책임귀속,”『형사정책연구』 제27권 제4호, 2016.
장연화·백경희, “왓슨의 진단조력에 대한 현행법상 형사책임에 관한 소고,”『형사법의 신동향』 제55호, 2017.
정정원, “인공지능발달에 따른 형법적 논의”,『과학기술과 법』 제7권 제2호, 충북대학교 법학연구소, 2016. 12.
최은창, “인공지능시대의 법적 윤리적 쟁점”,『Furture Horizon』 제28호, 과학기술정책연구원, 2016. 5.
허일태, “위험사회에 있어서 형법의 임무”,『비교형사법연구』 제5권 제2호, 2003.

Gabriel Hallevy, I, Robot-I, Criminal-- When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses, Syracuse Science & Technology Law Reporter, 2010.
Gabriel Hallevy, The Criminal Liability of Artificial Intelligence Entities-From Science Fiction to Legal Social Control, Akron Intellectual Property Journal, 2010.
George R. Cross & Cary G. Debessonet, An Artificial Intelligence Application in the Law: CCLIPS, A Computer Program that Processes Legal Information, High Technology Law Journal, 1986
James Barrat, Our Final Invention: Artificial Intelligence and the End of the Human Era, 2013

Matthew U. Scherer, Regulating Artificial Intelligence System: Risks, Challengers, Competencies, and Strategies, Harvard Journal of Law & Technology, Spring 2016.

Nils J. Nilson, The Quest for Artificial Intelligence: A History of Ideas and Achievements, Cambridge University Press, 2010.

Sabine Gless, Emily Silverman, Thomas Weigend, If Robots Cause Harm, Who Is To Blame? Self-Driving Cars and Criminality, New Criminal Law Review, 2016.

The European Parliament's Legal Affaires Committee, European Civil Law Rules In Robotics, 2016.

[Abstract]

A Study about Criminal Liability of Artificial Intelligence

Hwang, Man-Seong

Professor of Wonkwang Univ. Law School, Ph.d. in Laws

For some years, there has been significant controversy about the very essence of AI robots. Futurologists have proclaimed the birth of a new species, machina sapiens, which will share the human place as intelligent creatures on Earth

The fundamental question of criminal law is the question of criminal liability, i.e., whether the specific entity (human or corporation) bears criminal liability for a specific offense committed at a specific point in time and space. In order to impose criminal liability upon a person, two main

elements must exist. The first is the factual element, i.e., criminal conduct (*actus reus*), while the other is the mental element, i.e., knowledge or general intent in relation to the conduct element (*mens rea*).

AI robots are taking larger and larger parts in human activities, as do corporations. Offenses have already been committed by AI robots or through them. Thus, there is no substantive legal difference between the idea of criminal liability imposed on corporations and on AI robots. It would be outrageous not to subordinate them to human laws, as corporations have been.

When AI robots and humans are involved, directly or indirectly, in the perpetration of a specific offense, it will be far more difficult to evade criminal liability. All entities--human, legal or AI-- become subject to criminal law.

Let us assume an AI robot is criminally liable. Let us assume it is indicted, tried and convicted. After the conviction, the court is supposed to sentence that AI robot. If the most appropriate punishment under the specific circumstances is one year of imprisonment, for example, how can an AI robot practically serve such a sentence? The punishment adjustment considerations examine the theoretical foundations of any applied punishment.

Key words : AI, artificial intelligence, criminal Liability, criminal liability of corporations, criminal liability of robot