□ 특집논문 □ https://dx.doi.org/10.16936/theoph..45.202312.95

‘인공지능의 책임’ 쟁점에 대한 비판적 고찰

이상헌
서강대학교, 전인교육원 교수

들어가는 말

‘지적인 게임인 바둑에서 인공 ‘지능(intelligence)’ 알파고가 인간에게 승리했다. 5레벨 ‘자율(autonomous)’ 주행 자동차는 인간 운전자 없이 스스로 운전할 것이다. ‘자율형

(autonomous)' 살상 로봇이 시가전을 훌륭하게 수행할 날이 멀지 않았다. 정보통신기술(ICT) 기반으로 도시의 구석구석까지 신경망처럼 연결된 '스마트(smart)' 시티는 교통, 환경, 안전, 주거, 복지 등 다양한 서비스가 인공지능에 의해 최적의 상태로 유지되는 미래 도시이다. '학습(learning)'을 통해 스스로 결과를 수정할 수 있는 학습하는 알고리즘이 인간이 주입한 데이터에만 의존하던 기존의 알고리즘으로 해결하기 어려웠던 자율주행 자동차, 필기체 문자인식, 음성인식, 번역 등 다양한 영역에서 놀라운 성과를 내고 있다. ……' 인공지능과 관련하여 일반적으로 서술할 수 있는 내용이다.

오늘날 우리가 과학기술과 관련하여 가장 많이 듣는 수식어는 '지능', '자율적', '스마트', '학습' 등이다. 과거에는 사람에게만 적용되었던 이런 수식어들이 요즘은 기계와 인공지능에게도 적용된다. 그리고 이런 추세는 사람들로 하여금 새로운 윤리적, 철학적 문제 상황을 상상하게 만들고 있다. 그 가운데 하나가 책임에 관한 쟁점이다. 인공지능이나 로봇이 인간의 개입 없이 한 행위, 즉 자율적 행위에 대해서 누가 책임을 져야 하는가? 인공지능이나 로봇에게 책임을 물을 수 있는가? 그렇게 하기 위해서는 무엇이 바뀌어야 하는가? 인공지능이나 로봇에게 책임을 묻는 그럴듯한 방식은 무엇인가? 지금까지는 넌센스로 들렸던 이런 물음들이 의미 있는 것이며, 심지어 중대한 것이라는 주장이 제기되고 있고, 그에 관해 다수의 연구자들 사이에서 논쟁이 진행되고 있다.

아래는 자율적으로 판단하고 행동하는 인공지능이 전통적인 책임 이해에 교란을 가져온다고 주장되는 사례이다.

(사례 1)

자율형 무기 시스템(autonomous weapon systems: AWS)은 이미 현장에서 활용되고 있는데, 현재는 지극히 제한적인 자율성만 허용되고 있다. 하지만 미래의 AWS는 수집된 정보를 토대로 스스로 판단하고, 결정하고, 실행할 수 있는 정도의 자율성을 인정받을지 모른다. 이런 상황을 가정하고 실제로 일어날 법한 일을 상상해 보자. 매우 정교한 인공지능에 의해 제어되는 AWS가 있는데, 미사일과 폭탄을 싣고 있다. 이 AWS가 항복 의사를 분명히 밝힌 적군 대열을 의도적으로 폭격했다고 상상해 보자. 적군 병사들은 무기를 내려놓았으며 아군에게 어떤 즉각적인 위협도 가하지 않았다. 물론 AWS의 폭격은 전혀 실수가 아니었으며, AWS가 내린 결정에 따른 것이었다. ASW의 결정은 합리적인 고려에 따른 것으로, 우연히 발생한 프로그램상의 오류로 인한 것이 아니었다. AWS는 이유 있는 결정을 했다. 다시 말해, 적군들을 포로로 삼았을 때 수용소에 가두고 감시하는 비용이 상당하다고 판단했다. 혹은 적군 병

사들에게 공포심을 심어주어 얻을 수 있는 효과가 매우 크다고 판단했을 수도 있고, 또는 최근에 파괴된 로봇 동지들의 죽음을 복수하여 미래의 안전을 도모할 수 있다고 판단했을 수도 있다. 그러나 이 AWS의 행동은 도덕적으로 정당화될 수 없는 것이다. 인간이 그런 행위를 저질렀다면 전쟁 범죄로 여겨져 재판을 받아야 할 것이다.[1]

(사례 2)

누군가 가사 도우미 로봇을 구매했다고 가정하자. 이 로봇이 영화 <바이센테니얼 맨>에 등장하는 앤드류와 같은 모델의 로봇이라고 가정해도 좋다. 이 로봇은 아파트 10층에서 혼자 엘리베이터를 타고 내려와 횡단보도를 건너고, 길 건너의 커피숍에 가서 음료수를 사올 수 있는 정도로 지능이 높다. 이 로봇은 일반적인 행동 원칙을 인식하고 있는 것은 물론이고, 생활하면서 마주치는 사람들의 행동을 지켜보면서 대처 능력과 사회성을 학습할 수 있도록 설계되었다. 지속적인 학습과 행동 수정은 이 로봇의 장점이다. 어느 날 이 로봇이 흔치 않은 광경을 목격한다. 어떤 의로운 사람이 소매치기를 붙잡았는데, 경찰이 올 때까지 소매치기를 붙잡아두고 있으면서 행인들에게 둘러싸여 감탄과 존경을 한 몸에 받고 있는 모습이었다. 그로부터 얼마 뒤에 이 로봇이 커피를 사러 가는 도중에 한 남자가 어떤 여자와 티격태격하는 장면을 보게 되었다. 여자가 싫다는 의사를 분명히 보였는데도 남자가 여자의 가방을 억지로 가로채는 광경을 보게 된 것이다. 이 로봇은 프로그램 된 내용과 지난 번에 길에서 목격한 사건의 경험을 토대로 지금의 광경이 범죄의 현장이라고 판단했다. 로봇은 남자를 제압하고 경찰에 전화한 후에 경찰이 도착할 때까지 남자를 붙들고 있었다. 경찰이 도착한 후에 자초지종이 밝혀졌는데, 남자와 여자는 부부지간이었으며, 누가 운전할지를 놓고 옥신각신하며 서로 열쇠를 빼앗으려고 했던 것이었다. 로봇이 남자를 제압하는 과정에서 남자가 거세게 저항하는 바람에 작은 상처가 생겼다고 가정하자. 그리고 이 부부는 어디를 급히 가야 할 상황이었다고 가정하자. 부부는 매우 화가 났고, 경찰에게 로봇을 체포하라고 강력히 요구했다.[2]

위의 사례에 등장하는 AWS나 가사 도우미 로봇에게, 그것의 행동에 대해 위에서 언급한 그런 종류의 물음을 적용할 수 있을까? 그렇게 하는 것이 적절하거나, 효과적이거나, 정당한 것일까?

---

1) Robert Sparrow, “Killer Robot”, *Journal of Applied Philosophy* 24, 2007, 66. 이 사례는 스패로우가 논문에서 간략하게 제시한 것을 내용의 손상이나 변형 없이 재구성한 것임.

2) 제리 카플란, 『인간은 필요 없다』, 신동숙 옮김, 한스미디어, 122-123. 이 사례는 이해를 돕기 위해 이 책에 소개된 것을 내용의 훼손이나 변경 없이 일부 수정한 것임.

1. 책임 공백이란 무엇인가?

정치 철학자인 로버트 스패로우(Robert Sparrow)가 위의 시나리오를 상상한 것은 인간의 직접적인 통제 없이 자율적 지각과 판단에 기초해 중대한 결정을 내리고 실행하는 이른바 자율형 살상 로봇(일명 killer robot)을 실전에 배치하는 것이 정당하지 않다고, 그런 로봇의 사용을 허용할 근거가 전혀 없다고 주장하기 위해서였다. 하지만 그의 의도와 달리 스패로우의 시나리오는 로봇의 책임 문제를 부각시키는 데 일조한 듯하다.

스패로우는 위의 사례에서 AWS의 제조업체, 프로그래머, 지휘관 등 관련이 있어 보이는 어느 누구에게도, AWS의 행동이 전쟁 범죄에 해당한다고 하더라도, 책임을 묻기 어렵다고 말한다. 왜냐하면 그들 모두 범죄에 해당하는 AWS의 행동에 대한 통제력을 가지고 있지 않기 때문이다. AWS는 스스로 정보를 수집하고 분석할 수 있으며, 의사결정이 필요한 상황에서 자율적으로 판단하고 인간의 개입 없이 독자적으로 결정하고 행동할 수 있다. 전투 상황에서 적군을 폭격하거나 살상하는데 인간의 허락을 별도로 구하지 않는다. 그래서 위에 언급한 적군 살상 행동에 제조업체나 프로그래머, 심지어 지휘관까지도 개입되지 않았다고 할 수 있다. 사실상 이들은 위의 구체적인 상황에서 AWS가 어떻게 행동할지 정확하게 예측할 수 없다. 실제 전투 현장에서는 다양한 돌발 변수들이 있고, 그런 변수들을 모두 검토하는 빈틈없는 사전 테스트는 원칙적으로 불가능하기 때문이다. 대신에 이런 자율형 살상 로봇에게 일반적인 행동 원칙을 학습시키고, 기존의 경험과 데이터에 기초해 최적의 분석 및 판단 알고리즘을 구축한 것이다. 이런 이유들로 인해 제조업체나 프로그래머에게 문제의 AWS 행동에 대한 직접적인 책임을 추궁하는 것은 정당하지 않아 보인다. 그렇다고 지휘관에게 책임을 물을 수도 없다. 지휘관 역시 AWS의 행동에 대한 통제력을 행사할 수 없었다고 보는 것이 합리적이다. 그런데 전쟁 범죄에 해당하는 사건이 벌어진 것은 사실이다. 다수의 사람이 정당하지 않은 방식으로 목숨을 잃었다. 이 사건에 대한 우리의 상식적 이해는 누군가 책임을 져야 한다는 것이다. 사망자가 발생한 것은 AWS의 행위 때문인 것이 분명하지만 그렇다고 기계에 책임을 묻는 것은 적절하지 않아 보이기도 하다. 책임은 처벌받을 수 있는 존재, 책임질 수 있는 존재에게 부과되는 것이다. 기계가 책임을 진다는 말은 이상하게 들린다. 기계는 고통을 느끼지 않으며, 칭찬이나 비난에 영향받지 않는다. 처벌을 통해서만 기계의 행위가 바로잡히는 것도 아니다. 기계의 행위 교정은 처벌 여부와 관계 없다. 따라서 기계에게 책임을 물을 수 없다. 스패

로우는 위의 시나리오에 대해 이렇게 해석함으로써 AWS의 사용은 허용되지 않아야 하며, 처음부터 그런 무기체계의 구축 자체가 잘못이라고 주장한다. 왜냐하면 기계의 행동이 책임질 일을 만들어내지만 기계 자체에게 책임을 묻을 수 없기 때문이다.

스패로우의 시나리오와 그의 해석은 이른바 책임 공백(responsibility gap)에 대한 적절한 사례를 제시해준다. 자율형 로봇의 행동으로 인한 해로운 결과에 대해 인간에게 책임을 물을 수도 없고, 기계에게 책임을 물을 수도 없는 이러한 상태를 책임 공백 상태라고 한다. 책임 공백을 처음 거론한 것은 카셀대학교의 컴퓨터공학자 마티아스(Andreas Matthias)이다. 그는 기계학습을 거론하며 프로그래밍 알고리즘의 혁신으로 인간에게 전적으로 책임을 지울 수 없는 상황이 도래했음을 역설하고 인간이 충분히 통제할 수 없는 기계의 행동에 대해 인간에게 책임을 묻는 것이 부당(injustice)하다고 지적한다. 이런 부당함을 피하기 위해 우리의 도덕적 관행과 법률에서 책임 공백을 해소할 방법을 모색해야 한다고 주장한다.[3)]

일상적인 의미에서 책임은 누군가 저지른 잘못된 행위에 대해 그 사람에게 그 행위의 귀결에 대해 의무 혹은 부담을 지도록 하는 것을 말한다. 원칙적으로 책임은 그 행위를 한 행위자에게 귀속되는데, 행위자는 자신의 행위에 대해 통제력을 가진 사람이며, 자신의 행위와 그로부터 영향받는 귀결들에 대해 알고 있고 의식하고 있는 사람을 말한다. 행위자는 자신의 행위에 대해 통제력이 있으므로 모종의 행위를 할 수도 있고, 하지 않을 수도 있다. 자신의 행위와 그 결과에 대해 알고 있기 때문에 행위자는 모종의 행위와 관련하여 자유롭게 의사결정할 수 있으며 분별력을 지니고 있기 때문에 모종의 행위를 할 때의 결과는 물론이고 행위하지 않았을 때의 영향에 대해서도 이해하고 있을 것이다.

책임에 대한 일상적인 이해 방식은 아리스토텔레스에서 유래한다. 모종의 행위에 대해 누군가에게 책임을 귀속시키기 위한 조건으로 아리스토텔레스는 두 가지를 거론했다. 이른바 통제 조건(control condition)과 인식적 조건(epistemic condition)이다. 통제 조건이란 행위자가 자신의 행위를 충분히 통제할 수 있어야 한다는 것이고, 인식적 조건이란 행위자가 자신의 행위와 그 귀결에 대해 알고 있고 충분히 의식하고 있어야 한다는 것이다. 아리스토텔레스에 따르면 행위의 원천은 행위자이며, 행위자는 자신의 행위에 대해 모를

3) Andreas Matthias, “The responsibility gap: Ascribing responsibility for actions of learning automata”, *Ethics and Information Technology*, 6(2004), 183.

리 없다.[4]

책임 공백은 책임에 대한 이와 같은 전통적인 이해에 도전한다. 책임 공백이 도입되는 맥락은 프로그래머가 개발한 알고리즘에 기반해 작동하는 학습하는 인공지능에게 자율성이 주어졌을 때, 프로그래머나 사용자, 혹은 그 누구도 그 자율적 인공지능의 행위 결과에 대해 책임 추궁을 당하는 것이 부당하다는 의식이 생기는 상황이다.

> 기계 행위(machine actions)라고 부를 수 있는 부류가 증가하고 있는데, 이 경우에 책임 귀속의 전통적 방식이 우리의 정의 감각 및 사회의 도덕적 틀과 양립하지 않음을 보여줄 수 있다. 왜냐하면 기계의 행위에 대해서 책임이 있다고 가정할 만큼 통제력을 가진 사람이 아무도 없기 때문이다. 이런 경우들에서 우리가 책임 공백이라고 부르는 것이 발견된다.[5]

간단히 말하면, 마티아스의 주장은 자동화된 알고리즘에 기반해 이루어진 모종의 기계적인 행동이 해악을 발생시켰을 때, 우리는 해악이 발생했기 때문에 책임 귀속의 상황이 발생했다고 생각하지만 해당 기계적 행동에 대해 통제력을 가지고 있는 사람이 아무도 없었으므로 책임 추궁을 당해야 하는 사람이 존재하지 않는다는 것이다. 혹은 어떤 식으로라도 연루된 사람들을 찾는다면 다수를 떠올릴 수 있지만, 물론 그들 가운데 어느 누구에게도 해당 기계적 행동에 대한 직접적인 통제력이 없었기 때문에 책임 추궁을 당할 사람을 특정할 수 없다는 것이다. 사정이 이러하기 때문에 인공지능과 관련하여 책임 공백의 사례들이 발생한다. 책임 귀속의 일상적 요구와 전통적인 방식으로 책임 귀속의 대상을 특정할 수 없는 불능의 상황이 책임의 공백 상태를 만든다.[6]

2. 책임 공백 문제에 대응하는 다양한 방식

책임 공백은 다루기 까다로운 문제이다. 이것이 쟁점이 되고 나서 다양한 방식의 접근법과 해결책들이 제시되었다. 마티아스가 책임 공백의 존재를 주장한 이후 논쟁에 참여한

---

4) Aristotle, *Nicomachean Ethics*, 1111a3-1111a5.

5) Andreas Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and Information Technology*, 6(2004), 177.

6) 만일 책임 공백이 발생한다면, 그 원천은 기술에만 국한되지 않으며, 제도 또한 책임 공백의 또 다른 원천이다. 그러나 이 논문에서는 기술적 원천에만 국한해서 논의를 진행한다.

다수의 연구자들이 책임 공백의 존재를 인정했다. 스패로우는 이 논쟁에 참여하지는 않았지만, 그의 뜻과 무관하게 책임 공백을 다룰 때 빈번히 거론된다. 스패로우가 자신의 견해를 논증하기 위해 책임 공백을 가정했기 때문이다. 책임 공백을 인정하지 않으면 스패로우의 주장, 다시 말해 책임을 물어야 하는 상황에서 AWS에게 책임을 물을 수 없기 때문에 AWS를 허용해서는 안 된다는 주장이 성립하지 않는다.

책임 공백을 인정하는 연구자들은 모두 같은 의도로 이것을 인정하지는 않으며, 이 문제를 해결하기 위해 각양각색의 해결책을 제시한다. 가장 극단적인 형태는 미래의 인공지능을 가정하고 인간과 동등한 행위자로 취급할 것을 주장하는 것이다. 컴퓨터 윤리학자인 제임스 무어(James Moor)의 인공적 도덕 행위자(artificial moral agent: AMA)의 분류 방식에 따라,[7] 인공지능이 완전한 윤리적 행위자(full ethical agent)라면 인공지능에게 책임을 묻는 것이 인간에게 책임을 묻는 것과 마찬가지로 타당하다고 말할 수 있을 것이다. 완전한 윤리적 행위자는 의식, 의도성, 자유의지 능력을 갖추고 있어서 명시적인 도덕적 판단을 내릴 수 있다.[8] 명시적 윤리적 행위자(explicit ethical agent)가 프로그램된 윤리적 규범에 따라 행동하는 것에 제한된다는 점과 달리 완전한 윤리적 행위자는 스스로 윤리 규범을 생성할 수 있다. 우리가 쉽게 구할 수 있는 다수의 책들이 미래의 인공지능을 언급하면서 이런 식의 인공지능을 상상하고, 인공지능에의 책임 귀속을 당연한 듯이 기술하고 있는데, 이 논문에서는 이런 종류의 견해에 대해서는 거론하지 않는다. 이런 유의 인공지능은 단지 상상이기 때문이다.

다수의 연구자들은 다양한 분야와 응용 현장에서 책임 공백이 발생할 수 있음을 증언한다. 이들의 증언이 책임 공백을 기정사실처럼 받아들이기 쉽게 하는 경향이 있다. 책임 공백을 인정할 때, 우리가 취할 수 있는 반응은 제한적이다. 인간이 아닌 인공지능에게 책임을 물어야 한다는 주장을 제외하면, 전통적인 책임 개념이 오늘날 벌어지는 새로운 상황, 혁신적인 기술이 등장한 상황에 적합하지 않다는 견해가 많은 지지를 얻는 듯하다. 이런 상황에서 남은 것은 책임 개념을 어떻게 수정할 것인지, 혹은 어떤 새로운 책임 개념을 도입할 것인지를 결정하는 것이다.

---

7) 월러치 & 알렌, 『왜 로봇의 도덕인가』, 노태복 옮김, 메디치, 2014, 61-62. 무어는 AMA를 윤리적 영향 행위자(ethical impact agent), 내재적 윤리적 행위자(implicit ethical agent), 명시적 윤리적 행위자(explicit ethical agent), 완전한 윤리적 행위자로 분류했다.

8) 월러치 & 알렌, 『왜 로봇의 도덕인가』, 노태복 옮김, 메디치, 2014, 62.

오스트리아 빈대학의 기술철학자인 마크 코켈버그(Mark Coeckelbergh)는 로봇이 무어(Moor)의 분류 방식에 따라 완전한 도덕적 행위자(full moral agent)가 아니라면 로봇에게 책임을 물을 수 없다는 점을 인정하면서도 자율형 로봇의 개발자와 사용자가 자신의 행위에 대해 완전히 알지는 못하는 상황 역시 받아들인다. 코켈버그는 전통적인 책임 개념을 행위자 중심의 책임 개념으로 규정하고 책임에서 책무성(accountability)으로 무게 중심을 이동시키고 관계론적으로 책임을 이해할 것을 제안한다.[9)]

책임 개념 대신 책무성을 대안으로 제시하는 논의들은 설명 가능한 인공지능에 대한 논의로 연결될 수 있다. 인공지능의 행위에 대해 책임을 물을 수 없는 이유 가운데 하나는 인공지능의 의사결정의 불확실성에 있기 때문이다. 그러나 책무성이 책임 개념의 대안이 될 수 있을지는 의문이다. 책무성, 혹은 책무 있음(Being accountable)은 자신의 행위에 대해 그 행위가 나타나게 되는 데에 결정적이었던, 선행하는 인과적 사건들을 밝히는 것이 아니라 이유를 밝히는 것을 말한다. 우리가 보통 누군가의 어떤 행위에 대해 '왜 그랬어?'라고 말할 때, 우리가 기대하는 것은 '내적 상태 $x_{1....n}$, 입력값 $y_{1...n}$, 그리고 전자에서 작동하는 메커니즘 $z_{1...n}$ 의 조합이 출력값으로서 이러한 행동을 야기했어.'가 아니다. 이러한 답변은 기껏해야 그러한 행동을 하는데 관계있는 생물학적 인과관계(생리적, 심리적 등)를 설명한 것이지 그러한 행동을 한 이유를 밝힌 것이 아니다. 이러한 설명으로는 해당 행위의 정당성을 주장할 수 없다. 이런 설명은 문제의 행위를 우리가 상호 수용하는 행위 규범들의 그물 속에 위치시키지 못한다.[10)] 책무성과 설명가능성(explainability)은 바꿔 쓸 수 있는 개념이 아니다.

네덜란드 위트레흐트대학의 스벤 네이홈(Sven Nyholm)은 책임 공백의 존재를 인정하고 이 문제를 해결하기 위해 우리의 책임 귀속 관행을 수정할 것을 제안한다.[11)] 책임 공백이 발생하는 것은 전통적인 이해 방식이 적용되지 않는 상황이 발생하기 때문이고, 이런 상황은 다른 시각에서 바라볼 필요가 있다고 그는 주장하다. 네이홈은 기계를 인간과 무관하게 스스로 행동하는 존재로 간주하는 것이 잘못이라고 주장하지만, 그렇다고 전통

9) Mark Coeckelbergh, *AI Ethics*, MIT Press, 2020, ch.4, ch.8. 신상규, "인공지능 시대의 윤리학", 『지식의 지평』 21(2016), 1-17 참고.

10) Jan-Hendrik Heinrichs, "Responsibility assignment won't solve the moral issues of artificial intelligence", *AI and Ethics*, 2(2022), 730-731.

11) Sven Nyholm, "Attributing agency to automated systems: reflections on human-robot collaborations and responsibility-loci", *Sci. Eng. Ethics* 24(2018), 1201-1219.

적인 이해에 따라 인간에게 전적으로 책임을 물을 수 없다는 점도 인정한다. 그의 제안은 인간-로봇 협업의 관점에서 사태를 다시 바라보는 것이다. 네이홈은 행위자성의 분산(distributed agency)을 통해 문제를 해결하는 어려운 길을 택한다.

책임 공백 문제에 답하는 또 하나의 길이 분산된 책임(distributed responsibility)을 주장하는 것이다. 루트비히-막시밀리안대학의 아나 슈트라서(Anna Strasser)는 책임 공백의 문제를 도덕적 책임 분배의 문제로 이해한다. 슈트라서는 인공적 도덕행위자의 자격을 논하고 인간에 전적으로 책임을 귀속하는 것이 여전히 정당한지 논의한 뒤에 인공적 행위자에게도 어느 정도의 책임을 귀속시키는 것이 합리적이라고 결론내린다. 슈트라서는 인간 행위자와 인공적 행위자 사이에 도덕적 책임의 분배를 주장한다.[12)]

3. 책임 공백 지지 논변에 대한 분석과 비판

책임 공백을 인정하고 문제를 해결하는 방안을 모색하는 연구자들도 있지만, 책임 공백을 부정하는 논의들도 다수 있다. 물론 이들의 반응도 한 가지가 아니다. 호주의 CSIRO(Commonwealth Scientific and Industrial Research Organisation) 소속 연구원들이 공동으로 쓴 논문에서 데이비드 더글라스(David M. Douglas) 등은 책임 개념을 인과적 책임(causal responsibility)과 역량 책임(capacity responsibility)으로 구분하여 자율형 인공지능 시스템에 도덕적 책임을 물을 수 없기 때문에 책임 공백은 존재하지 않는다고 주장한다. 도덕적 책임은 책임질 수 있는 역량이 있는 존재에게만 귀속될 수 있는 것인데, 인공지능 시스템은 그런 역량이 없기 때문이다. 책임질 수 있는 능력이 없는 존재에게 책임을 묻는 것은 도덕적 책임 개념을 잘못 이해한 것이다. 인공지능 시스템은 인과적 책임은 있지만 역량 책임은 없으므로 인공지능 시스템에게 도덕적 책임은 없다. 그러나 더글라스 등의 견해는 지능형 시스템에 인과적 책임을 인정한 것이어서 인공지능을 행위자로 볼 여지를 남겼다. 더글라스 등의 논의를 따르면, 자율형 인공지능 시스템은 인과적인 관점에서는 행위자이지만 도덕적인 관점에서는 행위자가 아니라는 귀결이 생긴다.

독일의 아헨라인베스트팔렌공과대학의 얀-헨드릭 하인리히스(Jan-Hendrik Heinrichs)는

12) Anna Strasser, “Distributed responsibility in human–machine interactions”, *AI and Ethics*, 2(2022), 523–532.

책임 공백이 실제로 존재하는 것이 아니며, 도덕적 상황의 복잡성을 얼버무리려고 하거나 반사실적으로 사이비 행위자(pseudo-agent) 지위를 인공지능에게 귀속시키려고 할 때에만 책임 공백이 존재하는 듯한 인상을 받게 된다고 주장한다.[13] 또한 하인리히스는 책임 공백을 주장하는 상황을 분석하고 이해하는 데에는 책임이라는 용어보다 다른 윤리적 용어들이 더 적절할 것이라고 주장한다. 책임에 대한 언급이 문제의 도덕적 상황과 도덕적 행위자의 복잡성을 모호하게 만드는 경향이 있기 때문이다.

나는 책임 공백의 쟁점을 다룰 때 좀 더 원론적인 물음에서 시작할 것을 제안한다. 이 문제를 논의하는 다수의 연구자들이 책임 공백의 존재를 너무 쉽게 인정하는 듯한 인상을 받았다. 그래서 나는 다음과 같은 물음에서 시작하려고 한다. '정말로 책임 공백이 발생하는가?' 이 논문에서 나는 '책임 공백'이라는 문제 자체가 성립하지 않는다고 주장할 것이다. 그리고 이 결론은 책임 공백에 관한 논의를 다음과 같은 다른 물음으로 이끈다는 것을 보여준다. 그러면 왜 책임 공백을 주장하는가? 누가 책임 공백의 문제를 수용하는가?

아래에서 책임 공백의 문제 자체가 성립하지 않는 이유에 대한 설명부터 시작해서 책임 공백 문제를 제기하는 이유를 짐작해 보려고 한다. 책임 공백의 부재에 관한 논의는 책임 공백의 존재를 주장하는 사례와 논증들을 전통적인 책임 귀속의 두 가지 조건을 기준으로 분류한 쾰러 등의 분석틀을 차용하여 진행한다.[14]

아리스토텔레스에 따르면, 책임 귀속을 위해서는 두 가지 조건을 만족시켜야 한다. 먼저, 행위자는 자신의 행위에 대해 충분한 정도의 통제력을 지니고 있어야 한다는 통제 조건을 만족시켜야 한다. 그리고 행위자는 자신의 행위가 어떤 것인지, 그로 인해 어떤 결과가 발생할지 알아야 하고 또한 의식적으로 이해하고 있어야 한다는 인식적 조건을 만족시켜야 한다. 책임 공백에 대한 주장은 이 두 가지 조건을 모두 만족시킬 수 있는 행위자가 없기 때문에 책임 공백이 발생한다고 역설한다. 책임 공백의 사례들에서 어떤 경우에는 통제 조건을 만족시킬 수 없고, 또 어떤 경우는 인식적 조건을 만족시킬 수 없으며, 어떤 경우는 양자 모두를 만족시킬 수 없음을 발견할 수 있다는 것이다.

---

13) Jan-Hendrik Heinrichs, "Responsibility assignment won't solve the moral issues of artificial intelligence", *AI and Ethics*, 2(2022), 735.

14) Sebastian Köhler, Neil Roughley & Hanno Sauer, "Technologically blurred accountability?", in Cornelia Ulbert, et al.,(ed.), *Moral Agency and the Politics of Responsibility*, 2017, ch.4.

3.1. 책임 공백을 주장하는 인과적 논증에 대한 검토

첫 번째 논의할 것은 해로운 결과를 초래한 행위를 발생시킨 인과 구조에 초점을 둔 논증이다. 자율형 기계에 의해 해악이 발생했는데, 한편으로 기계에 책임을 돌리는 것이 적절하지 않으며, 다른 한편으로 그 기계의 행동을 인과적으로 유발한 인간 관련자가 없기 때문에 책임 공백 상태가 생긴다는 주장이다. 자율형 기계의 행동은 인간 행위자로부터 독립적으로 이루어진 것이므로 책임을 져야 하는 인간 행위자가 존재하지 않는다는 것이다. 이것은 앞에서 언급한 스패로우의 논증의 핵심이다.

스패로우의 가정을 그대로 인정하더라도 그의 논증에는 결함이 있다. B의 행동에 대한 A의 통제력 부족이 A가 B의 행동의 결과에 대해 책임이 없음을 의미하지 않는다. B가 책임 능력이 없는 경우는 물론이고 책임 능력이 있는 행위자라고 하더라도 B의 행위에 대해 A에게 책임을 물을 수 있는 경우가 있다. 예를 들어, 상급자는 부하 직원의 행동에 대해 책임을 져야 하는 경우가 있다. 부하 직원의 행위가 상급자의 지시에 의한 것이거나 상급자에게 부하 직원의 행위를 관리, 감독할 권리와 의무가 있을 때 그러하다. 책임 공백의 존재를 수용하려면, 이러한 상식적 책임 이해가 자율형 인공지능에는 적용되지 않는다고 가정해야 하는데, 그렇게 할 이유를 찾기 어렵다. 스패로우의 시나리오에서 자율형 살상 로봇은 군대의 명령 체계 안에 있으며, 해당 로봇이 행위자라는 가정을 수용한다고 해도 지휘관은 로봇의 행동에 대한 지휘, 관리, 감독의 의무가 있다고 말하는 것이 잘못이 아니기 때문이다. 인간이 인공적 행위자를 통제할 수 없다는 해석을 받아들일 수 없다. 인간 행위자는 적어도 두 가지 방법으로 인공적 행위자를 통제할 수 있다. 인간은 인공적 행위자의 임무를 제한하는 방식으로 통제할 수 있으며, 자율적으로 행동할 것임을 알고 있는 인공적 행위자를 특정 종류의 상황에 배치하는 것에 관해서도 통제력을 발휘할 수 있다. 자율형 살상 로봇을 사용하는 시나리오에서 통제해야 할 것은 그런 로봇의 의사결정과 행동이 아니라 그런 로봇을 사용하기로 하는 인간의 의사결정과 행위이다. 그래서 이런 경우에 로봇의 행동으로 인한 결과에 대해 인간 행위자(이 경우 로봇의 전투 배치를 명령한 지휘관이 핵심적인 책임 능력자이며, 관리, 감독의 권한이 있는 현장 요원이 있는 경우에는 그 역시 책임 능력자임)에게 책임을 묻는 것은 정당하고 적절하다.

우리는 이런 상황을 일상적으로 경험하기도 하며, 이와 비슷한 상황을 상상하는 것이 어렵지도 않다. 훈련된 동물을 특정한 임무에 활용하는 경우를 상상해 보자. 마약 탐지견

도 좋고, 대테러 임무를 훈련받은 경찰견도 좋다. 자율형 로봇이 인간의 개입 없이도 스스로 행동한다는 의미에서 자율적 행위자라면 마약 탐지견이나 경찰견 역시 그런 의미에서 자율적이다. 훈련된 개는 인간이 없는 상황에서도 스스로의 판단에 따라 행동할 것이다. 그러나 우리는 비교적 독립적으로 특정 작업을 수행하도록 훈련된 개라고 하더라도 개에게 책임을 묻지 않는다. 개가 책임을 묻기에 적합한 존재라고 생각하지 않기 때문이다. 인간의 개입 없이 개 혼자 한 행동이라고 하더라도 책임질 상황이 발생한다면 그런 결과를 초래한 개가 아니라 누군가 인간에게 책임을 묻는다. 그리고 이런 식의 책임 추궁은 기이한 것이 아니다. 따라서 상대적으로 자율성을 지닌 로봇이 초래한 해악에 대해 해당 행위에 대한 직접적 인과성을 발견할 수 없다는 이유로 인간 행위자에게 책임을 물을 수 없다는 해석은 받아들일 수 없다.

스패로우의 시나리오에서 우리가 문제 삼아야 하는 것은 자율형 살상 로봇을 사용하기로 한 인간의 결정이며, 실제로 전투 현장에 투입한 행위이다. 우리는 그런 행위에 대해 충분한 정도의 통제력을 지니고 있다. 더욱이 우리는 그 로봇이 무엇인지, 전투 현장에서 어떤 유형의 행동들을 할지, 그것으로 인해 우리가 어떤 이득을 얻게 될지, 혹은 어떤 잠재적인 위험들이 있는지도 알 수 있으며, 설령 정확히 의식하고 있지 않다고 해도 의지 여하에 따라 그러한 사실들에 대해 충분한 정보를 얻을 수 있다. 무모함이나 태만은 일상적인 책임 개념의 핵심을 이룬다. 군사적 목적으로 이용되는, 더 정확하게 말해 인간(적군)을 죽일 수 있는 기능을 갖춘 전투 로봇을 실전에 배치한다는 것은 그것이 어떠한 것인지, 그것의 활약으로 어떠한 결과가 초래될지, 가능한 잠재적 위험이 어떤 것이 있는지 등에 대해 충분히 알아야 보아야 하고, 또 그런 앎을 자신의 행위 결정에 반영해야 한다. 이와 같은 식으로 사고하고 행하지 않는 것은 무모함이거나 정신적 태만에 해당하며, 그로 인해 발생한 해악에 대해서는 그 무모함과 태만을 핑계로 책임을 면할 수 없다. 오히려 그 무모함과 태만함 때문에 책임을 져야 한다. 이것이 책임에 대한 우리의 일상적인 이해이다. 결론적으로, 스패로우가 제시한 시나리오에서 로봇의 행동에 대한 인간 행위자의 인과적 공백 때문에 책임져야 할 사람을 찾지 못하는 일은 없을 것이다.

인과적 논증의 또 다른 형태는 이른바 '다수 관여자(many hands)'의 문제를 근거로 삼고 있다. 오늘날 기술들이 한 사람에 의해 만들어지지도 않고 그 기술로 인해 어떤 사건이 발생할 때 관여된 사람이 다수여서 한 개인으로 보면 해당 사건의 발생에 인과적으로

기여한 바가 매우 적다는 지적을 많은 연구자들이 한다. 개개인의 인과적 기여가 너무 적기 때문에 그들 개개인에게 책임을 요구하기 어렵다는 주장이다. 니센바움이 소개한 사례는 이 맥락에서 자주 인용된다.[15] 그것은 1985년에서 1987년 사이에 컴퓨터로 제어되는 방사선 치료기인 테락-25 (Therac-25)의 오작동으로 최소 6명의 환자가 방사선에 과다 피폭된 사건이다. 니센바움은 수 개월간의 분석 결과로 테락-25의 오작동이 다수의 결함에서 발생한 것임이 밝혀졌다고 설명한다. 하드웨어 언더록의 부재, 모호한 오류 메시지, 부적절한 검사 및 품질보증, 시스템 신뢰성에 대한 과장된 주장, 해당 병원의 과실 등 테락-25의 오작동을 유발하는 데에 '많은 손들'이 관여되었다. 니센바움은 테락-25 오작동 피해 사건에 관련된 원인들의 복잡한 그물망을 풀어내는 것이 어렵다고 주장한다. 고의적 범죄가 아니라 태만이나 무모함에서 비롯된 것이라고 해도 해당 사건의 원인을 정확히 밝히는 것이 쉽지 않다. 그래서 테락-25의 오작동 사건에 대해서는 누군가를 책임질 사람으로 지정하거나 비난할 수 없다는 의미에서 '단지 우발적 재난'이라고 결론내리지 않을 수 없다고 설명한다.

니센바움의 사례는 해당 사건에 관해 인과적으로 관여된 사람들, 다시 말해 책임을 져야 할 사람들이 한 개인이 아니라 다수일 수 있음을 보여주는 것이지, 책임 공백이 있음을 입증하지 않는다. 어떤 사건에 있어서 책임을 져야 할 사람이 두 명 이상인 경우가 적지 않으며, 다수에게 책임을 묻는 것이 우리의 일상적 책임 개념에 위배되는 것도 아니다. 문제는 다수의 관여자가 있기 때문에 책임져야 할 사람을 분별하는 것이 어려운 복잡한 사건이라는 것이다. 사건의 발생에 연루된 복잡한 인과적 관계들을 규명해 내는 것이 어렵다는 점이 면책 이유를 제공하지 않는다. 다수의 관여자가 있는 경우에 더 큰 문제는 책임의 분배이다. 인과적 기여도를 따져 해당 사건에 대해 책임의 정도를 분배하는 것이 합당하기는 하지만 여기에 우리의 착각을 불러오는 요소가 있다. 다수의 관여 있을 때 책임의 분배가 인과적 기여 정도에 따라 분배되어야 하고, 또 관여된 사람들의 수에 의해 산술적으로 분배된다는 생각이다. 이렇게 생각하면 해당 사건에 대한 인과적 기여자가 수가 많을수록 한 개인의 책임의 양이 줄어들고, 인과적 기여자의 수가 대단히 많을 때는 개인의 책임이 너무 미미해서 책임을 묻는 것이 의미가 없을 정도라고 생각하게 된다. 그

15) Helen Nissenbaum, "Accountability in a Computerized Society", in B. Friedman (ed.), *Human values and the design of computer technology*, Center for the Study of Language and Information, 1997, 41-64.

러나 도덕적으로 책임져야 할 사건에 대한 책임은 인과적 기여자의 숫자에 의해 산술적으로 분배되지 않는다. 테락-25 오작동 사건의 경우에 책임져야 할 사람들을 가려내는 일이 어려울 수는 있겠지만 원칙적으로, 또 현실적으로 불가능한 일이 아닐 것이다.

3.2. 책임 공백을 주장하는 인식적 논증에 대한 검토

인식적 논증은 책임 귀속의 또 다른 조건인 인식적 조건에 관련된다. 책임 귀속의 인식적 조건은 인식(knowledge)과 의식함(awareness)이다. 이 조건에 따르면, 어떤 행위자에게 책임을 묻는 것은 그 행위자가 자신의 행위와 그 결과에 대해 알고 있어야 한다고 여기기 때문이며, 그 자신의 행위의 도덕적 관련성과 문제있는 상황으로 귀결될 가능성 등을 의식하고 있어야 한다고 믿기 때문이다. 여기에는 행위자가 자신의 행위와 문제 있는 사건 사이의 연결에 관해 분별 있는 사람이라면 누구나 예상할 수 있는 정도로 인지한다는 점, 그리고 문제 있는 상황의 발생 가능성이 식별 가능한 수준이라는 점이 전제된다. 그런데 오늘날의 기술들은 사람들로 하여금 자신의 행위의 결과를 예측할 수 없게 만드는 특징이 있으며, 이 때문에 책임 공백이 초래된다. 이런 식으로 주장되는 인식적 논증은 몇 가지 양태로 세분된다.

먼저, 혁신적 기술들이 사람들로 하여금 책임 귀속의 인식적 조건을 만족시킬 수 없게 만든다는 주장이 있다. 신기술에서는 그것의 사용이 어떤 귀결을 낳을지 우리가 그 잠재적 영향들을 예측할 수 없는 경우가 종종 있다. 예를 들어, 내연기관의 초기 발명자들과 다수의 사용자들은 그것의 광범위한 사용이 지구온난화에 크게 영향을 미칠 것이라고 예측하지 못했을 것이다. 그런데 지구온난화는 이미 일어난 사태이고 누군가에게 책임이 있다고 말하는 것이 불합리하지 않아 보인다. 하지만 내연기관의 발명가들과 발명 이후의 다수 사용자에게 책임을 묻는 것이 적절할까? 내연기관의 발명가와 초기의 다수 사용자들에게 지구온난화에 대한 도덕적 책임을 묻는 것은 상식적이지 않아 보인다. 이 경우에는 분명 책임 공백이 존재한다.

이런 주장은 그럴듯해 보일지 모르지만 정당하게 성립하지 않는다. 우리는 예방하거나 교정했으면 좋았겠지만 그렇게 되지 않은 나쁜 사태와 누군가에게 책임이 있는 사태 사이를 구분할 수 있다. 후회할 만한 일들, 심지어 어떤 행위자의 후회가 정당한 것인 경우에도 그런 일들 모두에 대해 누군가에게 책임을 묻는 것이 적절한 것은 아니다.[16] 내연

기관의 경우에 지구온난화와 결부시켜 책임을 묻는 것이 적절하지 않지만, 현대 사회에서 신기술을 사용할 때 충분한 테스트를 거치고 필요한 예방조치를 마련하는 것은 필요하다. 모종의 신기술이 심각한 위험을 수반할 수 있음을 알고 있으면서, 그럼에도 불구하고 잠재적으로 발생 가능한 결과를 정확히 예측할 수 없다는 식의 변명은 통용되지 않는다. 스패로우의 사례에서 AWS에 관여된 책임 있는 사람들이 AWS의 잠재적 위험에 대해 알고 있으며 알 수 있기 때문에 AWS의 문제의 행위를 누구도 정확히 알아차릴 수 없다는 변명은 효과를 발휘하지 못한다.

또 다른 논증은 모종의 기술들은 사용자에게 심리적인 영향을 미쳐 사용자가 자신의 행위가 불러올 결과를 예측하기 어렵게 만드는 경향이 있다는 경험적 주장에 근거한 논증이다. 기술적 장치의 사용이 우리의 인식 능력에 악영향을 미쳐 정상적인 이해를 불가능하게 만든다면 우리의 행위로 인해 벌어진 사건에 대해 우리에게 전적으로 책임을 지게 하는 관행은 정당하지 않은 듯하다. 오늘날 우리가 사용하게 되는 복잡한 기계 장치의 경우에 그것을 사용하다 보면 시간의 경과에 따라 기계에 과의존하게 되거나, 사용자에게 과도하게 부담을 지우는 정보의 출력에 직면해 그것을 이해하려는 인식적 노력을 포기(인식적 항복)하는 쪽으로 심리적 영향을 받을 수 있다.[17] 다시 말해, 복잡한 기계 장치는 이용자가 자율적이고 지능적인 해당 시스템에 지나치게 의존하거나, 반대로 충분히 의존하지 않는 심리적 성향을 활성화시킬 수 있다.

미국의 미사일 순양함 빈센즈(USS Vincennes)호 사건은 시스템에 대한 과의존 사례를 보여준다. 1988년 호르무즈 해협에 있던 빈센즈호는 이란의 A300 항공기를 F-14로 오인하여 격추시켰고, 탑승객 290명 전원이 사망했다. 순양함의 이지스 방어 시스템은 수색 반경 안으로 들어온 미사일과 적 항공기를 자동으로 추적하고 타격하는 능력을 갖추고 있었다. 사건을 분석한 결과, 승무원들의 이지스 시스템에 대한 과신이 있었으며, 이런 과신이 사람이 개입할 수 있을 때 개입하는 것을 차단했다.

복잡한 시스템 이용자의 인식적 항복 상황은 테락-25의 사례에서 발견된다. 1985년 6월과 1987년 1월 사이에 최소 6건의 과다 피폭 사고를 낸 테락-25는 적정량의 30배에서 100

16) S. Köhler, N. Roughley & H. Sauer, “Technologically blurred accountability?”, in Cornelia Ulbert, et al, (ed.), *Moral Agency and the Politics of Responsibility*, Routledge, 2017, 96.
17) S. Köhler, N. Roughley & H. Sauer, “Technologically blurred accountability?”, in Cornelia Ulbert, et al, (ed.), *Moral Agency and the Politics of Responsibility*, Routledge, 2017, 97.

배까지 환자에게 방사선을 투여했다. 테락-25는 이전 모델에서 업그레이드 된 신형 모델이며 안전장치를 하드웨어 인터록 대신 소프트웨어 인터록을 사용했는데, 결과적으로 인터록이 제대로 작동하지 않았다. 이 사례에서 기기 운영자 중 한 명의 증언 내용을 들어보면, 기기에서 다수의 암호 오류 메시지가 발생했지만 작동에 아무런 문제가 없었고, 이내 오류 메시지에 익숙해졌다고 한다. 그래서 그 운영자는 기기의 오류 메시지들을 무시하게 되었으며, 결과적으로 기기가 환자에게 과다 투여를 결정했을 때 알아차리지 못했다.

위의 사례들에서 우리는 행위자의 책임을 면제하거나 경감할 정도로 인식적 조건의 충족을 방해하는 요인을 발견할 수 없다. 어느 것도 책임 공백을 입증하는 사례가 아니다. 위의 사례에서 발견되는 행위자의 귀책 요인은 전형적인 태만이다. 두 사례에서 관련된 행위자들은 자신들의 행위가 가져올 도덕적으로 유관한 귀결들을 고려하지 않았거나 충분히 고려하지 않았다. 첫 번째 사례에서는 기계의 기능에 지나치게 의존한 데에서 인식적 의무의 태만이 발생했으며, 두 번째 사례에서는 기계가 내놓은 의견을 제대로 고려하지 않은 데에서 행위자의 인식적 의무의 태만이 발행했다.

앞에서 언급한 스패로우의 시나리오에서는 인과적 공백 이외에 인식적 공백도 발견할 수 있다. AWS와 같은 자율형 기계 장치를 사용할 때 책임의 올바른 할당을 불가능하게 하는 인식적 공백을 야기한다. 왜냐하면 자율형 살상 로봇은 스스로 정보를 수집 · 분석하고, 자율적으로 판단하여 결정하고 독자적으로 행동하는 시스템이어서 인간은 그것의 행동 방식을 예측할 수 없기 때문이다. 만일 인간 행위자가 해당 로봇이 어떤 결과를 야기할지 예측할 수 없다면, 인간은 로봇의 운영에 관련된 사람이라고 할지라도 로봇이 야기한 결과에 대한 책임을 묻기에 적절한 행위자가 아니다. 그렇지만 사건이 발생했고 그에 대한 책임을 누군가는 져야 한다고 생각하는 것이 우리의 관행이다. 그래서 이 경우에 책임 공백이 발생한다. 이것이 스패로우 시나리오의 전형적 해석이다.

이 경우에 로봇의 운영에 책임이 있는 사람에게 로봇의 행동 결과에 대한 책임을 묻는 것이 왜 부적절한가? 치명적 무기를 장착한 자율형 로봇을 전투 현장에 풀어 놓는 경우에 그로 인한 이득에 대해서 뿐만 아니라 해악에 대해서도 운영자에게 책임질 몫이 있다. 로봇의 운영자(지휘관)는 로봇이 어떤 식으로, 즉 자율적으로 행동할 것임을 알고 있으며, 그로 인해 예측하기 어려운 위험이 발생할 수 있다는 것도 알고 있다. 결국 운영자는 현실적 위험과 잠재적 위험에 대해서 충분히 이해하고 있으며, 그것을 감수하고라도 로봇을

운영함으로써 얻는 이득이 더 크다고 판단한 것이다. 따라서 이 경우에 로봇의 행동과 관련하여 인식적 공백은 없다. 특정 목적에 훈련된 동물을 이용하는 경우와 다시 비교해 보면 사정이 더욱 분명하게 이해될 것이다. 예컨대, 경찰견을 이용해 임무를 수행하는 인간 행위자는 아무리 잘 훈련된 경찰견이라고 해도 100퍼센트 사람의 뜻대로 경찰견이 행동하지 않는다는 점과 잠재적인 위험이 있을 수도 있다는 것을 알고 있다. 그래서 경찰견이 임무를 수행하는 중에 책임져야 할 일을 저질렀다면 그에 대한 책임은 인간 행위자의 몫이다. 여기에 책임 공백은 없다. 스패로우의 시나리오에서 AWS의 해악은 예측 가능한 범위에 있으며, AWS의 구체적인 특정한 행동의 예측 가능성 결여가 책임의 부재를 만들어내지 않는다.

인식적 공백을 근거로 책임 공백을 논증하는 방식은 우리가 자신의 '행위에 대한' 인식적 조건을 만족시키는 것과 우리가 사용하는 '기계에 대한' 인식적 조건을 만족시키는 것 사이를 혼동하고 있다. 책임은 행위자와 그 행위에 적용된다. 우리는 자동차에 대해 완벽하게 이해하고 자동차의 각 부품과 구조, 작동 방식 등에 대해 온전한 지식을 가진 사람만 자동차를 이용할 수 있다고, 다시 말해 자동차를 운전할 수 있다고 주장할 수 없다. 거의 모든 자동차 운전자가 이런 지식을 가지고 있지 않으며, 그렇다고 해서 자동차 사고에 대해 운전자에게 책임을 묻는 것이 부당하다고 말하지도 않는다. 그러나 자동차를 판매하는 제조업체는 이러한 지식을 반드시 갖고 있어야 한다. 만일 자동차 제조업체가 이해하지 못하는, 지식을 갖고 있지 않은, 자동차의 어떤 동작이 있고, 이로 인해 사고가 났다면, 이 사고의 책임은 전적으로 제조업체에 있다. 만일 제조업체가 이해할 수 없는 동작의 영역이 있다는 것을 알면서도 그 자동차를 판매했다면, 이에 대해 제조업체는 도덕적 책임은 물론 법적 책임도 져야 한다. 자동차에 대해서는 '상식인' 책임 귀속의 관행이 자율형 인공지능(로봇)에게는 적용하기 곤란하다고 주장하는 이유가 무엇인지 이해하기 어렵다. 자동차 운전자가 자동차에 대해 알아야 하는 정도를 자율형 로봇의 이용자는 해당 로봇에 대해 알면 될 것이며, 만일 자율형 인공지능의 경우에는 시스템의 특성상 자동차 운전자의 지식보다는 좀 더 많은 지식이 필요하다면 그것을 규정하고 사용의 조건으로 제한하면 될 것이다. 그리고 자율형 인공지능의 운용으로 인해 벌어진 사건에 대해서는 인간 행위자에게 책임을 물으면 될 것이다.

나가는 말: 왜 책임 공백을 주장하는가?

이제 좀 더 근본적인 의문으로 돌아가 보자. 도덕의 영역에서 우리의 상식적 개념들은 완벽한 것은 아니지만 지금까지 우리의 도덕적 삶을 적절하게 유지하는 데 문제가 없었으며, 기술이 발전하고 문화가 변화하는 상황에서도 여전히 정상적으로 작동하고 있다. 그런데 왜 혁신적 기술들로 인해 책임 공백이 발생한다는 주장이 대두되는 것일까? 위에서 살펴본 것처럼 책임 공백이 발생한다고 제시된 사례들에 대해 전통적인 책임 개념과 상식적인 책임 귀속 관행을 적용해서 답변할 수 있다. 책임 공백을 인정하는 쪽보다 기존의 책임 귀속 관행을 따르는 쪽이 더 논리적이며 설득력 있어 보인다. 그러면 왜 책임 공백의 문제가 제기되었으며, 왜 많은 사람들이 이에 민감하게 반응하는가?

기술의 발전에 따라 기존의 책임 귀속 관행에 변화를 요구하는 주장은 오늘날의 특별한 일이 아니다. 미국 메릴랜드대학의 도날드 기포드(Donald G. Gifford)는 미국에서 불법행위에 관한 법률이 기술의 변화 상황에 따라 어떠한 중대 변화를 겪어왔는지를 잘 보여줬다. 기포드에 따르면 산업혁명 이후 기계가 생산의 주요 동력으로 활용되면서 기존의 엄격한 불법행위의 표준이 과실 제도로 대체되는 것을 비롯해 변화를 겪었다. 그는 이런 변화를 기술 혁신이 개인 부상의 빈도, 부상의 심각성, 주장 입증의 어려움, 신기술의 사회적 유용성 등의 기준에 따라 설명하는 모델을 제안했다.[18]

이런 맥락에서 보면, 산업혁명 이후 기술 혁신에 따라 기존의 엄격한 책임 개념에서 책임의 경감, 책임의 분산, 책임의 면제 범위 확대 등의 방향으로 변화가 이루어지고 있다고 판단할 수 있다. 최근의 주장은 이런 방향을 좀 더 극단으로 몰고 간 것이다. 책임져야 할 존재를 인간 행위자, 이른바 책임의 주체에서 인공적 행위자로 확대하여 인간과 인공적 행위자 사이의 분산된 책임, 심지어 인공적 행위자의 단독 책임을 주장하기 때문에 극단이다.

현대의 기술들이 책임 귀속과 관련된 상황에 변화를 가져온다는 주장을 자주 듣는다. 우리 주변에서 흔히 접하는 인공지능을 활용한 시스템이 사람들의 합리적이고 자기통제적인 행동을 어렵게, 혹은 불가능하게 만든다는 것이다. 예를 들면, 건강관리 앱은 사용자의 생활양식에 대한 데이터를 기반으로 사용자에게 일정한 행동 유형을 제안한다. 온라인 판

18) Donald G. Gifford, “Technological triggers to tort revolutions: steam locomotives, autonomous vehicles, and accident compensation”, *Journal of Tort Law* 11(2018), 71-143.

매 사이트의 추천 알고리즘은 웹사이트에서 사용자의 검색 기록을 기반으로 사용자가 관심을 가질 만한 제품을 추천한다. 사람들은 이러한 기술적 시스템 때문에 자신의 행동에 대한 변명거리가 있다고, 또는 책임이 경감될 수 있다고 믿는다. 우리가 다른 사람에 의해 암묵적으로 조종되거나 강제로 강요되었을 때 우리의 행위로 인해 비롯된 결과에 대해 책임이 경감되거나 면제될 수 있는데, 오늘날 기술적 시스템들이 우리의 행위에 영향을 미치는 방식이 이와 유사하다고 생각하기 때문이다.

현대 사회는 유형의 기술뿐만 아니라 다양한 무형의 기술도 발전시켜왔다. 그리고 기술이 사회적으로 수용될 때 효용성의 관심과 윤리적 관심 사이에서 갈등이 발생한다. 혁신적인 기술이 도입되면 우리의 삶에 많은 변화, 때로는 급격한 변화를 불러올 수 있다는 예상은 어렵지 않다. 그런데 그런 변화에 대한 고려를 어디까지, 어떤 식으로 할 것인지를 판단하는 정해진 규칙이 없다. 반면에 기술의 효용은 상대적으로 예측하기 쉽다. 물론 기술의 도입으로 발생한 이득의 수혜자를 누구로 볼 것인지에 따라 다른 평가가 가능하지만 말이다. 일반적으로 신기술 도입의 최대 장애물은 윤리적 관심이다. 그래서 신기술의 도입은 기존의 윤리적 규범과 개념을 변형하고 약화시키는 방식으로 더 큰 효용을 얻을 수 있는 듯하다.

현대 사회는 유무형의 기술을 통해 책임을 분산하는 전략을 지속적으로 실행해 왔다. 인공적 도덕 행위자의 인정과 책임의 주체로서의 인공적 행위자라는 개념은 책임의 분산을 넘어 책임 회피의 전략이 아닌가 의심스럽다. 바로 앞에서 일종의 넛지 효과의 사례를 들어 일반 대중에게서 나타나는 책임경감에 대한 믿음의 경향을 언급했지만, 책임 분산과 책임 회피의 전략은 첨단 기술의 핵심 관련자들에게서 뚜렷하게 나타난다. 최초의 챗봇인 일라이자(Eliza)를 개발한 MIT의 컴퓨터 공학자 요제프 바이첸바움(Joseph Weizenbaum)은 이 점을 분명히 인식하고 있었다.

> 군사 기술 시스템에서 볼 때는 인간이 가장 약하고 불확실한 부분입니다. 최근 몇 십 년 동안 군사적 영향이 두드러지게 드러날 만한 사건이 몇몇 있었는데, 그때마다 개인이 무엇을 해야 할지 스스로 결정해야 했어요. 미국 해군이 페르시아만에서 격추시킨 이란 여객기가 생각나네요. 이 여객기를 적군의 공격용 비행기로 여겼기 때문에 벌어진 사건이었죠. 1988년의 일입니다. 그 순간 한 인간이 격추 여부를 결정해야 했어요. 그것도 대단히 빨리 말입니다. 좀 더 꼼꼼히 생각하거나 신중하게 검토할 시간이 도대체 없었으니까요. 결정이 어떻게 내려졌는가는 잘 알려져 있는 사실이에요. 그런데 이 사건에 대한 반응이나 이 사건으로 인해 수

반된 결과는 사격 명령의 책임에 관한 것들이 아니었어요. 오히려 명령 시스템을 자동화시킴으로써 이런 경우 '책임을 지는' 인간, 즉 지휘관을 명령 시스템에서 빼내는 것이었어요.[19]

오늘날 기술적 변화가 특정 사건에 대한 책임을 규명해야 할 때 책임져야 할 행위자를 가려내기 힘들 정도로 복잡하게 만드는 경향이 있는 것은 사실이다. 여기에는 무형의 기술(제도)도 중요한 역할을 하며, 앞에서 서술했듯이 통제하기 어려운 기술적 복잡성도 관련이 있다. 책임의 분산에 기여하는 복잡한 시스템은 해당 사건에 연루된 행위자의 수를 크게 늘려 놓았다. 이런 여러 정황이 책임 규명이 필요한 사건에서 책임의 주체를 밝혀내고 책임의 정도를 명확히 규정하기 어렵게 만든다. 그러나 이런 사정이 책임의 공백을 인정해야 할 충분한 이유가 되지 않는다. 오히려 책임의 공백을 인정하는 것은 결과적으로 도덕의 공백을 불러올 것이다. 책임 공백을 메우기 위해 로봇에게 행위자의 지위, 나아가 도덕적 행위자의 지위를 인정하고, 로봇에게 법적 책임을 묻고 도덕적으로 질책하는 상황을 상상해 보자. 이때 인간은 책임이라는 공기의 저항을 모두 물리치고 도덕의 무중력 상태에서 마음껏 허공을 날 수 있을 것이다. 하지만 도덕의 무중력 상태는 우리에게 허공을 마음껏 날 수 있는 자유를 선사하지 않을 것이다.

19) 요제프 바이첸바움 · 군나 벤트, 『이성의 섬』(모명숙 옮김), 양문, 2008, 39-40.

참고문헌

신상규, “인공지능 시대의 윤리학”, 『지식의 지평』 21, 2016.

요제프 바이첸바움 · 군나 벤트, 『이성의 섬』, 모명숙 옮김, 양문, 2008.

윌러치 & 알렌, 『왜 로봇의 도덕인가』, 노태복 옮김, 메디치, 2014.

제리 카플란, 『인간은 필요 없다』, 신동숙 옮김, 한스미디어, 2016.

Andreas Matthias, “The responsibility gap: Ascribing responsibility for the actions of learning automata”, *Ethics and Information Technology*, 6(2004), 175-183.

Anna Strasser, “Distributed responsibility in human–machine interactions”, *AI and Ethics*, 2(2022), 523–532.

Aristotle, *Nicomachean Ethics*, 1111a3-1111a5.

Donald G. Gifford, “Technological triggers to tort revolutions: steam locomotives, autonomous vehicles, and accident compensation”, *Journal of Tort Law* 11(2018), 71–143.

Helen Nissenbaum, “Accountability in a Computerized Society”, in B. Friedman (ed.), *Human values and the design of computer technology*, Center for the Study of Language and Information, 1997, 41–64.

Jan-Hendrik Heinrichs, “Responsibility assignment won’t solve the moral issues of artificial intelligence”, *AI and Ethics*, 2(2022), 727-736.

Mark Coeckelbergh, *AI Ethics*, MIT Press, 2020.

Robert Sparrow, “Killer Robot”, *Journal of Applied Philosophy*, 24(2007), 62-77.

Sebastian Köhler, Neil Roughley & Hanno Sauer, “Technologically blurred accountability?”, in Cornelia Ulbert, et al, (ed.), *Moral Agency and the Politics of Responsibility*, Routledge, 2017.

Sven Nyholm, “Attributing agency to automated systems: reflections on human-robot collaborations and responsibility-loci”, *Sci. Eng. Ethics*, 24(2018), 1201–1219.

‘인공지능의 책임’ 쟁점에 대한 비판적 고찰

이상헌

지능, 자율적 행위 능력, 학습 역량을 갖춘 기계가 등장하고 있고 그 쓰임이 확대되고 있는 상황에서 새로운 철학적, 윤리적 쟁점들이 대두되고 있다. 그 중 하나가 책임의 문제이다. 지능을 갖춘 자율형 로봇의 행동으로 인한 해로운 결과에 대해 누구에게 책임을 물어야 하는가? 누구도 해당 로봇의 특정 행위에 대해 직접적인 통제력을 갖고 있지 않기 때문에 사람에게 행위를 물을 수 없지만, 그렇다고 해서 로봇에게 책임을 묻는 것도 적절해 보이지 않는 상황을 책임 공백(responsibility gap)이라고 부른다. 이 논문은 책임 공백의 쟁점의 등장 배경을 소개하고, 책임 공백의 쟁점에 대응하는 다양한 방식을 살펴본다. 책임 공백의 쟁점에 대한 대응은 크게 두 부류로 나뉜다. 한 부류는 책임 공백을 실질적인 문제로 인정하지 않는다. 다른 부류는 책임 공백을 인정하고 이에 대응하는 방안을 모색하려고 한다. 이 논문에서는 특히 책임 공백을 인정하는 논의를 집중적으로 다룰 것이다. 이 쟁점에 관한 논의에 참여하는 다수의 연구자들이 책임 공백의 존재를 너무 쉽게 인정하는 듯한 인상을 받았기 때문이다. 나는 이들의 논의를 주의 깊게 살펴보면서 책임 공백의 문제를 좀 더 원론적인 물음에서 다루고 싶었다. 그래서 책임 공백을 인정하는 핵심 논증인 인과적 논증과 인식적 논증을 비판적으로 검토하고, 다수의 연구자들이 책임 공백을 주장하는 이유에 대한 나의 견해를 밝히려고 한다.

주제어: 인공적 도덕행위자(AMA), 인공지능의 책임, 책임 공백, 다수 관여자 문제, 로버트 스패로우

Abstract

A Critical Reflection on the Issue of 'the Responsibility of Artificial Intelligence'

Rheey, Sang-Hun

In the current context where machines with intelligence, autonomous action capabilities, and learning capacities are emerging and their applications are expanding, new philosophical and ethical issues are arising. One of these issues is the question of responsibility. Who should be held responsible for harmful outcomes resulting from the actions of intelligent autonomous robots? Since no one has direct control over specific actions of these robots, it seems inappropriate to attribute their actions to humans. This situation is referred to as the 'responsibility gap.' This paper introduces the background of the emergence of the responsibility gap and explores various ways to address it. Responses to the issue of the responsibility gap can be broadly categorized into two types. One type does not consider the responsibility gap a substantial problem, while the other acknowledges it and seeks ways to address it. This paper will particularly focus on discussions that recognize the existence of the responsibility gap. Many researchers participating in this discussion have given the impression of too easily acknowledging the existence of the responsibility gap. I wanted to delve into the issue more deeply from a fundamental perspective, so I critically examine the causal argument and the cognitive argument, which are the core arguments supporting the recognition of the responsibility gap, and present my perspective on the reasons why many researchers argue for the existence of the responsibility gap.

Key Words: artificial moral agent(AMA), responsibility of AI, responsibility gap, many hands problem, Robert Sparrow

논문 투고일	2023년 11월 5일
논문 수정일	2023년 12월 1일
논문게재 확정일	2023년 11월 27일