과학철학 26권 3호 (2023) pp. 71-94

인공지능으로부터 야기되는 윤리적 문제에 대한 사회적 합의는 가능한가? - 고인석 저, 『인공지능과 로봇의 윤리』에서 인공지능의 자율성과 책임귀속의 문제를 중심으로[†]

김 태 경 [‡]

고인석의 『인공지능과 로봇의 윤리』는 인공지능과 같은 인공물의 기능으로부터 나타나는 현상들에 대응하는 개념들—자율성, 행위, 의도, 지각, 주체 등—에 대한 철학적 검토를 통해 인간의 자율성과 인공물의 자율성을 구분하고, 인공물의 자율성은 인간의 그것과 다르므로, 인공물의 작동은 현상적으로 보기에 자율적인 움직임으로 보이지만 인간의 자율적 행위와 본질적으로 다르기 때문에 이러한 존재들에 대한 규제를 이들에게 위임하기 보다는 사회적 합의와 결단을 통해서 마련해야 함을 주장한다. 논자는 저자의 논의가 정당하게 받아들여지기 위해서는 다음의 두 가지가 전제되어야 함을 지적한다. 하나는 내적상태를 알 수 없는 존재들에 대해 내린 우리의 사회적 합의가 합리적이었는지의 여부에 대한 것이며, 다른 하나는 우리가 의사결정에 있어 진정한 의미의 자율적 존재에 해당하는지에 대한 것이다. 이를 통해 논자는 인공지능과 같은 인공물이 야기하게 될 여러 가지 유형의 윤리적 문제에 대한 논의는 인공물이 그 대상이 아니라, 본질적으로 우리의 윤리적 판단에 대한 논의임을 주장한다.

주요어 : 인공지능, 행위, 자율성, 윤리적 판단, 사회적 합의

† 이 논문은 2023년 한국과학철학회 정기학술대회의 북심포지엄 (고인석 저, 『인공지능과 로봇의 윤리』)에서 발표된 내용을 수정·보완한 것이다. 훌륭한 책을 써주신 고인석 선생님과 예리하지 못했던 토론자의 논의를 토론을 통해 섬세하게 다듬어 주신 신상규, 노형래 두 분 선생님들께 감사드린다. 그리고 부족한 논문을 심사하고 예리한 지적을 통해 미비된 점을 보완할 수 있도록 도와주신 익명의 심사자 분들께도 감사의 인사를 전한다.

‡ 제주대학교 윤리교육과 조교수 (ktkw21@jejunu.ac.kr)

1. 들어가는 말

고인석(이하 '저자')은 그의 책 『인공지능과 로봇의 윤리』에서 인공지능 기술이 우리의 예측보다 빠른 속도로 발전하고 이를 활용한 다양한 기계가 일상화되고 있는 이 시점에서 현안을 깊이 있게 검토하고 앞으로 우리가 어떤 자세를 취하는 것이 적절한 지에 대한 논의를 진행한다. 이 책의 여러 논의들 중 논자가 주의 깊게 바라본 것은 5장과 6장 그리고 8장에서 다루고 있는 인공지능의 자율성 여부와 책임귀속에 관한 것인데, 이는 인공지능과 같은 인공물이 우리의 일상생활에 빠르고 깊게 침투되고 있는 현재의 시점에서 비판적으로 검토되어야 할 중요한 문제는 무엇인지에 대해 고찰하는 의의를 지니고 있기 때문이다.

우리는 일반적으로 어떠한 행위로부터 피해가 발생하는 경우, 해당 행위의 주체에게 책임을 귀속시킨다. 예를 들어, 경이라는 사람이 철이를 살해하기 위해 전자동 소총을 한 건물의 옥상에 설치해 놓았다고 가정해보자. 이 소총은 경이가 세팅한대로 작동하며, 경이가 소총을 작동시키고 싶지 않은 경우, 무선 스위치를 누르면 소총은 작동되지 않는다. 경이는 평소에 자신의 험담을 하고 다니는 철이를 살해하고자 철이의 출근 시간에 맞춰 철이가 회사 입구에 다다르는 즉시 소총이 발사되도록 세팅하였다. 그런데 그 시간에 철이와 체격과 옷차림이 매우 흡사한 돌이가 철이보다 조금 앞서 회사 입구에 다다르게 되었고, 소총은 돌이를 철이로 인식하고 작동하여 안타깝게도 돌이가 살해되었다. 이 경우, 우리는 경이가 직접적으로 소총의 방아쇠를 당긴 것도 아니고, 그가 의도한 바대로 소총이 작동한 것도 아니므로 책임이 경이에게 없다고 하지 않는다. 경이의 직접적 행위가 아니라도, 소총이 오류를 일으켰어도, 소총을 세팅한 사람이 경이라는 것이 알려지면 이 사건에 대한 책임은 경이에게 있다. 다음과 같은 경우도 마찬가지다. 경이가 철이를 살해할 마음을 접고 손안의 무선 스위치를 눌렀지만, 소총이 설치된 옥상에 날아든 비둘기 한 마리가 소총의 무선시스템을 살짝 건드려 소총이 작동돼 철이나 돌이가 사망하게 되는 경우에도 우리는

비둘기에게 그 책임을 묻지 않는다.

전자동 소총을 설치하여 다른 사람에게 피해를 입힌 경이의 행위와 소총을 건드린 비둘기의 행위는 소총으로 인해 발생한 사건의 원인이 되는 존재라는 측면에서 같은 행위자로 인식될 수 있지만 책임에 대한 귀속은 서로 다르다. 우리는 인간에게 책임을 부여하고 동물에게는 부여하지 않는다. 그 이유는 인간과 달리 동물이 피해에 대한 보상을 할 능력이 없어서가 아니라 그들이 가지지 못하는 도덕적 지위를 인간이 가지기 때문이다. 즉 인간은 도덕적으로 대우받아야 할 권리를 지님과 동시에 다른 존재에게도 불필요한 해악을 가하지 말아야 할 의무도 동시에 지니는 도덕적 객체이자 주체이지만, 동물은 도덕적으로 대우받을 권리는 지니지만 우리가 가진 의무는 지니지 않는 도덕적 객체에만 해당하기 때문이다.[1)]

이와 마찬가지로 만 14세 미만의 소년들에게도 성인과 동일한 법적 책임을 부과하지 않는다. 이른바 촉법소년[2)]의 경우가 이에 해당하는데, 현행 헌법 제 9조에 따르면, 형사미성년자의 기준인 14세 미만의 소년의 경우 중범죄를 저질러도 보호처분의 대상이 될 뿐 형사처벌의 대상이 되지는 않는다.[3)] 우리가 이들에게 성인과 동일한 처벌을 하지 않는 이유는 이들이 도덕적 객체에만 해당해서나 자신의 행위로부터 발생할 결과를 인식하지 못해서도 아니며, 이들의 행위가 **순수한** 자율적 행위가 아니기 때문도 아니다. 범죄의 원인은 여러 환경적 요인들이 작용했을 수도 있지만 보다 중요한 것은 형법에 의해 그들의 지적·도덕적 성장 상태와 무관하게 **절대적 책임무능력자**로 규정되고 있기 때문이다. 이는 소년들이 실제 사물을 변별할 수 있는 능력이나 의사를 자유롭게 결정할 수 있는 능력이 있음에도 불구하고 사물을 변별할 수 있는 능력과 의사결정능력이 없는 것으로 간주하

1) 김효은 (2019), pp. 19-20 참고.

2) 2008년 6월 22일부터 개정된 소년법에 따라 "촉법소년은 소년법 제4조 1항 2호의 형벌 법령에 저촉되는 행위를 한 10세 이상 14세 미만의 소년이라고 규정하고 있다." 2008년 개정 이전까지는 소년법 적용대상 상한이 20세였지만, 19세로 하향되었고, 촉법소년도 12세였다가 10세로 하향되었다. 점승헌 (2022), p. 347 참고.

3) 같은 글, 349쪽.

는 규정이다. 이에 따라 만 14세 미만의 소년은 형법적으로 비난이나 처벌의 대상이 아닌 '교육 내지 보호의 대상'으로 규정되고 있는 것이다.[4] 물론 이러한 책임무능력자에 대한 형법적 규정은 임의적인 것이 아니라 모종의 **사회적 합의**를 기반으로 한 것이며 **합리적인 결정**에 해당한다.[5]

위의 두 사안은 다르지만 논자가 주목하는 저자의 흥미로운 논의와 매우 밀접하게 연결되어 있다. 바로 인공지능의 자율성 여부와 책임의 귀속에 대한 것이다. 저자는 인공지능의 행위를 지칭할 때 사용되는 '자율적'이라는 표현이 인간에게 사용되는 그것과 동일한 의미인지를 먼저 검토하고, 이를 토대로 인공물이 책임과 권한의 주체가 되는지의 여부를 논의한다. 이로부터 인공지능에 책임을 귀속시킬 수 없기 때문에 모종의 사회적 합의를 통해 이를 적절히 규제하는 방안을 마련해야 함을 주장한다. 논자는 앞서 소개된 두 예시가 인공물에 대해 저자가 가진 문제의식과 어떻게 연결되어 있으며, 동시에 저자의 논의에 담긴 문제점은 무엇인지, 그리고 이 논의를 통해 우리가 진정으로 주목해야 하는 것은 무엇인지를 말하고자 한다.

2. 인공물의 행위

인공지능과 같은 인공물의 움직임은 현상적 차원에서 행위주체에 해당되는 기능을 보여주기도 하지만 행위주체에게 부여되는 도덕적·법적 책임의 의무를 가지지는 않는다. 이 때문에 우리는 인간에 의해 설계되고 제작된 인공물이 *스스로* 일으킨 피해에 대한 책임을 귀속시키는 문제에 혼란을 겪는다. 그중 하나는 인공지능과 같은 인공물이 가지는 책임의 개념이나 행위의 자율성이 인간과 동일하지 않기 때문이 아니라, 행위의 결과에

4) 물론 형사처벌을 받지 않는다고 해서 아예 책임을 부여하지 않는 것은 아니다. 이들에게는 소년법에 의한 보호처분이 시행된다. *Ibid.*, p. 354 참고.

5) *Ibid.*

대한 책임을 부과했을 때, 해당 인공물이 인간처럼 그 책임을 다할 수 없기 때문[6]이다. 따라서 인공지능이 책임의 개념을 인식할 수 있다고 가정해도 책임을 행위의 주체인 인공지능에게 귀속시키는 것은 불가능하다.

다른 혼란은 인공지능의 행위와 인간의 행위 사이에 나타나는 차이로부터 발생한다. 인공지능은 인공물임에도 불구하고 그것이 가지는 기능에 책임을 함축하는 행위주체성 내지 자율성(agency or autonomy)의 개념을 적용시키기에[7] 발생하는 혼란이다. 저자는 이러한 개념들을 철학적으로 검토하고 정리함으로써 이로부터 발생할 수 있는 혼란을 먼저 방지하려 한다. 저자가 제시한 아래의 예시는 위에 대한 문제를 잘 드러낸다.

> 2019년 10월 태풍 하기비스가 몰고 온 기록적인 폭우는 일본 나가노현의 지쿠마강을 범람시켰고 이로 인해 인명과 재산의 피해가 발생했다. 이것이 2019년 가을에 발생한 사건의 일면에 대한 **참된 서술**[8]임을 전제하고 따져보자. 지쿠마강을 범람시킨 것은 누구, 혹은 무엇인가? 그것은 태풍이 몰고 온 폭우였다. 그렇다면 강의 범람으로 인한 피해의 책임이 태풍에게 있다고 할 수 있을까?[9]

저자에 따르면, 우리는 태풍과 같은 자연재해와 관련된 피해에 대해 자

6) 금전적, 물리적, 정신적 손해배상과 같은 응보주의의 관점에서 그러하다. 하지만 인공물의 책임 실현 가능성에 관한 최근 논의에서 전자인격의 부여와 같이 적어도 금전적 배상은 가능할 것이고, 사람들 간의 일에서도 대부분 금전 배상이 실질적인 해결 방안이 되는 점을 고려할 때 이 논의가 약화될 수 있다는 주장이 있을 수 있다. 그러나 이는 인간에게 인공지능의 행위에 대한 책임을 대리하게 하는 방식에 해당하며, 논자가 언급하는 철학적 의미의 자율적 존재의 개념에도 부합하지 않는다. 또한 앞의 예시와 같이 인공지능이 제작자나 설계자 혹은 사용자가 의도하지 않은 행위를 하는 경우, 이에 대한 책임을 부과하는 것 역시 정당하다고 보기 어려운 측면이 있다. 따라서, 인공지능은 행위에 대한 책임의 소임을 직접적으로 할 수 있는 존재에 해당되지 않으므로, 스스로 행위를 일으킬 수는 있어도 책임을 질 수 있는 존재는 아니다.

7) 변순용, 이연희 (2020), p. 65.

8) 논자의 강조.

9) 고인석 (2022), p. 135.

연에게 책임을 묻지는 않지만, 자연현상과 그로 인해 발생한 피해에 대한 진술은 행위와 결과 사이의 인과적 관계를 보여준다. 예를 들어 '태풍이 동반한 폭우가 지쿠마강의 범람을 유발했다'라는 진술은 사실적으로 참이며, 이로부터 강의 범람과 범람으로부터 발생한 피해의 인과적 원인을 태풍에게 돌릴 수 있다. 하지만, 이러한 진술이 참이라고 하는 것이 피해의 책임이 태풍에게 있다는 것을 의미하는 것은 아니다. 저자의 언급대로 우리는 태풍에게 보상을 요구하거나 태풍에게 책임을 묻거나 벌하지도 않는다. 이는 태풍이 책임을 질 수 없는 존재이기 때문이 아니라 책임의 개념이 본질적으로 인과의 개념과 다르다는 것을 의미한다. 저자가 위의 예시를 통해 보여주고자 하는 것은 인공물의 자율적인 행위(공학적인 개념에서)로 인해 피해가 발생하는 경우, 그러한 피해에 대한 책임 귀속을 행위와 결과에 대한 인과적 관계나 자율성의 개념에 의거하여 부여할 수 있는지 아니면 다른 조건에 따라 부여해야 하는지에 관한 것이다.

물론 저자는 적지 않은 부분에서 인공지능이 행위주체에 해당하는지에 대한 판단 여부는 우리가 이미 알고 있는 행위주체의 조건에 부합하는지에 따라 결정되는 것이 아니라 사회적 합의에 의해 결정되는 것임을 강조한다. 이것은 인공물이 행위주체의 조건을 완전히 만족시키지 않아서라기보다는 3인칭의 과학적 관점에 의해 행위주체의 조건을 일부 만족하더라도 행위주체로 불리우는 존재들이 갖는 다른 하나의 기능(책임부여 및 이에 대한 수용 여부)을 수행하지 못하기 때문이다. 즉, 저자가 글을 통해 논의하고자 하는 것은 "책임 귀속과 결부된 행위이고, 그러한 의미의 행위 주체성"[10]의 조건을 인공물의 작동이 만족할 수 있는지의 여부인 것이다. 물론 저자는 먼저 이러한 논의가 인공물이 행위주체성을 가지지 않는다는 것을 전제하는 것은 아니라고 말한다.[11] 저자가 주목하는 것은 자연재해의 피해보상을 자연이 하지 않고 국가와 같은 사회가 직접적으로 개입하는 것과 같이 인공물의 **작동**으로부터 발생한 피해에 대한 책임 귀속에 있어 모종의

10) *Ibid.*, p. 137.
11) *Ibid.*, p. 138.

정당화된 사회적 합의로서의 규범과 기준을 설립하는 것이다. 그러기 위해서는 먼저 책임귀속과 관련있는 행위주체성의 조건을 인공지능이 만족하는지 여부를 따져보아야 한다. 즉, 인공지능과 같은 인공물이 행위주체성의 조건을 만족하지 않는다면, 이들의 행위로부터 발생한 피해에 대한 책임 귀속의 문제는 정당화된 사회적 합의를 필요로 하기 때문이다.

2.1. 행위의 성립요건

저자는 행위주체성을 가지는 행위의 성립요건으로 의도, 조절, 메타 인지의 세 가지 기준을 검토한다. 논자는 이 세 가지 기준들 중 저자가 중요하게 고려하고 있는 의도와 조절을 검토한다.[12)13)]

먼저 의도에 대해 저자는 형법에서 범죄의 성립조건[14)]으로 범의(범죄를 행할 의사가 있음)가 전제되어야 함을 먼저 언급하고, 행위 주체의 의도가 행위의 성립과 귀속을 결정하는 조건이 된다고 주장한다. 그 이유는 행위와 행위 주체간의 대응은 주체의 의도에 의해서 결정되기 때문이다.[15)] 특

12) 저자는 의도와 조절을 행위를 성립시키는 핵심요건으로 보고 있지만, 메타인지가 없다고 해서 행위가 불가능하다는 결론을 내리는 것은 불합리하다고 말한다. 그 이유는 메타인지를 결여한 사람이라고 해도 일으킨 행위에 대한 부분적 책임까지 없는 것은 아니기 때문이다. 논자 역시 이에 동의하며, 이 부분은 글에서 다루지 않는다.

13) 이 논의는 저자가 5장에 이어 6장에서 다루고 있는 '인공지능의 자율적 존재 여부'와도 밀접한 관계가 있다. 저자는 단순히 스스로 움직일 수 있는 자율적 기계의 성향이 자율성의 개념에 해당한다고 보지 않으며, 따라서 인공지능이 가진 기능적 속성만으로는 자율적 존재의 지위를 부여받을 수 없다고 주장한다.

14) 저자의 글에 적지 않은 부분에서 현행 법률에서 정의한 개념이나 판단이 저자의 주장에 대한 근거로 사용된다. 저자가 목표로 하는 것이 행위에 대한 책임과 관련된 부분이라면 이는 법률적 문제이기도 하지만 그에 앞서 윤리적 문제에 해당한다. 법률적 판단이 윤리적 논증에 근거가 되는지에 대해 다양한 의견이 제시될 수 있겠지만 이는 논의하지 않으려 한다. 그 이유는 현재 논자가 주목하고 있는 인공지능의 자율성 문제와 행위에 대한 책임귀속 여부에 대해 저자는 사회적 합의와 공동체의 결단에 달린 것이며, 그러한 성격이 잘 반영된 것이 법이라고 판단하고 있는 것으로 보이기 때문이다.

15) *Ibid.*, p. 138.

정 행위에 대한 도덕적 판단에 있어서도 우리는 해당 행위가 어떠한 의도에 의해 실행되었는지를 고려하는 것은 일반적이다. 그렇다고 해서 의도만이 행위를 성립시키는 것은 아니다. 저자에 따르면, 그러한 의도가 실제로 실행될 수 있도록 주체는 자신의 신체를 조절할 수 있어야 한다. 즉, 의도한 바에 따라 신체를 조절하지 않으면, 행위 자체가 일어나지 않거나(무위), 의도한 바와 다른 행위, 즉 오류가 발생할 수 있다. 흥미로운 점은, 오류가 발생하는 경우, 행위에 대한 평가는 맥락-의존적(사회의 관행, 환경적 영향 등) 문제에 해당한다는 것이다. 하지만 비록 행위가 속해있는 맥락 안에 여러 가지 고려 사항이 있다고 하더라도 의도가 행위를 성립시키는 조건이 되지 못한다는 것은 아니다. 법적으로 문제가 되는 행위의 경우, 행위에 대한 판단 중 하나가 행위 주체의 의도이며 우리는 행위 주체에 직접 실현, 혹은 오류에 대해서도 책임을 묻기 때문이다. 실제 우리는 범죄에 대한 유·무죄를 판단하여 형량을 구형할 때 범행에 대한 의도, 즉 범의가 있었는지의 여부를 매우 중요하게 검토한다.

이 두 가지 검토는 저자가 다루고자 하는 인공물의 행위에 대한 책임 귀속 여부에 매우 중요한 질문을 던진다. 하나는 특정한 업무를 수행하는 '기계가 인간과 같이 내적 상태(의도)를 가지고 실현시킬 수 있는 존재인가?'이며, 다른 하나는, 설계된 인공물의 본래 목적과 다르게 오류 등으로 인한 피해가 발생하는 경우, 과연 '누구에게 책임을 귀속할 것인가?'이다. 첫 번째 질문은 우리의 마음이 인공지능의 알고리즘이 처리하는 정보 또는 정보처리과정과 동일하다는 입장을 가진 경우에 '그렇다'라는 답은 내놓을 수 있겠지만 저자의 대답은 '그렇게 보기 어렵다'이다. 두 번째 질문에 대해 저자는 '사회적 합의와 결단과 같은 공동체의 결정에 달린 문제'라고 답한다.

2.2. 내적상태와 행위에 대한 판단기준

논자는 첫 번째 질문에 대한 저자의 답변에 공감한다. 저자는 행위자의 의도에 따라 일관성있게 조절되는 통일성의 실현을 행위주체의 성립조건

으로 보고 인공지능과 같은 기계가 이러한 특성을 가진다기 보기 어렵다고 말한다.[16] 물론 저자의 이러한 답변 안에는 인공지능이 특정한 목표를 실행시키기 위한 동작을 끊임없이 반복하여도 목표를 달성하는 결과를 산출하지 못하기 때문에 위의 조건을 만족하지 않는다고 말하는 것은 아니다. 인간은 어떠한 의도에 따라 행위하려고 할 때, 주변 환경으로부터 발생하는 수많은 변수들(환경적 영향이나 유혹 등) 사이에서 자신의 지각과 행위를 조절하는 특징을 지닌다. 그리고 이것은 기계적 특성이라기보다 우리의 마음의 1인칭적 측면을 닮아있다.[17] 논자는 저자가 말하는 마음의 1인칭적 측면이 무엇인지에 대해 정확히 표현할 수는 없으나, 물리적인 것이나 알고리즘이 아닌 인간의 마음과 같은 내적 상태를 지칭하는 것이라면, 이는 인공지능과 같은 인공물에게서 확인하기 어렵다는 점에서 쉽게 납득이 될 수 있다. 인간의 행위는 인간의 내부(마음)으로부터 발생하지만 인공지능의 행위는 외부(인간의 목적에 따른 설계)로 부터 발생하므로 인공지능의 행위는 인간의 행위와 동일하지 않고 따라서 인공지능은 책임귀속과 관련된 행위를 하는 존재가 아니다.

2.3. 책임 귀속의 조건

하지만 여기서 우리가 중요하게 생각해야 할 것이 있다. 저자는 자율적 행위를 논하면서, 단순히 스스로 움직이는 행위를 지칭하는 것이 아니라, 책임 귀속과 관련된 행위를 인간의 내적상태에 기반하여 논의하고 있다. 하지만 만일 누군가 주어진 상황에서 악한 행위를 하려는 의도가 있고, 이를 실행시키기 위해 자신의 신체를 조절하고, 주어진 환경이나 현실적인 실천의 어려움 등에 따라 목표를 일부 변경한다고 하더라도, 문제는 행위

16) *Ibid.*, pp. 157-158 & pp. 162-165.

17) 저자는 이를 윅스킬의 도식을 변경하여 1인칭적 중심이 감수자 체계와 행위자 체계를 연결하여 감수자 체계와 행위자 체계를 포함하는 전체를 포함하면서 그것이 가진 의도를 실현하는 행위를 하는 부분으로 설명한다. 물리적인 관점에서 보자면 뇌나 중앙신경체계가 이러한 역할을 할 수 있지만, 저자는 뇌의 외연이 1인칭적 중심의 물리적 외연과 일치한다고 단정하지는 않는다. 같은 책, 151-153쪽 참고.

주체의 내적 상태가 무엇이냐가 중요한 것이 아니다. 왜냐하면 인간이 일으킨 범죄 행위에 대한 판단이나 책임 귀속은 행위의 결과나 외부적 요인들에 의해 결정되는 측면이 있기 때문이다.[18] 앞서 언급한 촉법소년의 경우 이를 잘 보여준다. 촉법소년은 그들이 가진 인지능력이나 도덕성과 같은 내적상태와 무관하게 오직 그들의 연령으로만 책임 무능력자로 규정되고 있다. 하지만 피아제와 콜버그의 도덕성 발달이론[19]에 따르면, 촉법소년에 해당하는 존재들이 자율적 행위의 주체이자 책임에 관한 사고능력을 지니고 있다고 볼 수 있는 측면이 있다.[20] 피아제에 따르면[21], 이들은 구체적이지 않은 사상도 검증이 가능하며 고차원적 사고 역시 가능하다. 약 10세에서 11세가 지난 시점의 소년은 자율적 도덕성 단계로 발달하고 결과보다는 의도를 고려하는 도덕판단을 하는 주관적 책임의 특성을 지니고, 각 개인이 처한 상황을 고려하여 권위의 명령에 따를지의 여부를 결정하는 능력 또한 지닌다.[22] 콜버그는 피아제의 이론을 보다 세밀화하여 도덕성의 발달에 대해 논의한다. 그에 따르면, 도덕성 발달 단계는 3수준 6단계로 이루어져 있다. 촉법소년에 해당하는 단계는 10세에서 13세 사이의 시기

18) 논자는 법적 판단이 행위 주체의 내적 상태를 고려하지 않은 채 내려진다고 생각하는 것은 아니다. 다른 사람에게 피해를 발생시킨 행위 주체의 내적 상태가 고려된다 하더라도, 해당 행위의 결과에 대한 책임을 전적으로 피할 수는 없다. 즉, 내적상태에 따라 정상참작이 되긴 하지만 피해에 대한 책임은 여전히 귀속된다. 논자가 강조하고 싶은 것은 이어 등장하는 촉법소년의 경우처럼 인간도 내적 상태를 고려하지 않은 채 오로지 외부적 조건으로만 도덕적 지위와 책임 귀속 여부가 결정된다는 것이다. 그리고 이것이 사회적 합의와 결단에 의해 만들어진 것이라는 점이다.

19) 피아제와 콜버그의 도덕성 발달에 대한 이론이 어린 소년들의 도덕성 여부에 대해 철학적으로 해결해야 할 많은 문제(대표적으로 연령으로 도덕성 발달여부를 판단하는 것이 정당한지의 여부)들이 있지만, 논자가 이들의 이론을 간략히 언급하며 논의하는 이유는 한 존재의 내적상태를 고려하는 여러 가지 요소들(연령과 같은 외부적 요건들)이 있음에도 불구하고, 이에 대한 우리의 합의나 결단에는 이러한 요소들이 중요한 근거로 작동하지 않는다는 점을 강조하기 위함이다.

20) 김태훈 (2004)

21) 김봉석 외 (2013), p. 67.

22) 김태훈 (2004), pp. 325-329.

에 도달되는 2수준 3단계와 4단계이다. 이 수준은 '인습적 수준 혹은 인습적 역할 동조(conventional role conformity) 수준'이라 불리며, 이 수준에 해당하는 소년은 자신의 역할을 인습에 맞게 수행할 수 있고 이전 수준과 단계에 비해 사회적 표준이나 가치를 더 많이 내면화할 수 있다. 특히 4단계의 소년은 사회, 집단, 제도에 공헌하는 것을 선으로 인식하며 스스로 동의한 현실적 의무를 준수하려고 하는 권위와 사회질서 유지의 도덕성을 지닌다.[23] 피아제와 콜버그의 도덕성 발달이론을 통해 우리는 촉법소년의 도덕성이 성인의 그것과는 일부 차이가 있다고 할지라도 책임 능력을 동반하는 도덕성을 전적으로 결여하고 있다고 보기 어렵다. 소년들 사이에서도 도덕성 발달에 있어 개인차가 나타날 수 있겠지만, 이러한 점이 이들이 도덕적 책임에 있어 전적으로 무능력한 존재라고 보는 것은 받아들이기 어렵다. 다시 말해, 존재의 내면 상태를 정확히 알 수 없는 소년의 경우 우리가 선택한 합의나 결단의 방식은 행위주체성을 가지는 존재에게 행위의 결과에 대한 책임을 묻지 않는 것이었다.[24] 이와 마찬가지로 우리는 어린아이들이 주체적 행위와 의지, 의도 등이 명확하다고 판단되는 경우에도 이들에게 성인과 동일한 법적·도덕적 책임을 묻지는 않는다. 일반적으로 우리는 성인이 소년에 비해 더 높은 차원의 도덕성을 가지고 있고, 연령이 인간의 도덕적 사고능력과 같은 내적상태를 규정하는 것이라고 생각하지 않는다. 이렇게 생각하는 것은 비합리적이다. 현재 소년들의 강력범죄 비율이 증가됨에 따라 형사미성년자 연령 상한선을 14세에서 12세로 하향 조정해야 한다는 사회적 논의가 시작되었는데, 이 역시 해당 소년들의 내적 상태에 따라 제시된 것이 아니라 해당 연령의 형사 범죄율이 증가한 것이 그 원

23) *Ibid*., pp. 337-338.

24) 사회가 촉법소년들에게 책임을 묻지 않는 현행의 이유 역시 그들이 책임능력이 없다는 생각이라기보다 '그런 능력이 일반적으로 부족하다'거나 '그들이 사회에 기여할 수 있는 미래의 잠재성을 고려해야 한다'는 사회적 인식에 근거한다는 주장이 있을 수 있다. 하지만 논자가 주목하는 것은 존재의 내적상태를 정확히 알 수 없을 때, 책임 귀속의 문제에 있어 우리가 의존하는 것이 행위 주체와 관련된 외부적 요인들(연령 등)인데, 이러한 태도가 과연 합리적인지의 여부에 관한 것이다.

인이다.[25] 행위에 대한 책임 귀속 부여에서 중요한 역할을 하는 것 중 하나는 앞에서 언급한 내재적 성격의 자율성이 아니라 그 소년 그리고 그러한 소년의 사회참여와 더불어 발생하는 여러 유형의 일어난 결과와 일어날 결과에 대한 잠재성이다. 이처럼 책임 귀속이 저자의 주장처럼 전적으로 존재의 내적상태에 의존되는 것은 아니다.

이처럼 법률적 판단의 기초가 되는 우리의 사회적 합의의 일부는 존재의 내적상태가 아니라 외부 조건들에 의해 결정되기도 한다. 저자는 인공지능이 인간과 같이 내적상태를 요구하는 철학적 의미의 자율성을 결여하기 때문에 책임 귀속과 관련된 행위를 할 수 있는 존재가 아님을 강조한다. 이러한 이유로 인공지능의 작동이 발생시킬 수 있는 예기치 못한 문제들에 대해 인공지능 스스로 자신을 규율하도록 내버려두기 보다는 외부에 의해 규율되는 것이 더 바람직하다고 생각한다. 이는 저자의 두 번째 질문 '과연 누구에게 책임을 귀속할 것인가?'에 대한 대답, 다시 말해 이에 대한 합의가 어떻게 가능한지에 대한 논의로 이어진다.

3. 자율성과 책임 귀속의 문제

앞서 검토한 저자의 입장을 따르게 되면, 인공지능은 스스로 행위할 수 있지만, 인간에게 부여되는 자율성의 개념이 적용되지 않는다. 이들은, 저자의 표현대로, '공학적 의미의 자율성'만을 지니게 된다. 저자는 자율성의 개념을 루소가 제시한 '자기결정성과 사회적 맥락에서의 다른 자율적 주체들과의 상호 동등한 권능'의 개념, 칸트가 제시한 '실천 이성의 능동성', 크리스만이 제시한 '자기 전체에 대한 반성 및 반성된 내용을 변경·실행할 수 있는 능력'과 견주어 검토[26]하고, 인공지능이 저자가 제시한 행위의(책임귀속 문

25) 점승헌 (2022), pp. 347-348.

26) 저자가 제시한 철학적 의미의 자율성에 대한 검토는 고인석 (2022), pp. 179-185을 참고하라.

제에 있어서) 자율성을 가지지 못하기 때문에 인공지능에게 책임을 부여할 근거가 부족함을 주장한다. 즉 자율성에 관한 철학적 개념에 따르면, 책임을 귀속시킬 수 있는 행위를 만족하는 자율성이란 존재의 속성에 따라 결정되는 것이 아니라 그 존재가 속한 **사회적 관계의 맥락** 안에서 결정[27], 이성이 부여하는 실천 법칙을 스스로 따르는 **인간** 존재[28][29], 그리고 스스로 작동(행위)의 **목표를 변경**할 수 있는 능력[30]이라는 조건들을 만족시켜야 한다.

저자의 이러한 논의에 따르면, 인공지능이 가진 기계적 속성만으로는 자율성의 개념에 포섭되지 않는다.[31] 우리 사회 구성원들이 합의를 통해 자율적 존재로 규정하는 방식을 택하거나, 인공지능 스스로 진화하여 인간과 같이 도덕 법칙을 스스로 따르는 존재가 된다면, 자율적 존재로 간주할 수 있겠지만 저자는 그런 일들이 과연 가능한지에 대해 회의적인 시각을 가진

27) *Ibid*., p. 174.

28) *Ibid*., p. 175.

29) 신상규는 칸트적인 생각을 통하여 자율성을 규정하려는 시도는 인간중심적 사고이며, 도덕 행위자의 범주를 제한하는 매우 축소적인 방식임을 지적한다. 그는 이러한 방식은 전통적 사고를 일상적 개념들에 녹이려 하는 습관적인 의미부여이므로, 새로운 삶의 양식에 맞추어 그 의미를 재구성해야 할 필요가 있다고 주장한다. 그는 '행위자' 개념에 대한 수정을 통해 행위자성에 걸맞은 책임 혹은 책무성을 인공지능에게 귀속할 수 있음을 논의한다. 이에 대한 하나의 방안으로 일어난 결과에 대해 책임일 귀속시키는 응보의 방식이 아니라 예상되는 행위 결과에 선제적으로 대응하는 방식을 제안한다(신상규 (2017), pp. 268-288 참고). 하지만 논자가 보기에 이러한 접근방식은 오히려 저자가 우려하는 부분에 대한 해결점을 제시할 수 있어 보이지는 않는다. 저자의 지적처럼, 인공지능이 어떠한 선택을 하게 될지는 여전히 불투명하며, 인공지능 스스로가 인간에게 피해를 가하지 않는 행위를 선택한다 하더라도 우리는 그로부터 발생할 위험이 어떠한 것인지는 여전히 알 수 없기 때문이다.

30) 고인석 (2022), p. 176.

31) 박찬국은 저자의 책 6장의 주요 논증을 인용하여, 인간은 자신의 유지를 궁극적인 목적으로 가지고 행동하며 이를 통해 자신의 독자적인 세계를 가진다고 주장한다. 인공지능이 아무리 자율성을 갖는다고 하더라도 그것은 인간이 인공지능에게 외부로부터 투입한 목적을 수행하는 한도 내에서의 자율성이며, 인공지능은 인간이 만들어낸 세계에서 그것이 가진 기능에 따라 그 목적이 정해지는 존재에 해당한다. 박찬국 (2018), pp. 143-144 참고.

것으로 보인다. 저자는 다음과 같이 말한다.

> 공학의 맥락에서 특정한 속성을 갖춘 인공지능이나 로봇, 특히 로보틱 시스템을 가리키는데 통용되고 있는 '자율적'이라는 개념은 사회적, 윤리적 차원에서 해당 인공물의 지위를 결정하는 일과 유관한 자율성의 존재를 함축하지 않는다. 또한 자기 목표를 결정하는 권능이라는 의미의 자율성은, 만일 그것이 공학적으로 가능하다고 해도, 인공물에게 허용되서는 안된다."[32]

물론 저자가 인공지능이 위에서 제시된 철학적 개념을 충족시킬 가능성이 전혀 없다는 것은 아니지만, 현재 시점에서는 이 이상의 판단은 불가능하다고 여기는 것 같다.

저자의 자율성에 대한 철학적 검토에 있어 논자가 주의 깊게 본 것은 크리스만이 제시한 반성능력인데, 저자는 "자율적 주체가 지닌 진정한 역량으로서의 자율성은 크리스만이 말하는 것처럼 자신을 규정하는 속성들을 [계속] 수용할 것인지 아니면 부정할 것인지를 결단하고 나아가 그것들을 자기 뜻에 따라 변경하는 능력을 의미한다"[33]고 말한다. 저자는 위와 마찬가지로 인공지능 스스로 목표를 변경할 수 있는 가능성이 있다는 것을 부정하지는 않는다. 저자는 알프스 지역을 비행하는 인공지능 부조종사가 스스로 목표를 변경하는 사고실험[34]을 통해 이러한 가능성을 보여준다. 이때

32) 고인석 (2022), p. 191쪽.

33) *Ibid.*, p. 185.

34) 이 사고실험의 내용은 다음과 같다: 항공 운항의 고비용 인력을 절감하기 위해 부조종사를 인공지능이나 지능형 로봇으로 대체하는 경우, 비행 과정에서 인공지능 부조종사가 특정한 지역을 운항 중 특중한 지역을 더 자세히 관찰하려는 새로운 목표를 가지게 되는 경우, 즉 본래의 목표인 '예정된 고도와 경로에 따른 비행'에서 '특정한 지역을 가까이 보려는' 목표를 가지게 되었을 때 인공지능 부조종사 스스로 승객들에게 피해를 주려는 의도나 목적이 없음에도 불구하고 비행 고도를 한껏 낮추는 결정을 통해 심각한 인명 피해가 우려되는 사고를 발생시킬 수 있다. 저자는 "스스로 목표를 변경할 수 있다는 의미의 자율성을 지닌 인공지능이 인간을 대행하도록 하는 경우, 예측 불가능하고 심각한 위험이 발생할 수 있다"고 주장한다. *Ibid.*, pp. 188-189 참고.

중요한 점은 인공지능이 인간에게 피해를 주려는 의도 없이 충분한 합리적 과정을 거쳐 스스로 목표를 변경하는 경우라 하더라도 여전히 인간에게 심각한 위험이 될 수 있다는 것이다. 그리고 그 위험은 우리가 알 수 없는 종류의 위험이 될 수도 있다. 즉 인공지능이 합리적 과정을 거쳐 의사결정을 한다고 하더라도 여전히 인공물의 합리성이 가져다줄 위험이 도사리고 있다는 것이다. 저자는 결국 자율적이고 합리적인 인공지능의 구현이 가능하다 하더라도 우리가 예측할 수 없는 위험을 가져올 수 있으므로 그것을 허용해서는 안된다는 입장을 취하고 있는 것으로 보인다.

여기서 우리가 주목해야 할 저자의 입장은 인공지능의 자율성이나 합리성(만일 있다면)은 우리의 그것과 동등한 관점에서 평가되어서는 안된다는 것이다. 이는 저자가 정의하는 인공지능의 존재론적 속성에서 매우 잘 드러난다. 인공지능의 존재론적 속성은 "사회적으로 합의된 정신의 외화(externalization of social mind)"[35]에 해당하며, 이때 사회는 국가나 국제사회와 같이 큰 단위와 더불어 작은 단위의 사회공동체를 포함한다. 따라서 인공지능이 자율적 존재인지 아닌지의 여부와 더불어 자율적 인공지능을 개발할 것인지의 여부는 결국 인공지능 자체로부터 결정되는 것이 아니라 결국 우리의 선택에 따라 결정될 수 있는 문제[36]에 해당한다.

논자가 보기에 저자의 이러한 입장은 인공지능이 만들어낼 사회의 변화와 발전과 더불어 나타나게 될 다양한 문제들에 있어 우리가 취해야 할 태도는 인공지능이 안전하게 우리 사회에 자리매김할 수 있도록 사회 구성원들의 합리적 의사결정에 따른 합의를 통해 적절한 방식으로 규제해야만 한

35) *Ibid*., p. 240.

36) 논자가 보기에 저자는 이 점에서 샌델(Sandel, M.)과 비슷한 관점을 가진 듯 보인다. 샌델은 '선물논증'을 통해 유전공학을 통한 '인간강화(human enhancement)'가 공학적, 기술적, 공정성, 정의 등의 문제가 아니며, 우리 사회가 그러한 것들을 지향하는 태도(샌델에 따르면, 자연을 정복하고자 하는 '프로메테우스적 태도')가 과연 적절한지를 묻고 있다 (마이클 샌델 (2016) 참고), 마찬가지로 저자는(직접적으로 언급하지는 않지만) 공학적 의미의 자율성이든 철학적 의미의 자율성이든 인공지능에 그러한 자율성을 부여하려는 것, 그리고 그러한 것들을 관대하게 허용하려는 우리 사회의 태도가 과연 정당한지를 묻고 있는 듯 보인다.

다는 것으로 보인다. 그리고 이 규제는 인공지능이 일으킬 문제적 행위에 대한 책임 귀속을 누구에게 할 것인지에 대한 것이 아니라, 그러한 문제를 야기하는 인공지능을 개발해야 하는가에 대한 논의로 이어진다. 논자의 이해가 옳다면 이는 두 가지 문제를 포함한다. 하나는 실천적인 차원에서 그것이 쉽지 않다는 것이고, 둘째는 인간의 사회적 합의가 매우 합리적이라는 지나친 가정에 의존하고 있다는 것이다.

4. 사회적 합의의 합리성 그리고 인간의 자율성

지금까지 살펴본 저자의 논의에 따르면, 공학적 개념의 자율성은 인간이 가진, 보다 정확히 말하자면, 철학적 개념의 자율성과 일치하지 않는다. 그렇다고 해서 인공지능이 철학적 의미의 자율성을 획득하지 못한다는 의미는 아니다. 현재의 기술 수준으로 볼 때, 철학적 의미의 자율성을 인공지능이 가지지 못한다는 것, 둘째는 그러한 자율성을 가진 인공지능의 제작과 개발이 가능하다 하더라도 이를 무분별하게 허용해서는 안 된다는 것이다. 저자의 이러한 주장은 논자가 보기에 인공지능을 설계할 때, 우리가 감당 가능한 수준에서 인공지능의 개발을 멈춰야만 한다거나, 인공지능은 자율적 존재가 아니므로 인공지능 스스로 선택하여 작동하는 것이 아닌 우리 사회가 공유해 온 모종의 원칙들을 반영하여 **그대로** 따르도록 만들어야 한다는 것으로 들린다. 즉 도덕적 주체에 해당하는 자율적 행위자인 우리의 합리적 판단에 따라 결정되어야 할 문제인 것이다. 이는 다음의 두 가지를 전제한다. 하나는 우리의 사회적 합의가 매우 합리적이라는 것이고, 다른 하나는 우리가 인공지능에 비해 월등히 자율적 존재라는 것이다. 이 두 전제는 다음을 함축한다. 자율적 존재인 우리의 합의에 의한 판단은 선(good)하다는 것이다.

4-1. 내적상태가 불분명한 존재에 대해 내리는 우리의 결정은 합리적인가?

먼저, 현재 빠른 속도로 일상의 다양한 분야에 침투하고 있는 인공지능

의 개발을 멈추거나 후퇴시킨다는 것은 편의성 측면에서도 경제적인 측면에서도 그다지 합리적인 방식으로 보이지 않는다. 이에 대한 사회 구성원들의 합의 역시 도출하기 쉽지 않아 보인다. 앞서 언급한 촉법소년의 예시가 그런 상황을 잘 보여준다. 현행 헌법에서 촉법소년을 책임무능력자로 규정하는 기준은 연령이다. 여기서 책임무능력자는 절대적 책임무능력자를 말한다. 헌법재판소는 "일정한 정신적 성숙의 정도와 사물의 변별능력이나 행동통제능력의 존부·정도를 각 개인마다 판단·추정하는 것은 곤란하거나 부적절하므로 일정한 연령을 기준으로 하여 일률적으로 형사책임연령을 정한 것은 합리적인 방법으로 보인다."[37]라고 밝혔다. 이는 논자가 보기에 마치 과학적 지식이 부족하여 인간의 생명이 정확히 언제 시작되는지 알 수 없어 잉태된 순간부터 행해지는 모든 임신중절을 살인으로 간주한 19세기 로마 교황청의 결정과 크게 다를 바 없어 보인다.[38] 위 판결 내용과 로마 교황청의 결정은 그 사안이 서로 다르지만 특정 존재의 내적 상태에 대한 판단 기준을 명확하게 확립하기 어려우므로 임의적으로 설정한 방안이라는 점에서 같다고 볼 수 있다.

또한, 위의 촉법소년에 관한 판결을 따를 경우, 범죄에 대한 인식이 있고 책임 능력이 있음에도 불구하고 처벌되지 않는다는 사실을 고려하여 범죄 행위를 저지르는 경우가 발생할 위험 역시 가지고 있다.[39] 따라서 저자가 제안하는 사회적 합의나 공동체의 결단, 그리고 저자가 제안하는 "고집의 원리(principle of tenacity)에 따라 새로운 확신이 지배하는 세상이 되기 전까지는 기존의 것을 고수하는 자세"[40]가 합리적이며 윤리적인 방법이라고 보기 어려운 측면이 있다. 또한 이러한 결단이 언제나 선한 것이라고 판단되지는 않는다. 즉, 부분적으로 선한 측면을 가지고 있을지 몰라도 보편적으로 선한 선택이라고 여겨지지 않는다면 기존의 것을 고수해야 하는지에 대한 비판적 검토가 요구된다.

37) 헌법재판소 2003. 9. 25, 2002헌마533 판결 참조.
38) 피터 싱어 (2023), p. 62.
39) 점승헌 (2022), p. 354.
40) 고인석 (2022), p. 333.

이는 저자가 매우 중요하게 생각하는 사회적 합의나 공동체의 결단이 인공물의 행위에 대한 책임귀속 문제에 대한 옳은 답변을 제시할 수 있을지에 대한 회의적인 시각을 가지게 만든다. 저자가 여러 차례 제안하고 있는 사회적 합의와 공동체의 결단에는 오히려 저자가 피하고자 하는 임의적이거나 비합리적인 결정들을 포함하고 있기 때문이다. 또한, 저자가 제시한 철학적 의미의 자율성의 개념에 기초한다면, 자율적 존재들의 합의된 결정은 선하다는 것을 받아들여야 하는데, 앞서 논의한 것처럼 이러한 결정이 보편적인 차원에서 선한 것이라고 여기기 어려운 측면 역시 존재하고 있다.

4-2. 현재 우리는 자율적 존재인가?

스스로에 대한 규율(self-rule)이나 자기 결정(self-determination)은 자율성의 본질적 성격에 해당한다.[41] 이것은 "가치에 대한 자신의 개념을 발전시키는 능력과 더불어 중요한 것이 무엇인지를 인식하고, 자신의 행동과 결정을 안내할 가치를 [개발]하고, 자신이 옳다고 생각하는 가치에 따라 자신의 삶에 대한 중요한 결정을 내리는 것"[42]을 말한다. 따라서 자율성은 이러한 유형의 주체적인 자기 결정의 한 형태이자, "인간의 사고와 행동의 모든 범위를 수식하는 부사(자율적으로)"[43]의 형태로 사용된다. 이러한 대상에는 신념이나 다른 행동을 지배하는 원칙 뿐만 아니라, 가능한 여러 대안들 중에서 하나를 선택하거나 다른 사람의 간섭이나 제안 등을 수용하는 것과 같은 행위도 모두 포함된다. 이러한 의미에서 자율성은 인지적 측면과 실천적 측면을 모두 포괄한다. 칸트에 따르면, 자신의 독립적인 관점을 형성하지 않고 전통이나 권위에 대한 맹목적인 복종이나 조작 혹은 강압에 의해 결정하는 것은 바로 자율적이지 않은, 타율적인 인간이 되는 길이다.[44]

41) Laitinen, A. and Sahlgren, O. (2021), p. 3.
42) Rubel et al. (2020), p. 550, Laitinen (2021), p.3에서 재인용.
43) Laitinen, A. and Sahlgren, O. (2021), p. 4.
44) Kant, I. (1996).

논자가 위의 자율성의 개념을 통해 지적하고자 하는 것은 저자가 제시한 철학적 의미의 자율성이 어떤 방식으로 이해되든지 간에 그 자체로 선한 것으로 간주하도록 우리를 이끈다는 점이다. 이러한 입장을 극단적으로 이끌게 되면, 자율적 존재가 스스로 선택한 행위 자체가 일부 비합리적이고 나쁜 결과를 초래한다 하더라고 그 행위에 일부 선한 점이 있다는 뜻으로 받아들여질 수 있다. 그럼에도 불구하고 자율성을 선한 것으로 인식하는 것에는 큰 문제는 없어 보인다. 왜냐하면, 자율성은 본질상 좋은 것이 아니라 다른 선을 제공할 수 있는 기초가 되기 때문에 윤리적 문제를 해결하는 데 있어 우선되어야 할 가치가 있다는 관점을 취할 수 있기 때문이다.[45]

우리가 인공지능에 관한 어떠한 선택을 하든 간에, 우리의 자율성에 기반한 선택은 선한 것이고, 따라서 우리의 공공선을 실현시키는 데에 매우 유용한 도구가 될 수 있다는 입장은 쉽게 받아들여질 수 있는 것처럼 보인다. 하지만 이러한 입장이 받아들여지기 위해서는 현재 우리가 이와 같은 의미에서의 자율적 존재라는 것이 먼저 전제되어야만 한다. 현재 다양한 의사결정의 과정에서 알고리즘 시스템에 대한 의존이 높아지게 되면서, 사용자가 이에 의존하여 내리는 선택이나 판단이 과연 사용자의 진정한 자기결정 능력에 의한 것인지에 대한 우려가 제기되고 있는 것이 현실이다.[46] 우리는 AI '추천' 시스템에 매우 익숙하며, 많은 부분에서 이들로부터 추천을 받아 상품 등을 선택한다. 하지만 이러한 시스템은 우리에게 적절한 선택지를 탐색할 기회를 빼앗고, 우리를 특정한 가치만을 받아들이도록 유도

45) 이를 도구주의적(instrumentalist) 관점이라고 한다. 다나허 (2023), p. 201 참고.

46) 다나허는 이러한 우려에 대해 특정 유형의 인공지능을 의사결정의 보조적 수단으로 사용하는 것이 인간에게 해롭고 인간의 인지적 능력이나 자율적으로 삶을 살아가는 일 등에 퇴행을 가져오는 영향을 미칠지에 대해서는 그 사용 대상과 "해당 작업과 관련된 인지적 탄력성에 대한 필요성(the possible need for cognitive resiliency with respect to that task)"에 따라 달라질 수 있다고 논증한다(Danaher, J. (2018), pp. 629–53 참고). 하지만 논자가 보기에 맥락적 조건에 따라 영향력에 대한 평가가 달라진다는 것이 우리가 우려하는 문제가 완전히 사라질 수 있음을 의미하는 것은 아니다.

하며, 우리의 의지에 반하는 일을 하도록 강요하는데 사용될 수 있다.[47] 이러한 시스템에 익숙한 우리가 과연 진정한 의미의 자율적 선택을 할 수 있는지의 여부는 안타깝게도 불투명하며, 이는 먼 미래의 이야기가 아닌 현재 우리의 모습이다.

현재 우리가 일상적으로 마주하고 의존하고 있는 인공지능에 의한 자동화 기술은 개인이 자신의 길을 스스로 선택할 수 있는 자율성을 침해할 여지가 있고, 이러한 상황에서 우리가 내리는 판단 등이 저자가 주장하는 진정한 철학적 의미의 자율성에 기반한 것인지에 대한 논의가 추가적으로 요구된다.[48]

5. 나가는 말

저자는 다음을 언급하였다:

> "우리는 여러 과학을 통해 인간 정신의 면면을 착실히 알아가는 중이지만, 인간의 인간 이해는 운명적으로 불완전하다."[49]
> "우리는 '인간의 지적 능력을 (모든 면에서) 고스란히 구현하는' 인공지능을 만들 수가 없다. '인간 지능의 모든 면'에 대한 지식을 손에 넣을 수 없기 때문이다."[50]

47) 다나허 (2023), pp. 203-204.

48) AI 의사 결정 지원 툴의 한 가지 특징인 '하이퍼 넛지'의 경우, 단순히 사람들에게 도움이 되는 지침을 제공하는 것이 아니라, 선택의 자유를 제한하여 특정 방향으로 사람들의 행동이 나타나도록 하는 것을 목표로 한다. 이는 주로 특정한 이들의 이익에 부합하거나 조직 또는 사회의 목표와 일치하는 결정을 유도하기 위해 사용된다. 이는 때론 지나치게 공격적이거나 불합리하거나 비윤리적으로 보이는 방식을 사람들이 선택하도록 유도하는 문제를 지닌다. *Ibid.*, pp. 206-207 참고.

49) 고인석 (2022), p. 45.

50) *Ibid.*, p. 46.

논자가 보기에 저자의 이와 같은 주장은 특정 업무의 수행에 있어 우리보다 뛰어난 인공지능을 만들 수 없다는 주장이기 보다는 인간의 가진 지적 능력의 다양한 측면을 인공적으로 구현할 수 없고, 그것의 가장 대표적인 것이 인간의 도덕성이라는 내용을 함축하는 것으로 보여진다. 실제 도덕적 인공지능 로봇을 설계하기 위한 다양한 방법들이 제시되었다. 사회가 공유하는 하나의 윤리적 원칙을 적용해 그 원칙에 따르는 출력의 방식으로 행위하도록 만드는 하향식과 특정 행동이 선택되는 환경을 제공해 인공지능이 스스로 경험을 통해 도덕적 행동을 배우도록 하는 상향식 방식 그리고 이 둘의 장점을 합친 혼합식 접근이 대표적이다.[51] 하지만 이러한 노력과 별개로 우리는 인공지능의 설계와 제작뿐만 아니라 인공지능이 학습하는 데이터 등에 윤리적 원칙을 적용하는 방식도 취하고 있다. 논자의 오독이거나 지나친 자의적 해석일 수 있겠지만, 저자는 인공지능 스스로 도덕성을 학습하여 도덕적인 인공지능 로봇이 되는 방식보다는 인공지능 로봇을 설계, 제작, 활용하는 단계에서 합의된 윤리적 원칙이 적용되어야 한다고 보는 것 같다. 그 이유는 바로 위에서 언급된 저자의 말처럼 우리 자신도 우리에 대한 이해가 충분하지 않아 인간의 지적 능력을 모든 면에서 구현하는 인공지능을 만들기 어렵기 때문이다. 하지만 우리가 이미 가지고 있는 원칙이나 새로운 합의를 통해 만들어진 원칙을 적용한다 하더라도 논자가 앞서 지적한 임의성과 비합리성 그리고 자율성의 부재로부터 자유로울 수 있다는 보장은 하기 어려울 듯 보인다.

물론 논자의 이러한 지적이 그간의 모든 인간의 도덕성에 대한 연구가 모두 허구에 불과하거나 인간의 현재 모습이 인공지능과 관련된 문제에 어떠한 역할도 하지 못한다는 뜻은 아니다. 우리는 이러한 연구들을 통해 도덕적 객체의 범위를 확대해왔고, 변화된 사회에 어울리는 윤리적 원칙들을 결정해 왔으며, 나름의 긍정적인 변화를 가져왔다. 그리고 이는 여전히 진행 중에 있다.

그렇다면, 인공지능과 로봇에 대한 우리 사회의 합의와 결단을 무엇을

51) 김효은 (2019), pp. 97-119 참고.

기반으로 이루어져야 할까? 도덕적 주체로서의 우리 자신? 도덕적 객체나 무능력자로서의 동물이나 어린아이? 사회적 개념으로서의 법인? 물론 저자는 위의 논자가 열거한 사항들이 불충분함을 논의하고 있다. 논자의 이러한 질문은 저자의 의도를 왜곡하려는 뜻이거나 새로운 시대의 윤리적 원칙의 필요성을 훼손하려는 것은 아니다. 어떤 사람들은 인간이란 존재가 스스로의 사유에 의해 행위를 선택하는 자율적 존재, 안정적이고 합리적인 존재가 아니라 극히 불안정하고 사회적 이념이나 종교에 대한 맹목적인 믿음으로 이와 다른 믿음을 가진 사람들을 무자비하게 살육하고 끊임없이 환경을 파괴하고 전쟁을 일으켜온 존재라고 생각한다.[52] 그렇기 때문에 인간의 특정한 능력을 모방한 인공지능을 만들거나, 인공지능의 작동방식에 인간이 만들어낸 원칙을 적용하는 것이 적절한지에 대한 근원적인 물음이 발생할 수밖에 없다고 말한다.

이런 측면에서 본다면, 인공지능과 관련된 문제를 통해 우리가 그동안 가졌던 윤리 원칙들이 저자의 표현대로 자율적 판단에 의거하고 보편적이며, 합리적인 결단에 기반했는가를 다시 묻게 된다. 인공지능이 일상의 대부분의 영역에 침투하는 시대를 맞이하여 이와 관련된 윤리적 원칙을 성립하는 일은 반드시 필요하고 당연한 일이겠지만, 저자는 이와 더불어 우리의 도덕성에 대한 지적탐구가 보편적이며 합리적이었는지를 먼저 묻고 있는 듯 보인다.

이 책을 접하면서 논자는 우리가 도덕적 인공지능의 개발이 가능한지의 여부나 인공지능의 윤리적 개발과 사용은 무엇인지에 대한 물음과 응답을 계속하는 동안 인공지능은 오히려 우리에게 다시 묻고 있는 듯함을 느꼈다. '당신들의 윤리는 옳은가?' 라고.

저자는 2017년 7월 과학철학회 여름 발표에서 다음과 같은 언급을 하였다. "인간이 좋은 대답을 제시할 수 없는 종류의 물음에 대해, 기계에 그 답을 요구해서는 안 된다." 우리는 이 물음에 대해 아직 좋은 답을 가지고 있지는 않은 것 같다.

52) 박찬국 (2018), p. 126.

참고문헌

고인석 (2022), 『인공지능과 로봇의 윤리』, 세창출판사.

김봉석, 박지웅, 황준원, 유희정, 곽영숙, 반건호 (2013), 「피아제의 인지발달학적 측면에서 영화가 아동 배우에게 어떠한 영향을 미치는가?」, 『Journal of the Korean Academy of Child and Adolescent Psychiatry』, 4권 2호, pp. 65-70.

김태훈 (2004), 『도덕성 발달이론과 교육』, 인간사랑.

김효은 (2019), 『인공지능과 윤리』, 커뮤니케이션 북스.

마이클 샌델 저/이수경 옮김 (2016), 『완벽에 대한 반론』, 와이즈베리.

박찬국 (2018), 「인공지능과 인간의 미래:인간과 인공지능의 존재론」, 『현대유럽철학연구』, pp. 119-166.

변순용, 이연희 (2020), 『인공지능 윤리하다』, 어문학사.

신상규 (2017), 「인공지능은 자율적 도덕행위자일 수 있는가?」, 『철학』, 제132집, pp. 265-292.

점승헌 (2022), 「촉법소년의 연령 하향」, 『법이론실무연구』, 제10권 3호, pp. 345-67.

존 다나허 저/김동현 옮김 (2023) 『우리 시대 가장 중요한 질문: 생각을 기계가 하면 인간은 무엇을 하나?』, 뜻있는 도서출판.

피터 싱어 저/구영모 역 (2023), 「인간의 생명은 언제 시작되는가」, 『생명의료윤리』, 제4개정판, 구영모 엮음, 동녘.

Danaher, J. (2018), “Toward an Ethics of AI Assistants: An Initial Framework”, *Philos. Technol.* 31 (4): pp. 629–53.

Laitinen, A. and Sahlgren, O., (2021), “AI Systems and Respect for Human Autonomy”. *Front. Artif. Intell*: pp. 1-14, doi: https://doi.org/10.3389/frai.2021.705164

Rubel, A., Castro, C., and Pham, A. (2020). “Algorithms, Agency, and Respect for Persons”, *Soc. Theor. Pract*. 46 (3): pp. 547–572. doi:10.5840/soctheorpract202062497.

논문 투고일	2023. 10. 15.
심사 완료일	2023. 11. 16.
게재 확정일	2023. 11. 16.

Is it possible to reach a social consensus on the ethical issues raised by artificial intelligence?

- Issues of autonomy and responsibility for artificial intelligence in The Ethics of Artificial Intelligence and Robots by Ko In-seok

Taekyung Kim

Ko In-seok in his The Ethics of Artificial Intelligence and Robots critically examines the functions of artificial intelligence and some philosophical issues that may arise from it, as well as the relevant concepts - autonomy, action, intention, perception, agency, etc. Through this, he discusses that the autonomy of humans and the autonomy of artifacts are distinguished, and since the autonomy of artifacts is different from that of humans and the operation of artifacts is different from the autonomous actions of humans. He argues that regulations on these beings must be prepared through social consensus and decision. I indicate that the following two things must be prerequisite for his discussion to be accepted. One is whether we are truly autonomous beings in decision-making, and the other is whether our social consensus on beings whose internal states are unknown is reasonable. Through this, I argue that the discussion of the various types of ethical problems that will be raised by artificial intelligence is not about artificial intelligence itself, rather it is essentially a discussion about our ethical judgments.

Keywords: Artificial intelligence, Action, Autonomy, Ethical judgment, Social consensus