https://doi.org/10.33156/philos.2020.59..007

인공지능, 도덕적 기계, 좋은 사람*

맹 주 만**

주제분류 윤리학, 응용윤리, 인간학

주 요 어 인공지능 행위자(AIA), 인공적 도덕 행위자(AMA), 인간지능, 자연지능, 로봇, 도덕적 기계, AI로봇, 좋은 사람, 동기, 의도, 목적, 자율성, 의사 자율성, 자동성

요 약 문

이 글은 인공지능은 행위의 동기와 의도 그리고 목적을 스스로 생성하고 산출할 수 있는 자율적 행위자가 아니기 때문에 도덕적 기계가 될 수 없다고 주장한다. 예상 가능한 인공지능 행위자가 갖게 될 공학적 의미의 자율성은 인간 행위자의 도덕적 자율성과 근본적인 차이를 가지며, 그것은 단지 의사(pseudo) 자율성에 불과하다. 따라서 그것은 그냥 놀랍도록 편리하거나 위험한 기계일 뿐이다. 동기와 의도 그리고 목적의 자기생성과 자기산출 및 자기변경이 가능한 자율성만이 진정한 의미의 도덕적 자율성이다. 반면에 인공지능 행위자는 기껏해야 형식적 자율성만을 가지며, 그러한 자율성은 한정된 목적에 제약되고, 그것만을 실행하도록 특화된 자동기계의 자동성에 다름 아니다. 아무리 복잡한 윤리적 문제의 해결을 위해 설계된 인공지능이라도 언제나 프로그램이 지시하는 일정한 그리고 한정된 행위만을 충실히 이행하는 기계인 한, 그것은 실제로는 도덕적 기계도 도덕적 행위자도 아니다. 도덕적 기계는 이론적으로도 실제적으로도 불가능하다. 만일 그런 존재가 있다면, 그것은 분명 인간종 이상의 다른 새로운 가공할 위력을 지닌 인공종일 것이다. 현재 가능한 인공지능 행위자를 둘러싸고 벌어지는 대부분의 전망들은 그릇된 상상의 산물이다. 도덕적 기계의 불가능성과 함께 윤리적 관점에서 의사 자율적 행위자에 불

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A6A3A01078538)

** 중앙대학교

https://doi.org/10.33156/philos.2020.59..007

과한 인공지능 로봇은 모든 면에서 위험성을 완전히 제거하도록 철저하게 기획되고 통제된 기계이지 않으면 안 된다.

1. 로봇의사 왓슨과 로봇변호사 로스

세계 의료계에 슈퍼컴퓨터 인공지능 의사 왓슨(Watson)이 암 전문의로 등장한 것이 2012년이었다. 이어서 법조계에도 인공지능 로봇변호사 로스(Ross)가 등장했다. 이는 2016년 5월 세계 최초의 인공지능 변호사가 미국의 대형 로펌에 고용되면서 알려졌다. 한국에서도 2018년 2월부터 인공지능 변호사 유렉스(Eurex)가 대형 로펌에서 일하고 있다. 실제로 사건에 관련된 법 조항 검토와 판례 분석에 비서들의 도움을 받은 여러 명의 변호사가 몇 달식 걸려 처리하던 일을 단 2~30초에 해내는 등 이 인공지능 변호사는 초당 1억 장의 문서를 검토해서 관련 사건에 가장 적합한 판례들을 찾아낼 수 있는 능력을 갖추고 있다.

인공지능 바둑선수 알파고와 같이 한 가지 기능에 특화된 단순 인공지능 혹은 약인공지능(Artificial narrow Intelligence)은 특수한 영역과 분야에서만 인간보다 일을 잘 하거나 대신한다는 측면에서 인간에게 봉사하거나 특수 분야의 직업인들과 경쟁하는 존재가 될 것이다. 현재도 충분히 예상되듯이 이러한 일들이 일반화된다면, 그것이 인간 사회에 미치는 파장은 엄청날 수 있다. 많은 직업군들이 사라지고 신생 직업들이 생겨날 것이고, 사회의 구조와 작동방식에 근본적인 변혁이 이루질 것이다. 하물며 인공지능의 범용화를 예고하는 '일반인공지능'(Artificial general Intelligence)에 속하는 로봇의사 왓슨이나 로봇변호사 로스 같은 AI로봇들의 지적 능력과 수준이 어디까지 진화할 것인지 예측하기는 쉽지 않다. 통합 지능형 시스템을 장착한 자율주행 자동차와 같은 수준으로까지 진행된다면, 이 기계 운전자에게 운전면허증을 내줄 수 있듯이, 로봇 왓슨이나 로봇 로스에게도 의사면허증과 변호사면허증을 발급해줘야 할 지 모른다.[1)]

일반인공지능으로 분류되는 기계들의 인간화는 그것들이 특정한 한두 가지 영역과 분야에 특화된 인공지능 로봇이라는 점에서 여전히 인간에게 봉사하거나 특수 분야의 직업인과 경쟁하는 정도의 존재에 지나지 않을지 모른다. 그리고 또한 여전히 이들을 인간의 통제 아래 두고 운영하거나 사용하는 것이 가능할 수 있다. 그러나 한두 가지 특수한 기능에 국한되지 않은 다기능의 '통합 초지능형 시스템', 즉 '자율형' 인공지능이 될 초인공지능(Artificial super Intelligence)은 모든 면에서 개별 인간지능들을 단순히 합쳐 놓은 것 이상의 능력을 갖게 될 것이며, 이미 그 자체가 현재의 인류가 내세울 수 있는 집단지성 이상의 능력을 발휘하리라 예측하는 것은 어렵지 않다. 예를 들어, 초지능형 로봇기계는 동시에 로봇의사이자 로봇변호사이면서 또한 자가운전자이면서 119 구조대원이기도 한 초지능형 초능력자일 것이다. 한두 가지가 아닌 그 이상의 통합지능을 갖춘 기계가 가능하다는 것은 인간을 닮은 일반인공지능과 그 이상의 초인공지능 존재의 가능성을 증명하는 것이 될 것이기 때문이다.

그런데 단순히 인간을 닮거나 그 이상의 존재로서 '인공지능 행위자'(AIA; artificial intelligence agent)는 실제의 인간과 동일한 존재론적 지위를 가질 수 있는가? 마찬가지로 이런 행위자를 인간과 같은 도덕적 행위자로 간주할 수 있을까? 그리고 이러한 자율적 능력을 갖춘 AIA를 좋은 의사나 혹은 정의로운 변호사, 즉 좋은 사람에 비견되는 도덕적 기계라고 할 수 있을까? 사람도 늘 좋은 사람이기만 한 것은 아니다. 때로는 누군가에게는 나쁜 사람이 되기도 한다. 그리고 그 행위자가 최소한 나쁜 사람일 경우에 우리는 그가 저지르는 행위를 예측할 수도 예방할 수도 있고, 또 처벌할 수도 있다. 하지만 슈퍼맨 같은 능력을 지닌 AIA에 대해서도 그러한 조치가 가능할까? 그런데 우리는 그와 같은 존재를

1) 실제로 로봇의사가 의료현장에서 활동하게 될 경우를 예상하면서 이와 관련한 윤리적 대응을 다루고 있는 글로는 맹주만, 「인공지능과 로봇의사윤리」, 『이성과 공감 : 포스트모던 칸트와 공감윤리』, 551-579쪽.

실제로 가져본 적이 없다. 우선 그와 같은 존재가 된다는 것이 어떤 것일지 상상하기 어렵다. 왜냐하면 우리가 아직도 좋음과 옳음이 무엇인지 합의에 이르지 못하고 있듯이 그런 힘과 능력을 지닌 존재에게 인간적 관점에서의 좋음과 옳음을 동일하게 적용할 수 있을지가 의문이기 때문이다. 그리고 최악의 경우에 나쁜 짓도 서슴지 않는 AIA라면 얼마나 끔찍한 일인가! 그것은 인류에게는 최악의 재앙이 될 것이며, 인간종을 능가하는 새로운 인공종의 출현을 의미하게 될 것이다. 그러나 실제로 제아무리 초지능의 AIA라 할지라도, 적어도 인간적 관점에서 그것이 도덕적 기계가 될 가능성은 없다.

이 글은 인공지능은 행위의 동기와 의도 그리고 목적을 스스로 의욕하고 생성하며 산출할 수 있는 자율적 행위자가 아니기 때문에 도덕적 기계가 될 수 없다고 주장한다. 예상 가능한 인공지능 행위자가 갖게 될 공학적 의미의 자율성, 즉 인공지능의 자율성은 인간 행위자의 도덕적 자율성과 근본적인 차이를 가지며, 그것은 단지 의사(pseudo) 자율성에 불과하다. 따라서 인공지능 행위자가 아무리 “윤리적 영향력을 지닌 행위자”(ethical impact agents)로 고안되고 제작되었다 하더라도 그것은 특정 목적을 염두에 두고 설계된 “기술적 행위자”(technological agents)에 불과하다.[2] 또한 윤리적 행위를 지지해주고 비윤리적 행위는 회피하도록 프로그램된 “암묵적인(implicit) 윤리적 영향력을 지닌 행위자”이든 윤리적 딜레마를 해결하도록 프로그램된 “명시적인(explicit) 윤리적 영향력을 지닌 행위자”이든 그 모든 것들은 정확한 의미에서 도덕적 기계도 도덕적 행위자도 아니다.[3]

도덕적 기계라는 말은 형용상의 모순이다. 어떤 인공지능 행위자이든

2) James Moor, “The Nature, Importance, and Difficulty of Machine Ethics”, in Wallach, W./Allen, C., *Moral Machines*, 14쪽.

3) James Moor, “The Nature, Importance, and Difficulty of Machine Ethics”, 15-18쪽 참조.

그것이 기계인 이상 그런 존재에 대해서는 정확한 의미에서 '도덕적'이라는 표현을 쓸 수 없다. 동기와 의도 그리고 목적의 자기생성과 자기수립 및 자기변경이 가능한 존재만이 진정한 의미의 도덕적 자율성을 지닌 도덕적 행위자이며, 반면에 인공지능 행위자는 기껏해야 형식적 자율성만을 가지며, 그러한 자율성은 한정된 목적에 제약되고, 그것만을 실행하도록 특화된 자동기계의 자동성(automaticity)에 불과하다. 언제나 일정한 행위를 수행하는 기계는 실제로는 도덕적 기계도 도덕적 행위자도 아니다. 인공지능은 결코 도덕적 기계가 될 수 없다. 도덕적 기계는 이론적으로도 실제적으로도 불가능하다. 만일 그런 존재가 있다면, 그것은 분명 인간종 이상의 가공할 위력을 지닌 인공종일 것이다. 그러므로 현재 가능한 인공지능 행위자를 둘러싸고 벌어지는 대부분의 전망들은 그릇된 상상의 산물이다. 도덕적 기계의 불가능성과 함께 윤리적 관점에서 인간의 의지에 의해서 악용될 수 있는 의사 자율적 행위자에 불과한 인공지능 로봇은 모든 면에서 철저하게 기획되고 통제된 기계이지 않으면 안 된다.

2. 인간지능과 인공지능

인간과 기계 혹은 인공지능(AI) 로봇의 지위에 관한 생물학적, 의식철학적, 존재론적, 정보철학적, 윤리학적 관점, 감정철학적 관점에서 다양한 논의들이 있어 왔다.[4] 특히 윤리학적 관점에서 자율적 행위자로서 도덕적 기계의 가능성에 대한 논의들은 그동안 대부분 인간과 동물 존재에 한정되었던 도덕적 행위자 문제들이 인공지능 중심의 4차 산업혁명 시대의 중요한 윤리적 쟁점이 되리라는 것을 보여주기 충분했다. 만일 AI로봇이나 각종 환경에 적합한 기기의 형태로 등장하게 될 인공기계가 도덕

4) 이에 대해서는 Wallach, W./Allen, C., *Moral Machines*, Oxford University Press, 2009(노태복 옮김, 『왜 로봇의 도덕인가』, 메디치, 2014); 이중원 외, 『인공지능의 존재론』, 한울아카데미, 2018.

적 행위자로서 인간적 자율성과 동일한 자율성을 가질 수 있다면, 그것은 도덕적 기계로서 최소한 인간과 동등한 인격체이자 존엄한 존재로 간주되어야 하며, 따라서 그에 준하는 대우를 받아야 할 것이다. 이런 생각과 주장에 대한 평가에 앞서 인공종 혹은 인공지능을 장착한 존재가 과연 어떤 점에서 인간종 혹은 인간지능(혹은 자연지능)을 지닌 존재와 유사하거나 혹은 근본적인 차이를 갖는지 살펴볼 필요가 있다. 그것이 도덕적 기계의 가능성에 대한 출발점이 되어야 할 것이기 때문이다.

간단히 말해서, 인공지능은 지능을 인공적으로 구현하는 것이다. 현재 개발되고 있는 인공지능의 기본 모델은 인간의 지능이다. 두뇌의 기능이나 능력으로 이해되는 인간지능은 자연지능인 반면에, 이 지능을 기계에 구현하는 것이 인공지능이다. 언젠가는 첨단기능이나 대체 장기를 장착한 기계인간의 출현을 넘어서 초인공지능을 장착한 제3의 존재로서 AI로봇의 출현이 예견된다. 상상 가능한 미래적 사실로서 인공지능은, 예상되듯이, 최소한의 특정한 분야에서는 인간의 능력을 뛰어넘는 가공할 위력을 보여줄 것이다. 그리고 최종적으로는 모든 면에서 인간의 능력을 훨씬 상회하는 인공종의 시대가 도래할 수도 있을 것이다. 그러나 어떤 경우든 그에 따른 부작용도 충분히 예단할 수 있다. 이를테면, 멀지 않은 시기에 아래로는 유인원에 속하면서 위로는 새로운 종에도 속하게 될 기계인간 내지는 초인공인간과 인공종의 구분이 필요할지 모른다. 이렇게 인간종이 현재인간(순수인간)과 사이보그처럼 인간과 인공지능의 결합체로서의 기계인간(AI인간)으로 구분되고, 순수기계이자 초인공지능의 AI로봇이 새로운 인공종인 초인공종으로 등극하게 된다면,[5] 현재의 순수인간인

5) 대체로 인공지능을 인간보다 1,000배 이상 높은 지능인 초인공지능(artificial super or ultra intelligence), 그 보다 조금 낮은 등급인 인간 수준의 강인공지능(artificial strong or general intelligence, 일반 인공지능, 범용 인공지능), 한 가지 일을 잘하는 약인공지능(artificial narrow intelligence)으로 구분하는 경우가 있는데, 이와 같은 구분은 그리 정확한 것이 아니다. 알파고와 같이 한 가지 일을 잘하는 인공지능의 경우에도 그 한 가지 기능의 수행을 어떤 방식으로 어느 정도까지 실행할

인간종은 물리적 능력과 기능면에서 이들 중에서 맨 하급계층에 속하게 될 것이다. 그러나 그렇다고 그것이 도덕적 존재의 서열을 가르는 척도가 될 수는 없다. 도덕적 평가가 물리적 조건 및 능력과 무관하다는 것이 오늘날 우리의 상식적인 도덕관이듯이 정신-물리적인 면에서 인간종보다 우월한 초인공종이 인간종보다 훌륭한 도덕적 존재라고는 단정할 수 없다.

무생물과 생물과 같이 상이한 종 범주에서 나타나는 물리적 차이가 도덕적 차별을 정당화하는 근거와 이유를 제공하는데 이용되기도 하지만, 이 역시 선결문제 요구의 오류를 범한다. 무엇을 도덕적 속성으로 볼 것인지가 먼저 정의되어야 하기 때문이다. 그런데 철학적 관점에서 어떤 존재들의 도덕적 지위를 그 존재의 존재론적 지위와 관련해서 평가할 때에는 현재에도 여전히 해결하기 어려운 철학적 문제들에 직면해 있다. 특히 그 무엇으로도 환원할 수 없어 보이는 인간의 의식이나 마음의 고유성과 특이성에 대한 믿음은 인간의 본질을 확정하는 데 걸림돌도 작용하며, 때문에 문제 해결은 더욱 난망한 일이 되곤 한다. 하지만 비록 현재로서는 문제 해결이 쉽지 않지만, 만일 어떤 존재가 최소한 인간 행위자가 갖는 도덕적 특성을 갖는다면, 그 또한 인간과 동일한 지위를 갖는 도덕적 존재로 간주되어야 한다는 것은 부인하기 어려울 것이다.

수 있느냐에 따라 강인공지능에 속할 수 있으며, 또 인간 수준을 무엇으로 그리고 어떻게 보느냐에 따라서 이 같은 구분 자체가 매우 임의적일 수 있다. 그 이유는 현재 기계가 인간 혹은 인간처럼, 혹은 그 이상이 된다는 것이 정확히 어떤 기준을 충족시켜야 하는지 엄격한 기준을 제시하기 어렵기 때문이다. 그러므로 특수한 한 가지 기능이나 다기능의 유무와 우월성에 기초하기 보다는 인간 자체와 유사한 능력을 갖는 기계, 그렇지 못한 기계, 그리고 인간의 능력을 넘어선 기계로 구분하는 것이 훨씬 편리할 수 있다. 자연종 내지는 인간종을 구분하는 방식과 동일한 기준으로 인공종을 구분할 수 있는지는 또 다른 문제다. 우리는 인간종을 물리적, 정신적, 도덕적 측면에서 구분해서 평가하는 경향이 있는데, 인공종에도 이러한 구분을 적용할 수 있는지는 의문이다. 이 글이 다루고 있는 문제는 이와 직접적인 관련이 있다.

기계로 업그레이드된 AI인간(기계인간)이나 초인공지능(인공종)의 AI로봇은 현행 인간에 속하는 순수인간과 비교하면 일견 단지 물리적 능력에서는 순수인간 보다 뛰어난 존재들로 간주할 수 있다. 그러면 도덕적 관점에서는 어떤가? 우리가 어떤 존재를 도덕적으로 고려하는 기준은 그가 어떤 피부색이나 능력을 갖고 있느냐가 아니라 단지 그가 현재 도덕적 존재로 간주되는 우리(순수인간)와 같은 존재인지 아닌지에 있다. 그런 인공종의 존재는 과연 가능한가? 더욱이 설사 그런 인공종이 가능하다 하더라도 우리가 상상할 수 없을 만큼의 아주 월등한 능력을 갖춘 그런 AI로봇이 언제나 도덕적으로 옳은 행위만을 하는 존재가 아니라면, 아마 심각하거나 끔찍한 일이 벌어질 수도 있다. 사전에 인공지능을 장착한 기계화된 AI인간이나 그 이상의 AI로봇의 출현을 어느 정도까지 제한하거나, 그에 부합하는 윤리적 및 법률적 기준을 만들어야 한다는 주장이 설득력이 있는 이유도 이런 가능한 사태를 예견할 수 있기 때문이다.

인공지능 행위자로서 도덕적 기계의 존재를 평가하고 그에 따른 가능한 예방조치들을 요구하기 위해서는 먼저 자율적 인공지능으로서 간주되는 도덕적 기계 자체의 가능성 및 예상되는 문제들에 대한 정확한 평가가 필요하다. 이와 관련해서 제일 먼저 필요한 작업은 인간을 도덕적 존재라 할 때, 그 도덕적 속성이 과연 무엇인지 여부이다. 그 경우에 한해서 우리는 AI로봇을 그와 같은 속성을 갖는 도덕적 존재로 만들 것인지 혹은 만들 수 있는지를 결정할 수 있으며, 또 궁극적으로는 가공할 존재가 될 수도 있는 초인공지능 로봇 혹은 도덕적 기계의 출현도 통제할 수 있다.

인간중심적인 윤리적 관점에서 도덕적 고려의 대상은, 그가 성인이든 어린 아이든, 남자든 여자든, 장애가 있든 없든, 어떤 물리적 제약과 관계없이 우리가 인간이라고 부르는 존재들이다. 이 경우 우리는 '인간은 도덕적 존재이다'라는 진술을 참인 명제로 간주한다. 인간의 본질을 도덕성에 두고 있는 것이다. 그러나 '인간임'과 '도덕적임'을 동일시하는 것

은, 인간의 인간임을 규정하는 도덕성의 정체가 무엇인지 규명되지 않는 한, 하나의 전제일 뿐 증명된 것은 아니다. 만일 그렇다면, 다시 말해서 무엇이 도덕적인지를 분명히 규정하지 않는 한, 어떤 것이 도덕적 기계인지도 말할 수 없다. 그러나 실질적으로 그리고 세부적으로 모든 면에서 인간의 도덕성이 무엇이며 어떻게 정의할 수 있는지에 대해서 의견일치가 이루어진 것은 아니다. 이러한 불일치에도 불구하고 도덕성에 대한 탐구는 도덕성에 대한 일말의 이해를 전제하고 있다. 소위 윤리학적 탐구 자체가 수행될 수 있는 근본 이유도 바로 이러한 '이해'를 공유하고 있기 때문이다. 바로 이 때문에 모두가 동의할 수 있는 결론에는 도달하지 못하더라도 실제로 어떤 식으로든 도덕에 대한 유의미한 진술을 할 수 있다. 즉, 이러한 이해와 상식의 관점에서 어떤 존재가 인간이기만 하다면, 그는 필연적으로 도덕적 존재가 된다.

그렇다면 여타의 존재와 본질적 차이를 갖는 인간의 '인간임'과 '도덕적임'을 동일시 할 수 있는 속성, 즉 인간적 도덕성의 본질은 무엇인가? 단 하나의 도덕성에 대한 정의가 마련되어 있지 않는 한, 그 도덕적 속성이 다른 존재들의 그것과 본질적 차이를 갖는 것인지, 아니면 정도의 차이에 불과한 것인지에 대한 고찰은 뒤로 미루어 두어야 할 것이다. 반면에 도덕성의 발생적 차이나 개념적 차이가 무엇이든 실제로 우리가 인간적 도덕성으로 간주하는 몇 가지 속성을 살펴볼 수 있다. 그러한 속성의 규명을 통해서 마찬가지로 인간종만이 아니라 인공지능에도 그러한 속성을 부여할 수 있는지를 판별할 수 있을 것이다. 내가 염두에 두고 있는 속성들은 '자율성', '좋음과 나쁨', '옳고 그름', '동기와 의도' 등이다. 이들 중에서 자율성은 다른 속성들에 수반되는 근본 속성으로서 인간과 여타의 존재들을 구분 짓는 근본 특성이다.

3. 자율적 인공지능 : 동기와 의도

도덕적 개념들로서 좋음과 나쁨, 그리고 옳음과 그름은 독립적 속성인가 아닌가? 철학적 관점에서 이 문제에 대해서는 역사적으로 플라톤과 아리스토텔레스의 전통이 대립하듯이 일방적인 결론을 내리기 어렵다.[6] 그러나 어떤 전통이든 이와 관련한 숱한 논의들에도 불구하고 그들 개념과 관련 있는 도덕적 속성은 동기와 의도의 행위 상관성, 즉 한 행위의 의지적 선택의 자율성과 결합되어 있다. 그것은 인간적 관점에서 한 행위는 도덕적 자율성과 자유의지와 내재적으로 결합되어 있다는 것을 방증한다. 모어는 윤리적 행위자로서 인간적 본성에 귀속시킬 수 있는 도덕적 속성들의 상관자로 "의식, 지향성, 자유의지" 등을 들고 있는데,[7] 이는 도덕적 자율성의 가능한 존재론적 조건들이라 할 수 있다.

상호 관련이 깊은 감정(motion)이나 동기(motive) 내지는 동기부여(motivation)라는 말은 일상적 맥락에서는 다소 일관적이지 못하며 모호하게 사용되는 경향이 있다.[8] '나는 너를 사랑한다'라고 말할 때 여기서 사랑은 특정한 감정 혹은 감정 상태일 수도 혹은 하나의 태도일 수도 있다. 어떤 (감정적) 느낌을 갖는다는 것은 정확히 어떤 상태를 이르는지 따져보아야 할 문제다. 동시에 이런 감정들과 결합되어 있는 동기 역시 마찬가지다. '그렇게 생각하게 된 동기(이유)' '그녀를 사랑하게 된 동기(계기)', '그렇게 행동한 동기(원인)' 등 '동기를 갖는다'는 것은 정확히

6) 이에 대해서, 지금 증명하려고 시도하지는 않지만, 나는 그것이 관계적 속성을 갖는다고 생각하고 있다. 이 문제와 관련해서는 다음을 참조. 맹주만, '이성적 공감과 윤리적 주체',『이성과 공감』, 62-72쪽.

7) James Moor, "The Nature, Importance, and Difficulty of Machine Ethics", in Wallach, W./Allen, C., *Moral Machines*, 18쪽; 신상규, '인공지능과 지향성', in 이중원 외,『인공지능의 존재론』, 215-244쪽.

8) Aaron Sloman, "Motives, Mechanisms, and Emotion", in Margaret A. Boden (ed.), *The Philosophy of Artificial Intelligence*, 231쪽.

어떤 의미인가? 이는 감정이나 동기의 문제와 관련해서 우리가 마음 상태 혹은 마음의 기제에 대한 하나의 이론이 필요하다는 것을 함축한다.[9] 마음의 상태 또는 마음의 항구적 성질들은 특수한 상황과 관련해서 사고방식이나 감정과 태도 및 동기들이 생기는 심리적 성향들로 구성되어 있는데,[10] 이와 관련해서 이러한 성질들이 어떻게 작용하는지에 대한 충분한 이론을 갖지 못한다면, 마찬가지로 인공지능에 어떤 마음이나 작용기제를 구현할 것인지 시도조차 할 수 없다. 그럼에도 '행위의 동기' 문제와 관련해서 분명한 것은 그것은 일종의 내적 표상임에는 분명하지만 대상화할 수 없는 '어떤 상태'로서 수동적 혹은 능동적 표상(상태)과 관계한다는 점이다. 수동적 표상이라면, 그것은 비자발적인 동기(배고픔과 같은 본성적 동기, 자비심과 같은 수동적인 도덕적 동기, 외적 강제에 의한 동기부여)일 것이며, 또 능동적 표상이라면, 그것은 자발적 동기(의지적 선택이나 의도 혹은 목적에 의한 도덕적 동기)일 것이다. 특히 이 중에서 자율적 인공지능의 문제와 관련해서 주목할 것은 도덕적 동기이다.

기본적으로 도덕적 동기의 생성은 의도와 밀접한 관계가 있는데, 그 원인이나 목적은 행위자 자신에 의해서 스스로 야기될 수도 있으며, 또는 외부로부터 강제될 수도 있다. 가령, '위험에 처한 사람을 도와주어야 한다' 그리고 '위기에 처한 사람을 도와주는 사람은 좋은 사람이다'는 도덕적 진술은 몇 가지 사실판단과 도덕판단을 포함하고 있다. '위험에 처해 있음'에 대한 인지적 판단이 선행되어야 하며, 그것이 도움을 필요로 하는 상황인지, 어떤 도움을 주어야 할 것인지, 심지어는 도움을 필요로 하는 사람이 누구인지, 만일 전쟁 중이라면 그가 적군인지 아군인지, 혹은 적군이면 도움을 주는 것이 옳은지 등을 구별할 수 있어야 한다. 아

9) 같은 글, 231-232 참조.

10) Jordan H. Sobel, *Walls and Vaults, A Natural Science of Morals*, 16쪽; Corliss G. Swain, "Passionate Objectivity", 480쪽; R. Cohon, "The Common Point of View in Hume's Ethics", 830쪽.

니면 그가 누구일지라도 위험에 처한 사람은 무조건적으로 도와주어야 할 수도 있다. 여기에는 이른바 “도덕적 선택”의 문제가 깊이 개입해 있다. 그리고 이 선택의 문제는 어떤 행위를 선택하고 행해야 할 동기와 의도에 의존적이다. 왜냐하면 동기는 선택적 원인이면서 의도와 목적의 제약을 받기 때문이다. 즉, ‘의도와 동기가 없다면, 도덕적 선택도 없다’고 할 수 있다. 게다가 비록 ‘행위의 동기(motives)’와 행위의 의도나 믿음과 관계있는 ‘행위의 태도(attitude)’는 논리적으로 구분된다.[11] 지각 가능한 행위는 도덕적 태도를 반영하지만, 그것이 본래의 의도에 부합하는지 여부는 별개의 문제이며, 따라서 의도와 행위 사이에는 필연적인 도덕적 관계가 성립하지만, 그렇다고 그것만으로 행위의 도덕성을 결정할 수는 없기 때문이다. 그러므로 '인공적 도덕 행위자'(AMA; artificial moral agent)의 가능성 여부를 판단하려면, 이 양자의 관계에 대한 별도의 독립적인 논구가 반드시 필요하다. 하지만 그것을 어떻게 정의하든 중요한 것은 어떤 동기를 갖게 되고 어떤 의도나 목적에 따라서 그에 부합하는 행위를 할 것인지 여부가 ‘사전에 미리’ 결정되어 있지 않다는 것이다. 그것은 전적으로 ‘사태의 발생 및 인지와 함께’ 작동하며, 그리고 ‘사후에 비로소’ 자신의 의지로부터 스스로 결정해야 한다는 의미에서 이는 칸트가 “자율성의 원리”라고 부르는 도덕적 자율성(Autonomy)을 전제한다.[12] 그리고 무엇보다도 이 점에서 도덕적 자율성과 공학적 자율성은 엄격히 구분되어야 하며, 후자는 기껏해야 자동성(Automaticity)의 의미밖에 갖지 못한다.

AIA가 본래적 의미에서 자율성이 아닌 자동성만을 갖는다고 주장하는 가장 중요한 이유는 기계적 자동성은 도덕적 속성으로서 의지의 자율성과 전혀 다른 개념이라는 데 있다. 동기의 원인성을 의지가 아닌 감정에 두는 도덕적 견해의 경우에도 마찬가지다. 동기로서의 감정을 정해진 매

11) Jordan H. Sobel, *Walls and Vaults, A Natural Science of Morals*, 17쪽 참조.
12) 맹주만, 『칸트의 윤리학』, 192쪽.

뉴얼이 아니라 스스로 생성하는 감정은 그것이 어떤 감정인지 그리고 어떤 감정이어야 하는지를 우리는 사전에 미리 특정할 수 없다. 다만 의지의 자율성에 더 주목하는 것은 행위의 선택과 결행은 의지적 요인에 의해서 최종적으로 결정된다고 보기 때문이다. 하지만 이와 같은 도덕심리학적 문제 혹은 그 근저에 있는 형이상학적 논쟁에 가담할 필요 없이 우리는 도덕적 자율성이 AMA에게는 원천적으로 불가능하다는 것을 이와 독립적으로 증명할 수 있다.

도덕적 관점에서 자율적 인공지능이 지녀야 할 도덕적 자율성과 동기 및 의도나 목적의 상관성은 실제로 우리에게 어려운 철학적 문제를 제기한다. 그것은 행위의 문제에 있어서 이성, 감정, 그리고 의지의 고유성과 상관성 문제인데, 오랜 철학적 논의들은 어느 일방의 손을 들어 주는 결말을 보여주지 않는다. 단적으로 누가 좋은 사람인지를 결정하고자 할 때, 이 문제는 먼저 좋음과 사람의 개념을 어떻게 규정하느냐에 따라서 그 대답이 달라질 수 있으며, 또 철학의 역사는 그에 대한 다양한 견해들을 내놓고 있다. 그러나 이런 쟁론들이 벌어지는 근본 요인은 주로 발생론적 측면에서 두드러진다. 이에 대한 입장과 해석이 다르다 하더라도 이러한 사정으로부터 두 가지 결론을 도출할 수 있다. 하나는 발생론적 요소가 무엇이든 도덕적 판단과 행위는 언제나 동기와 의도 그리고 목적과 관계한다는 사실이며, 다른 하나는 어느 하나의 발생론적 요소를 특정하지 못하는 한, 도덕적 기계의 생산은 원천적으로 불가능하거나 지극히 위험한 실험이 될 것이라는 사실이다.

도덕적 자율성에 기반을 둔 동기와 의도 혹은 그와 상관적인 특정한 믿음 등은 그것이 본능적 필연성이나 강제나 강압에 의한 것이 아닌 한 미리 결정되지도 주어지는 것도 아니다. 도덕적 판단과 상관하는 행위의 원인으로서 동기는 자연적 동기와 도덕적 동기로 나눌 수 있는데, 칸트의 경우에 이를 경향성과 이성(혹은 실천이성, 선의지, 이성적 존재자의 의지)의 동기로 구분한다. 칸트가 이렇게 하는 구분하는 이유는 도덕적

행위는 어떤 경우이든 자율적 선택의 문제이며, 그 밖의 행위는 자연적 필연성 즉 인간의 자연적 기질 혹은 경향성에서 기인하는 것으로 보았기 때문이다. 경향성의 동기가 수동적 동기라면, 의지의 동기는 자발적 및 능동적 동기다. 전자가 주어지는 동기라면, 후자는 생성하는 동기, 이른바 자율적 동기이며, 그것이 칸트가 말하는 "도덕적 동기"이다. 인공지능은 전자의 동기에 따라 행동하는 기계일 뿐, 결코 후자의 동기에 따라서 행위할 수 없다. 능동적 동기는 자발적으로 만들어내고 규제하는 능력인데 인공지능은 의도나 목적을 스스로 생성할 수 없을 것이기 때문이다. 만일 무조건적이며 독립적인 자기목적적 및 목적정립적 행위가 가능한 인공지능이 존재한다면 그것은 무한대의 능력을 지닌 신적인 존재가 될 것이다. 설사 가능하더라도 그런 존재가 사악한 존재이기까지 하다면?

칸트가 비록 내재적 동기의 차원을 이렇게 구분했지만, 칸트와 대척점에 있는 입장들도 도덕적 동기를 미리 정해져 있는 원인에 두지는 않는다. 즉, 도덕판단은 동기 없이 발생하지 않으며, 이러한 도덕적 동기는 의도나 믿음에 의존적이다. 동기 내재주의(motivational internalism)라 할 수 있는 이러한 입장에 따르면,[13] 내재적인 도덕적 동기 없이 도덕판단은 이루질 수 없다. 말하자면, 우리는 먼저 도덕적 개념을 갖고 있어야 한다. 그것이 무엇인지에 대해서는 다양한 도덕적 입장이 있을 수 있지만, 분명한 것은 도덕적 행위자는 어떤 식으로든 그러한 개념을 갖고 있어야 한다는 것이다. 그렇다면 AMA가 되기 위해서는 그것을 프로그램화할 수 있어야 하는데, 이는 순환 문제의 오류를 피할 수 없게 된다. 가령 인격성을 전제하지 않는 기능적 의미에 한정해서 AIA에 도덕적 행위자의 자격을 부여한다면,[14] 이미 그것은 그 자체로 어떤 도덕적 개념을 적용한 것이 되며, 오직 그러한 의미에서만 도덕적 행위자인 것이다. 그런데 우리는 그 도덕적 개념을 미리 확정할 수 없다.

---

13) Jesse J. Prinz, *The Emotional Construction of Moral,* 18-19, 42, 135-136쪽.
14) 신상규, 「인공지능은 자율적 도덕행위자일 수 있는가?」, 265-292쪽.

또한 자유의지론을 거부하는 대부분의 결정론자들의 경우에도 그들이 옹호하는 것은 사건 원인의 결정론이지 도덕적 동기의 결정론은 아니다. 도덕적 동기를 결정하는 것이 의지이든 감정이든, 행위의 의도와 목적은 동기에 대해서 숙고하게 만들며, 최종적으로는 동기의 변경 역시 의도나 목적 및 그에 수반하는 믿음에 따라 달라진다. 마치 우리가 어떤 인공지능을 만들 것인지 숙고하고 결정할 때, 그것은 원래부터 있었던 것이 아니라 우리가 만들어내야 하는 것이다. 그리고 어떻게 만들어야 할 것인지를 결정하지 않는 한, 결코 그런 것을 만들 수 없는데, 우리는 어떤 목적이나 의도를 가져야 으며, 그것이 도덕적 행위자라고 말할 수 있는가? 우리는 이와 같이 행동할 수 있는 인공지능 기계를 만들어낼 수 있는가?

만일 도덕적 자율성을 갖춘 인공지능이 존재한다면, 그것은 분명 인간과 다를 바 없는 도덕적 기계이다. 더욱이 이 도덕적 기계가 인간의 지적 · 물리적 능력 보다 뛰어난 행위자라면, 더불어 순수인간이 저지르기도 하는 끔찍한 해악들을 생각한다면, 인간 보다 더 뛰어난 도덕적 기계의 존재는 가히 예측할 수 없는 공포를 불러일으킨다. 만일 이런 AI로봇이 출현한다면, 그런 존재는 인간에게는 돌이킬 수 없는 재앙이 될 수 있다. 이러한 위험을 차단하려면 AMA는 언제나 좋은 일만 하는 도덕적 행위자로 제작되어야 한다. 그렇다면, 그것은 도덕적 기계인가? 그것은 논리적으로 그리고 개념적으로 불가능하다. 그런 존재는 도덕적 자율성 개념에 모순된다. 도덕적 자율성은 비윤리적 행위도 선택할 수 있다는 것을 함축하기 때문이다. 만일 언제나 도덕적으로 옳은 행위만을 하는 인공지능 기계가 가능하다면, 그때의 자율성은 기계적 자율성에 지나지 않으며, 그러한 자율성은 자동성에 불과한 의사 자율성(quasi-autonomy)에 지나지 않게 될 것이다.

그러나 특정한 행위에만 특화된 기계라면, 그러한 제한된 조건적 범위 내에서 가능한 도덕적 기계를 생각해 볼 수 있다. 이런 공학적 · 기계적 자율성을 갖춘 도덕적 자동기계는 특수한 목적에 한정된 '의사 도덕적

행위자'(quasi-moral agent), 즉 좋은 기계로서 특정한 목적에 봉사하는 목적형 혹은 기능형 인공지능에 해당될 것이다. 맞춤형 인공지능이라 할 수 있는 현행 로봇의사 왓슨과 로봇변호사 로스 수준의 인공지능도 엄밀히 말해서 단순노동형 기계에 지나지 않는다. 그들이 진정한 의미에서 인간의사나 인간변호사와 동일한 일을 대행하는 것이 아니기 때문이다. 현재의 인공지능은 모두 업무의 처리를 (놀라울 정도로) 수월하게 처리하는 도우미일 뿐이다. 따라서 인간 중심의 응용윤리의 한 분야이기도 하면서 인간윤리에 대비되는 성격을 갖는 기계윤리(Machine Ethics)의 분야에서 이루어지고 있는 최근의 시도들 중에서 마치 사람이 도덕적 문제를 해결하는 것과 같은 의미에서 인공지능에 "윤리적 딜레마를 해결하는 방법과 절차, 윤리적 원칙을 부가하려는"[15] 노력은 그것이 인간 재판관의 도우미 역할에 한정되어야 하며, 가능한 부작용을 예방하고 통제하는데 집중해야 한다.

4. 좋은 사람과 도덕적 기계

의사 자율성과 도덕적 자율성의 구분은 동기와 의도의 자기생성으로서 인간적 자율성을 설명해준다. 우리는 어떤 사람을 '언제나 좋은 사람'이라고 말할 수 없지만, 반대로 어떤 인공지능에 대해서는 '언제나 좋은 기계'라고 말할 수 있다. 의사 자율성이 맞춤형이나 특수목적형 인공지능의 경우에는 가능한 것처럼, 특정한 일이 특화된 인공지능은 '언제나 좋은 기계'라고 할 수 있다. 좋은 기계는 있어도 도덕적 의미의 좋은 기계, 즉 도덕적 기계는 있을 수 없다.

AIA와 AMA 연구를 선도하고 있는 웬델 월러치과 콜린 알렌은 그들의 저서에서 로봇의 도덕과 관련해 진행되고 있는 논의들을 개관하면서

15) Anderson, M./Anderson, S. L. (ed.), *Machine Ethics*, 1쪽.

여기에 뒤따르는 당연한 질문 세 가지를 들고 있다. 이에 의하면, “세상은 AMA를 필요로 할까?, 사람들은 컴퓨터가 도덕적 결정을 내리기를 원하는가? 그리고 컴퓨터가 도덕적 결정을 내리는 것이 필요하거나 불가피하다고 여긴다면, 공학자와 철학자는 AMA를 어떻게 설계해야 하는가?”[16] 앞서 제시했듯이 이에 대한 나의 대답은 각각 1) 세상은 그렇게 되어 가고 있으며, 2) 우리가 원하든 아니든 많은 의사결정에서 기계로봇의 참여는 이미 이루어지고 있으며, 따라서 선택적 결정과 그에 따른 책임은 최종적으로 인간의 몫이라 하더라도 이러한 결정에서 점차 인간과 기계의 역할 구분이 모호해지고 있으며, 그러나 3) AMA는 언제나 의사 자율성을 지닌 기계에 머물기 때문에 도덕적 기계가 아니라 단지 좋은 기계를 만드는 노력에 집중해야 한다는 것이다. 로봇의사와 로봇변호사의 존재가 시사하듯이 그들은 결코 인간의사나 인간변호사와 동일한 도덕적 존재가 될 수 없으며, 다만 특정한 일과 작업에 한정될 것이기 때문에 최종 선택과 결정에 있어서는 인간의 도움을 받아야 하며, 또 받도록 통제되어야 한다.

그러면 특정한 일에 종사하도록 설계된 맞춤형(특수목적형) AI로봇이 두 개 혹은 그 이상의, 나아가 인간이 할 수 있는 가능한 일과 행위 영역에 적합하게 설계된 만능 인공지능의 경우에, 그것은 진정한 의미에서 도덕적 자율 존재, 언제나 좋은 기계라고 할 수 있지 않은가? 가령 자율주행 자동차를 운전하면서 동시에 집안일을 하는 AI로봇, 혹은 그와 동시에 로봇의사이면서 도덕적 딜레마 해결을 도와주는 도덕적 기계가 되는 것이 동시에 모두 가능하다면 어떤가? 이에 대해 아니라고 답해야 한다. 도덕적 기계는 불가능하기 때문이다. 머지않아 상용될 될 것으로 예상되는 블록체인(BlockChain)을 자율적 규제가 가능한 시스템이라 부르기도 하지만, 이 또한 실제로는 기꺼해야 설정된 목적에만 봉사하는 자

16) 웬델 월러치, 콜린 알렌, 『왜 로봇의 도덕인가』, 22쪽.

동화된 시스템에 불과하다.

그렇다면 인간의 감정을 읽을 수 있거나 인간적 감정을 표현할 수 있는 컴퓨터, 이를테면 감정로봇은 어떻게 가능한가? 이를 위해서는 감정에 대한 정의가 필요하다. 그런데 감정을 어떻게 정의하든 인간의 감정은 자율적 행위자로서 행위의 선택과 결행에 있어서 능동적 동기 및 의도와 관계하기 때문에 인공지능 행위자는 특정한 의미에 기계적으로 반응하도록 프로그램된 감정 기호의 장치일 뿐, 자기생산적 의도와 목적과 관계하는 능동적 감정에 대해서는 전혀 무력할 것이다. 기껏해야 심리적 기제에 종속된 수동적 동기로부터 야기되는 감정에 특정한 방식으로 반응하는 자동인형 감정로봇에 불과할 것이다. 물론 이런 수준의 인공지능은 비록 도덕적 기계는 아니더라도 인간의 감정적 교감에 기여하는 유용한 기계는 될 것이다. 그러나 그렇다고 그것이 도덕적 기계나 도덕적 행위자라고 할 수는 없다.

만일 아마 우리가 통상적으로 좋은 사람이라고 부르는 특징이나 조건을 충족할 수 있다면, 그것은 우리가 도덕이라고 말할 수 있는 최소한의 조건을 충족시키는 일이 될 것이다. 그러면 좋은 사람이란 무엇이며, 또 누가 좋은 사람인가? 그런데 우리가 좋은 사람이라는 표현을 반드시 도덕적 맥락에서만 사용하는 것은 아니다. 도덕과 무관하게 특정한 일이나 행동을 잘 하는 사람, 예를 들면 청소를 잘하는 사람, 아픈 사람을 잘 돌보는 사람, 공을 잘 차는 사람, 그림을 잘 그리는 사람, 심지어는 수술을 잘하는 의사 등의 경우에도 우리는 그 사람을 좋은 사람이라고 부르기 때문이다. 이런 경우의 '좋음'은 기능적 의미에서의 좋음이라 할 수 있다. 이는 인공지능과 같은 기계에도 적용된다. 가령 좋은 기계의 좋음은 모두가 기능적 의미의 좋음이며, 이런 좋음은 조건적 좋음이다. 즉, 잘 드는 칼을 좋은 칼이라고 하지만, 이런 의미의 좋음은 사람을 죽이는 데도 쓰일 수 있으므로 특정한 기능에 한정해서만 타당하기 때문이다.

반면에 도덕적 의미에서의 좋음은 일반적으로 우리가 좋은 사람이라고

말할 때의 좋음이다. 기능적 의미에서의 좋음이나 조건적 좋음과 대비해서 이를 무조건적 좋음 혹은 자체적 좋음, 좋음 그 자체라 할 수 있다. 이런 입장을 견지하고 있는 관점으로 아리스토텔레스와 칸트를 꼽을 수 있다. 이 무조건적 좋음으로서 아리스토텔레스는 행복을, 칸트는 선의지를 제시한다.[17]

아리스토텔레스는 좋은 사람과 마찬가지로 좋은 삶 혹은 좋음 일반을 정의할 때, 기능(ergon)의 관점에서 접근한다. 이에 의하면, 좋은 사람이나 좋은 칼은 모두 기능적 의미에서 평가할 수 있다. 그런데 좋은(잘 드는) 칼의 기능이 도구적 목적에의 적합성에 따라서 평가되는 기능이라면, 인간적 좋음의 기능은 이성의 본래적 목적에의 적합성, 즉 이성적 탁월성/덕(arete)의 발휘에 있다. 인간적 탁월성을 규정하는 이 기능은 동시에 좋은 기계만이 아니라 좋은 삶을 규정하며, 마찬가지로 좋은 사람을 정의한다. 그런데 이 관점의 핵심은 '잘 함'(doing well)으로서의 기능에 있다. 좋은 칼은 그 도구적 목적에 맞는 '잘 듦'이며, 좋은 사람은 그 본성적 및 본래적 목적에 맞는 '잘 함' 즉 훌륭한 행위이며, 그것이 곧 아리스토텔레스적 의미에서 도덕적 행위이다. 그리고 이런 도덕적 행위는 무엇보다도 합리적 선택, 즉 자율적 판단에 기초해서 이루어진다.[18] 아니, 자율적 선택에 입각해서만 이루어질 수 있으며, 그것이 곧 도덕적 능력이며, 동시에 그런 능력과 성품을 갖춘 사람이 좋은 사람이다. 아리스토텔레스가 의미하는 자율성과 합리성을 지닌 도덕적 행위자는 미리 정해진 혹은 프로그램된 동기와 의도에 맞춰서 행위하는 사람의 특성이 아니라 벌어진 사태나 일어날 일에 대해서 어떤 선택적 행위를 '의욕하는 사람'이다. 이는 곧 의욕하지 않을 수도, 선택을 유보할 수도 있는 행위의 능력으로서의 도덕적 속성이다. 그리고 이 같은 도덕적 행위의 가능성을

17) Aristoteles, *The Nicomachean Ethics*, 1095b20-5, 1098a5; I. Kant, *Grundlegung zur Metaphysik der Sitten*, 18-19, 21-22쪽.
18) 맹주만, 「칸트의 덕과 덕-감정」, 34-38쪽; 『칸트의 윤리학』, 301-302쪽.

단적으로 보여주는 사례가 거짓말이다. 이를테면 자율적인 도덕적 행위자만이 거짓말을 의도할 수 있으며, 의욕할 수 있으며, 실행할 수 있다. 마찬가지로 좋은 사람이 되려고 노력할 수 있으며, 나쁜 사람이 될 수도 있다.

5. 도덕적 관점과 도덕적 기계

앞서 밝혔듯이 만일 도덕적 기계가 가능하다면, 아마 그것은 우리가 '도덕성' 또는 '도덕적 속성'을 의사 자율성에 부합하는 특정한 방식으로 정의하는 경우가 될 것이다. 그러나 어떤 경우에도 본래적 의미에서의 인간적 내지는 도덕적 자율성을 지닌 기계는 불가능할 것이다. 앞서 밝혔듯이 그것이 불가능한 이유는 도덕판단은 동기와 의도, 믿음, 태도 등과 관련한 도덕적 개념을 포함하고 있으며, 또한 이에 기초해서 좋은 사람과 나쁜 사람의 구별도 이루어지는데, AIA는 이러한 개념과 판단을 스스로 생성할 수 없기 때문이다. 한 존재가 도덕적 개념을 갖는다는 것은 곧 그가 도덕적 관점을 가질 수 있는 존재라는 것을 의미하며, 이것이 인간종과 인공종을 가르는 근본 조건이다.

동기와 의도를 갖거나 좋은 사람과 같은 행위자가 될 수 없는 인공지능은 결코 도덕적 기계가 될 수 없다. 도덕이라는 말의 본래적 의미에서 인공지능으로서의 도덕적 기계는 불가능하다. 어떤 한 존재가 도덕적 존재이냐 아니냐를 가르는 기준은 그가 도덕적 관점을 가질 수 있는 존재이냐 여부에 달려 있기 때문이다. 인간과 인공지능의 근본적 차이 역시 이에서 비롯된다. 인간은 도덕적 관점에 입각해서 행동하는 존재이지만 인공지능은 그럴 수 없다. 아리스토텔레스가 말하는 좋은 사람과 나쁜 사람을 가르는 척도 역시 하나의 도덕적 관점의 적용에 다름 아니다.

아리스토텔레스에 비견되는 칸트의 도덕적 관점은 도덕적 속성으로서

좋음을 더 엄격하게 규정한다. 무제한적으로 그리고 무조건적으로 선한 의지, 즉 의무에서 하는 행위를 의욕하는 의지만이 도덕적 선(좋음)이다. 이 의지는 "법칙의 표상에 따라 자기 자신의 행위를 규정하는 능력"[19]이다. 이른바 칸트가 말하는 선의지에서 하는 행위만이 도덕적 행위인 것이다. 이런 엄격한 규정의 정당성을 따지는 문제는 제쳐두고, 무엇보다도 행위의 도덕성을 규정하는 칸트의 도덕적 관점 역시 도덕적 자율성에 입각해 있다. 무엇을 어떻게 할 것인지가 먼저 또는 미리 주어져 있는 것이 아니다. 소위 칸트가 말하는 자율 도덕은 그것이 어떤 구체적 행위를 의욕하기 전까지는 형식적 자율성만을 갖지만, 이 자율성은 앞으로 일어날 혹은 의욕하게 될 행위와 함께 비로소 작용되며, 이 양자의 통일에 의해서 법칙적 자율성이 작동된다. 이것이 칸트의 도덕적 자율성의 요체이다.

그런데 간혹 칸트의 형식적 자율성을 그의 도덕법칙과 동일시하면서 기계적 법칙화가 가능한 인공지능의 자동성을 도덕적 자율성으로 오인한다. 하지만 자율성의 본래적 의미가 그렇듯이 그것은 전혀 사전에 기획된 것도 미리 주어져 있는 것도 아니다. 그러므로 칸트의 정언명법을 행위의 "도덕성을 확인하는 공식적인 도구로 AMA가 이용할 수 있는"[20] 가능성은 오직 사전에 설정된 특정한 목적에 한정되는 경우에만 적용될 수 있으며, 설사 그렇다 하더라도 목적 달성의 다양한 경우의 수들을 고려할 때 세부적인 하위의 도덕적 규칙들을 추론하는데 많은 어려움을 겪을 것이다. 이러한 과정과 절차를 프로그램화 하는 것도 실질적으로는 거의 불가능해 보인다. 오히려 진정한 윤리적 고민은 윤리적 문제나 딜레마 해결에 필요한 특정한 절차나 원칙을 적용하는 것이 아니라 과연 어떤 윤리적 이론이나 원칙들이 옳은지 또 타당한 도덕적 관점인지를 결정하는 데 있다.[21] 특정 목적을 위해 프로그램된 기계 이상으로 도덕적

19) I. Kant, *Grundlegung zur Metaphysik der Sitten,* 59쪽.
20) W. Wallach/C. Allen, 『왜 로봇의 도덕인가』, 165쪽.

기계의 제작이 원천적으로 불가능한 것은 그러한 목적을 위해 제작하려는 기계에 이미 특정한 도덕적 관점이 적용될 수밖에 없다는 점과 함께 그러한 관점을 결정하는 것 자체가 근본적인 윤리적 문제이자 딜레마이기 때문이다. 그런데 이것을 사전에 미리 고민할 수 있는 도덕적 기계의 제작이 어떻게 가능할 수 있는가!

공학적 의미에서 인공지능에 부여하는 자율성은 그냥 모든 기계적 장치에서 구현되는 자동성의 구현일 뿐이다. 마찬가지로 도덕적 의사결정 혹은 도덕적 판단을 위해서 "컴퓨터 기반 의사결정 지원 모델"로서 다양한 윤리이론에 기반을 둔 윤리시스템,[22] 이를테면 덕 기반이나 의무 기반 혹은 규칙 기반의 윤리시스템을 구축하더라도 그것은 인공지능 바둑 알파고처럼 바둑이라는 '한 가지의 주어진 고정된 관점'에 입각해서 과제를 수행하는 것에 불과 한 것이다. 따라서 도덕적 기계가 되려면 그것은 선행 조건 없이 '도덕적 관점'을 스스로 가질 수 있고 또 채택할 수 있는, 다시 말해 인간 본성 내지는 본성적 구조로부터 발원하는 도덕적 자율성을 가질 수 있는 인공지능의 '창조'가 가능해야 한다. 오직 그 경우에만 인공지능을 자율적 행위자 혹은 도덕적 행위자, 이른바 AMA 또는 자율적 인공지능 행위자로 간주할 수 있는 길이 열리게 될 것이다.

아리스토텔레스나 칸트의 경우만이 아니라 어떤 윤리이론이든 도덕적 행위를 판정하는 척도로서의 도덕적 관점은 도덕적 자율성과 함께 그 존재 혹은 행위자를 도덕적 존재로 규정해주는 근본 특성이다. 자율주행 자동차의 경우에도 그것은 이미 특정한 원리와 방식에 맞춰 기획된—칸트의 원칙주의든 계산적 공리주의든 덕 윤리적 규칙이든— 시스템으로서 그 자체가 특정한 도덕적 관점을 적용한 한 가지 실례가 될 뿐이다. 즉, 한 가지 방식의 행동만 가능한 자동 기계인 것이다. 그러나 도덕적 관점

21) Susan Leigh Anderson, "Machine Metaethics", in Wallach, W./Allen, C., *Moral Machines*, 23-24쪽 참조.

22) W. Wallach/C. Allen,『왜 로봇의 도덕인가』, 74쪽.

은 어떤 특정한 종류의 인공지능—그것이 인간을 완전히 닮은 도덕적 기계이든 아니면 그 이상의 가공할 기계이든—을 제작할 것인지를 의욕하는 행위 자체가 바로 도덕적 관점인 것이다. 다시 말해 인공지능 자체가 도덕적 관점을 가질 수는 없는 것이다.

6. 나오는 글 : 도덕적 자율성과 의사 자율성

AIA가 도덕적 기계가 될 수 있는 가능성은 전혀 없다. 그러한 가능성을 허용하는 유사 추론들은 논리적 상상일 뿐이다. AIA는 사전에 특정 조건을 설정해 두었을 경우에만 그에 부합하는 의미에서 도덕적 행위를 할 수 있을 뿐이며, 그 경우에만 좋은 사람과 동일한 의미의 '도덕적 기계' 혹은 AMA일 수 있다. 그러나 AIA는 그러한 조건 자체를 스스로 생성하고 설정할 수 있는 도덕적 관점을 가질 수 없기 때문에 결코 좋은 사람과 같은 존재가 될 수 없다.

도덕적 기계의 가능성은 AI로봇이 자동성이 아닌 자율성을 갖는 존재일 경우에만 성립한다. 공학적 인공지능 로봇이 갖는 의사 자율성 혹은 유사 자율성(semi-autonomy)은 도덕적 자율성이 아니다. 자기생성적, 자기정립적, 자기선택적인 실질적 자율성만이 진정한 의미의 자율성이며, AIA는 기껏해야 형식적 자율성만을 가지며, 그러한 자율성은 자동기계의 자동성에 다름 아니다. 마찬가지로 개념적으로 언제나 좋은 도덕적 행위자이기만 한 AIA는 한 가지 특정한 방식의 행위만을 하도록 설계된 좋은 기계에 불과하다. 그것은 놀랍도록 편리하거나 똑똑하거나 위험한 기계일 뿐이다.

인공지능을 도덕적 기계가 될 수 있거나 도덕적 행위자로 간주하거나 그에 따른 도덕적 행위의 가능성을 다루는 많은 논의들은 대부분 언어적 속임수이거나 실제로는 과대 포장된 수사에 불과하다. 그러한 논의들은

기본적으로 도덕성, 혹은 도덕적 속성을 오해하거나, 아니면 도덕성 자체를 상품 설명서의 사용지침과 같은 것으로 오인한 것이다. 그것은 마치 상품의 무분별 사용을 방지하거나 올바른 사용을 위한 매뉴얼을 윤리적 원칙과 같은 것으로 간주하는 것과 같다. 또한 AI로봇을 자율주행 자동차의 경우처럼 자율적 행위자로 호칭하는 것은 마치 인간과 유사한 행위자인양 부름으로써 인공지능 기계에 보다 많은 상품가치를 매기려는 과대광고일 뿐이다.

인간지능이나 초지능 AI로봇의 등장 보다 우려되는 것은 AI로봇이 갖게 될 예상되는 놀라운 능력—정확하게 말하면, 그것은 무기다—은 어떤 경우에도 자동기계인 이상, 결코 도덕적 행위자가 아니기 때문에 그로부터 빚어지는 위험을 예측하고 통제하려는 노력이 더 중요하다는 점이다. 도덕적 존재인 인간이 윤리적 행위와 마찬가지로 비윤리적 행위도 할 수 있듯이 진정한 의미에서 자율적인 도덕적 기계가 가능하다면, 그것은 곧 비윤리적인 행위도 할 수 있는 위험한 존재가 될 것이다. 그러나 살펴본 바와 같이 이런 인공존재의 출현은 불가능하다. 하지만 인간의 지시에 따르도록 프로그램 되거나 조작된 AIA는 가능할 것이기 때문에 사이보그처럼 인간-인공지능-기계의 결합체로서의 기계인간 혹은 초인공지능 로봇을 소유하거나 부리는 인간은 순수인간 이상의 존재가 될 수 있다. 게다가 모든 것을 컴퓨터 작동으로 통제할 수 있게 되고, 하나의 시스템으로 연결하고 연산하게 되는 인공지능 사회에서 그 위험은 가공할 것이 될 것이다. 이러한 가능한 위험을 예견하고 예방하는 것이 '기계윤리' 또는 '인공지능 윤리'에 맡겨진 과제다.

참고문헌

고인석,「인공지능의 존재 지위에 대한 두 물음」,『철학』 제136집, 2018.

도용태 · 김일곤 · 김종완 · 박창완,『인공지능 개념 및 응용』, 사이텍미디어, 2009.

맹주만,『이성과 공감 : 포스트모던 칸트와 공감윤리』, 어문학사, 2020.

_____,『칸트의 윤리학』, 어문학사, 2019.

_____,「인공지능과 로봇의사윤리」,『철학탐구』 제52집, 2018.

_____,「칸트의 덕과 덕-감정」,『칸트연구』 제28집, 2011.

변순용 · 송선영,『로봇윤리란 무엇인가?』, 어문학사, 2015.

_____,「로봇윤리의 이론적 기초를 위한 근본 과제 연구」,『윤리연구』 제88집, 한국윤리학회, 2013.

신상규,「인공지능은 자율적 도덕행위자일 수 있는가?」,『철학』 제132집, 2017.

이중원 외,『인공지능의 존재론』, 한울아카데미, 2018.

Anderson, M./Anderson, S. L. (ed.), *Machine Ethics,* Cambridge University Press, 2118.

Aristotle, *The Nicomachean Ethics,* Translated and edited by Roger Crisp, Cambridge University 2000.

_____,『니코마코스 윤리학』, 최명관 옮김, 서광사 1984.

_____,『니코마코스 윤리학』, 이창우 · 김재홍 · 강상진 옮김, 이제이북스 2006.

Asimov, I., *I, Robot,* New York: Spectra, 2008.

Boden, Margaret A. (ed.), *The Philosophy of Artificial Intelligence,* Oxford University Press, 1990.

Bostrom, N., *The Transhumanist FAQ,* www.nickbostrom.com (Version 2.1 2003)

_____, *Superintelligence: Paths, Dangers, Strategies,* Oxford University Press, 2014.

_____, "In defense of Posthuman Dignity", *Bioethics,* Vol. 19, No. 3, pp. 202-214, www.nickbostrom.com, 2005.

Cohon, R., "The Common Point of View in Hume's Ethics", in *Philosophy and Phenomenological Research* 57, 1997.

Edelman, Gerald M., *Second Nature: Brain Science and Human Knowledge,* Yale University Press, 2006.

Engelhardt, Jr. H. Tristram, *Foundations of Bioethics,* Oxford University, 1996.

Gelhaus, P., "Robot decisions: on the importance of virtuous judgment in clinical decision making", Journal of Education in Clinical Practice 17 (2011) 883-887, Blackwell Publishing Ltd.

Hume, D., A *Treatise of Human Nature,* edited by L.A. Selby-Bigge, Oxford: Oxford University Press, 1980.

_____, *An Enquiry concerning the Principles of Morals,* edited by J.B. Schneewind, Indianapolis: Hackett Publishing Company, 1983.

Jatinder N. D. Gupta/Guisseppi A. Forgionne/Manuel T. Mora (eds.), *Intelligent Decision-making Support Systems: Foundations, Applications and Challenges,* Springer, 2006.

Kant, I., *Grundlegung zur Metaphysik der Sitten,* Herausgegeben von Wilhelm Weischedel, Frankfurt am Main: Suhrkamp, 1968.

Kass, L. R., "Ageless Bodies, Happy Souls: Biotechnology and the Pursuit of Perfection", www.TheNewAtlantis.com, 2003.

Oliver, Richard W., *The Coming Biotech Age,* New York: McGraw-Hill, 2000.

Prinz, Jesse J., *The Emotional Construction of Moral,* Oxford University

Press, 2007.

Rifkin, J., *The Biotech Centry*, New York: Jeremy P. Tarcher/Putnam, 1998.

Russell, S. and Norvig, P., *Artificial Intelligence : A Modern Approch*, 3rd ed., Prentice Hall, 2010.

Sobel, J. H., *Walls and Vaults: a natural science of morals, virtue ethics according to David Hume*, New Jersey: John Wiley & Sons, Inc., 2009.

Sebeok, Thomas A. (ed. et al.), *Semiotica*, Journal of the International Association for Semiotic Studies, 2001 Vol. 134(1/4), Mouton de Gruyter · Berlin · New York.

Swain, Corliss G., "Passionate Objectivity", Nous 26, 1992.

Turing, Alan D., "Computing Machinery and Intelligence", in *Mind*, vol. LIX, No. 236, 1950.

Vallverdú, Jordi (ed.), *Thinking Machines and the Philosophy of Computer Science: Concepts and Principles*, Hershey · New York: IGI Global, 2010.

Wallach, W./Allen, C., *Moral Machines*, Oxford University Press, 2009.

_____, 『왜 로봇의 도덕인가』, 노태복 옮김, 메디치, 2014.

AI, Moral Machine, Good Person

Maeng, Jooman (Chung-Ang Univ.)

In this paper I will argue that artificial intelligence cannot be a moral machine because it is not an autonomous agent who can create and establish the motive, intention and purpose of the act on his own. The autonomy of the engineering meaning that an anticipated artificial intelligence agent will have, or the autonomy of artificial intelligence, has a fundamental difference from the moral autonomy of the human agent, and it is merely a pseudo-autonomy. So it's just an incredibly convenient or dangerous machine. Only self-creation and self-establishment and self-modification of motive, intention and purpose are the real moral autonomy, whereas artificial intelligence agents have only formal autonomy at best, and such autonomy is no less than the automation of automatic machines which is limited to a limited purpose, and specialized to implement only that. A machine that always performs certain acts is not really a moral machine or a moral agent. Artificial intelligence can never be a moral machine. Moral machines are theoretically and practically impossible. If there was such an existence, it would certainly be an artificial species with other formidable powers beyond human species. Thus, most of those prospects that take place around the current possible artificial intelligence agents are a product of false imagination. Artificial intelligence robots, which are nothing but pseudo-autonomous agents from an ethical standpoint, along with the impossibility of moral machines, must be thoroughly planned and controlled machines in all

respects.

Key words: Artificial Intelligence Agent(AIA), Artificial Moral Agent (AMA), Robot, Human Intelligence, Natural Intelligence, Moral Machine, Moral Agent, Good Person, Motive, Intention, Purpose, Autonomy, quasi-autonomy, Automaticity

맹주만 : maengjm@cau.ac.kr

투 고 일	2020년 7월 15일
심 사 일	2020년 8월 3일
게재확정	2020년 8월 17일