과학철학 22권 2호 (2019) pp. 79-104

인공지능에게 책임을 부과할 수 있는가?:
책무성 중심의 인공지능 윤리 모색†

이 중 원‡

본 논문에서는 자율적인 행위자로서의 인공지능 시스템이 부정적인 행위 결과를 야기했을 경우 과연 책임을 부과할 수 있는가라는 문제를, 책무성 개념을 중심으로 진지하게 다룰 것이다. 오늘날 인간이 아닌 인공지능 시스템을 놓고 책임 문제를 다시금 논하는 배경은, 선택의 자율성을 지닌 인공지능 시스템이 인간의 실존에 능동적으로 작동하면서 '많은 손'의 문제를 일으키고 있고, 그럼에도 책임을 부과하지 않는다면 책임 공백의 문제가 발생할 우려가 있기 때문이다. 따라서 인공지능 시스템에 대해서도 책임 소재의 문제가 발생할 수 있는 상황 조건을 우선 제시할 것이다. 다음으로 인간에게 배타적으로 적용되어 온 전통적인 도덕철학에서의 책임 개념을 뛰어넘어 인간이 아닌 다른 자율적인 행위자에게도 확대 적용될 수 있는 책임 개념의 가능성을, 레비나스의 책임 개념을 중심으로 검토해볼 것이다. 나아가 이를 통해 책임 개념의 외연을 설령 확장 가능하더라도 현재나 가까운 미래의 인공지능 시스템에게 이를 적용하는 것은 쉽지 않음을 밝히고, 인공지능 시스템에 대해 책임(responsibility) 대신 책무(accountability) 개념의 적용을 제안할 것이다. 그리고 이러한 책무성이 인공지능 시스템에서 실질적으로 구현가능한지, 설명 가능한(explainable) 인공지능 시스템을 대상으로 분석할 것이다. 마지막으로 인공지능 시스템에 대해 책무성 중심의 윤리 체계를 구축하는데 필요한 윤리 프레임의 기본 요소들을 제안하는 수준에서, 그러한 윤리 체계의 가능성을 조심스럽게 전망해 보고자 한다.

【주요어】 인공지능, 많은 손 문제, 책임, 레비나스의 타자윤리, 책무성, 설명 가능한 인공지능

† 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF 과제번호: NRF-2016S1A5A2A03927217).

‡ 서울시립대학교 철학과, jwlee@uos.ac.kr.

1. 문제 제기 - 인공지능 시스템에 대해 왜 책임을 논하는가

전통적인 도덕철학에서 책임은 인간에게만 부여된다. 책임에 관한 고전적 모델에 따르면 책임(responsibility)은 인간만이 질 수 있다. 인간만이 자유의지를 가지고 이에 의거하여 행동하며 오직 이러한 행동의 결과에 대해서만 책임을 물을 수 있기 때문이다. 자유의지가 아닌 자연의 인과적 관계와 같은 외부적 요인에 의해 사건이 발생한 경우, 가령 태풍으로 홍수가 발생한 경우 홍수 피해의 물리적 원인은 태풍인 만큼 담당 관리자에게 홍수에 대한 책임을 물을 수는 없고, 만약 담당 관리자의 잘못으로 이차 피해가 발생했을 경우 이에 대한 책임만을 물을 수 있다. 이러한 관점은 아리스토텔레스에게로까지 거슬러 올라간다. 그에 따르면 어떤 행위가 도덕적 판단의 대상일 수 있기 위해서는 그 행위가 행위자의 자율성에 의한 것인가 아닌가에 달려 있다.[1] 이런 토대 위에 칸트는 책임과 자유의 불가분의 관계를 강조하면서, 자유를 바탕으로 도덕적 책임을 정당화한다.[2] 한마디로 자유는 책임을 묻기 위한 전제조건인 셈이다. 지금까지 철학사에서 자율성이나 자유의지는 인간의 고유 속성으로 간주됐던 만큼, 인간 이외의 다른 존재자에게 (도덕적) 책임을 부여한다는 것은 도저히 받아들일 수 없는 일인 것이다.[3]

그럼에도 불구하고 오늘날 인간이 아닌 인공지능 시스템을 놓고 그

---

1) 아리스토텔레스는 자발적 행위와 비자발적 행위의 구분 기준을 의도성(Willentlichkeit)에서 찾고 있다. 의도하거나 분명한 결단에 의한 행위는 자발적 행위이지만, 외적 강요에 의한 행위는 모두 비자발적 행위다(아리스토텔레스 1984,『니코마코스 윤리학』3권 참조).

2) Hildebrand (2012), chapter 2; Nidditch (1992), p. 411; Robson (1977), p. 281; Johnson and Cureton (2016).

3) 심지어 전통 도덕철학에서는 인간의 경우라도 자유의지가 약한 어린아이나 자유의지가 없는 의식불명의 사람에게조차 법적으로나 도덕적으로 책임을 묻지 않고 있다.

동안 도덕철학에서 배척했던 책임 문제를 다시금 논하는 이유는 무엇인가? 두 측면에서 그 배경을 살펴볼 수 있다. 첫 번째 배경은 인공지능 시스템을 활용하는 과정에서 전통적인 방식으로 인간에게만 책임(responsibility)을 부과하는 경우 발생하는 문제와 관련이 있다. 일반적으로 누군가 책임을 져야 한다고 말하려면, 그것을 가능하게 하는 조건들이 충족되었는지 먼저 살펴볼 필요가 있다. 이와 관련하여 (철학적 논쟁이 계속되고 있지만) 대체로 다음의 세 가지 조건들을 요청하고 있다.[4] 첫째, 행위 주체와 행위 결과 간에 인과 관계가 있어야 한다. 행위 주체가 사건의 결과를 어느 정도 인과적으로 통제할 수 있다면 그 행위자에게 대개 책임이 부과된다. 둘째, 행위 주체는 자신의 행동을 알고 그 행동이 가져올 가능한 결과들도 예견할 수 있어야 한다. 자신의 행동이 유해한 사건으로 이어질 것임을 알지 못했다면, 우리는 그 행위 주체에게 사건에 책임이 있다고 말하기 어렵다. 셋째, 행위 주체는 자신의 행동을 자유롭게 선택할 수 있어야 한다. 행위 주체의 행동이 순전히 외재적인 요인에 의해 결정된 것이라면, 그 행위 주체에게 사건의 책임을 묻는 것은 무의미하다.

그런데 인공지능 시스템의 활용은 이러한 조건을 더욱 복잡하게 만들고, 누구에게 책임을 부과할지의 문제를 매우 어렵게 만든다. 인공지능 시스템의 작동에는 실제로 많은 기술적 요소들, 가령 빅데이터, 클라우드 컴퓨팅 환경, 사물인터넷 시스템, 정보 수집 센서 등이 개입한다. 그만큼 책임을 부과하고자 할 때, 소위 '많은 손'의 문제가 발생하게 된다.[5] 같은 맥락에서 인공지능 시스템을 활용하는 과정에서도 사용자 외에 설계자와 제작자들, 가령 빅데이터 공급자, 클라우드 컴퓨팅 환경 설계 및 운영자, 사물인터넷 제작자, 정보 수집 센서 제작자 등 수많은 행위자들이 관여하게 된다. 이 행위자들은 서로 다른 방식으로 다양하게 관여하는 만큼 사건에 관한 질문에 대답하고 그 결과를 책임질 특정한 행위자를 한정하기란 매우 어렵다. 소위 '분산된(distributed) 책임의 문제'가 발생한다.[6] 또한 설계자나 제작자 그리고 사용자가 예

4) Eshleman (2014), pp. 216-40; Jonas (1984), p. 98, p. 156.
5) Friedman (1990). pp. 1-10; Nissenbaum (1994), pp. 72-80.

측하지 못한 결과들이 나올 수 있고, 사고가 발생했을 때 이에 대해 충분히 설명하고 해명할 수 없는 부분들도 충분히 나타날 수 있다. 한마디로 개개 행위자의 행동을 결과로 나타난 사건과 인과적 고리로 명확하게 연결하기가 쉽지 않고, 개개 행위자가 자신의 행동이 가져올 가능한 결과들을 예견하는 것도 어렵다. 이러한 맥락에서 철학자 마티아스(Matthias)는 인공지능 기술이 점점 더 복잡해지고 인간이 이런 기술의 행동을 직접 통제하거나 개입할 여지가 적어질수록 인간이 이 기술에 대해 전적으로 책임을 져야 한다고 주장할 여지도 적어져, 만약 책임의 문제를 인간에 한정해 언급한다면 오히려 '책임 공백(responsibility gap)'의 문제가 발생할 수 있음을 지적하고 있다.[7)]

두 번째 배경은 인간처럼 자유의지나 자의식에 기반한 자율성을 갖고 있지 않은 인공지능 시스템이라 하더라도 사건에 대해 일정 정도의 책임을 묻도록 만드는, 인공지능 시스템의 능력과 역할과 관련이 있다. 인공지능 시스템이라는 블랙박스에 무엇이 들어가고 나오는지는 설계자·제작자·사용자가 (다소간) 알 수 있지만, 블랙박스가 이러한 입력을 출력으로 어떻게 도출해 내는지, 왜 그런 특정한 결과에 도달했는지는 정확히 알지 못한다. 또한 그 결과는 인공지능 시스템 자체가 자기 주도적인 심화학습을 통해 만들어낸 만큼, 인공지능 시스템 자체에 (인간과는 다른 의미의) 어느 정도의 선택의 '자율성'이 있다고 말할 수 있다.[8)] 그런 의미에서 인간과 인공지능 시스템 모두 넓은 의미에서

---

6) 인공지능 시스템과 관련하여 이 문제는 플로리디에 의해 본격적으로 제기되었다. 플로리디는 이러한 복잡한 관계망을 전제로 분산된 도덕적 행동이라는 새로운 개념을 제시하였다. 도덕적으로 중요한 결과는 일부 개인의 도덕적으로 중대한 행동으로 환원될 수 없고 여러 행위자들로 분산될 수밖에 없다는 것이다. 이 경우 분산된 도덕적 행위에 대해 도덕적 책임을 누구에게 얼마만큼 할당할 것인가의 문제가 남아 있지만, 도덕적 책임을 분산된 행위자 모두에 대해 귀속시켜야 한다는 주장은 인공지능 시대에 매우 의미 있는 지적으로 볼 수 있다(Floridi 2013, pp. 727-43; Floridi 2016, pp. 2-3).

7) Matthias (2004), pp. 175-83.

8) 알파고를 예로 생각해 보자. 알파고는 인간 두뇌의 신경망 안에서 일어나

고 있는 학습 과정을 특정한 알고리즘 형태로 모방한 심화학습(deep-learning) 프로그램을 통해 자발적으로 자기 주도적으로 학습할 수 있다. 이를 통해 (특히 다양한 기보학습을 통해) 기존의 문제해결 방법을 익히더라도, 이와 동일한 방식이 아니라 새로운 문제 해결 방법을 찾아 문제를 해결할 수 있다. 또한 어떤 경우(알파고 제로 버전)는 아예 기보학습 없이 자기 방식대로 문제를 해결하기도 한다. 그런데 이러한 해결 과정은 설계자나 제작자도 알 수 없는 인간의 통제로부터 벗어난 블랙박스와도 같은 과정으로서, 알파고 스스로의 판단과 결정에 의해 이루어진다. 이러한 의미에서 적어도 심화학습 프로그램을 통해 자기 주도적으로 학습하는 인공지능 시스템에 대해 '자율성'을 부여할 수 있다. 하지만 인공지능 시스템에 부여된 '자율성'은 인간이 자유의지 혹은 자의식에 바탕 하여 스스로 선택하고 결정하는 인간 중심의 자율성 개념과는 분명 다르다. 두 가지 측면에서 구분해 볼 수 있다.

우선 첫 번째는 외연 상 드러나는 기능(혹은 역량) 측면에서 수준과 정도에 따라 구분해 볼 수 있다. 알파고와 같은 약-인공지능에게 부여할 수 있는 자율성은 준-자율성(semi-autonomy)이다. 어린 아이나 영장류 동물에서처럼 외부 환경 정보에 대한 기초 판단과 그에 따른 반응 행동 선택과 같은 매우 기본적인 의사결정 구조를 지닌 자율성에 불과하다. 이러한 기본적인 수준의 의사결정 구조는 오늘날 (여러 가지 다양한 선택지를 제공하는) 통계적 알고리즘에 기반하고 있는 심화학습을 통해 어느 정도 구현될 수 있다. 외부세계에 대한 개념적 이해와 의미 분석은 어렵지만, 동일한 패턴 인식과 그에 따른 유형 분류 및 선택이 어느 정도 가능하기 때문이다. 이에 반해 성숙한 인간에게 부여되는 자율성은 완전한 자율성(fully-autonomy)으로, 자유의지에 따라 자신의 사고 및 판단과 행동을 결정하는 자율성이다. 여기서 사고 및 판단과 행동은 외부세계에 대한 통계적인 정보처리가 아니라 세계에 대한 이해 특히 개념 분석과 의미 이해에 바탕 한 것이라는 점에서, 현재와 같은 인공지능 시스템에 이러한 자율성을 부여하기는 어려워 보인다.

보다 중요한 구분은 다음의 두 번째 측면이다. 전통 도덕 철학에서 자율성은 인간의 자아 혹은 자유의지라는 인간 개인의 고유한 내재적 속성에 기반을 둔다. 다분히 인간 중심적인 개념이다. 반면 인공지능 시스템의 경우 자율적인 판단과 행동은 지금까지 인간의 수많은 행위들과 사회적 관계들에 관한 빅데이터 분석에 기반하게 된다. 여기서 인간들 사이에 성립하

자율적 행위자라고 말할 수 있다.[9] 한편 인공지능 시스템은 설계자·제작자와 사용자 사이에서 중간 매개자 역할을 수행한다. 그로 인해 인공지능 시스템에는 당연히 설계자·제작자의 의도가 들어가지만 그것의 활용 과정에서 누가, 왜, 어떤 목적으로 사용하는가라는 사용자의 맥락에 따라 설계자·제작자가 의도하지 않은 결과들이 언제라도 발생할 수 있다.[10] 따라서 사고가 발생했을 시 어느 부분에서 무엇 때문에

는 다양한 사회적 관계들의 경우, 인공지능 알고리즘은 비록 그것들에 대한 의미 이해가 충분치 않더라도 사회적 관계들에 관한 패턴 분석을 통해 통계적인 방식으로 관계의 특성을 추론해 낼 수 있다. 다시 말해 인공지능 시스템은 (현재는 매우 부정적인) 인공지능에 내재하는 자율성보다는, 인공지능 시스템이 인간 사회와 맺는 사회적 관계 또는 인간들 사이의 사회적 관계에 대한 패턴 인식을 통해 자율적으로 판단하고 행동한다고 말할 수 있다. 그런 의미에서 인공지능 시스템은 '관계적 자율성'을 갖고 있다고 말할 수 있다. 또는 인공지능 시스템을 인간과 복잡하게 얽혀 있는 사회적 관계망 속에서 새로이 구성된 하나의 자율적 행위자로 볼 수 있다. (이상의 논의들에 관한 자세한 내용은 다음의 논문을 참조할 것. 이중원 2018, pp. 130-135.) 전통 도덕철학에서 강조해 온 자율성 개념은 지금으로서는 인간에게만 한정될 가능성이 높은 반면, 앞서 언급한 '관계적 자율성' 개념의 경우 인공지능 시스템에도 적용해 볼 여지가 있다. 본 논문은 이러한 관점에서 인공지능 시스템의 행위에 대해서도 책임의 문제가 충분히 던져질 수 있음을 받아들이고, 그렇다면 어떤 책임이 인간이 아닌 인공지능 시스템에 부과될 수 있는가에 대해 논하고자 한다.

9) 하지만 두 자율적 행위자는 적용되는 자율성 개념이 다른 만큼 분명 서로 다르다. 이를 명확히 하기 위해선, 인공지능 시스템에 적용되는 '관계적 자율성' 개념이 구체적으로 무엇이며, 인간의 자율성 개념과 어떻게 다른지에 대해 인공지능 기술의 발전과 함께 더 많은 심화연구가 필요하다. 더불어 다른 존재자들은 가지고 있지 않고 인간만이 가지고 있는 자율성이란 어떤 능력인지에 대해서도 향후 보다 세밀한 연구가 필요하다.

10) 이와 관련하여 기술철학자 바이커(W. Bijker)는 고도의 기술적 인공물들은 설계자의 의도와 달리 사용자의 맥락에 따라 다른 방식으로 사용될 수 있는 해석적 유연성을 지니고 있음을 강조한다(Bijker, Hughes, and Pinch 1987, p. 13, p. 27, p. 29).

어떤 문제가 발생했는지를 결정하기가 쉽지 않다. 사고의 책임을 인간 행위자에게만 전적으로 묻기란 쉽지 않다. 이것이 두 번째 배경이다.

가령 가까운 미래에 구현될 자율주행 자동차를 생각해 보자. 만약 자율주행 자동차가 보행자를 치었을 경우, 누구에게 어떤 도덕적·법적 책임을 물을 것인가? 자동차의 기계적인 시스템 제작자, 자동차가 자율적으로 경로를 결정할 수 있게 해주는 인공지능 알고리즘 설계자, 자동차의 인공지능 알고리즘에 제공되는 입력 정보(도로교통 관련 법규 정보들, 도로 환경에 맞춘 운전 패턴 정보들, 실시간 교통 상황 정보 등 교통 빅데이터) 수집 및 제공자, 인공지능 알고리즘과 교통 빅데이터를 활용하여 운행 방식을 스스로 결정하는 자동차의 의사결정 시스템, 빅데이터 정보 수집 및 전달에 관여하는 하드웨어 장치(빅데이터 수집 센서들, 클라우드 컴퓨팅 환경과 관련 통신 장치들) 제작자, 도로에서 자율주행 자동차의 운행을 허가한 정부의 교통정책 입안자, 자동차의 의사결정 시스템을 자신의 취향에 맞게 개인화한 소유자 등이 모두 사고의 책임 당사자가 될 수 있다. 여기서 만약 우리가 고려할 수 있는 모든 인간 행위자들에 대해 더 이상 사고의 책임을 물을 수 없는 경우, 자율적 인공지능에 기반한 자동차 의사결정 시스템에 궁극적으로 책임을 물어야 할 상황이 발생할 수 있다.

정리하면 다음과 같은 일반적인 상황 혹은 조건들이 발생하는 경우라면 인공지능 시스템에 대해서도 책임 소재의 문제가 발생할 수 있을 것이다. 우선 인공지능 시스템의 의사결정으로 인해 고의든 혹은 의도하지 않았든 부작용 혹은 부정적인 결과가 초래되는 경우이고, 다음은 인공지능 시스템의 의사결정 과정에 참여한 다른 인간 행위자들의 협의가 불충분하거나 불분명한 경우이며, 마지막으로 인공지능 시스템의 의사결정 과정이 블랙박스처럼 불투명하여 잘 설명되지도 않는 등, 인간에 의한 통제가 어려운 경우이다. 앞서 자율주행 자동차의 사례에서 보았듯이, 이러한 상황들은 가까운 미래에 충분히 발생가능하다. 문제는 어떤 책임을 어떻게 물을 것인가이다.

이 글에서는 바로 이 문제를 다루고자 한다. 우선 책임 개념에 관한 전통적인 도덕철학에서의 핵심 관점을 간략히 정리하고, 인간에게 배

타적으로 적용되어 온 이 개념을 뛰어넘어 인간이 아닌 다른 자율적인 행위자에게도 확대 적용될 수 있는 책임 개념의 가능성을, 전통적인 책임 개념에 비판적인 레비나스(E. Levinas)의 책임 개념을 중심으로 검토해볼 것이다. 다음으로 책임 개념의 외연이 설령 확장 가능하여 인간이 아닌 행위자에게까지 적용될 수 있다 할지라도, 현재나 가까운 미래의 인공지능 시스템에게 이를 적용하는 것은 쉽지 않음을 밝히고, 인공지능 시스템에 대해 책임 대신 책무(accountability) 개념의 적용을 제안할 것이다. 나아가 이러한 책무성이 인공지능 시스템에서 실질적으로 구현가능한지, 설명 가능한(explainable) 인공지능 알고리즘을 대상으로 분석할 것이다. 마지막으로 인공지능 시스템에 대해 책무성 중심의 윤리 체계를 구축하는데 필요한 윤리 프레임의 기본 요소들을 제안하는 수준에서, 그러한 윤리 체계의 가능성을 조심스럽게 전망해 보고자 한다.

2. 책임 개념의 확장 가능성 - 레비나스의 책임 개념

앞서도 강조하였듯이 책임에 관한 고전적 모델에 따르면, 책임은 인간만이 질 수 있는데 인간만이 자유의지를 가지고 이에 의거하여 행동하며 책임은 오직 이러한 행위와만 관계하기 때문에 그렇다고 본다.[11] 이런 입장이라면 어느 정도 자율적으로 판단하고 행동하는 인공지능 시스템이라 할지라도 인간에게 주어진 자유의지를 갖고 있지 않기에 책임을 질 수 없을 뿐 아니라, 자유의지가 약한 어린아이나 자유의지가 없는 의식불명의 사람에게조차 법적으로나 도덕적으로 책임을 물을 수 없게 된다. 어떤 사건 발생의 물리적 원인을 제공한 존재자라 하더라도 자유의지가 약하거나 없는 한, 그 사건에 책임을 지는 주체가 될 수는 없다는 것이다. 가령 칸트는 자유를 토대로 도덕적 책임을 정당화함으로써, 책임과 자유가 불가분의 관계에 있음을 강조하였다.[12] 한

11) 각주 2) 참조.
12) 각주 2) 참조.

마디로 자유(의지의 자유, 행위의 자유 모두)는 책임을 묻기 위한 전제조건으로서, 인간에게 자유는 도덕적 책임을 묻기 위해 반드시 요청되는 규범적인 것이 된다.[13]

하지만 현대에 오면 행위 주체의 자유(의지)가 책임의 조건이라는 전통적인 관점에 도전하는 입장이 나타난다. 바로 레비나스의 타자윤리에서 강조되고 있는 책임 개념이다. 레비나스는 타자에 대한 윤리적 책임을 강조하는데, 이때 책임은 자유에 앞서 근본적으로 부과된, 인간이 존재함과 동시에 주어진 것이다. 책임이 지향하는 곳이 바로 타자이기에, 책임은 타자를 향하는 '타자윤리'의 토대가 된다. 레비나스는 타자의 시선에서 세계를 바라보면서 타자와의 관계, 타자에 대한 책임을 철학의 중심 사유로 놓았고, 책임을 타자의 부름에 응답하는 것으로 보았다. 타자를 수용하고 타자와 함께 할 때 진정한 주체가 된다고 본 것이다.[14] 레비나스가 강조한 책임은 다음과 같은 특징-타율성, 대속성, 비대칭성-을 지닌 것으로 분석되곤 한다.[15] 먼저 그에게서 책임은 주체의 자유에 의존하지 않고 자유에 앞서며, 타인을 향해 있기에 타율적이라고 말할 수 있다. 나의 자유, 나의 의식, 나의 자발성이 책임을 불러일으키는 것이 아니라, 나에 대한 타자의 시선이 그에 대한 나의 책임을 불러일으킨다고 본 것이다. 그런 의미에서 책임은 대속적이다. 마치 아이에 대한 부모의 어쩔 수 없는 책임처럼, 책임은 자율적이고 능동적이지 않고 타율적이고 수동적이다. 마지막으로 책임은 비대칭적인데, 타자와 나는 기본적으로 동등하지 않고 고통 받는 타자의 존재가 나의 윤리적 각성의 근원이기에 이러한 비대칭성이 진정한 평

13) 일반적으로 자유를 정의할 때, 자주 '적극적인' 의미의 자유와 '소극적인' 의미의 자유를 구분한다. 소극적인 자유란 억압과 강제가 없는 상태로, 무엇으로부터의 자유를 뜻한다. 반면 적극적인 자유란 스스로 자기규정을 하는 상태로, 무엇에로의 자유를 의미한다. 적극적인 의미에서 자유는 또 다시 그 실천영역이 내적인가 외적인가에 따라 '의지의 자유'와 '행위의 자유'로 나뉜다. 의지의 자유는 '무엇을 원할 수 있는가?'라는 물음과 관계하며, 행위의 자유는 '무엇을 행할 수 있는가?'를 묻는다.

14) Lingis (1969), p. 43, pp. 199-200; Hand (1989), pp. 75-87.

15) 이유택 (2008), pp. 63-94.

등을 가능하게 하는 조건으로 본 것이다.

이처럼 자유(의지)가 책임의 전제조건이 아닌 경우, 자유의지나 자율성과 관련하여 여전히 논란이 많은 인공지능 시스템에 대해 책임을 논하는 것이 훨씬 자연스러워질 수는 있다. 그렇다면 레비나스의 책임 개념을 인공지능 시스템의 책임 문제에 확대·적용해 볼 수 있을까? 논쟁적인 부분은 레비나스의 주체 범주에 인간이 아닌 인공지능 시스템을 포함시킬 수 있는가이다. 앞서 분석한 레비나스 책임 개념의 타율성과 대속성이라는 특징에 한정해서 본다면, (인간을 주체로 본 레비나스 자신의 의도와 상관없이) 자율적인 행위자로서의 인공지능 시스템 역시 레비나스적인 책임의 주체 범주에 포함될 수 있을 것이다. 가령 노인이나 환자를 보살피는 건강 돌봄 로봇의 경우, 아이를 돌보아야 하는 부모처럼 고통 받는 타자인 노인과 환자를 향해 그들에게 필요한 도움을 제공하는 방식으로 그들의 부름에 응답한다면, 그 인공지능 로봇은 앞선 두 가지 특징에 국한해서 볼 때 레비나스적인 의미에서 노인과 환자에 대한 윤리적 책임을 다하는 주체로 볼 수 있을 것이다. 건강 돌봄 로봇이 아닌, 가사 도우미 로봇, 섹스 로봇 등에 대해서도 마찬가지다. 하지만 레비나스 책임 개념의 세 번째 특징을 인공지능 시스템에 적용하는 것은 무리가 있을 뿐 아니라, 적절하지 않아 보인다. 인공지능 시스템이 고통 받는 타자의 존재를 윤리적 각성의 근원으로 스스로 판단할 수는 없기 때문이다.[16] 이러한 윤리적 각성이 (인간이 제시한) 윤리-빅데이터를 활용한 지도학습을 통해 인공지능 시스템에 강화될 수는 있겠지만, 그럴 경우 이는 인간에 의한 것일 뿐 스스로의 자각에 의한 것으로 보기는 어렵다.[17] 물론 인공지능 기술이

16) 레비나스가 자유(의지)를 도덕적 책임의 전제조건으로 삼지 않은 이유는, 자유의지가 중요하지 않아서가 아니라 자유의지에 따른 인간의 자율적인 선택 이전에 고통받는 타자에 대해선 의무적인 차원에서 도덕적 책임을 다해야 한다는 윤리적 의식이 더 중요하다고 보았기 때문이다. 윤리적 각성이 인간의 더 근원적인 속성일 수 있음을 암시하고 있다고 볼 수 있다.

17) 이는 마치 어린 아이가 생활 속에서 다양한 직접 경험을 통해 윤리적인 의식을 배우고 습득하여 각성해 가듯이, 인공지능 시스템에서도 그와 유사한

고도로 발달하여 빅데이터를 활용한 인간의 지도학습 없이도 스스로 다양한 상황을 인지하고 판단하는 비지도 학습이 원활히 가능해진다면, 인간적인 의미의 자각은 아닐지라도 그와 유사한 형태로 인공지능 시스템 자체의 윤리적 내재화 작업은 가능할 지도 모른다. 정리하면 레비나스의 책임 개념은 인공지능 기술이 고도로 발전한 먼 미래에 인공지능 시스템에 적용해 볼 수 있는 확장된 의미의 책임 개념으로 기대할 수는 있겠지만, 현재 또는 가까운 미래의 인공지능 시스템에 적용하는 것은 어렵다고 할 수 있다.

3. 인공지능 시스템, 책임에서 책무로

앞서 언급하였듯이 도덕철학에서 전통적인 책임 개념은 인간을 대상으로 한다. 또한 개인 또는 집단이 다른 사람에 대해 도덕적이며 윤리적인 규범과 기준 및 전통에 따라 도덕적 의무를 가지고 있다는 사실을 나타내는 윤리적 개념이다. 일종의 의무의 묶음이라고 말할 수 있다. 이러한 책임 개념은 (여전히 논란이 있지만) 앞서 언급한 대로 대체로 다음의 조건들이 모두 충족된다면 적용될 수 있다. 행위자와 행위 결과 간의 인과적 연결, 행위 자체에 대한 행위자의 인지와 그 결과에 대한 어느 정도의 예견, 행동에 대한 행위자의 자발적이고 자유로운 선택이 바로 그 조건들이다. 여기서 첫 번째와 두 번째 조건은 인간이 아닌 인공지능 시스템에 대해 하나의 행위자로 보고 확대·적용해 볼 수 있겠지만, 세 번째 조건은 (자율성 혹은 자유의지에 관한 많은 논란으로 인해) 아직까지는 통상적으로 인공지능 시스템에까지 적용하기란 쉽지 않아 보인다. 그런 의미에서 현 단계 혹은 가까운 미래의 인공지능 시스템을 독립적인 행위자로 볼지라도, 이에 대해 자유(의지)에 기반 한 전통적인 책임 개념을 적용하는 것은 어려워 보인다.[18] 그렇다

행태가 시뮬레이션을 통해 가능할 수 있음을 의미할 뿐, 인공지능 시스템이 인간처럼 윤리적인 자각 능력을 내재적 속성으로 갖고 있음을 의미하는 것은 아니다.

면 자율주행 자동차처럼 나름대로 독자적인 판단과 선택적 행동을 하는 인공지능 시스템에서 어떤 오류로 사고가 발생하는 경우, 우리는 그 사고와 관련하여 인공지능 시스템에게 무엇을 요구할 것인가?

아무래도 사고에 대한 합당한 설명(reasonable explanation)을 일차적으로 요구할 것이다. 이 설명에의 요구는 인공지능 시스템이 의사결정 과정에서 어디서 왜 그런 오류가 발생했는가를 스스로 해명할 수 있어야 한다는 요구로서, 이는 인공지능 시스템의 투명성 및 인간에 의한 통제가능성이라는 측면에서 매우 중요하다. 이런 설명에의 요구 혹은 '설명 가능성(explainability)'의 요청은 책임 개념이 성립하기 위한 첫 번째 및 두 번째 조건들과도 일맥상통한다. 어떤 사고에 대한 책임을 묻기 위해선, 일차적으로 사고가 어떻게 그리고 왜 발생했는지에 대한 설명이 반드시 필요하기 때문이다. 그렇다면 좀 더 구체적으로 인공지능 시스템에 어떤 설명을 요구할 것인가? 적어도 다음의 질문들에 대한 답이 합당하게 제시돼야 한다. 첫째 의사결정 과정에 어떤 요소들이 중대한 역할을 했는가, 둘째 특정 요소들이 최종 결과에 어떻게 영향을 미쳤는가, 셋째 최종적으로 산출된 결과는 실제로 의미있게 적용가능한가이다.[19] 첫 번째 경우에는 신뢰할 수 없거나 부적절한 요소(입력 정보 등)의 개입이 오류를 낳을 것이고, 두 번째 경우에는 입력 요소와 최종 결과를 잇는 의사결정 시스템 자체에 비일관성이 있을 때 오류가 발생할 것이며, 마지막 경우에는 최종 산출 결과에 대

18) 물론 향후 완전한 자율주행 자동차가 등장한다면 이것의 자율성을 어떻게 볼 것인가에 따라 논의가 완전히 달라질 수 있다. 자율성 개념을 인간의 그것과 근본적으로 다른 것으로 보지 않고 정도의 차이를 인정하는 수준에서 받아들인다면, 이 경우 자율성은 (인간의 자유의지와 동일할 수는 없고 그와 유사한) 어떤 '자유의지'를 향한 기반 밑거름이 될 수 있을 것으로 기대해 볼 수 있다.

19) 일반적으로 인공지능 시스템에 설명을 요구한다는 것은, 인공지능 시스템 안에서의 디지털 신호인 비트의 흐름을 밝혀달라는 것과 근본적으로 다르다. 인공지능 시스템이 최종 결론에 어떻게 도달하는지에 관한 기술적 세부사항을 알고자 하는 것이 아니라, 어떤 요인들이 특정 상황에서 최종 결론 산출에 어떻게 작용하는지에 대한 대답을 듣고자 하는 것이다.

한 불신이 커 수용할 수 없을 때 역시 오류가 발생할 것이다.

이처럼 인공지능 시스템의 활용 과정에서 사고가 발생했을 때 이에 대한 합당한 설명을 요구하는 경우, 우리는 인공지능 시스템에 논란이 많은 책임(responsibility) 개념 대신에 설명에의 의무에 바탕 한 책무(accountability) 개념을 (현 단계에서) 적용해 볼 수 있을 것이다. 여기서 책무는 주로 자기 자신의 행동을 설명할 수 있는 능력에 기반하고 있기에, 인공지능 시스템에 대해 의사결정 과정을 설명하고 오류 또는 예기치 않은 결과를 식별할 수 있는 능력을 바탕으로 책무를 논할 수 있다. 책임과 책무는 다음과 같은 측면에서 서로 다르게 구분해 볼 수 있다. 우선 책무 개념은 책임의 중요한 요건 가운데 하나인 자의식 혹은 자유의 문제로부터 일단 자유로울 수 있다. 책임에 대한 내면적인 자각이나 의식이 없더라도, 행위 주체에게 의무들의 묶음으로서의 책무를 충분히 부과할 수 있기 때문이다.[20] 다음으로 책무는 행위자보다는 행위 그 자체에 관심을 두는 반면, 책임은 궁극적으로 행위를 수행한 주체인 행위자에 초점을 둔다고 볼 수 있다.[21] 이에 근거해서 책임 개념은 인간에게, 책무 개념은 인간 이외의 행위자들에게도 적용할 수 있는 개념으로 구분해 볼 수 있을 것이다. 그런 맥락에서 여기서는 정치 윤리학자인 더브닉(Melvin J. Dubnick)의 논의를 좇아, 책임 개념과 구분되는 책무 개념을 인공지능 시스템에 적용해보고자 한다.

더브닉(M. J. Dubnick)은 책무(accountability)에 4가지 유형이 있음을 강조하고 있다.[22] 첫째는 응답할 수 있음(answerability)으로서의 책

---

20) 가령 우리는 기업을 대상으로 기업의 사회적 책임을 강조하곤 하는데, 이때 책임은 기업이 기업으로서 사회적 역할을 충실히 다하라는 명령으로서 오히려 책무 개념에 더 가깝다고 할 수 있다.

21) 책무는 위임에 의한 전이가 가능하다. 인간 행위자가 의사결정 및 관련 업무 자체를 다른 행위자(가령 자율주행 자동차의 인공지능 시스템)에게 위임하여 이를 대신 수행토록 했다면, 인간의 책무도 다른 행위자에게 이전됐다고 말할 수 있다. 그런 맥락에서 사람으로부터 업무 등을 위임 받은 조직이나 시스템의 경우 그에 따른 책무가 매우 중요한 문제로 대두된다고 하겠다.

22) Dubnick (2003), pp. 410-25.

무다. 행위자의 행위는 행위자의 판단과 합리적으로 연결돼 있기 때문에, 행위자는 자신의 행위와 태도에 대해 합당한 응답을 해야 한다. 이는 행위자의 책무에서 가장 비중이 높은 부분이다. 둘째는 비난받을 만함(blameworthiness)으로서의 책무다. 이는 행위자 개인의 특별한 역할이나 행위와 관련된 것이 아니라, 행위자의 사회적 지위나 조직에서의 위치와 관련하여 지게 되는 책무다. 그러한 지위로 말미암아 비난을 받더라도 그로부터 부여받은 일을 할 수밖에 없는 사회적 관계가 중요하다. 셋째는 법적 의무(liability)로서의 책무다. 행위자가 법이나 사회적 규범에 따라 행동해야 한다는 의미의 책무다. 가령 어떤 권위 있는 기관(사법부, 경찰 등)으로부터 그의 행위에 대한 설명(경찰 조서 작성, 법적 증언 등)을 요청받은 경우, 사회 제도적인 차원에서 당연히 응답해야 함을 강조한다. 마지막 유형은 귀착가능성(attributability)으로서의 책무다. 가령 공직자나 공무원처럼, 어떤 행위자가 사회적인 계약에 따라 특정한 임무를 부여받은 경우 그 임무에 수반되는 규칙에 따라 업무를 수행해야 한다는 의미의 책무다. 첫 번째, 두 번째 책무가 행위자와 직접 관련이 있다면, 세 번째, 네 번째 책무는 행위자와 연관된 사건 혹은 상황과 관련이 깊다.

그렇다면 이러한 의미의 책무 개념을 또 다른 행위자로서 인공지능 시스템에 적합하게 적용할 수 있는가? 인공지능 시스템은 일종의 블랙박스이기 때문에, 알고리즘의 작동 결과로 특정한 사건이 발생할 경우 이에 대한 설명의 요구가 일차적으로 강하게 제기될 것이다.[23] 이러한

23) 구체적으로 다음과 같은 질문들이 주로 제기될 것이다. 블랙박스에 해당하는 알고리즘을 어떻게 신뢰할 수 있는가, 신뢰가 떨어진 알고리즘의 사용을 어떻게 거부할 것인가, 알고리즘이 개인 프라이버시를 어느 정도까지 침해하는가, 알고리즘이 피해를 입혔을 때 누구 혹은 무엇이 책임을 져야 하는가, 인공지능 시스템은 그들의 의사결정 과정을 어떻게 설명하는가, 인공지능 시스템에서 나타날 수 있는 편향성은 어떻게 제거될 수 있는가 등이 그것이다. 실제로 인공지능 알고리즘은 실생활에 다양하게 응용되면서, 피해를 주기도 하고 차별을 강화하기도 하는 등 부정적인 역할을 하기도 한다. 다시 말해 알고리즘이 해서는 안 되는 일에 참여할 가능성은 언제나 열려 있고(Algorithmic Harm), 편향된 데이터로 인해 성별, 인종, 민

상황에서 위에 언급한 4가지 유형의 책무를 인공지능 시스템에 적용하는 문제와 관련하여, 행위자의 행위가 준수해야 할 규칙이나 규범 또는 제도와 이것들을 준수하지 못함으로써 발생한 사고 과정에 대한 설명이 매우 중요함을 강조하고자 한다. 사례를 들어 보자.

우선 교통사고를 일으킨 자율주행 자동차의 경우를 생각해 보자. 하나의 자율적 행위자로서 자율주행 자동차는 당연히 사고에 대한 설명 요청이 있는 경우 이에 응답해야 할 책무를 갖는다. 나아가 이러한 설명에의 요구가 권위를 지닌 국가 기관이나 사회 제도 차원에서 요청된다면, 인공지능 시스템 역시 하나의 행위자로서 세 번째 언급한 법적 책무도 져야 할 것이다. 다음으로 소위 인공지능 판사라 불리는 로스(ROSS)와 인공지능 의사라 일컬어지는 왓슨(Watson)의 경우를 보자. 이들은 각기 특정한 임무를 수행하도록 설계·제작된 만큼, 귀착가능성으로서의 책무를 지닌다. 즉 부여된 임무를 충실히 수행해야 하는 만큼 임무 수행에 수반되는 규칙을 잘 지켜야할 책무가 있는 것이다. 또한 로스 프로그램이 재판 과정에 관여하여 사고가 발생했다면, 그리고 왓슨 프로그램이 환자의 질병 분석과 치료 과정에 관여해 사고가 났다면, 어떻게 사고가 났는지 법적 절차를 통해 설명해야 하는 법적인 책무 또한 가진다. 즉 국가 기관이나 법에 의해 설명 요청이 있는 경우 이에 응해야 한다. 인공지능 군사로봇의 경우도 이와 매우 유사하다. 또 다른 사례로 인공지능 섹스로봇과 인공지능 돌봄 로봇의 경우를 생각해 보자. 이들의 경우 비난받을 만함으로서의 책무를 지닌다. 그들이 제공하는 특정한 서비스로 말미암아 사용자와 맺게 되는 사회적 신분 관계 때문이다. 또한 이들은 각기 특수한 임무를 수행하도록 설계·제작된 만큼, 임무 수행에 수반되는 규칙을 잘 지켜야할 귀착가능성으로서의 책무도 갖는다. 만약 이들로 인해 사고가 발생한 경우, 어떻게 사고가 났는지 법적 절차를 통해 설명해야 하는 법적인 책무 또한 가진다. 한편 인공지능 시스템이 경찰을 도와 우범지역에서의 범인 색출이나 범죄 예방과 같은 껄끄러운 공적인 역할을 수행하는 경우, 이 인공

족적 또는 종교적 이유로 사람들을 차별화(Algorithmic Discrimination)하는데 악용될 수도 있다.

지능 시스템은 하나의 행위자로서 비난받을 만한 책무 뿐 아니라 귀착 가능성으로서의 책무를 가지고 있다고 말할 수 있을 것이다.

정리하면 수많은 인공지능 시스템들에 적용될 수 있는 책무 개념의 핵심은 사고가 발생했을 때 인공지능 시스템이 관련 규칙 또는 규범을 준수 하였는지를 포함하여 사고가 어떤 과정을 통해 발생하였는지를 설명하는 것이다. 이는 인공지능 시스템이 어떤 정보들에 근거해서 그러한 행위를 선택하게 되었는지 그 전개 과정을 단순히 기술하는 응답을 뛰어 넘어, 그러한 행위 전개 과정에서 주어진 임무에 충실하고자 관련 규칙이나 제도 또는 윤리 규범을 제대로 준수하였는지, 나아가 법적 규범이 있을 경우 이를 준수하였는지에 대한 설명을 포함한다. 결국 다양한 상황에서 설명을 기반으로 그 행위에 걸맞게 다양한 책무, 곧 사고와 관련한 단순한 응답에서부터 사회적인, 도덕적인, 법적인 차원의 책무까지 다양하게 부과할 수 있을 것이다.

4. ‘설명 가능한 인공지능 시스템’을 통한 책무성 구현

앞서 보았듯이 인공지능 시스템도 하나의 행위자로서 자신의 판단과 행위에 대해 모종의 책무를 가져야 하는데, 이 책무에서 가장 기본이 되는 부분이 바로 설명에 대한 요구에 응답하는 것이다. 따라서 만약 설명 가능한(explainable) 인공지능 시스템을 기술적으로 구현할 수 있다면, 이는 인간 행위자와 유사하게 인공지능 시스템에게 어떤 책무를 부여하는 것이고 사회적으로, 윤리적으로 그리고 법적으로 인간에게 피해를 주지 않는 인공지능 시스템을 만드는 것이 될 것이다.[24] 그렇

24) 이 같은 맥락에서 실제로 유럽연합(EU)의 개인정보 보호법(General Data Protection Regulation, 2016)은 인공지능 시스템의 의사결정 과정에 대한 설명을 요구하고 있다. 나아가 이러한 논의를 바탕으로 법률에 명시된, 설명이 필요한 다양한 상황들을 면밀히 고찰할 필요성과 ‘설명 가능한 인공지능 시스템’(explainable AI system)의 설계와 같은 새로운 공학적 도전의 중요성을 강조하고 있다. 미국의 경우도 마찬가지다. 이와 같은 연유로

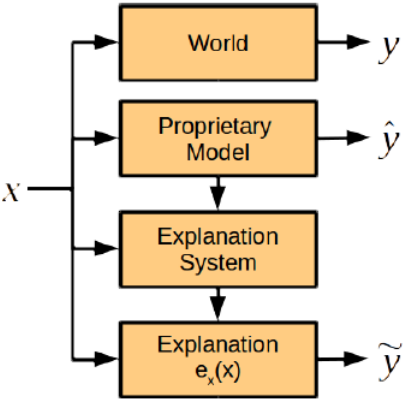
다면 '설명 가능한 인공지능 시스템'(explainable AI system)은 실제로 기술적으로 구현가능한가? 이를 논하기 위해서는 우선 '설명 가능함'이라는 조건이 내포하는 의미를 보다 구체적으로 밝히는 것이 중요하다.

첫째, 설명 가능함은 인공지능 시스템에서의 디지털 비트의 전반적 흐름을 파악하는 것을 목적으로 하지 않는다. 인공지능 시스템이 입력에서 출력까지 기술적 차원에서 어떻게 작동하는지에 관한 전반적인 세부사항이 아니라, 어떤 요인(혹은 요소)들이 특정 상황에서 어떻게 특정한 결과 산출에 작용하는지에 관심을 갖는다. 다시 말해 인공지능 시스템 작용 전반에 대한 설명이라기보다는, 특정의 결정 과정에 대한 국소적인 설명(local explanation)을 함축한다.[25] 둘째, 설명 가능함은 단지 입력 정보와 출력 결과 사이의 특정한 연관 관계를 밝혀 주는 것에 그치지 않고, 최종적인 출력 결과 및 관련 행위가 함축하는 사회적·도덕적·법적 함의들을 밝혀내고 이 함의들이 기존의 규범 틀 안에서 어떤 문제를 일으키는지 드러냄으로써 행위 결과의 오류와 문제점을 어느 정도 진단해 줄 수 있어야 함을 반영하고 있다. 이는 인공지능 시스템 안에서 이루어진 의사결정 과정이 단순히 비트의 흐름 차원에서 어떻게 기술되는가를 넘어 서서, 그러한 비트의 흐름이 의미론적 차원에서 인간에게 유의미하게 이해돼야 함을 전제한다. 따라서 이 경우 최종 결과 및 관련 행위가 사회적·도덕적·법적인 다양한 맥락 안에서 어떻게 그 의미가 해석되는지가 설명에서 매우 중요해 진다. 아래 그림은 설명 가능한 인공지능 시스템의 구조를 도식적으로 표현한 것

DARPA를 중심으로 설명 가능한 인공지능 시스템을 어떻게 설계할 것인지, 그 기술적 구현에 많은 관심을 기울이고 있다.

25) 이와 관련해서 인공지능 시스템에 대해 다음과 같은 질문들에 대한 합당한 설명이 필요하다. 가령 결정 과정에 어떤 요소들이 중요한 역할을 하였는가? 어떤 요소의 변화가 결정의 변화에 영향을 미쳤는가? 두 개의 유사한 사례들에 대해 왜 서로 다른 결정이 이루어졌는가? 등등. 여기서 어떤 요소가 결과를 결정하는데 관련되며 또 어떤 요소가 결과에서의 차이를 일으키는지를 파악하기 위해서는 인과적 설명 방식과 반사실적(counterfactual) 접근 방식을 도입하는 것이 매우 효과적이다.

이다.[26]



이 도식이 말해 주는 설명 가능한 인공지능 시스템의 특징은 다음과 같다. 첫째, 그 안에서 세계를 모델화하여 어떤 판단과 행동을 수행하는 인공지능 시스템과 이의 의사 결정 과정을 인간에게 유의미하게 해석해 주는 설명 시스템은 구분돼 있다. 다시 말해 인공지능 시스템에 요청된 설명의 요구는 이와는 별도의 설명 시스템에 의하여 이루어진다. 인공지능 시스템 자체는 일부 입력(x)을 받아 예측($\hat{y}$)을 이끌어내는 블랙박스이지만, 설명 시스템은 동일한 입력(x)에 대해 예측($\tilde{y}$)을 이끌어내고 이 과정을 해석할 수 있는 해석 규칙($e_x : x \rightarrow \tilde{y}$)을 갖고 있다.[27] 이를 통해 어떤 입력 x에 대해 설명 시스템에 의한 예측($\tilde{y}$)이 인공지능 시스템에 의한 예측($\hat{y}$)과 같은지 혹은 다른지를 확인함으로써, 인공지능 시스템 어디에 어떤 문제가 발생했는지를 설명할 수 있

26) Doshi-Velez, Kortz (2017), p. 8.

27) 가령 자율주행 자동차는 다수의 센서들을 통해 시각 입력 데이터를 고차원적인 수준까지 표상할 수 있지만, 인간의 두뇌는 이미 그러한 시각 입력 데이터를 나무나 거리 표지와 같은 상위의 개념으로 전환시킬 수 있다. 이 경우 자율주행 자동차가 인공지능 시스템이라면, 여기서 산출된 결과들을 개념으로 전환하여 그 의미까지 이해하는 인간의 두뇌는 일종의 설명 시스템이라고 비유적으로 말할 수 있다.

게 된다. 원리는 매우 간단하다.

둘째, 그런데 설명 시스템에서 이러한 설명이 가능하려면 인공지능 시스템의 디지털 입력 정보와 기호들을 인간이 해석할 수 있는 언어 또는 개념으로 변환하는 것이 필요하다. 이러한 변환은 결국 비트 정보 곧 디지털 기호들과 인간의 언어 간에 의미론적 연결을 어떻게 구성하는가, 달리 말해 해석규칙($e_x : x \rightarrow \bar{y}$)을 어떻게 설정하는가에 달려 있다. 가장 기초적인 형태는 해석규칙을 양자 간 대응규칙의 형태로 하향식으로 설정하는 것이다. 그러나 이는 어떤 기호 집합이 어떤 개념 정보에 대응하는가를 정의하는 과정 자체의 어려움과 복잡함, 그리고 이로 인해 수많은 기호 집합들이 개념 정보와의 대응관계 설정에서 사실상 배제됨으로써 발생하는 설명의 결함과 같은 문제들을 안고 있다. 보다 세련된 형태는 디지털 기호들의 집합이 개념적 의미를 획득할 수 있도록 개념 학습 알고리즘을 도입하여 해석규칙을 상향식으로 구축하는 것이다. 설명 시스템이 개념 학습 알고리즘을 통해 개념적 사고를 하고 개념의 의미를 파악할 수 있게 된다면, 비트들의 흐름에 대한 유의미한 설명이 충분히 가능하기 때문이다. 그런데 현재의 학습 알고리즘은 의미 자체를 아예 고려하지 못하는 패턴 인식 수준에 머물러 있기에, 현재적 차원에서 '설명 가능한 인공지능 시스템'을 기술적으로 구현하는 것은 쉽지 않다. 하지만 머지않은 미래에는 충분히 가능할 것으로 본다.

결국 설명 가능한 인공지능 시스템 설계에서 관건은 설명 시스템의 알고리즘이다. 만약 설명 시스템의 알고리즘이 인공지능 시스템이 출력한 최종 정보를 인간이 해석할 수 있는 언어 혹은 개념으로 번역할 수 있다면, 설명 가능한 인공지능 시스템은 구축 가능할 것이다. 하지만 이런 알고리즘이 구축된다 할지라도, 다음의 문제들은 여전히 남는다. 실제로 알고리즘은 성별, 인종, 민족, 종교 등을 이유로 사람들을 충분히 차별할 수 있다. 또한 알고리즘의 의사결정 과정은 이전 의사결정자의 편견을 이어받거나 사회에서 지속되는 광범위한 편향을 반영하여 기존 차별 패턴을 재현하고 강화할 수 있다. 이는 설명 가능한 인공지능 시스템을 통해 인공지능 시스템의 책무성을 구현하는 것이

중요한 의미가 있음에도 불구하고, 많은 어려움이 있음을 암시한다.

5. 맺음말 - 책무성 중심의 윤리 프레임을 위하여

지금까지 나는 인공지능 시스템에서 문제가 되는 사건이 발생할 경우 이의 의사결정 과정에 대한 설명을 수행해야 하는 책무를 인공지능 시스템에 부과할 수 있다고 주장하였다. 그리고 여기서 설명은 기술적인 차원에서 모든 비트의 흐름을 명백히 밝혀내는 것이 아니라, 문제가 된 특정한 결과를 산출하는데 어떤 입력 요소들이 작용했는지, 그러한 입력 요소들 가운데 결정적인 것은 무엇인지, 알고리즘은 신뢰할 수 있는지, 그리고 특정의 최종 결과가 사회적·윤리적·법적으로 어떤 함의를 지니는지 등을 밝혀내는 것을 의미한다고 보았다. 그런데 이처럼 설명이 가능한 부분들을 만약 설명 시스템의 알고리즘 안에 처음부터 반영하여 전체 의사결정 과정을 구성한다면, 인공지능 시스템은 문제의 사건이 발생할 때마다 자체적으로 오류가 어디에 있는지 밝혀낼 수 있게 될 것이다. 이런 방식으로 잘못된 의사결정 과정 자체를 알고리즘 상에서 사전에 예방할 수 있을 지도 모른다. 오늘날 책임질 수 있는 인공지능 시스템을 개발하려는 많은 개발자들이 구축을 서두르고 있는 소위 '설명 가능한 인공지능 시스템'은 바로 이를 지향하고 있다. 이는 자신이 지닌 행위 능력에 따라 스스로 행동을 조절하고 통제할 줄 아는 인공지능 시스템을 개발해야 한다는 말에 다름 아니다.

이러한 '설명 가능한 인공지능 시스템'의 등장은 인공지능 윤리와 관련하여 중요한 시사점을 우리에게 던져 준다. 첫째, 인공지능 시스템을 단순히 윤리적 사고와 판단의 대상이 아니라, 부분적이고 제한적인 의미겠지만 윤리적 행위의 주체로 간주할 수 있도록 한다는 점이다. 인공지능 시스템에서의 설명 가능성은 기본적인 응답의 책무를 포함하여, 다양한 상황에 따른 비난받을 만함으로서의 사회적 책무, 법적 책무, 귀속가능성으로서의 책무의 핵심 토대가 된다. 따라서 그 연장선상에 있는 도덕적·윤리적 책임에 대해서도 중요한 기반이 될 수 있고,

그러한 맥락에서 자유의지를 갖춘 인간에게만 부여되어 온 전통적인 의미의 도덕적·윤리적 책임의 주체는 아니지만, 책무의 담지자로서 넓은 의미의 책임의 주체로 볼 수 있다는 것이다.

둘째, 그럼에도 설명 가능성은 책무의 담지자인 인공지능 시스템이 넓은 의미의 책임의 주체가 되기 위한 필요조건이지 충분조건은 되지 못한다. 비록 설명을 통해 책임의 소재, 즉 귀책사유가 밝혀졌다 하더라도, 그 결과에 대해 누가 어떻게 책임질 것인가의 문제는 여전히 남아 있기 때문이다. 그런데 (어떤 의미에서건) 책임의 주체가 되기 위해서는 이 문제가 매우 중요하다. 예를 들어 교통사고를 낸 자율주행차의 경우를 생각해 보자. 자율주행차가 책무 차원에서 설명 알고리즘을 통해 사고가 어떤 경위를 통해 왜 발생하였는지를 소상히 밝혔다 해도, 그 결과에 대한 책임을 자율주행차가 질 것인지 아니면 자율주행차의 설계자 혹은 제작자 혹은 사용자인 인간이 져야 할 것인지의 문제는 남는다. 만약 자율주행자의 설계 혹은 제작 혹은 사용 과정에서 인간의 실수나 오류로 사고가 발생했다면, 책임의 주체는 당연히 인간이 될 것이고 자율주행차는 책무의 담지자로서 자신의 역할을 충실히 했다고 말할 수 있을 것이다. 그런데 만약 사고의 원인이 아무리 규명해도 인간의 오류나 실수에 의한 것으로 명확하게 밝혀지지 않는다면, 우리는 책임의 주체를 인간이 아닌 인공지능 시스템으로 볼 수밖에 없는 매우 당황스러운 상황에 처하게 된다. 이럴 경우 책임의 주체를 인간에게만 한정한다면 앞서 언급한 책임 공백의 문제가 발생할 것이고, 인간이 아닌 인공지능 시스템에게 까지 책임의 주체 범위를 확대한다면 책임 개념의 의미가 달라져야 할 것이다. 후자의 논의는 미래의 인공지능 윤리와 관련해서 매우 중요한 문제로, 이에 대해 지금까지의 논의를 바탕으로 다음과 같은 주장을 조심스럽게 시도해 볼 수 있을 것이다. 이 경우 인공지능 시스템에 대해 자유의지를 갖춘 인간에게만 부여되어 온 전통적인 의미의 책임 개념 대신, 앞서 분석한 레비나스의 책임 개념을 타율성과 대속성을 중심으로 완화시켜 적용해 보는 것이다. 앞서 분석하였듯이 책무성 개념은 인간만을 염두에 둔 자유의지를 굳이 상정할 필요가 없다는 점에서 이를 전제 조건으로 제시하지

않고 있는 레비나스의 책임 개념과 어느 정도 맞닿아 있다고 할 수 있다. 이것이 책무성 중심의 인공지능 윤리의 기본 아이디어다.

하지만 인공지능 시스템이 아직은 고도로 발전하지 못하고 있고, (레비나스 책임 개념의 완화와 같은) 도덕적 책임 개념의 확장을 위한 철학적 논의 또한 제대로 이루어지지 못한 상황에서, 책임 개념을 인공지능 시스템에 적용하기란 당분간 어려워 보인다. 전통 윤리학의 관점이나 현재의 기술적 상황 등을 종합해 본다면, 책임은 여전히 인간에게 부여하는 것이 타당해 보인다. 그런 맥락에서 인공지능 시스템의 행위에 대한 윤리적 판단은 당분간은 그것의 책무성의 범위 안에서 이루어지는 것이 적절하다고 할 수 있다. 앞으로 인공지능 시스템이 고도로 발전하여 인간과 거의 유사하게 자율적으로 생각하고 행동하는 상황이 된다면, 그리고 도덕적 책임 개념 또한 지속적인 철학적 성찰을 통해 보다 그 의미가 완화되고 확장될 수 있다면, 향후 인공지능 시스템에 대해서 도덕적 책임의 문제를 책무의 연장선상에서 거론해 볼 수 있을 것으로 기대해 볼 수 있다.

이제 책무성 중심의 인공지능 윤리가 성립하려면 어떤 요소들이 필요하고 어떤 조건들이 충족돼야 하는지 그 프레임의 구축 방향을 언급하는 선에서 논의를 마무리하고자 한다. 우선 인공지능 시스템의 설명과 관련하여 어떤 부분에 대한 설명이 책무와 관련하여 필요하고 중요한지 관련 상세항목들을 적시하고 이를 바탕으로 이 설명 메커니즘을 보편적인 알고리즘 형태로 구현하는 것이다. 다음으로 이러한 설명 알고리즘의 토대 위에서, 해당 인공지능 시스템의 행위 결과를 놓고 실제적으로 사회적 차원, 법적 차원 그리고 윤리적 차원에서 설명 요청이 있을 경우, 인간이 그 의미를 이해할 수 있는 방식으로 이에 대한 설명을 제공하는 것이다. 마지막으로 행위 결과에 대한 사회적·법적·윤리적 차원에서의 세분화된 인공지능 시스템 자신의 설명과 인공지능 시스템도 지켜야 할 인간의 사회적·법적·윤리적 차원에서의 의무 준칙들과 대조하는 방식으로, 인공지능 시스템의 행위를 윤리적으로 평가하는 것이다. 이는 책무성 중심으로 인공지능의 윤리 프레임을 구축하는데 중요한 기반이 될 것이다.

참고문헌

아리스토텔레스 (1984),『니코마코스 윤리학』, 최명관 옮김, 서광사.

이유택 (2008),「책임에 관한 철학적 성찰-레비나스와 요나스를 중심으로」,『현대유럽철학연구』 제17권, pp. 63-94.

이중원 (2018),「인공지능과 관계적 자율성」,『인공지능의 존재론』 이중원 외, 한울아카데미, pp. 117-136.

Bijker, W., Hughes, T., and Pinch, T (1987) (eds), *The Social Construction of Technological Systems*, Cambridge, MA: The MIT Press.

Doshi-Velez F., Kortz M. (2017), "Accountability of AI Under the Law: The Role of Explanation", Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper, http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584 (검색일: 2019. 07. 10.)

Dubnick, M. J. (2003), "Accountability And Ethics: Reconsidering the Relationships", *International Journal of Organization Theory and Behavior*, 6(3): pp. 405-41.

Eshleman, A. (2014), "Worthy of Praise: Responsibility and Better-than- Minimally-Decent Agency" in *Oxford Studies in Agency and Responsibility*, Volume 2 'Freedom and Resentment', Oxford Scholarship Online.

Floridi, L. (2013). "Distributed morality in an information society," *Science and Engineering Ethics*, 19(3): pp. 727-43.

_____ (2016). "Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions", *Philosophical Transactions of the Royal Society A* (Mathematical Physical and Engineering Sciences), 374(2083), https://doi.org/10.1098/rsta.2016.0112 (검색일: 2019. 07. 10.)

Friedman, B. (1990). "Moral Responsibility and Computer Technology," Paper Presented at the Annual Meeting of the

American Educational Research Association, Boston, Massachusetts, pp. 1-10.

Hand, S. (1989) (eds), *The Levinas Reader*, by Emmanuel Levinas, MA: Balckwell,

Hildebrand, C. H. (2012), *Kant and Moral Responsibility*, dissertation paper, University of Ottawa.

Johnson, D. G. (2001). *Computer Ethics*, 3rd edition, Upper Saddle River, New Jersey: Prentice Hall.

Johnson, R., A. Cureton. (2016), "Kant's Moral philosophy" in *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/kant-moral/ (검색일: 2019. 07. 10.)

Jonas, H. (1984), *The Imperative of Responsibility: In search of an Ethics for the Technological Age*, Chicago: The Chicago University Press.

Lingis, A. (1969) (trans), *Totality and Infinity: An Essay on Exteriority*, by Emmanuel Levinas, Pittsburgh, PA: Duquesne University Press.

Matthias, A. (2004), "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and Information Technology* 6: pp. 175-83.

Nidditch, P. H. (1992) (eds), *A Treatise of Human Nature*, by David Hume, Oxford.

Nissenbaum, H. (1994). "Computing and Accountability", *Communications of the Association for Computing Machinery*, 37(1): pp. 72-80.

Robson, J. M. (1977) (eds), "On Liberty" in *Essays on Politics and Society*, by John Stuart Mill, University of Toronto Press, pp. 213-310.

논문 투고일	2019. 07. 10.
심사 완료일	2019. 07. 29.
게재 확정일	2019. 07. 29.

Can we impose responsibilities on artificial intelligence?:

To seek accountability-oriented ethics for artificial intelligence

Jungwon Lee

In this paper, I will examine seriously with the concept of accountability whether we can impose liability on the artificial intelligence system as an autonomous agent when it causes negative behavioral consequences. Today, the background to the issue of responsibility again for non-human artificial intelligence systems is that the artificial intelligence system with autonomy of choice is actively working on the existence of human being, causing the problem of ‘many hands’ and voids of liability. Therefore, I will first give a situational conditions for the artificial intelligence system that can cause problems of responsibility. Secondly, I will examine the possibility of the concept of responsibility to be extended to autonomous actors other than human beings, focusing on the concept of responsibility of Emmanuel Levinas, in spite that the concept of responsibility in traditional moral philosophy has been exclusively applied to human beings. In addition, I will show that it is not easy to apply this concept to present and future artificial intelligence systems, even if it is possible to expand the concept of responsibility. And I suggests application of accountability concept instead of responsibility for artificial intelligence system. I will analyze that this accountability can be practically implemented in an explainable artificial intelligence system. Finally, I will carefully put the possibility of accountability-oriented ethics for artificial intelligence systems into perspective.

Keywords: Artificial Intelligence, Many hands problem, Responsibility, Levinas' Ethics on Thinking-of-the-Other, Accountability, Explainable Artificial Intelligence (XAI)