https://doi.org/10.5392/JKCA.2018.18.03.592

포스트휴먼 시대의 로봇과 인간의 윤리

The Ethics of Robots and Humans in the Post-Human Age

유은순, 조미라
중앙대학교 인문브릿지사업단 공동연구원

Eun-Soon You(tesniere@naver.com), Mi-Ra Cho(goho38@hanmail.net)

요약

로봇의 영역이 인간의 정신적, 감정적 노동까지 대신하는 지능형 로봇으로 진화하면서 인간과 로봇 관계에서 발생할 수 있는 '로봇윤리'가 중요한 이슈로 떠오르고 있다. 본 연구는 포스트휴먼 시대에 필요한 인간과 로봇의 윤리 성찰을 고찰하고자 하며, 그 중심 내용은 다음과 같다. 첫째, 로봇의 윤리적 실천 가능성에 도전하는 윤리 소프트웨어 개발 사례를 통해 오로지 강제 입력된 윤리 코드만으로 로봇이 과연 옳고 그름을 판단할 수 있는가라는 문제의식에서 출발한다. 둘째, 로봇윤리는 인간의 편향성이 내재된 데이터를 학습했을 때 발생할 수 있는 비윤리적 문제들을 고려하고, 더불어 국가와 문화 간의 윤리적 상대주의를 인정해야 한다. 셋째, 로봇윤리는 로봇을 위한 윤리 강령만이 아니라, 인간과 로봇이 서로 공진화할 수 있는 새로운 개념의 '인간 윤리'가 전제되어야 한다.

■ 중심어 : | 로봇윤리 | 포스트휴먼 | 아이작 아시모프의 로봇 3원칙 | 윤리 소프트웨어 | 인공적 도덕 행위자 |

Abstract

As the field of robots is evolving to intelligent robots that can replace even humans' mental or emotional labor, 'robot ethics' needed in relationship between humans and robots is becoming a crucial issue these days. The purpose of this study is to consider the ethics of robots and humans that is essential in this post-human age. It will deal with the followings as the main contents. First, with the cases of developing ethics software intended to make robots practice ethics, the authors begin this research being conscious about the matter of whether robots can really judge what is right or wrong only with the ethics codes entered forcibly. Second, regarding robot ethics, we should consider unethicality that might arise from learning data internalizing human biasness and also reflect ethical differences between countries or between cultures, that is, ethical relativism. Third, robot ethics should not be just about ethics codes intended for robots but reflect the new concept of 'human ethics' that allows humans and robots to coevolve.

■ keyword : | Robot Ethics | Post-human | Issac Asimov' s three laws of Robotics | Software Engineering Ethics | Artificial Moral Agent |

I. 서 론

기술은 오랫동안 도구적 유용성과 효용성 차원에서 사유되어 왔다. 하지만 첨단 과학이 기술과 결합하면서, 단순한 도구적 차원이 아니라 전통적인 '인간'의 개념마저 뒤흔들 정도로 인간의 삶 전체를 지배하는 절대적인

* 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2016S1A6A9931352)
접수일자 : 2018년 01월 19일
수정일자 : 2018년 02월 09일
심사완료일 : 2018년 02월 11일
교신저자 : 조미라, e-mail : goho38@hanmail.net

요소로 자리하게 되었다. 한스 요나스(Hans Jonas) 역시 현대 기술이 갖는 특이한 양상을 포착, 인간과 기술의 관계를 다음과 같이 정의하였다. 첫째, 현대 기술은 그 변화 속도와 규모가 너무 크기 때문에 그 결과를 예측하기 힘들며, 설령 선하고 정당한 목적을 위해 사용하더라도 장기간에 걸쳐 사용할 경우 위협적인 요소가 들어있다. 둘째, 목적에 대한 더 나은 만족을 위해 새로운 기술을 양산하는 것이 아니라, 거꾸로 새로운 기술이 새로운 목적을 부여하거나 그것을 강요할 수도 있다. 따라서 기술의 진보는 우리의 의지에 의해 선택할 수 있는 것이 아니라 현대 기술 자체 안에 자리 잡고 있는 동인으로서 우리의 의지 너머에서 작용하고 있다는 것을 세 번째 특징으로 지적한다[1]. 나아가 자크 엘륄(Jacque Ellul)은 "현대 기술은 그 자체의 동력을 가지고 스스로 이루어지기 때문에 더 이상 인간의 개입이나 결정이 무의미하다[2]"라며 기술이 인간의 통제를 벗어나 자율성을 갖게 될 때 인류에게 미칠 위험성을 경고하기도 했다. 이처럼 요나스는 그동안의 전통 윤리학[1]만으로는 현대 기술의 "묵시록적 위협[1]"을 통제할 수 없기에 "윤리적 진공상태[3]"를 야기할 것임을 강조, '책임'이라는 개념으로 현대 기술 시대를 위한 새로운 윤리적 실천을 모색했다.

4차 산업혁명을 이끌어 갈 현대 기술 분야 중에서도 로봇 공학은 그 어느 때보다도 윤리적 사유를 필요로 한다. 이러한 현실적 요구들을 반영하듯 최근에는 로봇권과 로봇윤리와 관련된 일련의 중요한 발표들이 있었다. 2017년 유럽연합(EU)의회에서는 인공지능 로봇의 법적 지위를 '전자인간'으로 인정하고 전자인간이 지켜야 할 윤리의무를 규정하였으며, 국내에서도 한국로봇산업진흥원이 2017년 12월 31일까지 로봇윤리헌장을 마련할 것임을 발표한 바 있다. 이처럼 국내외에서 로봇 시대를 대비한 윤리적 방침을 세우고 있는 이유는 로봇기술이 이전의 것과는 다른 양상으로 우리의 일상을 침투하면서 인간과 로봇 간의 관계에 새로운 성찰을 요구하고 있기 때문이다. '로봇(robot)'의 어원이 '노동'에서 유래하듯 공상과학(SF)소설이나 영화에서 로봇의 대부분은 인간의 힘든 노동을 대신하거나 인간의 명령에 무조건 복종하는 인간을 위한 도구로 그려졌다. 이렇게 상상으로만 존재하던 로봇은 1960년대부터 실제 산업 현장에 투입되어 인간의 고된 육체노동을 대신하는 충실한 도구로서의 역할을 수행해왔다.

하지만 인공지능과 결합된 오늘날의 로봇은 더 이상 단순한 수동적 기계에 머무르지 않고 인간의 정신적, 감정적 노동까지 대신하는 "인조지능(synthetic intellect)[4]"을 가진 형태로 진화하고 있다. 의료 로봇, 전투 로봇, 교육용 로봇, 애완견 로봇, 그리고 인간과 감정적 소통이 가능한 사회적 로봇의 등장까지 로봇은 눈부신 진화를 거듭하고 있다. 문제는 이전에는 볼 수 없었던 다양한 지능형 로봇들이 인간의 모든 활동 영역 안으로 들어오게 되면 로봇 스스로 어떤 윤리적 판단이나 결정을 내려야 하는 상황들이 발생할 가능성이 커지게 되는데, 그 판단이나 결정이 인간에게 어떤 영향을 미칠지 알 수 없다는 것이다. 예를 들어 전투 로봇은 그것을 사용하는 과정에서 복잡한 윤리적 문제를 야기한다. 로봇이 잘못된 판단을 하거나 기술 자체에 오류가 있을 경우 무고한 인간들이 목숨을 잃을 수 있으며, 혹은 테러리스트가 전투 로봇을 사용할 경우 인류에게 큰 재앙을 가져올 수 있다. 로봇윤리가 필요한 이유는 바로 이 때문이다. 로봇의 판단이 인간의 생사에 영향을 끼칠 수 있는 상황이 되면서 로봇의 모든 판단과 결정에는 윤리적 요소가 들어갈 수밖에 없다. 로봇윤리는 로봇이 인간의 통제를 벗어날 경우 일어날지도 모르는 재앙을 예방하기 위한 수단으로, 그리고 인간과 로봇 간의 관계에서 발생할 수 있는 수많은 윤리적 문제들을 해결하기 위한 지침서로서의 역할을 수행한다. 그렇다면 로봇은 어떻게 윤리를 학습하고 실천할 수 있을까? 다시 말해서 로봇에게 인간의 도덕적 인성에 기초한 '윤리'를 '교육'시키고 '공유'하도록 하는 것이 가능한가? 만약 로봇이 프로그래밍 된 윤리적 지침에 따라 행동한다 하더라도 그 행동의 결과가 인간에게 해가 될

1. 요나스의 주장에 따르면 전통 윤리학은 윤리적 주체와 대상을 '인간'으로 국한했다. 다시 말해서 인간들 간의 관계를 통해 야기되는 행위만을 고려했고 자연, 기술과 같은 인간 이외의 대상과의 상호 행위는 도덕적 판단에서 제외되었다. 뿐만 아니라 행위의 결과를 현재에 귀속시키고 미래적인 책임은 다루어지지 않았다[1]. 따라서 전통적인 윤리학은 자연과 인간의 본성에 대한 '불가침'의 원칙을 깨고 어떤 특수한 목적을 갖고 그것을 변형시킬 수 있는 잠재적인 힘을 갖고 있는 현대 기술의 본질을 포착하는데 한계가 있다.

가능성은 없는가? 만일 해가 된다면 그 책임을 누구에게 돌려야 하는가? 그리고 로봇을 포함한 모든 지능적 기계의 윤리와 인간의 윤리가 충돌했을 때 어떻게 해결해야 할 것인가? 이처럼 로봇윤리는 인간의 삶과 관련한 긴박하고 중요한 질문으로 이어지게 된다. 본 논문은 바로 이 일련의 질문들에 대한 고민을 통해 현대 기술 시대에 필요한 인간과 로봇의 윤리가 무엇인지를 성찰하는 것을 목적으로 한다. 본 논문의 구성은 다음과 같다. Ⅱ장은 로봇윤리 개념에 대한 다양한 시각을 살펴보고, Ⅲ장에서는 윤리 소프트웨어의 개발 사례를 통한 로봇의 윤리적 실천가능 여부와 그 문제점을 검토한다. Ⅳ에서는 편향성이 내재된 인간의 데이터를 로봇이 학습했을 때 발생하는 여러 문제점들을 고찰한다. 마지막 Ⅴ장에서는 로봇의 윤리 이전에 인간의 윤리를 되물음으로써, 인간과 로봇의 바람직한 공존을 모색하는 것으로 결론을 맺는다.

Ⅱ. 로봇윤리의 개념에 대한 고찰

1. 아이작 아시모프(Issac Asimov)의 로봇 3원칙과 그 한계

사실 로봇이 인간에게 위험한 존재가 되지 않도록 준수해야 할 원칙을 시도한 것은 최근의 일이 아니다. 사이언스 픽션 소설가인 아이작 아시모프가 그의 소설에서 다양하게 적용한 바 있는 '로봇의 3원칙'[2]은 대표적인 사례이다. 아시모프는 그의 소설에서 인간을 즐겁게 해주기 위해 인간이 듣고 싶어 하는 거짓말을 반복하는 로봇, 인간에게 피해를 주는 것은 인간이라 판단하여 인간을 통제하려는 로봇, 절대 다수의 인간을 구하기 위해서는 소수의 인간에게 해를 끼쳐야 하는 모순에 직면하는 로봇 등을 다양하게 등장시키며 로봇 3원칙이 어떻게 작동하고 개념들이 서로 어떻게 충돌하는지 끊임없이 실험했다.

그렇다면 아시모프의 3원칙은 현실에서 실현될 수 있을까? 과학기술이 무서운 속도로 발전됨에 따라 전통적인 '인간' 개념이 해체되고, 미래 인간을 지칭하는 다양한 용어들이 등장하기 시작했다. 트랜스 휴먼(Trans Human), 네오 휴먼(Neo Human), 그리고 포스트 휴먼(Post Human)등이 바로 그것이다. 이 용어들은 모두 인간의 생물학적 육체와 정신의 한계를 뛰어넘는 인간 이후의 인간' 혹은 '인간을 벗어난 인간'을 의미한다. 이처럼 전통적인 인간의 개념을 벗어난 '포스트 휴먼' 시대가 앞으로 어떤 형태로 등장할 지는 정확히 알 수 없으나, 지금의 생물학적 인간과는 분명 '다른 종'일 것임은 틀림없다. 그리고 이 새로운 종으로서의 인간을 가능하게 만드는 결정적인 요소는 바로 '테크놀로지'가 될 것이다. 따라서 아시모프의 3원칙이 현실에 적용되기 위해서는 '인간'의 개념에 대한 새로운 성찰이 필요하다. 또 다른 문제는 인간의 명령에 대한 복종이다. 인간의 명령이 항상 옳은 것은 아니기 때문이다. 윤리적 가치가 서로 다른 명령들이 동시에 이루어졌을 때 로봇은 누구의 명령에 복종해야 하는지도 어려운 문제이다. 이처럼 아시모프 3원칙은 여러 가지 측면에서 한계를 가지며 다시 한 번 재고할 필요가 있다.

2. 로봇윤리의 개념

한 작가의 상상력에서 만들어진 로봇 3원칙은 그 한계점에도 불구하고 2002년 로봇공학자 지안마르코 베루지오(Gianmarco Veruggio)가 '로봇윤리(roboethics)'라는 용어를 처음 사용하면서 현실화되기 시작했다. 베루지오는 로봇윤리란 "로봇공학이 인간의 삶에 적용될 때 나타날 수 있는 윤리적 문제를 다루는 것이며, 로봇기술의 긍정적 확산을 위해 우리가 지금 무엇을 해야 하는지 논의하는 것[6]"이라고 정의하였다. 그리고 유럽연합이 명시한 전자인간으로서의 로봇이 지켜야 할 윤리의무는 바로 아시모프의 '로봇의 3원칙'을 바탕으

2. 아시모프는 로봇 3원칙을 다음과 같이 제안한다. 제1원칙 : 로봇은 인간에게 해를 끼쳐서는 안 되며, 위험에 처한 인간을 방관해서도 안 된다. 제2원칙 : 1원칙에 위배되지 않는 한, 로봇은 인간의 명령에 복종해야만 한다. 제3원칙 : 제1원칙과 제2원칙에 위배되지 않는 한, 로봇은 자기 자신을 보호해야만 한다. 아시모프는 나중에 제0원칙을 추가하는데 '로봇은 인류가 위험에 처하도록 해서는 안 된다'라는 내용이다. 이는 인간 개개인이 아닌 집단으로서 인류 전체의 안전에 대해 로봇에게 경각심을 심어주기 위한 것이다. 예를 들어 로봇들에게 아마존의 밀림을 모두 개간하라는 명령을 내린다면 당장 누군가의 생명을 위태롭게 하지는 않겠지만 인류의 생존 환경에는 큰 악영향을 미칠 수도 있다. 이 0원칙은 다른 3원칙보다도 상위에 있는 가장 중요한 법칙으로 자리 잡게 된다[5].

로 만들어졌다. 그 의무에는 “로봇이 인간에게 저항하는 것을 막기 위해 로봇의 움직임을 멈출 수 있는 킬 스위치를 마련해야 한다[7]”는 내용이 포함되어 있다. 하지만 이것은 ‘로봇윤리’라기 보다는 ‘로봇의 강제조항’에 더 가깝다고 할 수 있다.

따라서 현재의 로봇윤리는 로봇이 인간에게 지켜야 할 의무뿐만 아니라 로봇을 설계하고 그것을 사용하는 사람들의 윤리까지 포괄하는 방향으로 나아가고 있다. 로봇 공학자 키스 애브니(Keith Abney)는 로봇 안에 프로그램된 ‘모럴 코드(moral code)’까지도 로봇윤리에 포함시켜야 한다고 주장하기도 하였다[8][9].

고인석은 로봇윤리를 크게 두 가지 차원에서 정의한다. 첫 번째는 로봇을 설계, 제작, 관리하는 사람의 관점이다. 여기에는 로봇을 어떻게 만들어야 하는가에 대한 문제의식이 들어가는데, 결국 건전한 도덕적 판단을 갖춘 로봇을 만들어 인간에게 위협이 되지 않아야 한다는 것이다. 두 번째는 로봇이 실현하는 행위의 도덕적 함의를 판단하는 것이다. 로봇의 행위가 오작동으로 인한 실수였는지, 아니면 고의적으로 인간을 해하려는 행위였는지, 그리고 그 행위로 인한 결과의 책임은 누구에게 있는가를 판단하는 것이다[10]. 하지만 로봇의 행위에 대한 도덕적 판단은 물론 판단에 따른 법률적 책임을 따지는 것은 매우 복잡한 문제이다. 영화 <아이, 로봇(I, Robot)>(2004)에서는 “로봇에게도 살인죄가 적용될 수 있는가”라는 복잡한 상황을 제시한 바 있다. 영화 속 변호사가 주장하듯이 살인죄는 사람이 사람을 해하였을 때만 적용되기 때문에 로봇에게는 살인죄가 적용될 수 없는 것인가? 설령 로봇이 사람을 해하였다 하더라도 그것은 살인죄가 아닌 산업 재해에 불과하며 고장난 로봇은 해체해 버리면 끝나는 문제인가? 등이 바로 그것이다. 영화 속 로봇처럼 고도의 지능과 자유의지를 가진 로봇의 출현이 현실적으로 가능하지 않다고 해도 로봇은 계속해서 다양한 형태로 진화해 나갈 것은 분명하기 때문에 그러한 상황을 대비하지 않으면 안 된다. 로봇윤리는 요나스의 표현을 빌리자면 “미래를 고려하는 현재의 윤리[1]”인 것이다.

III. 로봇도 윤리적 실천이 가능한가?

진(眞)·선(善)·미(美)는 인간의 역사에서 가장 오래되고 첨예한 인문학적 고민이다. 그 중에서도 선(윤리)은 인간의 행위 규범과 관련한 근원적이고 총괄적인 분야이다. 아리스토텔레스의 『니코마코스 윤리학』 제1장도 “모든 행위와 선택이 추구하는 것은 어떤 좋음인 것 같다...좋음(선)이야말로 당연히 모든 것이 추구하는 목표[11]”라는 문장으로 시작하고 있다.

윤리와 도덕은 모두 오랜 세월 인간의 경험과 학습을 통해 체화된 것이라는 공통점이 있지만 여러 가지 측면에서 구분되어 사용되는 개념이기도 한다. ‘도덕’이 외부에서 주입되는 것이라면, ‘윤리’는 주체 스스로의 실천이다. 옳고 그름을 이해하는 것은 무미건조한 산술적 계산이 아니라, 인간의 정서, 느낌에 뿌리를 박고 있는 것이다. 가령, 우리가 ‘양심을 따른다’라고 했을 때 그것은 밖에 있는 것이 안으로 들어온 것이지만(도덕), ‘내면을 따른다’는 것은 나 자신으로 살라는 소명이다(윤리). 도덕이 집단의 것인 것과는 달리 윤리는 개별 주체의 것이다[12].

그렇다면 인간처럼 로봇에게 윤리를 입력시켜 학습시키는 것만으로 윤리적 실천이 가능한가? 이 질문에 대한 해답을 얻기 위해 여러 학자들이 도전적인 연구들을 진행하고 있다.

웬델 월러치(Wendell Wallach)와 콜린 알렌(Colin Allen)은 그들의 저서 『왜 로봇의 도덕인가』에서 도덕 행위자의 주체를 인간을 넘어 로봇과 같은 인공물까지 확대하고 이를 "인공적 도덕 행위자(artificial moral agent, AMA)[13]“라고 불렀다. 아직 초기 단계이기는 하지만 “인공적 도덕 행위자”를 구현하기 위한 다양한 접근법이 소개되고 있다. 대표적인 구현 방법으로는 크게 전통적인 공리주의와 같이 어떤 특정 윤리 이론에 기반을 둔 하향식(top-down)과 다양한 기계 학습을 통해 도덕적인 추론을 배워나가도록 하는 상향식(bottom-up)이 있다[13]. 각각의 내용을 아래에서 살펴보자.

1. 하향식(top-down) 접근법

하향식은 다양한 윤리 이론을 선택하고 큰 원칙을 미리 로봇에게 프로그래밍 한다. 대표적인 사례는 셀머 브링스요드(Selmer Bringsjord)가 개발한 윤리 프로그램들이다. 그는 '생명은 소중하다' '최대다수의 최대 행복'과 같은 공리주의 원칙을 프로그래밍 해주고 로봇 스스로 판단하게 한다[6]. 만일 로봇이 공리주의 원칙에 따라 환자의 생명유지 장치 여부를 결정할 경우 그 판단은 과연 어떤 결과를 가져올 것인가? 인간도 결정을 내리기 어려운 복잡한 상황에서 과연 로봇의 결정을 옳은 것이라고 생각할 수 있을까? 인간의 삶이 그렇듯이 늘 완벽한 답을 갖고 살아가는 게 아니기 때문이다. 이 문제와 관련해서는 영화 <아이, 로봇>에서도 잘 드러나고 있다. 주인공 스프너(Spooner) 형사는 차 사고로 어린 소녀와 함께 물에 빠진 경험이 있다. 이때 로봇은 스프너 형사와 어린 소녀 중 어린 아이보다 어른의 생존율이 더 높다는 합리적 연산과 판단에 따라 스프너만 구조한다. 로봇은 인간에게 해를 끼치지 않는다는 '로봇 제1원칙'에 따른 결론이었지만, 스프너는 오히려 이 사건이 트라우마가 되어 인간적인 괴로움에 시달린다. 그는 혼자만 구출된 기억을 떠올리며 "생존율이 낮더라도 여자아이를 먼저 구했어야 했어. 로봇은 단순한 고철덩어리"에 불과하다며 로봇에 대해 강한 불신과 분노를 표출한다. 이 에피소드는 옳고 그름의 가치 판단이 로봇과 인간 모두에게 얼마나 어려운지를 단적으로 보여주고 있다. 또한 소녀를 먼저 구해야 한다는 스프너의 윤리적 가치와 로봇의 윤리적 판단이 충돌하고 있음을 상징적으로 표현하고 있다. 이처럼 하향식 방식은 상황에 따라 여러 변수가 존재하고 서로 다른 답을 내릴 수 있다는 데에 문제가 있다. 더 결정적인 문제는 로봇에게 윤리를 '가르친' 인간조차도 이를 잘 지키지 않는다는 것이다. 인간이라고 해서 로봇보다 올바른 판단과 실천으로 살아가는 종이 아니기 때문이다.

2 상향식(bottom-up) 접근법

상향식 방법은 로봇에게 윤리와 도덕을 꾸준히 학습시키는 것이다. 그런데 이 접근법에서는 위에서 기술한 로봇의 행위에 대한 법률적 책임에 대한 문제가 제기된다. 인간에게는 '실수'와 '용서'라는 것이 존재한다. 아이가 잘못을 저질렀을 때는 어린아이이기 때문에 혹은 잘 몰라서 저지른 실수라는 이유로 처벌과 용서가 가능해진다. 반면 이와 동일한 사건이 '로봇'에게 벌어졌을 때는 또 다른 딜레마가 발생한다. 미국의 한 대형 쇼핑몰에서 경비 역할을 하는 1미터 가량의 작은 '로봇'이 생후 1년 4개월 된 된 아이의 발을 다치게 한 적이 있다. 이 로봇은 경비 일을 수행하던 중 아이를 넘어뜨렸는데, 멈추지 않고 계속 움직이면서 벌어진 사건이었다. 로봇 개발업체와 쇼핑몰 관리사무소는 일단 경비 로봇을 철수시켰다고 하지만, 이런 문제가 발생했을 때 사람을 다치게 한 '로봇'에게 그 책임을 물을 것이지, 아니면 '로봇'의 소유주에게 책임을 물을 것인지에 대한 법률적, 철학적 과제는 여전히 요원하다[14]. 그런데 여기서 한 가지 의문이 생긴다. 만약 저 경비용 로봇이 넘어진 아이를 곧바로 일으켜 세운 후, 정중하게 사과를 했다면 아이의 부모는 로봇의 실수가 고의가 없었음을 인정하고 너그러이 용서해줄 수 있을까. 누군가를 용서한다는 것은 상대방의 '인권'을 인정하고 그것을 존중한다는 의미이다. 결국 로봇이 윤리를 실천하기 위해서는 인권(혹은 로봇권)이 전제되어야하기 때문이다.

위에서 살펴보았듯이, 인공적 도덕 행위자를 구현하기 위한 두 개의 접근법 모두 한계를 갖고 있다. 그런데 더 중요한 문제는 다른 데 있다. 그것은 지금 현대의 기술로도 구현할 수 없는 특별한 것, 즉 로봇의 '자유의지'이다. 로봇이 '윤리' 의지를 갖기 위해서는 무엇보다 '자유의지'가 전제되어야 한다. 로봇 스스로 무엇이 옳고 그른 것인지를 판단할 수 있을 때 윤리적 실천은 가능하기 때문이다. 그렇다면 다양한 변수가 발생하는 상황에서 로봇은 과연 강제로 입력된 윤리코드만으로 올바른 판단을 내릴 수 있을까?

IV. 누구의 윤리를 학습하는가?

로봇의 윤리적 실천 가능성과 함께 제기되는 또 하나의 문제는 로봇이 과연 누구의 윤리를 학습하는가에 대한 질문이다. 이 문제는 윤리는 과연 객관적인 것인가라는 질문과도 일맥상통하다. 최근 기계학습에서 인간의

신경망을 모방한 인공 신경망, 딥러닝(deep learning)이라고도 불리는 기술이 큰 주목을 받고 있다. 인간과의 바둑 대결에서 승리한 인공지능 알파고(AlphaGo)와 인간을 누르고 퀴즈쇼에서 우승한 인공지능 컴퓨터 왓슨(Watson)의 뒤에는 방대한 양의 데이터에 대한 딥러닝을 이용한 기계학습이 있었다. 문제는 딥러닝 기술이 아니라 인간이 만든 데이터에 있다. 2017년 4월 '사이언스(Science)'지에 흥미로운 논문이 발표되었다. 논문에 의하면 수집된 코퍼스를 이용하여 기계학습을 한 인공지능이 '여성'이라는 단어를 예술과 인문에 연결한 반면, '남성'은 수학과 공학과 높은 상관성이 있다고 판단했다. 또한 유럽계 미국인 이름은 긍정적인 단어와 관계있는 것으로, 아프리카계 미국인 이름은 부정적인 단어와 연결하였다. 수 천 년 동안 인간이 쌓아 놓은 엄청난 양의 데이터에 숨겨진 편향성을 기계가 학습한 것이다[15][16]. 현재 구글을 비롯한 영향력 있는 기업들이 앞 다투어 만들고 있는 기계학습 알고리즘은 기업의 이익을 위해 설계된 것이다. 이러한 이유로 로봇윤리의 개념 안에 개발자의 윤리와 심지어 로봇 안에 프로그램된 모럴 코드까지 포함해야 한다는 주장이 제기되는 것이다. 더 극단적인 사례는 마이크로소프트(MS)가 만든 채팅봇 테이(Tay)를 들 수 있다. 테이는 트위터 사용자들이 악의적으로 입력한 데이터를 학습한 결과 인종차별적이고 여성혐오적인 발언을 쏟아내면서 큰 파문을 일으켰다. MS사가 16시간 만에 서비스를 중단했지만 백인 우월주의자들의 편향되고 왜곡된 데이터를 기반으로 학습한 인공지능이 얼마나 비윤리적으로 변할 수 있는지를 단적으로 보여주었다는 점에서 테이의 사례는 많은 것을 시사한다[17].

로봇이 누구의 윤리, 어떤 윤리를 학습했느냐에 따라 그 선택과 판단은 달라질 수밖에 없다. 영화 <채피(Chappie)>(2015)에서 로봇은 스스로 생각하고 판단하는 자율적인 존재라고 생각하는 로봇 개발자 디온과 디온에 의해 만들어진 로봇 '채피'가 등장한다. "누군가를 죽이면 안 됩니다. 나쁜 짓을 하면 안 됩니다"라며 인간보다 더 인간적인 모습의 채피는 자신의 권력욕을 위해 로봇을 이용하는 비정한 무기 개발자 빈센트(Vincent)와 돈을 벌기 위해 채피를 이용하려는 탐욕스러운 인간들의 모습과 대조를 이룬다. 영화는 마치 인간의 경우처럼 로봇도 누구에 의해 어떤 교육을 받느냐에 따라 어떤 로봇으로 성장할 수 있는지를 보여준다.

데이터에 숨겨진 편향성뿐만 아니라 도덕과 윤리에 대한 상대주의도 로봇의 윤리 학습에서 중요한 문제가 된다. '좋음'과 '나쁨'의 문제는 개인뿐만 아니라 국가와 종교, 문화마다 서로 다르고 동일한 문제에 대해서도 상이하게 해석할 수 있기 때문이다. 현재 윤리 프로그램에서 가장 많이 활용되고 있는 칸트(Kant)의 정언명령과 벤담(Bentham)과 밀(Mill)의 공리주의를 포함하여 수많은 윤리 이론이 존재한다. 도덕적 출발점을 선의지로 보는 칸트의 정언명령과 인간의 윤리적 동기를 이익과 쾌락에 두는 공리주의 중 개발자가 어떤 윤리를 선택하느냐에 따라 똑같은 상황이라도 로봇은 다른 선택을 할 것이다. 이렇게 볼 때 윤리는 객관적이라고 말할 수 없다. 로봇윤리는 결국 국가와 문화 그리고 종교 간의 윤리적 상대주의를 고려할 수밖에 없다.

V. 다시 인간의 윤리로

본 연구는 로봇윤리는 결국 인간의 윤리를 다시 생각하는 것이라는 결론에 이르게 된다. 정지훈의 『호모사피엔스의 위험한 고민』에는 '지뢰 제거 로봇'과 관련한 에피소드가 소개된다. 1인1로봇 1조로 지뢰 제거 업무를 수행하는 이 지뢰제거로봇은 동료 병사와 함께 지뢰 제거를 성공적으로 수행하던 중, 큰 부상(고장)을 당하고 만다. 업무 능력 복원이 힘들 정도로 크게 파손되자 결국은 지뢰제거를 폐기 처분하는 것으로 결정된다. 그러자 그동안 지뢰제거로봇과 한 조가 되어 임무를 수행하던 병사가 울부짖으며 자신의 동료였던 지뢰 로봇을 살려달라고 애원했다는 것이다[14]. 이처럼 지뢰제거로봇의 폐기 처분에 측은지심을 느끼고, 로봇을 생명체로 동일화하는 현상은 낯선 감정이 아니다. 인간은 누구나 자신과 오랜 시간을 함께 해온 존재를 따뜻한 동반자로 여기기 마련이다. 심지어 미국의 한 로봇 제작 회사에서 자신들이 만든 로봇의 균형 감각을 증명하기 위해 발로 차도 넘어지지 않는 장면을 공개했다가 여론

으로부터 로봇 학대 논란에 휩싸인 적도 있다. 그렇다면 그동안 로봇이 '인간'이 아니라는 이유만으로 '인간'보다 못한 존재로 취급해오던 이 관습적 사고를 다시 생각해봐야 할 것이다.

영화 <매트릭스(The Matrix)>(1999)의 프리퀄(prequel)에 해당하는 <애니 매트릭스>(2003)의 "두 번째 르네상스"는 로봇에게 멸망한 인류가 매트릭스 공간에 갇히기까지의 과정을 보여주면서 인간과 기계, 즉 인간과 비-인간의 관계를 섬세하게 파헤친다. 이 에피소드에서 '로봇 E1-66ER'이 인간의 부당한 대우와 폭력을 견디다 못해 자신의 사용자(인간)를 살해하는 사건이 일어난다. 분노한 인류는 자신의 '재산'을 파괴할 권리가 있음을 주장하며 주인을 살해한 로봇을 비롯하여 전 세계 모든 기계인간을 파괴할 것을 결정한다. 이에 반발한 '로봇 E1-66ER'은 인간의 정신이 주어진 기계에게도 공정한 재판을 받을 자격과 권리를 주장하지만, 로봇E1-66ER에게 '폐기'처분이 내려진다. 인간이 스스로의 필요에 의해 창조한 기계인간을 모두 파괴하기에 이른 것이다[18]. 모든 로봇과 기계는 인간의 노동력이나 능률적인 생산 활동의 필요에 의해 창조된 피조물이다. 그럼에도 인간을 위협하는 존재로 생각하는 이유는 무엇일까. 단순히 그들의 기계성(물리적 능력)때문만은 아닐 것이다.

루이스 멈퍼드(Lewis mumford)는 『기계의 신화』에서 '거대 기계(Megamachine)'라는 개념을 제시하면서 "최초의 기계는 차가운 금속으로 만들어져 계기판과 점멸등이 장착된 덩어리 즉, 물질기계가 아니라 인간사회 자체"[19]라고 주장한다. 무엇보다 그는 기술(기계)을 인간 존재에 외적인 하나의 도구로서 이해하는 시각을 거부하고 그 발전 과정이 인간의 외면적·내면적 존재의 변화과정과 동일하다는 것으로 이해한다. 따라서 인간 자신은 변하지 않는 상태에서 어떤 기술만 나타나 그 기술을 인간이 그저 이용만 한다는 식의 생각은 환상이라는 것이다. 이것이 그가 신석기 혁명 이후 기술과 사회 발전에 따라 인간 자체가 어떻게 변화했는가를 논한 『인간의 전환』의 핵심 주제이자, 모든 것을 자신의 발아래에 두고 절대적인 '신'처럼 군림하고자 하는 '권력'의 문제이기도 하다[20]. 따라서 도래할 미래에 기계와 인간을 구별하는 경계가 사라지고 기계 스스로 자유의지를 갖게 된다면, 인간과 기계의 갈등은 필연적일 수밖에 없을 것이다.

VI. 결론 : 로봇은 인간을 닮아간다

로봇과 인간의 대립은 기계성과 인간성의 충돌 과정에서 비롯되는 것 같지만, 실제로는 인간이 자신의 권위와 통치를 정당화하기 위한 '열등한' 타자의 필요성에서 기인한다고도 할 수 있다. 미셸 푸코에 따르면 타자 앞에 선 동일자의 전략은 두 가지 뿐이다. "타자가 차이를 동일시하거나 아니면 무화시키는 것이다. 전자를 위해서는 지식이, 후자를 위해서는 무력이 동원된다. 동일자의 궁극 목표는 남김없이 자기 영토에 편입시켜 완전한 동일성의 세계를 구축하는 것이다. 그러나 타자가 동일성의 영역에 편입된다고 곧 동일자가 되는 것은 아니다. '같은 영토 안의 타자'로 남아 있어야 한다[21]." 이것은 마치 기계 스스로 인간이 아니라는 사실을 자각하는 순간, '타자의 영역'에 들어섰음을 인식하는 것과도 같다. 이렇게 타자의 영역에 들어서게 된 기계는 자신이 속한 영역에 거부감을 가질 수밖에 없고 동시에 '타자'로서는 갖기 힘든, '인간'이 가진 무언가를 지향할 수밖에 없다. 여기에서 기계가 갖고자 하는 '무엇'은 피와 살을 가진 육체로서의 '인간'이 아니다. "인간이 된다는 것, 혹은 어떤 주체, 어떤 인물이 된다는 것은 단지 신체를 옷처럼 걸쳐서 살아가는 것을 뜻하지 않는다. 그 의식이 자기 정체성과 하나로 된다는 것을 의미한다[22]."

오늘의 현실은 테크놀로지를 싫어하는 것이 인간 자체를 싫어하는 것에 다름 아닌 것이 될 정도로 테크놀로지적이다[23]. 그리하여 테크놀로지는 인간에게 더 이상 다른 인간이 필요하지 않다는 환상을 더욱 확고하게 다져갈 지 모른다. 하지만 인간은 물질인 동시에 정신이다. 과학의 놀라운 성취와는 별도로 정신 혹은 영혼이라 부르는 인간의 복잡성을 인정하고 이와 관련한 심도 깊은 고민으로 이어지지 않으면 안 된다.

2015년 캐나다의 '오리아 재단'이 세계적 석학들을 초

청하여 “인류는 과연 진보 하는가”라는 주제로 격론을 벌였을 때 스티븐 핑커(Steven Pinker)는 인간 존재를 물질주의적 관점에 근거해 인류의 안녕이 10가지 차원(인간의 생명연장, 건강, 물질적 번영, 평화, 안전, 자유, 지식, 인권, 성 평등, 지능)에서 ‘진보’했다고 자신 있게 답했다. 하지만 이에 반론을 제기한 알랭 드 보통(Alain de Botton)과 말콤 글래드웰(Malcolm Gladwell)은 인간에게는 과학적 데이터로는 환원되지 않는 인간정신의 복잡성이 존재 한다 라며 우리가 직면하게 되는 미래는 ‘나은 미래’가 아니라 ‘다른 미래’라고 바로잡기도 했다[24]. 과학적 수치나 통계만으로 파악할 수 없는 인간의 실존적 문제들을 고려했을 때, 로봇의 윤리 역시 이러저러한 윤리코드를 강제로 입력한다고 해서 해결될 문제는 아닐 것이다. 아니 어쩌면 인간과 로봇의 평화로운 공존을 위한 해결책은 의외로 간단할지 모른다. 그것은 로봇(인공지능)이 인간의 수단이나 도구를 넘어 인간과 공존하게 될 새로운 ‘타자’들이자 “동반 종들(companion species)[25]”로서의 인정이다. 영화 <아이, 로봇>에서 스프너 형사는 자유의지를 가진 로봇 ‘써니’에게 “네 스스로 결정해. 그것이 네가 자유로운 존재가 되는 것”이라고 말하면서 로봇의 자유의지를 인정하고 그들과의 평화로운 공존을 선택하는 것으로 마무리 한다. 하지만 이러한 평화로운 공존이 영화 속 이야기로만 머물지 않기 위해서는 동일한 유전자를 가진 ‘인간종’ 끼리와도 반목과 대립을 반복해온 인간의 역사를 기억하고 반성할 필요가 있다. 우리가 로봇의 윤리만큼 인간의 윤리를 고민해야 하는 이유도 바로 여기에 있다. 인간이 만든 로봇은 결국 인간을 닮아가기 때문이다.

참 고 문 헌

[1] 한스 요나스, 이유택 옮김, *기술 의학 윤리*, 솔출판사, 2005.

[2] 손화철, *토플러 & 엘륄*, 김영사, 2006.

[3] 한스 요나스, 이진우 옮김, *책임의 원칙*, 서광사, 1994.

[4] 제리 카플란, 신동숙 옮김, *인간은 필요 없다*, 한스미디어, 2016.

[5] 아이작 아시모프, 김옥수 옮김, *아이, 로봇*, 우리교육, 2008.

[6] http://terms.naver.com/entry.nhn?docId=3578789&cid=58941&categoryId=58960

[7] http://techm.kr/bbs/board.php?bo_table=article&wr_id=4531

[8] http://slownews.kr/54060

[9] K Abney, *The Trends of Contents Techology* in *Robot Ethics: The Ethical and Social Implication of Robotic,* The MIT Press, 2012.

[10] 고인석, “로봇윤리의 기본 원칙,” 汎韓哲學. Vol.75, pp.401-426, 2014.

[11] 아리스토텔레스, 천병희 옮김, *니코마코스 윤리학*, 숲, 2013.

[12] 신영철, *문학의 윤리*, 문학동네, 2005.

[13] 웬델 월러치, 콜린 알렌, 노태복 옮김, *왜 로봇의 도덕인가*, 메디치미디어, 2014.

[14] 원종우, 이명현, 정지훈, 이창무, 권복규, 홍성욱, 이필렬, 이정모, *호모사피엔스씨의 위험한 고민*, 메디치미디어, 2015.

[15] http://asadal.bloter.net/20950

[16] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” Science, Vo.356, No.6334, pp.183-186, 2017.

[17] http://www.hankookilbo.com/v/a88e15f19eb94bf7a2f87e74f0ef2def

[18] 조미라, *애니메이션, 이미지의 모든 것*, 한국학술정보, 2014.

[19] 루이스 멈포드, 유명기 옮김, *기계의 신화*, 아카넷, 2013.

[20] 홍기빈, *인간의 위기와 자치 기획*, 문학동네, 제17권, 제2호(통권 63호), 2010.

[21] 미셸 푸코, 이정우 옮김, *지식의 고고학*, 민음사, 2000.

[22] 이왕주, *철학, 영화를 캐스팅하다*, 효형출판, 2005.

[23] 데이비드 겔런터, 현준만 옮김, *기계의 아름다움*, 해냄출판사, 1999.

[24] 알랭 드 보통, 말콤 글래드웰, 스티븐 핑커, 매트

리들리, 전병근 옮김, *사피엔스의 미래*, 모던 아카이브, 2016.

[25] D. Haraway, *The Companion Species Manifesto: Dogs, People, and Significant Otherness*, Prickly Paradigm, 2003.

저 자 소 개

유 은 순(Eun-Soon You) 정회원

- 1995년 2월 : 인하대학교 불어불문학과(문학사)
- 2007년 7월 : 프랑스 브장송 대학교 언어학(박사)
- 2016년 ~ 현재 : 중앙대학교 인문브릿지사업단 공동연구원
- 2017년 ~ 현재 : 인하대학교 인공지능 콘텐츠 창작연구센터 연구교수

<관심분야> : 기계번역, 인공지능, 빅데이터, 스토리텔링

조 미 라(Mi-Ra Cho) 정회원

- 2005년 : 중앙대학교 첨단영상대학원 영상예술학과(애니메이션 이론 박사)
- 2018년 현재 : 중앙대학교 인문브릿지사업단 공동연구원

<관심분야> : 영상미학, 애니메이션 이론, 서사 이론