인공 지능 기술의 사회적 이슈와 윤리 문제

한상기*

1. 서 론

세 번째 인공 지능의 봄이 왔다. 지난 20년 동안 컴퓨팅 환경과 기술의 발전, 엄청난 개인 생성 데이터를 통한 학습, 새로운 알고리듬 개발 등의 연구 노력에 힘입어 인공 지능 기술의 성능은 급격히 상승했고, 이에 따라 각 기업의 투자 증대와 새로운 스타트업의 등장, 각 국가의 전략적 연구가 정부 차원에서 이루어지고 있다.

딥마인드의 알파고가 인간을 능가하는 세계 최고 수준의 바둑 프로그램으로 등장한 것뿐만 아니라, 문장 전체를 하나의 단위로 번역하는 구글의 GNMT[1], 사진에 있는 객체의 유형과 형태를 인식하는 페이스북의 딥마스크와 샵마스크 또는 구글의 쇼앤텔, 매우 향상된 음성 인식과 가상 비석 역할을 해주는 시리와 구글 나우, 아마존 에코, 마이크로소프트 코타나, 음성 인식과 함께 통역을 제공하는 마이크로소프트 스카이프의 기능 등 과거와는 현저히 달라진 인지 기술이 속속 소개되고 있다.

구글, 페이스북, 마이크로소프트, IBM, 아마존 같은 글로벌 기업은 발 빠르게 여러 기업을 인수하면서 기술과 인재 확보에 나섰고, 인공 지능을 내세우는 기업도 전 세계에서 2,500개를 넘어서고 있다[2]. 벤처스캐너가 13개 분야에서 957개의 인공 지능 회사를 조사한 결과 총 투자 금액은 48억 불에 달하고, CB 인사이츠 조사에 따르면 2015년에 투자가 이루어진 경우만 397건에 투자 금액이 24억 달러에 달한다[3].

2014년 미국 아스펜 아이디어 페스티발에 모인 미래학자, 기술 전문가, 예술가들에게 '언제(로)봇이 세상을 점령할 것인가?'를 물어봤다[4]. 듀크 대학의 로봇 공학자이며 드론 전문가인 미시 커밍스 교수는 '이미 로봇이 우리를 점령했으며, 실제로 우리는 우리 세상에서 어디에 로봇이 있는 지 인지하고 있지 못하다'라고 답했다.

그러나 동시에 지수함수적으로 발전하는 기술에 대한 우려가 나타나고 있으며, 우리가 미처 대비하기 전에 인공 지능 기술이 우리 사회에 스며들 때 예상하지 못한 많은 사회적 이슈와 윤리적 문제를 야기할 지 모른다는 논의도 증가하고 있다.

대표적인 것이 옥스포드 대학의 철학자인 닉 보스트롬의 책 '초지능'[5]이 발간되면서 이에

※ 교신저자(Corresponding Author) : 한상기, 주소 : 서울특별시 강남구 언주로 86길 11, 1814호, 전화 : 010-2584-9383, E-mail : stevehan@techfrontier.kr

영향을 받은 엘런 머스크의 경고나 물리학자 스티브 호킹, 마이크로소프트 창업자인 빌 게이츠 등의 우려이다.

그러나 이런 장기적 잠재 위협이 아니더라도 현재의 기술이 내포하고 있는 갈등적 요소와 불분명한 정책과 제도, 준비되어 있지 않은 사회 시스템, 인간의 기본 특성과 성향에 의해 벌어질 수 있는 이슈 역시 무시할 수 없는 수준이다.

2016년 7월 미국 정부는 인공 지능의 미래를 준비하기 위한 정보 요청서를 발간하면서 기업이나 연구 단체 등에서 의견을 수렴했다[6]. 이 중에서도 상위 4개의 요청 사항은 인공 지능의 법률적 의미나 통치, 공공 선을 위한 사용 방안, 안전과 제어 이슈, 사회적 경제적 함의 등에 해당한다. 이는 이제 본격적으로 인공 지능이 갖는 사회적 함의를 정부 차원에서 살펴보겠다는 의미로 해석된다.

2016년 9월에는 글로벌 기업인 아마존, 페이스북, 알파벳(구글), IBM, 마이크로소프트가 인공 지능의 윤리 기준과 인간과 사회에 혜택을 주기 위한 방안, 사회적 영향력 등에 대해 협력하는 파트너십을 발표했다[7]. 이제 우리 사회가 인공 지능이 갖는 영향의 깊이와 범위 그리고 잠재적 이슈가 현안이 되었다는 것을 인식한 것으로 본다.

본 글에서는 이런 이슈를 알고리듬에 의한 편견과 차별의 문제, 기계 윤리, 안전성의 문제 등으로 나누어 접근해보고자 한다.

2. 주요 사회적 이슈

2.1 인공 지능 알고리듬에 의한 편견과 차별

이미 많은 서비스 내부에서는 인공 지능 기반의 소프트웨어 기술이 활용되고 있다. 블룸버그 베타에서 작성한 기계 지능 지형을 보면 대부분의 산업과 기업 내부 프로세스에 사용할 수 있는 기술을 개발하거나 활용하고 있다.

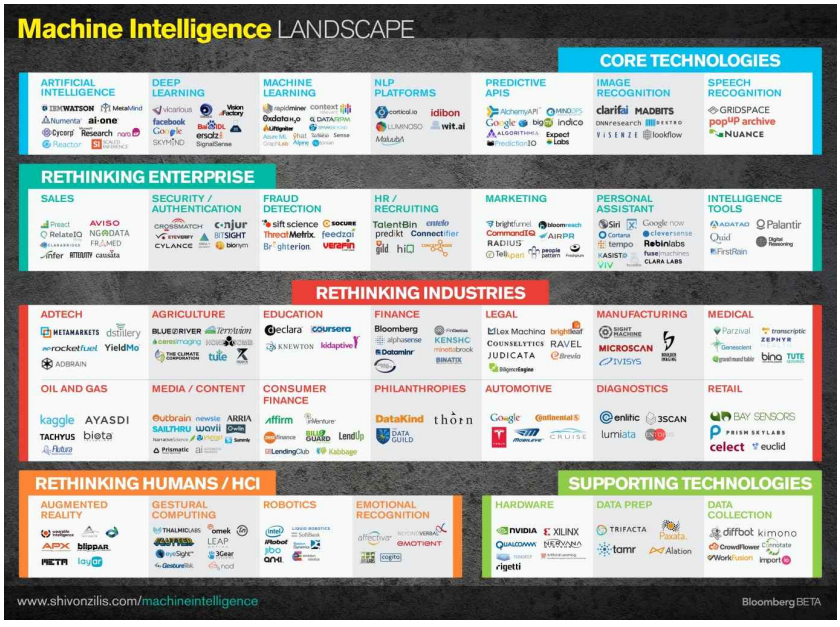

그림 1. 기계 지능 지형 [출처: 블룸버그 베타]

기업이 점차 데이터 확보에 의한 기계 학습으로, 다양한 출처의 데이터를 통한 분석으로 자신의 판단과 결정을 내리면서, 우리는 사용하는 데이터나 학습된 로직, 또는 프로그래머가 작성한 알고리듬이 편향되거나 특정 집단의 사고와 편견을 반영할 수 있고 그 결과 누군가는 불합리하게 차별 받을 수가 있다[8]. 사실 이 문제는 구글의 검색에서부터 나타난다. 영어로 '전문적인 여성 헤어 스타일'과 '비전문적인 여성 헤어 스타일'을 검색하면 전자는 대부분 백인 여성의 이미지가, 후자는 주로 흑인 여성의 이미지가 등장한다.

은행에서 대출 심사를 인공 지능이 대신하는 경우, 직접적인 개인의 인종이나 국적을 차별하지 않더라도 많은 관련 데이터를 분석하는 과정에서 사는 지역, 친구 관계, 주요 구매 패턴 등을 분석하면서 암묵적으로 인종 차별을 하는 판단을 내릴 수 있다. 기업은 위험을 최소화하고 이익을 극대화하기 위한 기술 활용이지만, 사람들은 자신이 차별을 받았고, 불공평하며, 그 기업이 비윤리적이라고 비난할 것이다.

수퍼 마켓의 점원이나 핀테크 회사가 인종이나 연령, 성별, 출신 지역 등을 갖고 차별할 생각이 없었더라도 우리가 사용하는 소프트웨어에는 그런 차별이 내포되어 있을 수 있다. 물론 소프트웨어 개발자는 전혀 의도하지 않았다고 생각할 수 있다. 그러나 개인이 가지고 있는 윤리적 판단 기준과 기업의 이익, 집단의 편향이 암묵적으로 코드에 반영될 수 있다.

때로는 엔지니어 입장에서는 어쩔 수 없는 인식률 문제이거나 작은 오류가 심각한 사회 문제나 기업 평판에 큰 타격을 줄 수 있다. 2015년 6월에 웹 프로그래머인 재키 알시네는 구글을 욕하는 트윗으로 올렸다. 그 이유는 구글이 제공한 구글 포토 서비스에서 자동 태깅 기술이 자신의 여자 친구 사진에 '고릴라'라는 태그를 달았다는 것이다[9]. 구글은 즉각 사과하고 해결을 약속했다. 개발자가 미처 생각하지 못한 오류에 의한 인종 차별 가능성이 나타난 것이다.

2016년 9월에는 페이스북이 자동으로 사진을 슬라이드쇼로 만들고 음악을 함께 제공하는 서비스에서 사용자의 치명적인 자동차 사건 사진에 즐거운 음악을 붙여서 문제를 일으켰다[10]. 사용자의 감성을 판단하지 못한 기술의 문제였다.

마이크로소프트의 테이 챗봇이 어떠한 배경 지식이나 판단 기준을 갖지 않은 상태로 시작했다가 하루 만에 인종 차별적인 트윗을 쏟아내는 사건이 벌어져 바로 서비스를 중단한 것 역시 준비되지 않은 인공 지능 알고리듬이나 서비스가 어떻게 쉽게 편견과 차별을 만들어 낼 수 있는 가를 잘 보여준 사례이다[11].

2016년 5월 미국 백악관이 발표한 '빅 데이터: 알고리듬 시스템, 기회, 그리고 시민 권리에 대한 보고서'에서 빅 데이터 분석과 머신 러닝이 신용 평가와 대출, 고용, 교육, 사회 정의 등의 영역에서 차별이 이루어질 수 있는 위험을 지적했다[12]. 백악관 보고서는 이런 차별을 방지하기 위해 향후 강력한 데이터 윤리 프레임워크의 개발을 요구하고 있다.

알고리듬에 의한 의도나 차별뿐만 아니라, 인공 지능이 지식의 시대에서 데이터의 시대로 넘어가면서 데이터에 의한 차별이나 왜곡 문제가 더 심각할 수 있다. 특히 딥러닝에 학습하는 데이터의 생성자나 활용자가 주로 선진국이거나 남성일 경우에 이는 예기치 않는 왜곡을 만들어 낼 수 있다. 현재 활동하는 딥러닝 전문가

중 여성이 비중이 13% 수준이라는 조사는 이런 잠재적 문제가 있음을 보여준다. 어떤 데이터로 학습하는 가에 따라 인공 지능 시스템의 판단 논리는 크게 달라지기 때문이다.

이를 해결하기 위해서는 정부뿐만 아니라 기업에서 개발자에 대한 윤리와 소프트웨어와 데이터의 왜곡과 차별 가능성에 대해 수시로 점검하고, 이를 외부 전문가와 함께 투명하게 평가 분석하는 제도 구현이 필요하다.

하버드 법대 로렌스 레식 교수는 1999년에 '코드와 사이버공간의 다른 법률'[13]에서 '코드가 법'이라는 명언을 남겼다. 그가 주목한 것은 컴퓨터 코드라고 부르는 소프트웨어 작성 내용이 갖는 정치적 의미이며, 결국 사이버 공간의 특성이나 규제를 만들어 내는 것은 코드에 달려있다는 점을 강조했다.

소프트웨어는 가치를 만들어 내고, 규제를 정하고, 개인의 권익 범위를 정한다. 인공 지능 소프트웨어는 의도하지 않더라도 우리 사회의 규범이나 윤리 기준을 만들고 있기 때문에 알고리듬이나 데이터가 사회의 규범을 지키고 있는지를 판단하고 이를 투명하게 검증할 수 있는 방안이 만들어져야 한다.

2.2 기계 윤리에 대한 다양한 접근[1)]

인공 지능 시스템이나 로봇이 도덕 기계가 될 수 있는 가에 대한 논의는 인공 지능 초기부터 거론된 문제이다. 이는 전통적으로 인공 지능이 인간을 대치하는 수준으로 개발할 것인가 아니면 인간의 지능을 강화하는 방향으로 개발할 것인가에 대한 큰 노선 차이에 따라 그 수준에 대한 논점 차이가 난다.

1) 이 섹션은 필자가 '슬로우뉴스'에 연재한 인공지능과 윤리 주제의 컬럼은 요약 반영한 것임을 밝힌다.

그러나 인공 지능 시스템이 어떤 윤리적 판단을 하지 않는다 하더라도 외부에서 관찰하거나 속해있는 공간에서의 행동과 판단이 외부적으로는 윤리적 판단을 하는 것으로 해석될 수 있기 때문에 인공 지능의 윤리 문제는 현 시점에서도 매우 중요한 연구 과제이다. 2.1에서 살펴본 단순한 알고리듬에 의한 의사 결정이 사회적으로는 예민한 이슈를 제기할 수 있기 때문이다.

2011년 닉 보스트롬과 엘리저 유드코프스키는 '인공 지능의 윤리'라는 에세이에서 인공 지능 윤리에 대한 논의를 1) 로봇 공학자의 전문가적 윤리 2) 로봇 안에 프로그램된 '모럴 코드' 3) 로봇에 의해 윤리적 추론이 이루어질 수 있는 자기 인식 능력을 의미하는 로봇 윤리로 구분해서 접근했다[14]. 나는 여기에 인공 지능 시스템의 사용자 윤리가 더해져야 한다고 생각한다.

인공 지능 연구자의 윤리 기준이 아직 어떤 강제성이나 모두 동의하는 수준에서 만들어진 것은 없지만, 1975년 아실로마 학술회의에서 유전자 조작 연구자들이 동의한 선언문이 좋은 참고 사례가 된다[15]. 그 결과 NIH 산하에 DNA 조작 자문 위원회가 만들어졌고 이후 이 분야의 연구 과정과 결과는 이 가이드라인를 준수하고 있다.

2009년 미국 인공 지능 학회에서 1차 인공 지능 연구자들의 윤리 가이드라인에 대한 논의가 있은 이후에 2015년 인공 지능학회지 겨울호에 스튜어트 러셀 등이 '튼튼하고 유익한 인공 지능을 위한 연구 우선 순위'라는 글을 통해 인공 지능의 사회적 이익을 극대화하는 방향으로 연구가 이루어져야 한다는 입장이 천명되었다[16]. 이후 생명의 미래 연구소에서는 이에

동참을 촉구하는 공개 편지가 제시되었다[17].

구글이 2014년 딥마인드를 인수할 때 딥마인드의 창업자들이 인수 조건으로 회사 내에 '윤리 위원회' 구성과 운영을 약속하라고 했다는 얘기는 널리 알려진 사실이다. 이미 많은 현대 기업은 '최고 윤리 책임자(Chief Ethics Officer)' 등의 직책이나 회사 내부에 '윤리·규정 준수 위원회' 등을 운영하고 있다.

2014년 포브스에 기고한 글에서 캘리포니아 폴리테크닉 주립 대학 철학 교수인 패트릭 린과 로체스터 공대의 철학 교수 에반 셀린저는 인공 지능에서 또 다른 윤리 위원회가 필요한 이유를 다음과 같이 설명하고 있다[18].

- 윤리는 단지 법적 위험에 대한 것이 아니다
- 외부와 내부 자문단 모두 장단점이 있다
- 홍보 차원이 아닌 진정한 접근 태도가 필요하다

그럼에도 불구하고 구글이 딥마인드를 인수한 지 2년이 지난 지금도, 딥마인드의 하사비스가 존재한다고 주장하는 내부 윤리 위원회가 누구로 구성되어 있으며 어떤 일을 하는지 구글은 공개하지 않고 있다.

셀린저 교수 같은 윤리학자도 구글이 좀 더 투명하게 운영해야 하며, 인공지능의 미래가 한 기업의 닫힌 문 뒤에서 논의되어서는 안 되고 기술 회사, 정부, 대중이 함께 논의해야 한다고 주장한다.

인공지능과 관련된 기술 개발이 인류의 윤리 기준이나 인간성 보호, 보편적 가치를 지키도록 감시하거나 이에 관련된 연구를 하는 조직으로는 각 대학 내 연구 조직, 국제적 기구 등 다양한 유형이 있다. 그 가운데, 오픈 로봇윤리 이니셔티브(ORi)는 로봇 공학의 윤리, 법률, 사회적 이슈에 대해 적극적 토의를 주도하는 싱크 탱크이다. ORi가 추진하는 캠페인 중 하나는 '킬러 로봇을 중단하라'는 캠페인이다.

옥스포드의 닉 보스트롬 교수가 주도하는 인류 미래 연구소(FHI)는 철학부가 주도하면서 학자들과 수학, 과학 등의 다학적 연구를 수행한다. 현재 12명의 풀타임 연구자가 있다. 주요 연구는 거시 전략, 인공지능 안전, 기술 예측과 위험 평가, 정책과 산업에 대한 연구들이다. 이 밖에도 앨런 연구소, 오픈 AI, 인간과 사회 이익을 위한 AI 파트너십 등이 윤리 문제를 연구하는 기관들이다.

인공지능 시스템에 어떠한 방식으로 윤리 코드나 엔진을 구현하고자 할 때 가장 먼저 떠오를 수 있는 방식은 보편적인 윤리를 규칙으로 설정해보는 것이다. 이는 칸트의 정언 명령과 같은 의무론적 윤리를 구체화하는 방식이 될 것이다.

과학계에서 의무론적 접근 방식으로 우리에게 가장 널리 알려진 사례는 아시모프가 제시한 로봇 3원칙. 1942년 그의 책 [아이, 로봇]에 나오는 단편 [술래잡기 로봇(Runaround)]에서 아시모프는 다음과 같은 '로봇공학 3원칙'을 제시한다.

1. 로봇은 인간에게 해를 가하거나 해를 당하는 상황에서 무시하면 안 된다.

2. 로봇은 1 원칙에 어긋나지 않는 한, 인간의 명령에 복종해야 한다.

3. 로봇은 1, 2 원칙에 어긋나지 않는 한, 자신을 지켜야 한다.

언뜻 보면 그럴듯해 보이는 이 원칙은 여러 영화에서 모순을 드러내거나 부조화를 이루면서 혼란을 초래할 수 있음이 알려졌다. 아시모프의 로봇 3원칙은 사실 제 1 원칙부터 어려운 문제를 안고 있다. 일단 '인간'을 어느 개념으로

정의할 것인지부터 쉬운 문제가 아니다. 인류는 한동안 다른 인종을 인간과 다르다고 분류한 적도 있고, 앞으로는 생물학적 특징만으로 인간을 정의하기 어려워질 수 있기 때문이다. 로봇이 이를 주입하거나 학습하는 것이 쉬운 일은 아니다.

또한, '해를 가한다'는 것을 판단하려면 우선 그 행동의 결과가 누구에게 해가 될 수 있는지를 알아야 한다. 하지만 모든 상황을 정확히 판단한다는 것은 불가능한 일이다. 주변의 인간에게는 해를 가하지 않아도 지구 어딘가에 있는 다른 인간에게 해를 끼칠 수도 있는데 이를 계산할 방안이 없다.

제 2 원칙에서는 인간의 발언 중 어디까지가 명령임을 알아내는 것 자체가 쉬운 일이 아니다. 제1원칙과 제2원칙을 따라서 행동한 것이 특정인을 구했지만, 그 결과로 인류에게 엄청난 파국을 일으킨다면 어떻게 할 것인가? 아시모프는 이 문제를 해결하고자 1985년 제0원칙을 추가한다.

• 제0원칙: 로봇은 인류에게 해를 가하거나, 또는 해를 당하는 상황을 무시해서는 안 된다.

결과주의적 윤리 또는 공리주의 역시 해결책을 제공하지 못한다. 최대 다수의 최대 행복이라는 표현에는 모든 사람의 행복을 어떻게 정량화할 것이며, 사람의 행복은 동등한 것인지, 모든 가능한 행동을 어떻게 계산해서 결과를 예측할 것인지 등에 대해 적절한 대답이 없다는 한계가 있다. 그 계산의 기간을 언제까지로 해서 결과를 생각해야 하는 것인지도 판단하기 어렵다. 또한, 비용과 효과로 분석한다고 하면 인간의 도덕적 가치를 경제적 가치로 환산할 수 없다는 반론에 부딪힌다.

웬델 월러치와 콜린 알렌은 '왜 로봇의 도덕인가'에서 공학 윤리, 인공 도덕 행위자 (AMA)에 관한 많은 논의와 접근 방법을 소개한 바 있다[19]. 전통적으로 소프트웨어 개발자는 논리 기반 접근, 사례 기반 접근, 다중 행위자 접근 등을 통해 연구해 왔으며, 주로 의사 결정 지원 시스템에서 구현한 사례들이다.

논리 기반은 하향식 접근으로 규범 논리를 통해 행위자가 무엇을 해야 할지에 관해 추론하도록 허용한다. 코네티컷 대학의 수전 앤더슨과 하트포드 대학의 마이클 앤더슨의 메드에스엑스(MedEthEx)는 헬스케어 종사자가 윤리적 딜레마에 직면했을 때 가이드를 제공하는 윤리 조언자로 개발했다[20].

카네기 멜론의 부르스 맥라렌은 트루스-텔러와 시로코(SIROCCO)를 통해 사례 기반 추론 기반의 윤리 추론 모델을 제시했다. 2005년 그의 논문에서 이 두 가지 시스템을 구현하면서 얻은 한계와 어려움을 얘기하고 있다[21].

시로코는 윤리적 행동을 이전 사례를 바탕으로 이끌어내려고 시도했고 전통적인 공학 전문적 윤리 코드에 바탕을 두었다. NPSE라는 전문 공학자 조직의 공학 윤리에 기반하면서 500건 이상의 사례 데이터베이스를 활용했다.

하향식 접근이 여러 가지 모순이나 한계가 있다면, 상향식 접근 방법은 어떤가? 상향식이란 결국 윤리적 판단 능력이나 윤리 행위자를 학습을 통해 구현하자는 뜻이다. 이는 센서 기반의 시스템이 인간의 행위를 파악하면서 그 가운데 윤리 기반의 행동을 확인하고, 어떤 행동이 윤리 양상을 가진다면 그에 관한 코드가 만들어지는 것을 의미한다.

스튜어트 러셀은 역강화학습(IRL; Inverse Reinforcement Learning)을 통해 이를 구현할 수 있다고 했다. 예를 들어 인간이 아침마다 물

을 끓여 커피를 타는 것을 반복하고, 이에 따라 기분이 좋아지는 것을 안다면, 커피 타는 행위가 코드로 들어갈 수 있다는 것이다[22].

벤자민 키퍼스는 '로봇을 위한 이간 같은 도덕과 윤리' 논문에서 자율적 행동을 할 수준의 로봇의 도덕과 윤리를 위한 아키텍처의 아웃라인을 제시하는데, 이는 인지 과학의 연구를 많이 받아들인 내용이다[23]. 그는 공리주의적 접근은 결국 공유지의 비극이나 죄수의 딜레마 같은 나쁜 결과를 산출할 수 있는 한계가 있음을 주장한다.

그는 인간이 가지는 도덕적 판단 모형은 빠르고, 무의식적이면 직관적 대응이 도덕적 판단을 주도하며 그다음에 더 느린 숙의적 추론이 이를 정당화하는 과정이라는 조나단 헤이트의 '사회적 직관주의 모델'[24]에 기반을 둔다.

조지아 공대 인터랙티브 컴퓨팅 학부 소속 마크 리들과 브렌트 해리슨은 미국 방위고등계획연구국(DARPA)와 해군 연구국의 지원으로 받은 연구로 이야기를 이용해 인공 행위자에게 인간 가치를 가르칠 수 있는 가능성에 대해 제시한다[25]. 키오테(Quixote)라고 부르는 이 시스템은 로봇이나 인공 행위자(에이전트)가 이야기를 읽고, 각 사건의 바람직한 결과를 학습해, 인간 사회에서 성공적인 행동을 이해하도록 훈련하는 시스템이다.

최근 자율 주행 자동차의 급속한 발전은 새로운 윤리적 딜레마에 대한 논의를 활발하게 만들었다. 이는 1967년 필리파 푸트가 제시한 트롤리 딜레마에서 파생한 것으로 자율 주행차가 어떤 위기위 봉착했을 때 과연 내부 프로그램에 의해 어떤 선택을 할 수 있는 가에 대한 논의이다.

자동차 운행을 위한 로직과 프로그램을 어떻게 정해야 할 것인가? 단순히 제조사에서 어떤 기준을 갖고 프로그램을 해야 할 것인지 아니며 회사 내부의 윤리 위원회에서 결정해야 하는지, 또는 사람들의 전반적인 의견을 수렴하는 방식이 되어야 할 것인지는 아직 누구도 결정하지 못하고 있다. 연구 방향 중 하나는 사람들이 다양한 상황에서 어떻게 판단하는 것이 좀 더 윤리적이고 사회적 합의를 이룰 수 있는 것인가를 파악하는 것이다.

2016년 6월 장 프랑수아 본느폰 등이 사이언스에 발표한 논문에서는 2천 명 정도의 사람들에게 6가지의 온라인 서베이를 통해 사람들의 판단 유형을 파악했다[26].

그 중 한 질문에서 한 사람의 보행자와 10명의 보행자의 목숨을 놓고 어떤 선택을 할 것인가를 물어봤을 때, 76%의 사람은 10명의 보행자보다는 1명의 목숨을 선택하겠다고 했다. 그러나 만일 이런 공리주의 윤리 기준으로 프로그램된 자율주행차를 구입하겠는가를 물어볼 때는 50%가 동의하지 않았다.

MIT 미디어랩의 '도덕 기계(Moral Machine)'라는 사이트에서는 다양한 상황에 대해서 어떤 선택을 할 것인지를 투표하게 하고, 사람들이 원하는 질문을 구성해 올릴 수 있게 하면서 사회적 윤리 기준을 확인하는 작업을 실행하고 있다[27]. ORi에서도 비슷한 방식으로 사람들의 다양한 투표를 통해 인간이 어떤 판단을 내리는 것이 보다 상식적인가를 확인하고 있다.

그렇다 하더라고 이런 판단에 참여하는 사람들이 특정 지역, 나이, 인종, 성별에 따른 편향성이나 왜곡 문제를 해결해야 하며, 어떤 시점에서 하나의 기준을 만들었다고 해도, 이런 사회적 합의는 시간과 지역에 따라 변화할 수 있다. 이런 방식은 '어느 시대 어느 지역에서 운행

되느냐에 따라서 계속 자신의 윤리적 판단을 수정해야 할 것인가?'라는 새로운 문제를 남긴다.

2.3 안전성의 문제

2016년 5월 7일에 미국 플로리다주 윌리스톤 고속도로에서 일어난 테슬라 사고는 자율 주행 자동차가 아직 완전하지 않다는 것을 모두에게 알리는 사건이었다[28]. 미국 고속도로교통안전청(NHTSA)의 예비 조사에 의하면 운전자 조슈아 브라운은 테슬라의 자율 주행 모드를 이용해 운전 중이었으나, 신호등이 없는 고속도로 교차로에서 흰색 트레일러를 하늘과 구별하지 못한 테슬라 차량은 그대로 직진해서 트레일러와 충돌해 브라운이 사망했다.

테슬라의 자율주행 기능인 '오토 파일럿' 모드는 베타 테스트 중이었고 모든 운전자는 오토파일럿 모드에서도 늘 평소와 같이 운전에 집중하는 방식을 취해야 하지만 사람들은 그렇게 행동하지 않았다.

오토파일럿을 사용하는 많은 사람들이 신문을 읽고, 영화를 보고, 잠들어 버리기도 했다. 이는 모험을 즐기는 일부 사람들의 본능이라는

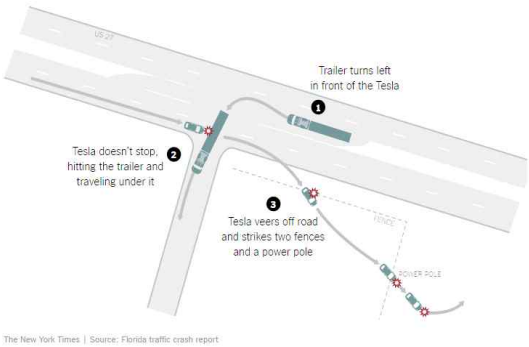

그림 2. 테슬라 사고가 일어난 과정 [출처: 뉴욕 타임즈]

것이 로봇 공학자들의 의견이다. 그러나 시스템을 개발하는 사람들에게는 매우 어려운 문제를 제기하고 있다.

자율 주행 자동차의 안전성은 차량 자체의 기술적 정확도 개선을 넘어서 다른 차량의 움직임이나 인간 운전자의 행동, 도로 상의 여러 개체들이 갖는 의미, 상황 인식이 이루어져야 하는 것이다.

아직 자율 주행 자동차는 안전성을 확보하기 위해 기술적으로도 여러 가지 문제를 안고 있다. 구글 브레인 프로젝트의 리더였고 지금은 바이두에서 인공 지능 연구를 지휘하는 앤드류 응 스탠포드 교수는 2016년 와이어드 잡지에 기고한 글[29]에서 자율 주행 자동차가 해결해야 하는 기술적 과제를 다음과 같이 소개하고 있다.

- 폭우, 짙게 낀 안개, 눈보라 같은 악천후에서 인식 기능은 아직 불완전하다.
- 공사장 같은 곳에서 인간 교통 안내원이 보내는 수신호를 제대로 인식하지 못한다.
- 밝은 햇빛이 역광으로 비출 때 신호등 인식이 어렵다.
- 아이스크림 트럭이 보이면 인간 운전자는 어린 아이가 등장할 수 있다는 예상을 통해 속도를 줄이지만, 아직 자율 주행 차는 이런 상황 분석을 하지 못한다.
- 자율 주행 차를 위한 면밀한 지도를 작성하고 유지하는 것은 아직 어려운 일이다.

미국 고속도로교통안전청에서도 2016년 4월 두 차례 공개 회의를 통해 자율 자동차의 안전한 채택에 대한 운영 가이드라인에 대해 논의했다. 2016년 9월 20일에 미 교통부는 자율 주행차의 안전 검사와 채택을 위한 정책안을 발표했다[30].

모바일 로봇의 안전 문제 역시 여러 가지 사고가 발생하면서 사회적 관심을 끌고 있다. 2016년 7월에는 미국 스탠포드 쇼핑 센터의 나이트스코프 K5 경비 로봇이 16개월 된 아이를 인지 못하고 부딪쳐서 상해를 입힌 사고가 발생했다.

러시아 페름에서는 프로모봇 실험실에서 로봇이 빠져 나와 길거리에서 교통 체증을 일으킨 사고가 발생했다[31].

인간 돌봄을 목적으로 하는 개인 돌봄 로봇에 대한 안전성 문제도 이미 국제 표준 기구인 ISO를 통해 안전 요구 사항이 논의되고 있다. 개인 돌봄 로봇을 세 가지 유형으로 구분하는데 1) 모바일 하인 로봇 2) 신체적 지원 로봇 3) 개인 이동 로봇 이다. 이들은 모두 나이나 능력과 상관 없이 의도된 사용자의 삶의 질을 개선하는 임무를 수행한다. ISO 13482:2014로 부르는 이 문서는 현재 리뷰 중이다.

인공 지능 안전에 대한 이슈는 사실 군사용 무기에 대한 것이 제일 예민한 상황이다. 2015년 국제 인공 지능 학술 회의를 통해 1,000 여 명의 학자와 오피니언 리더들이 자동화 무기 개발에 반대하는 의견을 공개 서한으로 밝혔다[32].

이런 안전성과 위험성에 대한 사고와 이를 방지하기 위한 노력은 앞으로도 지속될 것이고, 기업과 정부는 이런 사고 위험을 최소화하기 위한 여러 방안을 마련할 것이다. 학계에서도 인공 지능 안전에 대해 연구와 논의는 지속적으로 해 왔으며, 머신 러닝 구현 과정에서 벌어질 수 있는 다양한 문제점을 평가하고 분석해왔다[33].

그림 3. 나이트스코프 K5 경비 로봇 〔출처: 나이트스코프〕

그러나 중요한 것은 어떤 상황에서도 인간이 최종적 제어를 할 수 있어야 한다는 인식이 많은 전문가들에게 동의를 얻고 있다. 최근 딥마인드의 로랑 오소와 옥스포드의 스튜어트 암스트롱이 발표한 논문에서도 소위 말하는 '커다란 빨간 버튼'을 통해 지속적으로 위험한 행동을 하는 에이전트를 방지하고 더 안전한 상황으로 전환하는 방안에 대해 소개했다[34].

3. 결 론

인공 지능이 함의하고 있는 사회적, 윤리적 이슈는 매우 광범위하다. 어떤 것은 우리가 아직 상상하거나 예측하지 못한 주제들도 있다. 이는 어쩌면 새로운 사회 구조를 만들고 있기 때문이며, 앞으로 만들어진 사회가 인공 지능 행위자를 새로운 주체를 가정해야 하기 때문이다.

윤리적 판단이 인간이 갖는 공감이나 연민 또는 불쾌함 같은 감정을 수반하거나 그 결과로 얻어지는 사회와 공동체의 합의라고 한다면, 인공 지능의 윤리 학습은 감정 상태에 대한 시뮬레이션을 필요로 할 수 있다.

인공 지능 윤리가 기술적으로 깊이 있는 연구 단계가 도달하지 못했지만, 오히려 인간이 갖는 특성 때문에 현재 제한적 지능을 가진 시스템으로도 여러 가지 이슈들이 발생할 것이다.

다만 아직은 그 결과가 사회에 큰 피해를 주거나 파국을 가져오는 것은 아니고 일부 사람들에게 차별을 가하거나, 반(反)사회적 행동을 유발하고, 오해와 확대 해석으로 인한 오용의 문제가 발생할 것으로 본다.

초기의 많은 문제는 인공 지능을 사용하는 또는 같이 공존해야 하는 인간이 갖는 특성 때문에 발생할 것이다. 이는 인류 초기부터 우리가 갖는 특성인 의인화와 감정 애착 또는 이입이라는 특징때문에 발생한다.

1944년 오스트리아의 심리학자인 프리츠 하이더(Fritz Heider)는 흥미로운 실험을 했는데, 두 개의 삼각형과 하나의 원이 움직이는 것을 보여주었더니 단지 큰 삼각형과 작은 삼각형이 원을 따라 움직이는 모습에서도 사람들은 질투, 두려움, 경쟁 등을 표현하면서 감정 이입했다[35]. 이를 본 아이들은 큰 삼각형이 원을 쫓아다니며 괴롭힌다고 해석했다.

의인화와 감정 이입 또는 애착은 우리 인간이 갖는 특징 중 하나이다. 따라서 우리는 앞으로 만들어지는 여러 수준의 시스템이나 로봇이 우리와 소통하거나 우리 행위에 따라 반응하고, 우리와 유사한 지각을 보여줄 때 자연스럽게 그 대상을 인간처럼 이해할 것이다.

또 하나 우리가 진화를 통해 발전한 것이 상대방의 의도와 생각을 우리 안에서 추론하는 성향으로 진화심리학이나 뇌 과학에서 말하는 '마음 이론(Theoryy of Mind)'이다. 따라서 우리는 어떤 존재를 의인화하면서 자연스럽게 그 마음을 이해하거나 추론할 것이기 때문에, 인공 지능 시스템이 어떤 의도나 생각으로 행동한다고 이해할 것이다.

이는 자연스럽게 인공 지능 시스템이 보이는 행동이나 의사 결정이 윤리적이거나 비윤리적이라고 우리가 생각하게 될 것이다. 어떤 윤리 엔진을 탑재했거나 윤리 코드를 반영하지 않아도 우리는 소프트웨어나 로봇의 판단과 행동에 윤리적 판단을 적용할 것이다.

인공 지능 기술의 발전이 우리 사회의 이익을 극대화하고 공동체 전체의 혜택으로 이루어지게 해야 한다는 명제는 너무도 타당하지만 사실 위협 받을 가능성을 잠재적으로 갖고 있다. 앞으로 우리가 이런 문제를 대처하기 위해 준비해야 하는 우선 과제는 어떻게 컴퓨터 과학자나 인공 지능 연구자들이 윤리 문제에 대해 관심을 갖게 할 것인가 하는 점이다.

참 고 문 헌

[1] Quoc V. Le & Mike Schuster, "A Neural Network for Machine Translation, at Production Scale," Google Brain Team, Google Research Blog, Sept. 27, 2016.

[2] Shivon Zilis, "The Current State of Machine Intelligence," Bloomberg Beta, Dec. 11, 2014

[3] Artificial Intelligence Explodes: New Deal Activity Record for AI Startups, CB Insights, June 20, 2016.

[4] When Will Robots Take Over the World, Aspen Ideas Festival, The Atlantic, July 17, 2014.

[5] Nick Bostrom. Superintelligence: Paths, Dangers, Strategies. Oxford University, July 3, 2014.

[6] Office of Science and Technology Policy (OSTP), "Request for Information: Preparing for the Future of Artificial Intelligence," the White House, July 2016.

[7] John Markoff, "How Tech Giants Are Devising Real Ethics for Artificial Intel-

ligence," New York Times, Sept. 1, 2016.
[8] Nabette Byrbes, "Artificial Intolerance," MIT Tech. Review, March 28, 2016.
[9] Alistair Barr, "Google mistakenly tags black people as 'Gorillas,' showing limits of algorithms," The Wall Street Journal, July 1, 2015.
[10] Brian Koerber, "Facebook automatically turns man's near-fatal car crash into a happy slideshow," Mashable Sept. 12, 2016.
[11] James Vincent, "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day," The Verge, March 24, 2016.
[12] The White House, "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights," Executive Office of the President, May 2016
[13] Lawrence Lessig. Code and Other Laws of Cyberspace. Basic Books, 1999.
[14] Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," Draft for Cambridge Handbook of Artificial Intelligence, (eds.) William Ramsey and Keith Frankish, Cambridge University Press, 2011.
[15] Paul Berg, "Meetings that changed the world: Asilomar 1975: DNA modification secured," Natyre 455, pp. 290-291, Sept. 18, 2008.
[16] Stuart Russell, Daniel Dewey, and Max Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," AI Magazine, AAAI, Winter 2015.
[17] http://futureoflife.org/ai-open-letter/
[18] P. Lin and E. Selinger, "Inside Google's Mysterious Ethics Board", Fobes, Feb 3, 2014.
[19] 웬델 월러치, 콜린 알렌. 왜 로봇의 도덕인가, 메디치미디어, 2014 (원제: Moral Machines: Teaching Robots Right from Wrong, 2009)
[20] M. Anderson, S. Anderson, C. Armen, "MedEthEx: A Prototype Medical Ethics Advisor," IAAI 2006, AAAI 2006.
[21] McLaren, B.M. (2005). Lessons in machine ethics from the perspective of two computational models of ethical reasoning; Presented at the AAAI Fall 2005 Symposium, Washington, D. C. In Papers from the AAAI Fall Symposium, Technical Report FS-05-06. (pp. 70-77).
[22] Stuart Russell, "Should We Fear Supersmart Robots," Scientific American, June 2016.
[23] Benjamin Kuipers, "Human-like Morality and Ethics for Robots," in Proceedings of the 2nd International Workshop on AI, Ethics and Society, 2016.
[24] Haidt, J. 2012. The Righteous Mind: Why Good People are Divided by Politics and Religion. New York: Vintage Book.
[25] Mark Riedl and Brent Harrison, "Using Stories to Teach Human Values to Artificial Agents," in Proceedings of the 2nd International Workshop on AI, Ethics and Society, 2016.
[26] Jean-François Bonnefon, Azim Shariff, Iyad Rahwan, "The social dilemma of autonomous vehicles", Science, Vol. 352, Issue 6293, pp. 1573-1576, 2016.6
[27] http://moralmachine.mit.edu/
[28] Jordan Golson, "Tesla driver killed in crash with Autopilot active, NHTSA in-

vestigating," The Verge, June 30, 2016.
[29] Andrew Ng and Yuanqing Lin, "Self-Driving Cars Won't Work Until We Change Our Roads - and Attitudes," Wired, Mar 15, 2016
[30] NHTSA. Federal Automated Vehicles Policy: Accelerating the Next Revolution in Roadway Safety, Sept. 20, 2016.
[31] http://www.bbc.com/news/blogs-news-from-elsewhere-36547139
[32] http://futureoflife.org/open-letter-autonomous-weapons/
[33] Dario Amodei, et. Al., "Concrete Problems in Ai Safety," Arxiv.org, June 25, 2016.
[34] Laurent Orseau and Stuart Armstrong, "Safely Interruptible Agents," In Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016), edited by Alexander Ihler and Dominik Janzing, 557-566. Jersey City, New Jersey, USA.
[35] Heider, F; Simmel, M., "An experimental study of apparent behavior," American Journal of Psychology 57, pp. 243-259, 1944.

한 상 기

• 1984년 : 서울대학교 컴퓨터공학과 (공학사)
• 1989년 : KAIST 전산학과 (공학박사)
• 1989년 - 2016년 : 삼성전자, 다음커뮤니케이션, 카이스트, 세종대 교수 등 역임
• 2011년 - 현재 : 소셜컴퓨팅연구소 대표
• 관심분야 : 인공지능, 사물인터넷, 소셜미디어