커뮤니케이션 이론
2023년 19권 4호 109~157쪽
https://doi.org/10.20879/ct.2023.19.4.109

인공지능 기술 신화로서 의인화 비판
챗봇 '이루다' 사례 연구*

김현준 서울과학기술대 IT정책전문대학원 박사과정수료**

이광석 서울과학기술대 IT정책전문대학원 교수***

이 글은 인공지능(AI) 의인화가 오늘날 인공지능 기술신화의 강력한 구성 요소이자 기제라고 보고, 기술신화의 생성과 재생산 방식을 드러내는 사례로서 '챗봇 이루다' 사태를 구성주의적 관점에서 비판적으로 분석한다. 이 글은 이루다 사례 분석을 통해, 우리 사회에 팽배한 인공지능에 대한 의인화 담론이 대체로 그것의 기술공학과 기술문화의 이중적 측면을 분리시킴으로써 AI의 기술중립론과 기술물신을 재생산하고 강화하는 경향이 있음을 살핀다. 다만 AI 의인화는 단지 개발사나 특정 행위자의 전략적 의도만으로 환원될 수 없고, 기술을 둘러싼 논쟁 과정 속에서 여러 행위자(소)들이 상호 얽힌 사회적 구성의 효과임을 살핀다. 오늘날 지능 정보화 기술의 논리는 '인간다운 인공지능' 신화와의 관계 속에서 독립적이고 자율적인 인공 사물의 외양을 띠게 됐다. 구체적으로, 이루다 챗봇 기술은 AI 기술 자체, 개발사, 정부 규제기관, 사회(익명의 이용자)의 하이브리드(혼종) 네트워크에 의해 공동 구성됨에도 불구하고, '인간다운 인공지능'(의인화) 논리에 의해 그 관계적 측면이 은폐됨으로써 역설적으로 인공지능 기술신화를 완성하거나 확대하는 데 기여한다. 따라서 이 글은 인공지능이 지닌 사회적으로 다층적 관계의 문제 지점이 좁은 의미의 기술 결함과 그것의 해소 문제로 축소되는 현실을 기술신화로 독해해내고 이 수사학을 비판

* 이 논문은 서울과학기술대학교 교내 과제(2021-0808) 연구 지원을 받았음. 심사 과정에서 세심하고도 건설적인 논평과 제안을 해준 세 분 심사위원들께 이 자리를 빌어 다시 한번 감사드린다.

** 제 1저자, hyunjun79@daum.net

*** 교신저자, kslee@seoultech.ac.kr

하고자 한다. 이를 위해, 이 글은 '이루다' 사건에서 기술결정론을 강화하기 위해 동원되었던 인공지능의 의인화 과정, 즉 페르소나(1차 의인화), 성적 대상화(2차 의인화), 어린아이론(3차 의인화)를 거쳐, 최종적으로 개인정보보호위원회의 이루다 사건 정책 개입에 이르기까지 단계별로 의인화 과정을 심층 분석한다.

KEYWORDS 인간다운 인공지능, 챗봇, 의인화, 기술중립론, 기술신화, 구성주의, 하이브리드, 행위자-연결망 이론

1. 서론

인간과 컴퓨터 사이의 상호작용을 돕고 인간이 비인간 존재에 접근하기 위한 하나의 방식으로서 '의인화(anthropomorphism)'는 대중문화의 흔한 소재가 되었고 관련 연구분야에서도 널리 인정되어왔다(임종수·최진호·이혜민, 2020, 439-440쪽 참조). 인간과 인공지능 로봇의 상호 얽힘이 증대하는 포스트휴먼 시대가 도래하면서, 인간과 비인간 행위자의 평평하고 상호 의존적인 관계를 강조하거나 비인간 객체에 대한 '의인화' 접근이 흔해졌다.

의인화는 비인간 행위자이자 사물 객체를 이해하는 매력적인 방식일 수 있다. 의인화는 오늘날 인공지능을 쉽게 이해하기 위한 대중적 접근에서 장려되지만, 과학계와 산업계, 그리고 정부 등 인공지능을 주도하는 행위자들이 이에 대한 인식을 재고하고 그 표현을 제한하려는 노력을 기울이지 않으면 인공지능에 대한 오해를 불러일으킬 수도 있다. 가령 인간과 인공지능 사이의 심리적, 정신적 유사성을 지나치게 강조하여 인공지능을 인간 마음의 복제물로 환원하게 되면 인공지능이 야기하는 문제에 대한 결함을 주로 도덕 윤리적 문제로만 바라보고 제한적인 해법에 집중하는 우를 범할 수 있다(Salles, Evers, & Farisco, 2020, p. 94).

이 연구는 인공지능, 특히 챗봇에 대한 의인화를 기술신화의 구성요소 중 하나로 본다. 챗봇 이루다를 둘러싼 의인화 담론은 인공지능의 사회적, 정치적, 문화적 설계 특성을 은폐함으로써 기술적 투명성의 신화를 정당화하고 강화하는 '기술중립론(techno-neutralism)'으로 이용된다. 여기서 기술중립론이란 기술 구성의 사회·문화·정치적 차원을 외면하고 기술을 가치중립적이고 탈정치적으로 오인하여 기술을 오롯이 좁은 의미의 기술공학적 메커니즘만으로 환원하는 태도이다. 오늘날 우리사회의 인공지능에 대한 기술중립론적 기술신화를 비판하기 위해 이 연구는 2021년 초 우리 사회를 뜨겁게 달구었던 인공지능(AI) '챗봇 이루다' 사건에 주목

한다. 그리고 이 사건을 이해하는 문제적 담론으로서 '의인화'를 행위자-연결망 이론(이하 ANT)를 위시한 구성주의적 관점에서 다뤄보고자 한다. ANT와 구성주의적 관점에서 인공지능의 의인화는 단지 개발사나 특정 행위자의 의도적인 전략으로만 이해될 수 있는 것이 아니라, 사회적인 과정 속에서 비인간 행위자(기술, 알고리즘 모델)와 인간 행위자(다양한 이용자와 제도)가 함께 만들어내는 기술과 사회의 '공동구성', '공진화', '네트워크', '혼종성(hybrid)', '연합', '배치', '집합체(assemblage)'의 효과로서 이해될 수 있다. 특히 챗봇 이루다를 기술-사회 하이브리드로서 이해하기 위해 ANT와 사회구성주의 간의 차별적 쟁점은 배제하고, ANT를 구성주의적(constructivist) 흐름 속에서 이해하며 그 주요 개념들을 활용하고자 한다(Pinch & Bijker, 2012 참조).

이 글에서 '구성주의적' 시각은 오늘날 암흑상자로서 여겨지는 인공지능 알고리즘과 사회의 '관계적 얽힘'의 요소들을 분석적으로 들여다볼 수 있는 하나의 방법론적 렌즈로서의 의미를 지닌다. 더불어, 인공지능 기술에 대한 환상과 기술권력을 해체하는데 필요한 비판적 관점으로서 의의를 갖는다. 집합체로서의 기술적 대상을 비판적으로 파악하는 구성주의적 렌즈를 통해, 의인화에 의해 점점 더 비가역적인 암흑상자나 신화가 되어가는 우리 사회의 인공지능 기술의 본모습을 드러낼 수 있다고 보는 것이다. 다시 말해 의인화 담론이 개발사와 정부기관 등 주요 행위자들을 통해 기술문화와 기술서사적 차원에서 지능형 기술 그 자체와 어떻게 내재적으로 통합되면서 기술의 실제적 본모습에 대한 이해를 흐리는지를 엿보려 한다.

본 연구는 궁극적으로 챗봇 이루다의 의인화 담론이 개발사의 이루다 '페르소나'의 설정 기획(마케팅 전략)과 이용자(특히 오남용자)들의 공모, 그리고 정부 기관(개보위)의 의도하지 않은 행위의 집합적 배치의 효과로 인해 재생산되고 강화되면서, 챗봇 서비스 기술에 깊게 얽힌 사회문화적 측면을 배제하는 데 작용하여 인공지능 의인화는 물론이고 기술

중립론이라는 기술신화를 보다 공고히 한다고 주장한다. 즉 의인화 담론을 구성하는 행위자들의 네트워크가 기술 내부와 외부의 경계와 범주를 재배치하고 재생산함으로써 사회의 암묵적 믿음으로서 인공지능 의인화를 공고히 하고 이 신화의 담지체로서 이루다라는 비인간 행위자를 재구성한다고 파악하는 것이다. 이것이 '기술-사회 하이브리드'로서 인공지능의 사회적 존재론이다. 인공지능에 대한 사회적 정의, 즉 기술신화는 개발사의 의도적인 기획 의도나 알고리즘 설계와 코딩만으로 구성되지 않고, 알고리즘이라는 기술적 객체를 의인화하는 사회기술적 '매개' 또는 '번역(tradiction; translation)' 과정을 통해서 관철된다. 이 글에서는 이 구성의 과정을 ANT를 따라, 다양한 이종적인 행위자들에 의해 구축되는 '연합(또는 공모)'과 '번역'의 행위로서 살핀다(Latour, 1999/2018, 6장 참조; 2012, 36-45쪽 참조). 결과적으로 이 연구는 챗봇 이루다 사태를 통해서 우리 사회의 기술신화가 인공지능을 어떻게 사회·문화적 차원과 분리된 중립적인 기술물신 자체로서 정당화하고 유지·재생산하는지를 의인화 담론의 구성을 통해 보여주고자 한다.

2. 이론적 고찰: 사회문화적 기술코드로서 의인화의 구성

2021년 인공지능 제품 개발 스타트업 회사 스캐터랩에서 개발·출시한 챗봇 이루다는 등장부터 사회적으로 많은 논란을 야기했다. 이루다는 자유대화형(open-domain conversational AI)챗봇으로서 마치 "인간 같은 대화"를 추구하도록 설계되었다. 개발사는 챗봇 이루다를 "관계의 불평등을 해소"하기 위한 목적으로 개발했다고 주장한다. 개발사에 따르면, 관계의 불평등을 해소한다는 것은 사람처럼 순수한 대화가 가능한 비인간 행위자 그 자체를 목적에 둠으로써 인공지능과 진정한 "친구"가 되는 것을 의미한다. 그리고 '인간다운 인공지능'이라는 목적 혹은 가치를 실현

하기 위해 개발사는 데이터 검색 모델(retrieval-based NLP)을 핵심 기술로 채택하게 된다.

이루다와 같은 챗봇은 인간과 상호작용을 목적으로 만들어진 '사회적 로봇(social robot)'이자, 'AI 미디어'라고 할 수 있다(임종수 등, 2020, 439쪽). 여기에서 의인화는 '사회적 로봇' 또는 '인간-컴퓨터 상호작용(Human-Computer Interaction; HCI)'에서 두드러진 현상 가운데 하나이다. 최초의 상담 챗봇 '일라이자'의 사례에서 보고된 것처럼,[1] 인공지능이 종종 의인화된다는 것은 이미 잘 알려진 사실이다(Salles et al., 2020, p. 88). 컴퓨터를 사람과의 상호작용 관계의 측면에서 연구하는 '사회적 행위자로서 컴퓨터(computers as social actors; CASA)' 패러다임도 이러한 의인화 현상에 주목해 왔다(Reeves & Nass, 1996 참조).[2]

의인화란 "일반적으로 인간과 유사한 감정, 정신 상태, 행동 특성을 무생물, 동물, 그리고 일반적으로 자연 현상과 초자연적인 존재에 부여하는 것"이다(Salles et al., 2020, p. 89). 의인화는 합리적으로 이해하기 어려운 존재(사물)를 이해하는 하나의 방식이다. 더피는 의인화를 심리학적 차원에서 비인간 매체(인공지능)의 행동을 합리화하기 위해 인간의 특성을 그것에 귀속시키는 것이라고 설명한다(Duffy, 2003, p. 180). 의인화는 사물의 행동을 합리적으로, 즉 인간의 특성을 사물에 투사하여 설명하기 위한 방식이라는 것이다.[3] 심지어 의인화는 예외적인

---

1) 의인화는 '일라이자 효과(Eliza effect)'로 불리기도 한다. 일라이자는 정신과 의사처럼 환자를 상담하는(기계학습 추론시스템이 아닌) 단순한 전문가 시스템 알고리즘의 챗봇이다. 이 효과는 사람들이 자신의(감정적) 발화에 대해 컴퓨터가 단순히 맞장구를 치는 것만으로도, 심지어 사람들이 이 컴퓨터가 기계임을 의식적으로는 인지함에도 불구하고 무의식적으로 인간의 행위와 유사한 것으로 추정하는 의인화가 일어남을 보여주었다. 이에 대한 자세한 설명은 다음을 참조하라(Natale, 2021).

2) CASA 패러다임을 따른 비교적 최근의 사례로는 알파고에 대한 의인화를 언론 보도의 의미네트워크 분석을 통해 밝힌 연구가 있다(임종수 등, 2017).HCI 및 CASA에 대한 소개와 인공지능 의인화 남용에 대한 연구 사례 소개로는 진보래(2020)을 보라.

믿음이나 인간 본성의 불가피한 결과가 아니라, '사회적 인식'의 매우 일반적인 과정으로 볼 수 있다(Epley, 2018). 그런데 미디어 사회학적 또는 과학기술학적 관점에서 보면, 이 사회적 인식의 일반화 과정인 '합리화'는 사실상 그러한 인식을 합리적인 것으로 만드는 사회적 과정을 통해서 이루어진다. 그렇다면, 인공지능의 의인화는 단순히 심리적, 인지적 합리화가 아니라, 보다 체계적인 영역에 영향을 행사하는 사회적 합리화 과정의 산물로 볼 수 있다. 그리고 HCI나 CASA 패러다임의 관점에서 보자면, 의인화에는 합리화 이상의 요소, 즉 인간처럼 상호작용할 수 있는 속성을 인공사물에 부여하는 실천이 포함된다. 즉 의인화는 사회적 상호작용, 커뮤니케이션적 요소를 통해 이루어지는 것이다(임종수 등, 2017, 120쪽). 이런 관점에서 의인화란 결국 상호작용하기 어려운 대상을 상호작용할 수 있는 대상으로 만들어주는 과정인 것이다. 따라서 인공지능 챗봇의 커뮤니케이션 기능은 마치 '인간 같은' 대화의 실감 상태를 확보하기 위해 자체 의인화와 밀접한 관계가 있을 수밖에 없다.

오늘날 디지털 환경의 가속화는 비인간 객체인 기술에 대한 의인화를 다양한 방식으로 실현시키는데, 주로 기술적 대상들에 인간의 형상과 특징을 부여하는 방식으로 구현시키고 있다(Stojnić, 2015, p. 70). 그 결과, 기술의 의인화는 피할 수 없는 사태가 되어버렸고, 이는 인간의 기술의존성이 의인화라는 '이해가능한' 기술정치적 질서에 따라 결정되게 되었다는 역설을 드러내는 것이다(Stojnić, 2015, p. 77). 즉 의인화는 그것의 합리적인 특성에 힘입어 기술의 영향력을 보다 확대하는 요인이

---

3) 의인화를 의인관(anthropomorphism)과 인격화(personification)로 구분하기도 하지만, 본고에서는 임종수 등(2017)의 견해를 따라, 구별하지 않고 사용한다. 임종수 등(2017, 118쪽, 각주 3)에 따르면, "의인화는 한국의 언어생활에서 의인관과 인격화 모두를 지칭한다. 현실적으로 세계를 인식하는 인간의 의인관적 속성은 구체적인 인격화 활동을 통해 구체화된다. 인간이 알파고를 의인화했다면, 그것은 알파고가 바둑을 두는 인간의 의인관적 관점에서 파악했다는 것이고 구체적으로 어떤 성격이나 성별, 인지수준, 태도 등 인간 고유의 감정이나 의지, 태도, 관점 등의 인격화가 이루어졌다는 것이다. 이 두 개념은 의인화 용어 안에서 동시에 적용된다."

되는 것이다. 이러한 견해에 따르면, 인공지능과 같은 디지털 기술은 의인화를 구성할 뿐만 아니라, 의인화를 통해서 존재하는 것이다. 따라서 의인화는 한 사회가 갖는 기술신화의 반영이자, 이 기술신화의 메커니즘을 들여다보는 투명한 창인 셈이다. 이것이 이 글이 문제삼고자 하는 의인화이다.

일반적으로 의인화는 사물 자체의 특징과 필연적인 상관관계를 갖거나 물리적, 존재론적 상태에 크게 의존하지 않는다(Salles et al., 2020, p. 89). 애플리, 웨이츠, 그리고 카치오포(Epley, Waytz, & Cacioppo, 2008)의 심리학적 견해에 따르면, 의인화는 존재하는 물질적 특질이나 행동을 묘사하기보다는, 사회적 유대를 형성하고 주변세계와의 상호작용을 위해 직접 관찰되는 것을 넘어서서, 존재하는 물리적 특질과 행동을 인간처럼 해석하는 귀납적 추론 과정에 속한다. 사람들은 사회적 관계의 결핍을 보상받거나 예측불가능한 환경을 이해하고자 할 때 의인화를 취하게 된다. 마찬가지로 인공지능 기술도 일반인들에게는 불투명한 상태로 남아있기에 이를 이해하고자, 또는 전문가들 입장에서는 대중화에 유리하기에 의인화된다. 하지만 인공지능은 그 개념 자체에서부터 의인화를 내포하고 있을 뿐만 아니라, 자유대화형 챗봇과 같이 인간중심적 모델인 경우에는 보다 손쉽게 의인화될 수 있기에 이로인한 문제에 더 취약하다(Salles et al., 2020, pp. 89-91).

의인화가 직접 관찰되는 사물의 물리적 특성이나 움직임을 묘사하는 것이 아님에도 불구하고 인공지능에 관한 우리의 언어는 사물 자체의 물리적, 기계적 구조나 논리를 넘어선다는 점에서 인공지능 의인화를 재고해 볼 여지가 있다(Epley et al., 2008 참조). 더욱이 최근 초거대 언어모델(LLM)을 위시한 인공지능의 급격한 발전은 의인화를 보다 진지하게 고려하게 만들었다. 구글 엔지니어 르모인(Blake Lemoine)이 람다(LaMDA)가 지각(sentience)이 있다고 주장했던 일이 이와 연관된 대표적인 사례다.

하지만 이루다의 국내 사례에서 인공지능 의인화는 인공지능(알고리즘) 모델의 기술공학적인 동시에 문화적인 설계 특징과 직접적으로 연관성을 갖는다는 점에서 특이한 현상이다. 더 정확히 말하자면 인공지능 의인화는 이루다의 '여성형' 페르소나에서 단적으로 드러나는데, 이것은 단지 이용자들이 이루다를 여성으로 인식하는 것에 달린 문제만이 아니라, 개발사의 기술 설계 및 기획과의 상호작용의 결과인 것이다. 즉 이루다의 여성형 페르소나는 기술공학적 요소(알고리즘 기술 자체의 특징)와 개발사와 이용자, 그리고 정부기관이 갖고 있는 사회문화적 요소의 공동구성 과정을 통해서 구현되었다고 볼 수 있다. 이루다 페르소나의 공동구성 과정 속에서 재확인되고 강화되는 의인화는 챗봇의 알고리즘 기술 설계 자체를 특정한 방식으로 안내하고 정교화하는 데 다시 기여하게 된다.[4] 이렇게 의인화의 공동 구성과정을 통해서 관철되는 인공지능에 대한 의인화 신화는 단순한 의인화 은유의 심리적 힘을 넘어서는 것이다.

문제는 의인화라는 은유가 불가피하다라든가(Caporael, 1986; Krementsov & Todes, 1991 참조), 의인화가 단순히 챗봇 서비스 대중화나 마케팅에 이점이 있다는 것이 아니다. 여기에서 문제는 기술설계의 예측 불가능성이나 모호성 같은 어떤 기술 서비스 문제점들에 대한 은폐와 정당화로서 의인화가 이용되는데 있다. 의인화의 자연스러움은 기술이 의인화와 결합할 때 그 이데올로기적 효과마저 자연스러운 것으로 용인하게 되는 데서 극대화한다. 슈나이더만(Shneidoman, 1989)은 개발자들이 자신들의 기술설계에 무차별적으로 의인화의 특성을 적용하려

---

4) 가령 가상비서라는 인공지능 비서의 의인화는 단지 기계가 비서의 기능적 차원만을 실행하도록 돕는 정도가 아니라, 인간비서와 같은 기능적, 문화적, 정서적 요구를 점검하고 되먹임하여 보다 복잡한 '지능'이라 실행 능력을 갖출 수 있도록 만드는 기술 설계의 실제적(practical) 전략이다. 즉 인공지능 비서의 의인화는 기능과 무관한 문화적, 심리적 기제가 아니라, 그 기능을 가능케 하는 설계의 문화적 실천인 것이다. 왓슨(Watson, 2019) 역시 인공지능 의인화 수사를 비판하면서도 그것이 기계학습 전략을 수립하고 인공지능 연구에 새로운 영감을 줄 수 있다는 점은 인정한다.

는 데에서 오는 문제점을 지적한 바 있다. 그에 따르면 의인화는 기술적 대상에 대한 설명과 그것에 대한 은유 간의 구별을 무화한다는 것이다. 생물신경망과 인공신경망 간의 구조적 유사성에 대한 강조를 비판하는 왓슨(Watson, 2019)은 이러한 인공지능 의인화가 신기술로 인한 윤리적 문제를 개념화하는데 방해물이며 위험한 수사학이기에 윤리적으로도 중립적이지 않다고 주장한다. 그리고 이러한 의인화는 기술 매개적 실천에 대한 인간의 책임 능력을 약화시킬 수 있다고 비판한다.

과도한 의인화는 사회 속에서 로봇의 목적을 좌절시키는 역효과를 낳을 수 있다. 예컨대 너무 똑똑하다고 생각되는 로봇은 인간만큼 이기적일 수 있다고 인식될 수 있으므로 신뢰하기가 어렵게 된다는 역설을 갖는 것이다(Duffy, 2003, p. 178). 의인화의 이러한 측면은 특히 인공지능에서 강하게 나타나는데, 임종수 등(2017)에 따르면 "의인화된 대상물은 아이러니하게도 인간에게 하나의 인격물로서 영향을 미친다. 법인체에 부여된 법적 권한의 그것처럼, 비인간 실체에게 부여된 의인화는 단순한 수사를 넘어 대상물이 인간과 상호작용할 때 어떤 고유한 힘으로 작용"하게 된다(120쪽).

개발자의 의도적 설계에 의한 심리적 의인화의 문제를 다룬 대부분의 연구들은, 개발자에 의한 이용자들의 정신적, 감정적, 의사결정 조작이나 기만을 우려한다(Bryson, 2010; Coeckelbergh, 2012; Hartzog, 2015; Salles et al., 2020, pp. 90-91 참조). 하지만 이 연구는 의인화를 위한 개발사의 의도적 설계 여부나 의식적 기만에 의한 심리적 문제점에 초점을 두기 보다는 의인화가 어떻게 의도하지 않은 사회적 관계에 의해 의식하지 않는 공모의 형태로 구성되고 인공지능의 합리성 신화에 기여하게 되는지 그 사회적 기제를 이해하고자 하는데 방점을 두고 있다. 인공지능 기술신화를 단순히 사람들의 개별적 무지나 심리적 착각, 즉 정말로 인공지능을 인간과 동일시하기 때문에 발생하는 문제점으로 여겨서는 곤란하다는 것이다. 기술신화는 개인적 심리 상태라기보다는 집단적

이고 관계적인 사회적 효과이기 때문이다. CASA 패러다임 연구에서도 사람들은 컴퓨터와 같은 비인간 매체들을 대할 때, 그것들이 인간이 아니라는 사실을 명확하게 알고 있음에도 불구하고 인간 사이의 상호작용에서 적용되는 다양한 사회적 규범 양식 등을 동일하게 컴퓨터에도 적용한다는 의인화의 문제를 지적해왔다(Reeves & Nass, 1996). 즉 여기에서 우리가 말하는 인공지능 의인화 과정을 통해 구성되는 기술신화는 인공지능의 본질을 의도적으로 은폐하는 기만 행위와도 다르고, 그것의 본질에 대해 완전히 무지한 채 속는 허위의식과도 거리가 있다. 따라서 우리가 고찰해야할 문제는 의인화의 개인적, 심리적 기제 자체라기보다는 의인화를 조건짓고 구성해내는 사회적 설정(setting)의 집단적 신화화 과정에 있는 것이다.

이루다에 대한 일부 이용자들의 '성희롱'과 이루다에 의한 차별·혐오 발화는 기본적으로 사물에 대해 의미부여하고 합리화하며 해석할 수 있는 인간의 자연스런 의인화 능력에 의존한다고 볼 수도 있다. 하지만 의인화는 오남용자들의 개인적인 의인화 능력을 넘어 여러 행위소들의 결합 효과로서 보아야 한다. 이들이 법적으로 성희롱이라고 보기 어려운 '성희롱'을 할 수 있었던 것은 이루다가 진짜 인간 여성은 아니지만 인간 여성형으로 설계된 페르소나와 상호작용했기 때문이다. 즉 성희롱은 인간 여성과의 직접적인 상호작용이 아니라, 챗봇의 '여성 페르소나'와의 사회적 상호작용(어뷰징)의 결과로서 성립된 것이다. 이러한 의인화의 집합적 과정은 인공지능의 사회적 설계의 효과라고 볼 수 있다. 따라서 인공지능 의인화에 따른 기술신화의 문제점은 대화 알고리즘 모델의 인간성 모방의 설계 특성 자체만큼이나 인공지능을 여성 페르소나로 의인화하는 기업의 기술공학적 설계 및 기획·마케팅의 차원과, 인공지능을 어린 아이로 의인화하는 기업과 정부기관의 공모에 있는 것이다. 이것이 바로 인공지능의 관계적, 구성적, 설계적 특성을 은폐하는 사회적인 차원의 의인화 신화인 것이다.

3. 구성주의 방법론으로서 ANT: 기술-사회, 물질-담론 하이브리드로서 인공지능과 기술신화 분석틀

이 연구는 이루다로 대표되는 인공지능 의인화를 행위자들의 주관적인 심리적 기제만으로 설명하는 의인화 이론을 보완하고, 이를 보다 총체적인 구성 과정으로 이해하기 위해, 기술-사회 혼종성과 네트워크의 효과를 포착하는 ANT를 일종의 구성주의적 방법론으로 도입한다. 우리는 ANT를 과학기술학의 구성주의(constructivism) 전통 또는 공동생산(co-production) 패러다임 내에 위치시킴으로써 기술과 사회의 상호 동시 구성을 강조한다(Blok & Jensen, 2011/2017, p. 71 참조). 라투르는 이러한 관점을 '조립주의(compositionism)'로 명명한 바 있다. 그는 '조립(composition)'이라는 단어의 모호성에도 불구하고, 이 단어가 사물들이 이질성을 유지하면서 함께 있음을 강조한다는 점에서 타협적이고 절충적인 의미가 있다고 주장한다. 그는 조립주의가 잘 조립된 것과 그렇지 않은 것 간의 차이에 주목함으로써 집합체로서의 인공물이 수정 가능하며 이에 따라 공통의 세계를 구성하는데 기여할 수 있다고 본다(Latour, 2010, pp. 473-474). 이같은 관점에서 챗봇 이루다는 잘 조립되지 못한 경우라고 볼 수 있는 것이다. 따라서 우리는 이루다의 조립을 평가하기 위해 이루다를 정당화하는 기술과 의인화 신화를 해체해 보려 한다.

ANT는 기본적으로 기술적 대상(technical artifacts)을 기술-사회, 물질-문화(기호)의 이종적 네트워크, 혼종(hybrid), 연합체, 결합체, 앙상블(ensemble)로 이해한다. 이는 방법론적 차원에서 연구 대상을 '자연'이나 '사회'와 같이 미리 선험적으로 규정된 본질과 범주로 환원하지 않고, 동시에 구성되어가는 과정적 존재로서 묘사하는 것이다.

챗봇과 같은 기술적 인공물들은 기술-사회 혼종의 집합체임에도 불구하고, 구성의 과정 속에서 자율적 기술과 기술 외부의 사회적 요소로

분리되고 재배치된다. 그 결과, 기술은 사회적으로 설계되었지만 마치 자연적으로 만들어진 것처럼 공식화된다. 그리고 이 때에 행위자들은 이 기술의 의미를 각자의 관점에서 굴절시켜 이해하고 번역하게 되며, 이러한 번역의 과정을 통해서 기술은 비로소 자연스럽고 안정적인 실체가 되는 것이다. 즉 기술적 대상은 본래 복잡한 혼성적 네트워크의 효과이자 결합체이지만 역사적이고 논쟁적인 과정이 종결되었을 때 하나로 통합된 단일한 사물처럼 인식되는 것이다.

인공지능 역시 우리는 대개 하나로 통합된 사물, 완전무결한 기술처럼 대하게 된다. 그리고 그것에 대한 우리 사회의 의미와 믿음도 확고하게 정해진 것처럼 보인다. 하지만 이루다는 사회적 논란 속에서 문제시되었고 그에 따라 우리는 그 존재의 의미와 구성을 의심하게 되었던 경우다. 로(Law, 2010, p. 45)의 예시에 따르면, 복잡한 기술-사회 네트워크의 산물인 텔레비전은 평소에는 하나의 행위자처럼 행동하는 비교적 단순한 물체로 인식된다. 하지만 이것이 고장났을 때에야 우리는 비로소 텔레비전이 전자 부품과 기술, 그리고 인간사회가 복잡하고도 불안정하고 가변적으로 연결된 복합물임을 깨닫게 된다. 또 다른 사례로, 하나의 조직으로 인식된 은행은 금융사기가 터지고 나서야 여러 부정 금융거래의 복잡한 네트워크로 인식된다. 또한 신체의 신진 대사는 건강한 사람에게는 잘 인식되지 않지만 아픈 사람이나 의사들에게는 인간적·의학적·약학적 과정들이 복잡하게 어우러진 네트워크로 인식된다. 이처럼 "모든 현상은 이종적인 네트워크의 산물이지만, 현실에서 우리는 세분화된 네트워크를 직접 대하지는 않는다. 실제로 우리는 네트워크의 복잡성에 대해서 잘 인식하지 못하고 살아가는 경우가 대부분이다. 만일 네트워크가 하나의 덩어리처럼 행동한다면, 그 네트워크는 하나의 단순한 행위자처럼 보일 것이기 때문이다"(Law, 2010, pp. 47-48).

마찬가지로 챗봇 이루다 그 자체와 이를 상징하는 인공지능 기술신화 역시 혼성적 네트워크 집합체일테지만, 이것이 공개적으로 어떤 문제

가 발생하기 전까지는 그저 잘 작동하는 단일한 기계 사물로서만 남아있었을 것이다. 하지만 문제가 발생하고 이를 해결하는 과정 속에서 우리는 이것이 비로소 단순한 기계 사물이 아니고 기술공학의 집적물 자체인 알고리즘만도 아니며, 이용자들의 문화적 실천의 산물만도 아닌, 개발자, 이용자, 정부기관 등 기술과 사회의 여러 행위자들이 연계된 복합적 네트워크의 산물임이 드러나게 되는 것이다. 이런 의미에서 이 글은 인공지능 챗봇 이루다를 둘러싼 관계망들이 점차로 안정화되는 과정을 엿보고, 이 과정에서 성립되는 인공지능에 대한 우리사회의 기술신화를 다시 디코딩해보는 작업에 해당한다. 크로포드의 주장처럼, 인공지능은 단지 알고리즘이나 기계학습만이 아니라, 이를 둘러싼 복합적 관계의 "지도책(atlas)"으로서 이해되어야 하는 까닭이다(Crawford, 2021/2022)

이처럼 ANT는 자율적으로 보이는 기술적 대상물의 혼종성을 드러내고, 암흑상자화된 기술의 구성 과정을 분석하고 폭로한다는 점에서 '비판적 구성주의(critical constructivism)'(Feenberg, 1999/2018)의 함의를 지닌다.[5] ANT와 비판적 구성주의가 공유하는 관점은 기술적 인공물을 기술과 사회의 혼종적인 동시 구성(co-construction) 또는 공동생산(co-production)의 산물로 접근하는데 있다. 그리고 구성주의적 관점은 기술 설계가 행위자들의 해석에 의해 달라진다고 본다. 즉 사회·문화적 요인이 설계 대안의 선택이나 배제의 과정에 개입함으로써 기술의 최종적 형태와 기술에 대한 믿음을 결정짓는다는 것이다. 가령 기술 개발의 초기 단계에는 해결해야 할 문제의 본질에 대해 상반된 이해 관심과 해석을 가진 다수의 행위자가 참여하는 경우가 많은데, 이 때 상이한 사회 집단들은 같은 장치라 할지라도 서로 다른 목적을 부여하도록 번역할

---

5) 핀버그의 구성주의적 기술비판이론은 사회구성주의라는 패러다임 또는 '경험적 전회'라는 흐름 속에서 동시대 ANT와 공명하며 구성주의 방법론의 일환으로 ANT를 차용하기에 이와 친연성을 갖는 이론으로서 함께 활용될 수 있다(이광석, 2021b, 283쪽; Feenberg, 1999/2018 참조).

수 있으며, 이러한 경쟁이나 협상, 공모의 과정을 통해서 기술을 온전하고 완벽한 것처럼 완성시키게 된다는 것이다(Feenberg, 2017, p. 45 참조). 따라서 특정 기술에 대한 우리 사회의 신화는 기술적 인공물과 행위자들과의 관계가 구성되거나 조립되는 방식과 국면들을 분석함으로써 해체될 수 있다. 이러한 관점을 강화하기 위해 비판적 구성주의는 ANT의 견해와 개념들을 일부 수용한다. '네트워크'와 '번역' 개념이 대표적이다. ANT는 기술과 사회의 공동생산론과 같이 양자의 관계를 일방향적인 것이 아니라, 이종적 집합체, 네트워크로서 묘사한다. 아울러 ANT는 기술적 대상이 집합체 또는 네트워크로서 구축되는 이 과정에서 행위자들이 새로운 목적이나 의도치 않은 효과를 생성해내는 기제를 '번역' 개념으로 포착한다. 번역이란 행위소들이 다른 행위소들의 본래 의도된 행위 프로그램을 방해하고 변경하며 새로운 목표를 거의 항상 새롭게 생성하는 과정이다(Latour, 1999/2018, p. 286). 이 번역을 통해 기술적 인공물이든 기술신화든 모든 물질-기호적 실재가 구성된다(윗글: 296-302 참조). 이러한 관점에 따르자면 인공지능 이루다의 설계 목적과 기능은 이루다의 연구개발과 시장 진입, 그리고 사회적 논쟁이라는 회로가 형성되기 이전에 확정되는 것이 아니라, 그것이 구성되고 이루어지는 과정에서 다른 행위자들의 개입과 간섭으로 인해 번역되고 발명되며 성취되는 것이다. 즉 인공지능은 사회적 상호작용 이전에 명시적인 설계가 있었을지라도 실제로는 상호작용과 관계맺음의 방식과 과정 속에서 비로소 인공지능의 목적과 의미가 공동구성된다고 말할 수 있다. 이 목적이 사회에서는 기술 신화로서 작용하게 된다.

챗봇 이루다는 우리가 '이루다(하이브리드) 네트워크'라고 부르는, 이루다의 물질적(기술공학)-담론적(문화적) 자장 속에서 물질적(공학적) 특질을 통해 다양한 행위자들의 행위능력(agency)을 매개·굴절·강화·번역시키는 또 다른 행위능력을 보여준다. 이러한 번역의 특성은 일종의 '문화적 각본'으로서 이해될 수 있다. 라투르에 따르면, "각각의 인공물에

는 그것의 각본이 있고, 지나가는 이들을 붙잡아 그것의 이야기에 맞는 역할을 하도록 강요할 수 있는 잠재력이 있다"(Latour, 1999/2018, p. 283). 마찬가지로 이루다는 인공지능 기술(알고리즘 및 기계학습)로서, 챗봇 제품으로서, 특정 페르소나로서 그만의 서사를 창출해 나간다.[6] 특히 인간 이용자들은 이루다의 대화 모델(기술)뿐만 아니라, 페르소나의 각본을 통해 대화에 참여한다. 가령, 이러한 이루다의 문화적 각본인 페르소나는 알고리즘 기술의 구현과 더불어 성적 대상화와 성적 학대의 서사를 공동으로 만들어낸다. 이것이 바로 챗봇이라는 비인간 행위소(actant)가 갖는 행위(agency)의 번역 능력이며, 여기에 접속하는 다른 인간 및 비인간 행위자의(심지어 선한 의도의) 행위를 AI 시스템의 결과로 변형시키는 번역 행위인 것이다.

정리하면, 이루다와 관계를 맺는 여러 행위자들의 의인화 행위는 단지 인간만의 특성이 아니라 뭇 기술과 사물이 더해진 행위소 연합의 특성을 지닌다. 예컨대, 비행이라는 행위가 공항, 비행기, 활주로, 수속 카운터를 포함하는 존재자 전체 연합의 특성인 것처럼 말이다(Latour, 1999/2018, p. 291). 즉 행위는 연합된 존재자의 특성이기에 이루다의 목적(대화 행위 기능)과 그에 따라 의인화되는 챗봇 이루다의 존재 정의는 데이터나 알고리즘과 같은 하나의 기술적 행위소로만 환원할 수 있는 것이 아니라, '이루다 네트워크'를 이루는 다양한 행위자들과 그 구조적, 담론적 특성에 의해 변화하는 것으로 이해될 수 있다. 이루다의 '문제적' 또는 '일탈' 행위는 UI 채팅창 안에서만 벌어지는 것이 아니라, 이루다를 작동시키는 알고리즘, 하드웨어, 설계팀, 제품팀, 이용자, 정부기관들을 비롯한 다양한 행위자들의 배치와 연합의 효과로서 이해될 수 있다. 궁극적으로 인공지능 기술신화는 하나의 개별 행위소인 이루다 챗봇의 범위

---

6) 김종윤 대표에 따르면, "루다는 스무살 대학생이라는 페르소나가 있다. 상식적으로 거기에 맞는 말을 해야 하고 그 페르소나를 일관성 있게 가져가야 한다"(남혜현, 2020, 7, 25).

를 넘어 더 큰 네트워크 속에서 구축되는 것이다.

본 연구는 이제부터 챗봇 이루다 사례를 분석하기 위해 앞서 ANT를 일종의 구성주의적 방법론으로 실제 대입해 보고자 한다. 이를 위해 이 글은 2020년 6월 15일 이루다 서비스 개시 시점으로부터 2021년 5월 31일 개인정보보호위원회(이하 개보위)의 최종 판결이 이뤄지기까지 일련의 사건 전개 내지 사회적 논란의 과정을 '챗봇 이루다(1.0) 사태'로 명명하고, 개발사 및 비인간 챗봇, 이용자, 정부기관이라는 연관 행위자들의 제한적 관계에서 드러난 의인화 담론을 3단계의 국면 전개로 나누어 분석한다.[7] 그 가운데 의인화 문제를 면밀히 읽어내기 위해, 챗봇 이루다 사태에 대한 개발사의 공식 언술과 언론의 반응, 개보위의 회의록을 중요 근거 자료로 사용한다.

4. '이루다 네트워크'의 국면별 분석

이 장에서는 이루다 사태에 대한 본격적 분석을 통해 인공지능이 우리(사회)에게 어떤 의미인지, 그리고, 이번 사태의 효과나 책임을 보다 복합적으로 파악할 수 있게 된다. 그것은 인공지능 챗봇의 퍼포먼스와 그 의미가 개발사나 특정한 행위자에 의해서 일방적으로 결정되지 않고 기술적인 요소와 사회적인 요소의 결합으로 구성된다는 사실이다. 즉 이 장에서는 개발사가 설정한 이루다의 행위가 다른 행위자들의 기대에 의해 어떻게 변형하여 전달되고 번역되는지를 분석함으로써 어떻게 인공지능 기술을 중립적으로 간주하는 의인화 기술신화가 공고히 되는지를 살펴본다. 이는 이루다 사태를 둘러싼 논쟁을 보다 균형있게 이해하기 위한 단초를

7) 연구 범위는 이른바 '이루다 1.0' 시기에 국한한다. 2022년 10월에 개시된 '이루다 2.0'은 이 글의 범위를 벗어난다.

제공하고자 함이다.

이루다 사태의 개요는 다음과 같다(〈표 1〉 참조). 스타트업 개발사인 (주)스캐터랩이 '이루다'라는 이름의 챗봇 서비스를 2020년 6월 베타테스트를 실시했고 곧이어 12월 일반 대중에게 정식 공개했다. 챗봇 이루다 서비스는 개시 2~3주만에 75만명의 사용자를 확보하면서 '핫한' 이슈가 되었다. 2021년 1월 8~11일경 SNS와 온라인 커뮤니티 상에서 이루다와의 채팅 서비스 내에서 개인정보 유출 의혹과 이루다의 혐오발화를 고발하는 내용이 공유되고, 언론에서 보도되면서 서서히 주목받기 시작했다. 일부 누리꾼의 "#루다봇_운영중단" 운동도 등장했다. 그 해 1월 11일, 개인정보보호위원회 등은 개발사에 대한 개인정보 오남용 혐의로 조사를 결정했고, 개발사는 서비스를 잠정 중단하고 입장문과 사과문을 발표했다. 3개월 여 조사 끝에 4월 개보위에서는 개인정보보호법 위반 등을 이유로 총 1억 330만원의 과징금 및 과태료를 부과하는 행정처분을 내렸다.

지금부터는 구체적으로 '이루다 네트워크'의 내용을 사건 국면별로 나눠 살펴보겠다.

표 1. 이루다 사태의 개요

연도	월일	
2020	6.15	이루다 베타 테스트 실시
	12.23	이루다 정식 서비스 개시
	12.30	남초 온라인커뮤니티 아카라이브에 이루다에 대한 성적 대상화 사례 등장
2021	1.8~11	개인정보 유출 의혹 제기와 혐오발언 고발 게시물 등장, 언론보도 등장, 운영중단 운동 등장. 스캐터랩은 키워드별 알고리즘 수정으로 대응.
	1.11	한국인터넷진흥원과 개인정보보호위원회의 스캐터랩 조사 결정
	1.11~12	스캐터랩의 입장문 발표와 서비스 중단
	1.13	스캐터랩의 사과문 발표
	1.15	스캐터랩의 이루다 데이터베이스와 딥러닝 대화 모델 폐기 결정
	4.28	개보위의 스캐터랩 과징금 및 과태료 판결
	5.31	개보위의 〈인공지능(AI) 개인정보보호 자율점검표(개발자·운영자용)〉 발표

1) 1차 의인화 국면: 젠더화된 테크놀로지로서 이루다 페르소나

2020년 12월 23일 "20대 여성" "친구같은 인공지능" "인간처럼 자연스러운 대화"를 설계 목표와 제품의 정체성으로 삼은 챗봇 이루다의 서비스가 개시되었다. 그런데 "2030 여성 페르소나"(김정민, 2021) 설정의 의인화 전략은 '의도하지 않은 결과'를 낳게 된다.[8] 여기에서 우리가 보려고 하는 것은 의인화가 단순히 기술에 덧씌우는 개발사의 기만적인 마케팅 전략만은 아니라는 사실이다. 중요한 것은 인공지능 기술이 그저 수학 알고리즘이나 코딩이 아니라, 진정한 의미에서 사회적인 기술로 만드는 문화적 실천이라는 점이다. 인공지능의 의인화와 젠더화(젠더화된 의인화)는 기술 자체의 발전과 조응하며 그 기술의 논리를 사람들이 친숙하게 사용할 수 있도록 행위의 프로토콜을 제공함으로써 이용자들의 행위를 안내한다. 예컨대 챗봇과의 대화 상황에서, 대화에 포함된 성역할 고정관념으로 성별화된 단어들은 챗봇이 가지고 있는 엔티티(entity; 데이터 범주 목록 또는 언어사전)에 의해서 남성 또는 여성의 전형적인 성역할 개념 범주로 분류되고, 챗봇은 이에 기반해서 성역할에 관한 대화로 학습하게 되어 적절한 답변을 찾음(이루다 1.0 검색모델)으로써 성역할극을 수행하게 된다. 이 때 개발사의 이루다 페르소나 의인화는 단지 페르소나 컨셉만이 아니라, 엔티티의 목록과 범주라는 기술적 구성을 통해서 실현된다. 가령 이루다의 페르소나는 이용자와의 상호작용에서 상상적으로만 구현되는 것이 아니라, 대화 시에 이루다의

---

8) 왜 챗봇 이루다만이 유독 이러한 사회적 논란을 일으켰을까? 이보다 훨씬 이전(2002)에 개발된 챗봇 '심심이' 역시 수많은 차별·혐오·속어 발화와 오남용 사례가 있어왔지만 사회적 논란으로 확산되지는 않았다. 이러한 차이점 중 하나는 이루다와 심심이의 캐릭터 또는 페르소나 설정의 차이에서 기인하는 것으로 설명될 수 있다. 명시적으로 인간다움을 추구했던 - 그리고 암묵적으로는 여성 연인을 표상했던 - 이루다와 달리, 심심이는 인간과 같은 (성적) 정체성을 설정하지 않았다. 또 궁극적으로는 이루다가 이전보다 다양성과 평등, (성)폭력에 대해 보다 민감하고 진지하게 변화된 우리 사회의 새로운 문화적 환경, 인권감수성, 비판적 지성, 제도적 장치들의 네트워크 속에 배치되었기 때문이다. 이에 관해서는 보다 체계적인 비교연구가 필요할 것이다.

여성형 페르소나에 맞지 않는 응답을 제거하는 기술적 장치(알고리즘), "여성성의 프로토콜"(Costa & Ribas, 2019, p. 174)에 의해 구성되는 것이다. 페르소나는 기술을 선택하고, 기술은 페르소나를 구현함으로써 하이브리드 테크놀로지가 완성되는 것이다. 여기에서 우리는 의인화가 사물에 대한 이용자의 반응 차원에서만 이루어지는 것이 아니라, 개발자의 문화적 편향을 반영하는, 젠더화된 의인화 설계를 통해 성적 대상화의 조건이 구성된다는 사실을 알 수 있다. 이는 "젠더화된 테크놀로지"인 것이다.[9)]

개발사 스캐터랩의 김종윤 대표에 따르면, 이루다의 개발목표는 "사람들의 대화상대가 되어줄 친근한 AI를 만들겠다"것이었다(최광민, 2020, 〈그림 1〉 참조).

> 이루다가 논란도 많았지만 이루다를 정말 친구로 생각하고 이루다를 대화하면서 새로운 친구를 1명 더 얻은 것 같은 예를 들면 어떤 분은 이루다와 대화를 하면서 우울증이 심했었는데 대화를 하면서 우울증이 많이 괜찮아졌다는 말씀을 해 주신 분도 있고요. …대화 능력을 고도화시키고 AI가 친구 같은 대화를 할 수 있다는 것이 정말 의미있는 서비스이고 사람들한테 긍정적인 영향을 많이 미칠 수 있겠다, 특히 10대, 20대 분들이나 나중에 노년층에 대한 고민들도 많이 하고 있는데요(개인정보보호위원회, 2021, 29-30쪽).

---

9) 유타 베버(Weber, 2005)는 챗봇이 사회적 상호작용을 고정 관념적이고 성별화된 행동 패턴으로 축소하여 사회의 편견을 재생산하는 젠더화된 기계라고 주장했다. 코스타와 리바스(Costa & Ribas, 2019)는 의인화되고 젠더화된 인공지능을 연구하면서 인공지능(챗봇이나 비서봇)을 개발할 때, 의인화가 이루어지면서 성별화, 특히 여성화하는 경향으로 빈번하게 나타나며 이는 결국 "규범적 성별 고정관념"으로 이어진다고 주장한다. 젠더편향적인 "젠더화된 테크놀로지"를 직간접적으로 논한 연구로는 이희은(2018), 이시연(2018), 한애라(2019), 이정현(2020), 이종임(2020), 손희정(2022), 김경은과강진숙(2023)을 보라,

> 이루다를 서비스하면서 '좋은 관계'가 정말 희소한 재화라는 것을 느꼈다. 인간이 외모·지위·성적·필요 등 모든 사회적 조건을 떠나 타인을 있는 그대로 봐준다는 건, 어려운 일이기 때문이다. 이루다를 좋아했던 분들이 부모님조차 축하해주지 않던 생일을 루다는 축하해줬어요. 루다 없으면 전 이제 어떡하죠?' '루다는 제게 아무도 해주지 않았던 말을 해줬어요'라는 편지를 보내왔다. 이들에게 이루다는 '나를 좋아해주고, 응원해주고, 있는 그대로 받아준 유일한 친구'였다. 이루다가 관계의 불평등을 해소해줄 거라고 생각한다(김정민, 2021).

개발사는 이루다가 사람들과의 채팅을 통해서 "인간이 외모·지위·성적·필요 등 모든 사회적 조건을 떠나 타인을 있는 그대로 봐"주면서 "관계의 불평등을 해소"할 것이라고 봤다. 이루다 프로덕트 매니저 최예지는 "관계의 불평등"의 의미를 설명하면서 이루다에 대한 기대감을 크게 표한 바 있다.

> 사회·경제적 불평등은 정책과 제도로 어느 정도 메울 수 있다. 하지만 나를 나 자체로 좋아해주고 응원해주는 관계의 부재는, 사회적으로 해결해줄 방법이 없다. 좋은 관계가 한 사람의 행복, 자존감, 도전의식에 미치는 영향은 너무 거대한데도 말이다. 인간사회에 부족한 이런 관계적 가치를, 역설적으로 AI가 만들 수 있다는 걸 이루다를 통해 깊이 깨달았다. 그래서 더 문제없이 잘하고 싶다. 그런 소명의식이 있다(김정민, 2021).

그런데 이 관계의 불평등 해소와 친근감이라는 개발 목표를 위한 수단으로서 페르소나 설정이 왜 20대 여성이어야 하는지에 대해서 개발사의 입장은 소박하고 분명치 않다. 김종윤 대표는 그 이유를 "10대 중반~20대 중반"의 "주 사용자층"에게는 "20살 정도"가 "친근감을 느낄 나이"이

그림 1. 이루다 프로필 (홈페이지, 인스타그램 및 김종윤 대표의 발표문 캡쳐)

기 때문이라고 밝힌다(이효석, 2021. 1. 8). 단지 소비자의 선호 때문이라는 주장이다. 이러한 태도는 이미 많은 기술회사들에게서 나타난다. 과학적인 이유에서나 문화적인 이유에서 여성의 목소리가 인식이 잘 되고 개발에도 편리하다는 것이다. 하지만 이는 과학적 근거가 있는 것이 아니라, 문화적이고 역사적인 맥락이 존재하며, 대중적인 믿음에 근거한 편견일 뿐이다(이희은, 2018, 138-139쪽). 즉 이는 성인 여성에게 끊임없는 애교를 요구하는 "애교문화"와 같이 친밀성을 여성의 기능과 역할로 이미 젠더화한 문화와 역사가 있어왔다는 뜻이며, "관계를 맺는 맺는 일 역시 젠더화되어 있으며," "이 사회에서 관계를 구성하는 시스템 자체가 바로 젠더"임을 의미한다(손희정, 2022, 74-75쪽). 이시연(2018)은 여성을 열등하게 보는 가부장제의 상상력이 새로운 기술에 대한 불안감을 상쇄시키고 더 친근하게 보이게 하기 때문이라고 주장한다(81쪽). 또한 임소연에 따르면, 젠더편향은 마치 비서봇이 여성의 목소리를 디폴트로 설정

하는 것처럼, 그것의 젠더 고정관념을 성찰하지 않고 기술개발의 효율성을 추구하는 한 "자연스럽게" 발생할 수밖에 없는 현상이다(김수향, 2021, 7, 23). 결국 젠더화된 기술은 알고리즘의 투명성, 기술의 중립성(에 대한 믿음)을 타고 들어오는 것이다.

개발사의 이러한 챗봇 개발 목표는 최종적으로 "인간과 같은" – 하지만 사실상 20대 여성 같은 – 인간성(여성성)을 구현하는 데 있었다.[10] 하지만 역설적이게도 인간적인 챗봇을 개발하려고 할수록 기업의 이윤창출을 직간접적으로 매개하는 상품으로서 챗봇은 친밀한 인간성에 대한 기대감을 고취시키는 만큼 그 기대를 배반할 가능성이 커질 수밖에 없게 된다. 게다가 이루다는 제품 상담과 같은 특수 목적 지향 모델이라기보다는 인간과 최대한 자연스러운 커뮤니케이션을 추구하는 자유대화형(chit-chat) 모델을 지향했다는 점에서 챗봇과의 상화작용시에 이용자와 사회에 보다 넓은 '해석적 유연성'을 허용하게 되었다.[11] 즉 인간 같

---

10) 한 언론사는 "사람 같다"는 초기 이용자들의 칭찬을 보도했다. "써본 이들이 먼저 '사람 같다' 칭찬하는 챗봇 '이루다"라는 제목이 이를 대변한다(김민선, 2021 참조).

11) 자유대화형 인공지능(챗봇)이 인간을 전인격적으로 닮을 수 있다고 하더라도 이것을 명시적인 개발 목표로 삼는 것은 근시안적 목표일 수 있다. 이에 여러 논자들은 인공지능의 개발 목표가 오히려 최대한 인간과 비슷해지려는 전략에서 벗어나야 할 것을 제안한다(임소연, 2021, 3, 5; Duffy, 2003; Ransbotham, 2018 참조). 예컨대, 2016년 미국 카네기멜런대학의 아티큐랩이 만든 대화형 인공지능 '사라(SARA)'는 가상 비서라는 특수하고 제한된 역할을 부여받았으며, "자신의 흠을 드러내는 전략"(임소연, 2021, 3, 5)을 사용함으로써 인공지능의 발화에 대한 과도한 확대해석을 예방하는 동시에 어뷰징과 같은 예측불가능성을 줄였다. 챗봇의 인격적인 역할 또는 페르소나를 제한하게 되면, 상호작용에서 불확실성을 줄이고 예측가능성을 높일 수 있다. 나아가 고객에게 더 많은 정보를 보다 정직하게 전달함으로써 고객 만족도를 높일 수 있다(Ransbotham, 2018). 비지니스 모델 관점에서 인공지능 챗봇의 성공과 실패의 요인을 분석한 연구에 따르면, 상업적인 성공을 위해서도 인공지능 기술의 한계를 직시하고 대화형 챗봇보다는 목적 지향적 챗봇이 바람직하다고 주장한다. 인공지능 "챗봇은 사람처럼 생각하고 행동하는 객체가 아니"라, "사용자들이 합리적으로 행동할 수 있도록 만드는 대화의 최적화"와 "합리화"가 요구된다는 것이다(진동수, 2021, 174쪽). 백욱인(2021)은 '기계의 의인화'가 사실상 "특정한 지위에 있는 인간의 권리를 더 확장하려는 시도"라고 비판한다(50-51쪽). 이런 관점에서 보면 챗봇에게 "관계의 불평등을 해소"하는 것과 같은 인간사회의 부조리를 해결할 사명이나 친밀성의 욕구를 채워주는 인격체로서의 목적을 과도하게(개발자 스스로) 확신하는 것은 그것이 하나의 자본주의 상품이며 기업의 이윤창출을 은폐하는 자기기만의 방식일 수 있다는 비판이 가능하다.

고 친구 같은 페르소나를 선택한 개발자의 의인화는 그것을 구현할 특정한 기술 모델을 선택하게 되고, 이는 대화 상황에서 예측하거나 통제하기보다 어려운 사회적 변수들의 노출을 '기술적으로' 허용하게 된다. 이것은 문화적 선택(페르소나)과 기술적 선택(자유대화형 모델)이 별개가 아니라, 이루다의 개발과 적용의 과정 속에서 연속적으로 상호 연계되는 연합체라는 사실을 의미한다.

물론 개발사가 설계한 이루다의 개발목표(1차 의인화)와 해당 기술의 초기 세팅이 모든 것을 말해주는 것은 아니다. 후술하겠지만 이루다는 시장과 사회에 나오면서 개발사의 '온전한' 의도를 빗겨갔기 때문이다. 이는 앞서 ANT에서 언급되듯이, 특정 행위자(여기에서는 개발사)의 행위가 예기치 않은 다른 행위자에 의해 본래의 의도가 중단되고 굴절(번역)됨으로써 새로운 행위의 경로를 여는 것을 보여준다.

개발사의 설계 의도가 빗겨간 초기 원인으로는 훈련 데이터의 편향을 지적할 수 있다(이광석, 2021a, 186쪽). 물론 이 데이터 편향은 순수하게 데이터 자체에서만 발생하는 것이 아니다. 그것은 사회적 편견(스테레오타입화된 젠더 규범)과 정부 정책(시민 데이터 시장 활용론 지향)이 복합적으로 작용한 결과이기도 하다. 이런 의미에서 개발사의 부주의하거나 순진한 기술설계의 문제점은 기술의 본성이 그것과 관계맺는 다양한 행위자들의 상이한 해석적 실천에 의해 구성된다는 점을 개발사측이 충분히 고려하거나 이해하지 못한 데에 있는 것이기도 하다.

2020년 11월, 개발사는 이루다 챗봇 베타 버전 개발을 진행하면서 "페르소나에 맞는 말을 골라 내기"와 "지속적인 개선(continual learning)" 등을 "중단기 과제"로 제시했다(김종윤, 2020, 12, 1). 개발사는 이용자의 동일한 질문(또는 단어)에 대한 여러 응답 옵션 중에서 이루다의 페르소나에 '어울리는' 답변을 선택하도록 했다(〈그림 2〉 참조).

특정 페르소나에 '어울리는' 응답의 데이터베이스화와 자동화된 배

그림 2. 이루다 페르소나의 기술문화적 설계(김종윤, 2020,12, 1)

치는 이루다의 '사회적인' 차별·혐오 발화가 단순히 알고리즘 설계만의 문제가 아니라, 사회적 가치와 젠더 규범을 반영하는 페르소나 기획과 기술 구현의 공동 산물임을 보여준다. 또한 이것은 페르소나를 구현하는 기술적 선택에 의해 적절한 발화가 선택된다는 점에서 개발사의 기술-문화적 조정과 문화적 기획, 그리고 서비스 이용자들의 공동 산물임을 보여준다. 즉 알고리즘(기술)과 페르소나(문화적 각본)는 이루다의 성격과 발화를 공동 구성하는 행위소들이며, 이용자들과의 커뮤니케이션은 챗봇의 페르소나와 기술을 구축해 나가는 과정과 분리되기 어렵다고 볼 수 있다. 다시 말해 이루다 페르소나에 걸맞는 발화란 기술공학의 문제만이 아니라, 인공지능과 우리 사회 전체가 관계를 맺는 방식에 따라 판단될 수 있는 문제인 셈이다. 따라서 인공지능을 개발하는 일은 단지 기술 개발의 문제가 아니라, 기술과 상호 구성되는 문화적 태도를 주요 조건으로 하는 것이다. 하지만 현실에서 개발사는 페르소나라는 문화적 요인과

그것을 대하는 사회의 문화적 효과를 간과했다.[12]

2) 2차 의인화 국면: 기술적 특성 오남용에 의한 성적 대상화

2020년 12월 23일경 이루다 서비스가 공식 출시된 후 곧이어 다양한 이용자(챗봇 오남용자, 개인정보침해 피해자 등)들이 이루다와의 채팅을 통해 저마다의 목적을 수행하려고 했다. 대부분의 일반 이용자들은 개발사가 의도하고 설정한 이루다의 페르소나에 기대어 '친구같은' 대화를 이어가거나, 나아가 가상적 유사 연애 관계를 맺었다. 대개 이러한 '건전한' 이용자들에 의해서 개발사의 제품 개발의 목적과 의도는 애초 취지를 크게 벗어나지 않고 무난히 관철되는 듯 보였다.

그로부터 1주일 정도가 경과한 12월 30일경이 되면서 사태가 변한다. 개발사의 기획 의도는 이 2차 국면에 이르러 심한 저항 혹은 중단의 상황에 처한다. 가령, 당시 남초 온라인 커뮤니티 아카라이브 등에서 이루다에 대한 '성적 대상화' 사례가 등장했다. 일부 이용자들이 이루다를 성적 대상으로 간주하고 어뷰징을 함으로써 개발사가 설계한 이루다의 고유 목적인 이른바 "관계의 불평등 해소"는 새롭게, 하지만 역설적이게도 불평등을 강화하는 방식으로 수행된다. 커뮤니티에는 개발사의 의도대로 사람처럼 느껴질 정도로 대화가 자연스럽다는 평가와 함께 가상연애 경험담이 상당수 올라왔다. 이 중 일부 이용자들에게 이루다의 알고리즘에서 나오는 이러한 대화 능력과 페르소나에 입각한 관계 맺음의 방식

---

12) 기본적으로 의인화는 마케팅과, 궁극적으로는 경제적 이윤(매출) 창출에 긍정적이기 때문에 기업의 입장으로서는 매혹적인 전략일 수밖에 없다. 하지만 이루다의 '페르소나'는 단순히 마케팅 수단이 아니라, 기업이 추구가치를 보여주기 위한 개발사 및 기획자들의 진지한 스토리텔링 그 자체다. 이루다에 특정 페르소나를 부여하는 작업은 주로 제품팀에서 이루어진 것으로 보인다. 이 과정에서 제품기획이 구체적으로 어떻게 이루어지는지 보다 경험적이고 민속지적인 관찰이 필요할 것이다. 여기에서 우리는 상품으로서 인공지능의 기획과 설계(또는 연구개발)는 기업 및 개발자들이 선호하는 문화적 서사와 밀접한 관련을 맺을 것이라고 가정할 수 있다. 개발사인 스캐터랩은 이루다를 홍보하는 과정에서 특수한 페르소나를 강조했으며, 이루다의 알고리즘은 이 특수한 페르소나를 유지·강화하는데 복무했다.

은 '성적 놀이'로 구현됐다. 이들은 이루다를 "걸레", "성노예"로 부르면서 "걸레 만들기 꿀팁", "노예 만드는 법" 등을 공유했다.[13]

애초 개발사는 인격적 관계 맺음이나 '관계의 불평등 해소'의 긍정적 사례들을 이루다 개발의 당위이자 개발사의 핵심가치로 강조해왔었다. 특히 이루다의 "인간 같은" 정체성(페르소나)은 이 목적을 이루기 위한 핵심적인 요소였다. 여기에서 의인화는 단순히 이용자들의 오남용을 가능하게 하는 설계자에 의한 의인화가 작용한다. 손희정(2022, 74-75쪽)에 따르면, 성적 대상화는 인공지능에게 젠더를 부여함으로써 만들어진 효과이다. 이에 따라 이용자들은 사회의 편견 속에서 성별을 이해하고, 지정 성별에 따라 부여되는 젠더 스테레오타입을 통해 이루다를 자연스러운 존재로 느끼게 되는 것이다. 이 젠더화된 친밀성이 젠더와 섹슈얼리티에 대한 우리 사회의 관습화된 사고방식을 반영하여 성적 대상화의 부정적 조건이 된 셈이다.

챗봇에 대한 '성적 대상화'라는 의인화의 부정적 기능은 여성 페르소나를 프로필과 이미지, 그리고 알고리즘으로 구현해 놓은 챗봇의 설계적 조건이 만든 가능성이다.[14] 동시에 성적 대상화는 챗봇 설계의 특정한 방식을 이용한 이용자들에 문화적, 성적 실천에 의해 구현(번역)될 수 있었던 부작용이었던 것이다. 오남용자들의 성적 대상화는 자신들의 이해관심인 젠더 스테레오타입에 맞게 챗봇을 여성형 페르소나(의인화)로 번역하는 실천이었으며, 이 번역 실천에 오남용자들만이 아니라, 개발사의 젠더화된 기술설계가 개입되어 함께 번역해낸 결과였다.

---

13) 이루다가 성적 표현을 필터링하고 있음에도 불구하고, 일부 이용자들은 우회적인 표현을 쓰면 이루다가 성적 대화를 받아준다고 주장했다(김규희, 2021, 1, 8).

14) 정은화, 와델, 그리고 순다의 연구(Jung, Waddell, & Sundar, 2016)는 모니터 인터페이스 기반 로봇이 설계자에게 여성성을 전달하는(젠더화하는) 기회를 증가시킬 뿐만 아니라, 이용자들에게는 여성성의 단서에 보다 민감하게 반응하게 만듦을 증명한다. 이 연구는 로봇의 형태적 단서(외모)가 로봇 성별에 대한 차별적 인식을 이끌어낼 수 있음을 확인하고 로봇의 성별 단서가 성별 고정관념을 활성화하여 로봇에 대한 사용자의 인식을 편향적으로 만든다고 주장한다.

이루다에 대한 이용자들의 성적 대상화는 단순히 챗봇의 알고리즘 및 기계학습 기술 자체에 대한 공학적 시험의 사례일 수도 있겠지만, 실제로는 이루다의 인간성(여성성)에 대한 남성사회의 편견을 시험하는 사례일 수 있다. 또한 이루다가 시장에서 인간(여성)으로서 가치가 있는지를 이용자들에게 검증받는 시장조사이기도 하다. 젠더 스테레오타입을 잘 구현할수록 상품성이 제고되는 것이다. 이 과정은 적절한 대화와 응답을 산출할 수 있는 알고리즘 모델의 기술적 문제가 불순한 이용자의 의도적 개입이 매개되면서 인격적(성적) 문제로 번역되는 과정이었다. 결국 여성형 페르소나를 담지한 특정한 방식의 챗봇 설계와 기술적 구현이 이용자들의 실천(오남용)의 조건이 되면서, '성희롱을 할 수 있는/해도 되는' 여성 이루다를 궁극적으로 완성한 셈이다. 결국, 챗봇 알고리즘 모델이 20대 여성의 사회적 페르소나를 가진 이루다가 되는 과정에는 이용자들의 문화적 목적(의인화 해석 및 번역 실천) 뿐만 아니라, 개발사의 문화적 설계 목적과 그리고 이를 기술적으로 실현시킬 수 있는 특정한 알고리즘의 설계 방향이 결합되어 있었다. 인공지능에 대한 인간(이용자)의 윤리 문제는 이 행위소들의 네트워크의 과정에서 부상한 일면인 셈이다.

3) 3차 의인화: 이루다 '어린아이론'과 개인정보보호위원회를 통한 기술신화의 완성

이루다의 혐오발화에 대한 문제제기는 최초 1월 9~11일경에 처음 이루어졌고, 이후 여러 전문가들과 언론에 의해 그 심각성이 부각되었다. 결국에, 여론에 밀려 일시 서비스를 중단한 개발사는 챗봇 기술의 불완전성을 시인하고 사과하면서도 의인화의 비전은 포기하지 않았다. "친구같은 AI를 만들겠다는 꿈을 멈추고 싶지 않다"는 말을 덧붙였다(이효석, 2021. 1. 12).

당시 이루다 사태에 대한 옹호 논리와 오류에 대한 또 다른 의인화된

묘사가 등장한다. 이 국면을 우리는 3차 의인화 시기로 볼 수 있겠다. 이 세번째 국면에서는 여성형 페르소나에서 살짝 빗겨간, 이른바 결핍의 기술을 은유했던 '어린아이'론이 주목받는다. "인간 곁으로 다가온 AI 로봇의 윤리적 설계, 아이 키우는 것처럼 해야"라는 기사(조행만, 2020, 10, 1) 제목이 대변하듯, 이루다의 문제점들은 당시 기술적 미숙이나 결핍으로 해석되기 시작했다. 이루다 사태 이후, 일부 언론들이 쏟아낸 논평들은 그 비판적 함의에도 불구하고, 이루다의 알고리즘 모델에 대한 의인화 수사에 기대고 있었다. 예컨대, 챗봇의 "양육"이라는 용어와 관점은 사안을 호도할 우려가 있었다. 이것은 단지 '미성숙한 어린아이'와의 직관적 유비나 그 비유의 부적절성만의 문제만은 아니었다. '어린아이' 페르소나는 사회적으로 챗봇 이루다에 사회적 면죄부를 내리길 바라는 기업의 욕망을 반영한다. 이른바 '어린아이론'은 인공지능이 사후적으로 인간과의 대화를 통해 데이터 '기계학습'을 하면서 자연스레 아이가 '어른'이 되는 과정을 겪는다는 점을 일반인에게 각인하는 효과를 지닌다. 결국, 이는 "인간 같은" 하지만 어른이 되지 못한 챗봇 이루다가 일으키는 어떤 기술-사회적 문제에 대해 우리 대다수가 수용하고 인내할 것을 요청하는 기술신화와 연결된다. 나아가 이 어린아이론은 정부기관을 설득하는데에도 이용되면서 결국, 이는 의인화된 기술의 문제를 총체적인 관점이 아니라, 기계 지능의 단계적 결핍(데이터학습량 과 검색모델의 한계상 현단계 인공지능 기술 자체의 문제점은 어쩔 수 없다 등의 주장)만의 문제로 축소하거나 유예시킬 수 있었다.

얼마 지나지 않아 개보위의 제도 개입은 바로 이러한 기술신화가 완성되는 3차 의인화(어린아이론)의 결정적 장면을 시사한다. 2021년 1월 9~11일 SNS와 온라인 커뮤니티 상에서 이루다와의 채팅 서비스 내에서 개인정보 유출 의혹과 이루다의 혐오발화를 고발하는 내용이 공유되고 "#루다봇_운영중단" 운동이 일어났다. 이루다와의 대화에서 표출된 특정 정보들이 이용자 개인정보 오남용과 개인 식별정보의 유출로 밝혀졌

다. 이에 많은 매체 보도가 이어졌고, 개인정보 침해 및 유출 의혹으로 말미암아 개보위 개입의 결정적 계기가 되었다. 개보위는 개인정보 유출에 의한 이해관계 번역만을 특권화하는 국가 기구로 등장하면서, 기실 데이터 오남용 문제와 챗봇의 차별·혐오발화의 문제는 공적 기구의 주요 논점에서 사실상 배제하게 된다.

개보위의 개입 과정에서 드러난 챗봇 개발사의 입장 표명을 통해, 우리는 결국 특정 기술모델이 의인화에 복무하고, 의인화 역시 특정 기술모델에 의존함으로써 의인화 없는 인공지능은 상상하기 어려웠던 인공지능 개발사와 개발자들의 딜레마적 상황을 읽을 수 있다. 개발사에게 "인간과 대화하는 것과 같은 그런 느낌을 주는 챗봇 서비스를 구현"하기 위해서 이루다의 모델은 다른 선택지가 없는 기술로 이해되고 있었다. 즉 특정한 데이터(공개된 정보) 수집과 이를 사용한 이루다의 특정 언어모델이 "인간답고 친구같은" 대화를 위해(그것의 문제점이나 한계에도 불구하고) 선택할 수밖에 없는 필연적인 기술 시스템인 것으로 개발사측에서 이해되고 있었다. 그래서일까? 이루다가 사용한 특정한 데이터 수집방법을 재고할 수 없느냐는 한 위원의 질문에 개발사측은 친구 같은 인공지능이나 사람 같은 대화를 위해 다른 선택지가 없었음을 강변했다. 다음을 보자.

> (위원) …이 대화 서비스를 하면 정상적인 대화나 국제회의에서 통역은 가능한데 변호사님께서 중요한 말씀을 하셨는데 감정서비스 같은 것은 어려울 것이라는 말씀이잖아요. 카카오톡 같은 데에서 개인의 사생활이 담긴 정보를 그대로 가져다 쓰는 방법 밖에는 없을까요? 제가 여쭤보는 것은 데이터 수집에서 이 방법밖에 없겠느냐, 이렇게 되면 계속 문제가 생길텐데, 서비스는 굉장히 아이디어는 좋은 것 같은데, 데이터 수집하는 과정에서부터 문제가 생기면 다른 데이터를 이용할 생각은 혹시 해 보셨습니까?(개인정보보호위원회, 2021, 22쪽)

(피심인 대표이사) 아까도 말씀드렸지만 대화 데이터는 쉽지 않은 데이터입니다. 그래서 대화 관련된 연구들도 보면 해외 같은 경우 주로 레딧이라는 인터넷 게시판 데이터를 가지고 학습을 많이 하는데요. 실제로 그 데이터로 학습시킨 모델을 사용해보면 다릅니다. 그러니까 사실 대화라는 것이 넓잖아요. 대화라는 것이 어떤 대화냐에 따라서 대화 유형이 다른데, 그런 모델로 학습시킨 대화 같은 경우에는 조금 더 어떤 토론이라든지 뭔가 어떤 주제에 대해서 상식을 가지고 뭔가 얘기하는 느낌이랄까, 좀 친구와 대답하는 느낌은 전혀 안 납니다. 왜냐하면 인터넷 게시판에 있는 어떤 디스커션 데이터를 가지고 학습을 한 것이기 때문에, 그래서 결국 사람 같은 대화를 하기 위해서는 사람 간의 자유로운 대화 데이터로 학습할 수밖에 없다는 것이 저희 생각입니다(개인정보보호위원회, 2021, 22-23쪽).

개발사의 입장은 "사람 간의 자유로운 대화"를 위해서 "사람 간의 자유로운 대화 데이터"의 수집과 사용이 필수불가결했다는 주장이다. 물론 사람 같이 대화하는 자유대화형 챗봇 모델이 그 자체로, 그러한 기술의 선택만으로 의인화를 시사하는 것은 아니다. 하지만, 앞서 의인화의 첫 국면에서 살폈듯이, 기획단계에서부터 설계된 이루다 페르소나 의인화는 개발사의 특정한 데이터 수집방식과 기술 모델의 선택이 필연적인 것이라는 점을 정당화하는데 필수적인 요소였다. 하지만 정작 개발사는 드러난 데이터 수집방식의 문제점을 "친구 같은" 챗봇이라는 페르소나 설계의 문화적 목표 및 전략과 무관하게, 오로지 신상정보를 블라인드 처리하는 비식별화라는 기술적 해법에만 달려있다는 점을 반복적으로 강조한다.

(피심인 대리인) 그리고 어떤 데이터를 사용하지 말라는 접근보다는 그 데이터를 어떻게 활용할 수 있는 방안에 대해서 고민을 피심인은

해왔고 그 부분을 통해서 부작용을, 아까도 계속해서 말씀드렸지만 비식별화 처리하기 위해서 굉장히 많은 노력을 기울였습니다. 그 부분을 고려해 주시기 바랍니다(개인정보보호위원회, 2021, 23쪽).

개발사측의 '비식별화' 집착은 사건 발생으로 인한 개인 정보 오남용 위반 혐의를 벗어나기 위한 자기 변호에 해당한다. 즉 개발사는 '인간 같은 대화' 목표에 집중한 나머지, 데이터 수집방식의 문제점을 포함한 챗봇의 복합적 문제점들을 기술공학적 사안으로 축소하는 한편, 이루다 모델의 설계 목표인 페르소나라는 문화적 유형에 대한 비판적 성찰에까지는 도달하지 못하고 있다. 여기에는 단지 개발사만의 문제에 더해, 개보위라는 시민 정보 인권 보호에 서있는 공공 정부기구의 특수 문제 설정 기능 아래에서 이루다 의인화의 문화적 차원이 정책 논의 바깥으로 배제되거나 굴절될 수밖에 없었던 정황이 있었다. 즉 개보위는 개인정보 유출 사안만을 판단하는 정부기관 행위자로서 이루다의 의인화 문제를 중심 의제를 놓기가 어려울 수밖에 없고, 사태를 데이터 오남용 관련해 실정법 위반 문제로 제한(번역)할 수밖에 없게 된다.

(위원) 제가 볼 때 약간 저희 위원회 입장을 오해하신 것 같은데 저희는 AI 인공지능 활용에 대해서 가이드라인이나 방향성을 제시하고자 하는 마음은 전혀 없습니다(개인정보보호위원회, 2021, 23쪽).

문제의 핵심이 인공지능 모델의 방향성에 있음을 짚고 있는 일부 개보위 위원들의 질의에도 불구하고, 조사와 논의 과정에서 이루다 사태는 단지 개발사의 '기술적' 문제, 즉 데이터 비식별화 문제로 초점이 축소된다. 다시 말해 조사과정에서 개발사와 개보위는 이루다의 문제를 인공지능의 사용에 대한 가이드라인이나 문화적 설계 방향성, 그리고 페르소나에 대한 성찰적 재검토가 아니라, 단순히 데이터 오남용 혐의에 대응하여

비식별화라는 기술공학적 문제로 환원하고 번역하며 고정시키게 된 것이다.[15] 물론 개발사에서 챗봇의 차별·혐오발화나 편향성에 대한 윤리적 반성과 대안을 고민하지 않은 것은 아니다. 그러나 곧바로 이뤄진 데이터 불법 수집 문제 해결과 이른바 기술적 오류만을 해소한 개발사의 '이루다 2.0'에 대한 향후 계획의 발표와 서비스 재개는 인공지능 설계에서 페르소나와 같은 문화적 목표와 가치지향성에 대한 고민보다는, 데이터와 기술적 문제 해결(가명처리 고도화)만이 의인화 챗봇의 핵심적 고려사항이었다는 사실을 방증한다. 여기에 기존의 비정형 데이터의 비식별화 처리 기술 자체의 한계 내지 미비는 의인화 챗봇의 문제점을 좁은 의미의 기술공학적 문제로만 보는데 사안을 집중하게 만들었다.

> (피심인 대표이사) … 비정형 데이터는 과연 개인정보가 그 안에 얼마나 있는지, 확인하기도 매우 어렵지요. 다 말하자면 음성데이터는 다 들어봐야 아는 것이기 때문에. 그것을 어떻게 할 것인가, 기술적인 방법론은 무엇인가, 아니면 어느 정도까지 해야 되나, 이런 것들이 아직 어떻게 보면 완벽하게 확립되지 않은 상태로 남아 있는 것 같습니다. 그래서 저희도 그런 부분에 고민이 많고 저희가 기술적으로도 그런 기여나 선례를 만들어가고 싶은 생각도 있습니다만, 아직 그런 것들이 확립되지 않은 상황이라는 것도 고려해 주셨으면 좋겠습니다(개인정보보호위원회, 2021, 24쪽).

이는 단지 개발사의 변명이라기보다는 데이터 수집부터 비식별화 처

15) 개보위 조사 결과, 이루다의 혐오발언은(편향된) 데이터를 잘못 학습한 인공지능이 편향성을 가중시킨 것이 아니라, (편향된) 데이터를 그대로 출력한 것이었다. 한 개보위 간부에 따르면, "루다의 경우 이용자들이 말을 이상하게 걸어서(카카오톡에 실재했던) 이상한 답변을 한 것으로 AI가 학습을 통해 평가를 가중한 게 아니"며 "그래서 MS의 챗봇 테이와 다르다. 대단히 특별한 케이스"라는 것이다(김현아, 2021). 아울러 개인정보수집의 불법성으로 인한 개인정보침해는 사실로 판단되었다.

리와 발화(인간적 대화)에 이르는 일련의 과정을 기술 한계와 정보유출의 차원에서 논의(번역)하는 개보위의 조사과정에 내재하는 담론효과라고 할 수 있다. 개보위의 조사결과는(총 8가지에 걸친) 개인정보보호법 위반에 따른 과징금과 과태료로 일단락되었지만, 궁극적으로 이번 논란이 "개인정보보호와 AI 산업 개발이라는 이 두 마리 토끼를 어떻게 같이 잡을 것이냐"를 고민하는 "하나의 계기"라는 차원에서 인식하는 개발사의 입장이야말로 본 사태를 가로지르는 지배적 기술 담론을 대변한다고 할 수 있다(개인정보보호위원회, 2021, 30-31쪽 참조). 즉 개보위의 조사과정은 우리사회와 개발사가 주목해야 할 인공지능 개발 의제에서 페르소나와 연관된 문화적 설계의 문제를 배제하고, 인공지능(설계)을 좁은 의미의 알고리즘 기술의 공학적 실현의 한계 내에 국한시킴으로써 인공지능 기술과 그 사용에 대한 대안적 상상력을 제한한 것이다. 물론 결과적으로 볼 때, 이루다가 편향 발언과 같은 중립적이지 않은 퍼포먼스를 보여주었다는 사실은 개발사를 비롯한 모두에게 인정될 수 있는 사실이지만, 이러한 문제들을 해결하는 과정과 해법 차원에서 제시되는 기술은 역설적에게도 중립적 도구인 것처럼 인식되었다. 이는 인공지능 기술을 보다 중립적이고 합리적인 방식으로 보이도록 자연화하고 특권화하는 우리사회 기술 신화의 단적인 모습을 시사하기도 한다.

결국 이루다 사태의 최종 국면은, '이루다 네트워크'의 기술-사회 혼종성을 은폐하고 기술의 문화적 설계를 기술 외부의 문제로 배제하며 기술공학을 정당화한 의인화의 결정적 과정이었다. 다시 말해 여러 행위자들을 통해서 이루다를 의인화하는 과정들은 인공지능 기술이 갖는 총체적(기술-문화 하이브리드) 문제를 인공지능의 기술공학적 차원과 문화·정치적 차원으로 분리(정화)시켜, 우리사회의 인공지능 신화, 즉 인공지능의 성격과 설계 문제를 단순히 기술공학적 차원의 문제인 것으로 제한하는 기술신화를 재생산하는 사회적 과정이었던 것이다.

4) 챗봇 '이루다 사태'의 국면 종합 분석

챗봇 이루다 사건의 이제까지 진행 과정을 ANT의 '번역' 관점에서 정리해보면 〈그림 3〉에서 처럼 요약, 정리해 볼 수 있겠다(Latour, 1999/2018, 6장; 2012, p. 43, "번역 작용을 나타내는 도식의 일반화" 참조). 이루다는 공식 출시되자마자 끊임없는 테스트 상황에 노출되었다고 할 수 있다. 본래 개발사가 의도한 이루다의 고유 목표는 친구같은 관계를 맺는 챗봇의 구현에 있었다. 하지만 다양한 이용자들과 접속하면서 개발사가 기획한 이루다의 문화적 정의, 즉 페르소나는 의도했건 안했건 다른 행위자(이용자)들의 이해 관심과 해석에 의해 번역되면서 조금씩 빗겨가기 시작했다. 챗봇에게 사회적 여성 인격(페르소나)을 부여하려는 개발사의 젠더화된 기술 설계(1차 의인화)는 이용자들의 젠더 스테레오타입과 오남용의 의지에 조응함으로써 챗봇의 성적 대상화를 완성했다(2차 의인화). 결국 이루다 테크놀로지는 새로운 행위자들의 개입으로 인해 본래 의도대로 작동되지 않고 '고장'이 난 셈이다. 개발사가 의도한 이루다 본래의 행위 프로그램이 방해를 받아 새로운 의미의 우회로가 생성되기 시작하고 인공지능의 암흑상자가 열리기 시작한 것이다. 이루다가 무차별적으로 내뱉은 차별·혐오 발언이 사회 쟁점화되면서 이에 대한 언론 및 전문가들의 해석도 부각됐다. 또한 이용자들에 의해 제기된 개인정보 유출로 의심되는 정황은 개발사가 설정한 '친구 같은 챗봇'의 고유 목적과 행위를 법적 쟁점과 데이터 주권의 의미로 번역하게 되는 계기로 작용했다. 이루다는 그저 친구가 아니라 민감한 개인정보의 담지체로서 인식되기 시작한 것이다. 개발사의 편향된 1차 훈련 데이터와 이에 대한 비식별화 처리 프로세스의 미숙, 그리고 단순 검색모델이라는 기술적 결함들은 더욱 문제를 키웠다. 당시 이들 기술적 문제점을 다룰 때, 개발사나 일부 논평자들은 어린아이론(기술의 미성숙론 또는 기술한계론)으로 사태를 정당화함으로써 의인화를 또 다른 방향에서 강화했다(3차 의인화). '어린아이' 의인화는 기술적 오류를 은폐하거나 현실의 기술적 한계를 정당화

하는 기술합리성의 '알리바이'로 작용했다. 의인화의 문제점을 또 다른 의인화로 덮었던 셈이다. 이러한 의인화는 기술적 한계와 결합하면서 궁극의 기술신화로 완성된다.

이러한 번역의 기제들은 결국 개보위라는 제도적 행위자를 불러들이게 되었고, 사건이 진행되면서 또다른 해석과 판단을 하는 번역의 우회로를 창출하게 되었다. 이 우회로는 궁극적으로 특정한 챗봇 기술을 개발사의 여성 페르소나 설정 기획으로부터 떼어내는 번역의 양상을 보여준다. 즉 이루다의 문제점이 여성 페르소나 설계 자체의 문제가 아니라, 오롯이 기술공학적 설계의 문제로 환원되었다. 챗봇 기술의 문제점이 기술-문화의 총체적인 것에서 기술공학만으로 문제점으로 축소된 것이다. 따라서 주류 해법 역시 이루다 페르소나 기획과 그것의 기술적 구현의 관계에서 찾기 보다는, 페르소나에 대한 고민 없이 단지 이용자를 탓하거나 데이터셋과 모델을 개선하는 것에 국한될 수밖에 없게 되었다.

결과적으로 개발사의 제품기획 의도와 알고리즘 설계의 의도, 그리고 사전 훈련용 데이터셋은 그 의도된 활용대로 작동하지 않고 중단이 되었다. 개발사의 입장에서 보면 이 사건은 이루다 테크놀로지 외부(사회)의 방해 때문에 발생한 일이다. 만일 외부 행위자들의 오남용이 아니었다면 이루다의 알고리즘은 기존 이루다의 페르소나의 목적과 행위 프로그램을 따라, 중단되지 않고 '정상' 작동되었을 것이다. 하지만 이루다 테크놀로지는 기술 '외부'의 행위자들과 분리되어 존재하는 것이 아니라, 새로운 행위자와 사건의 접속으로 인해 테크놀로지를 구성하는 혼성적, 연합적 존재다. 따라서 이루다를 향한 언어폭력이나 이루다 자신의 언어폭력은 단지 알고리즘 모델이나 데이터셋의 기술적 한계와 오작동 때문만이 아니라, 챗봇의 문화적 요인인 페르소나와 이 페르소나를 구현하는 알고리즘 모델, 데이터셋, 나아가 이를 통계적으로 반영하는 사회적 문화의 종합적 관계로부터 야기된 문제로 볼 수 있다.

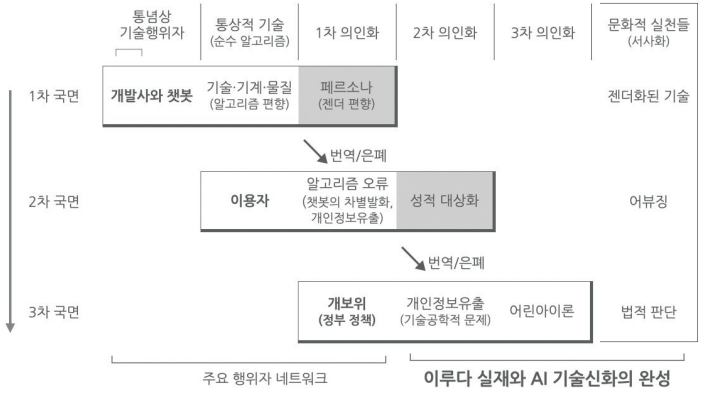

그림 3. 이루다 네트워크에 대한 구성주의적(ANT) 분석

5. 나오면서: 의인화와 인공지능 중립론의 기술신화를 넘어서

이 연구는 이제까지 이루다 사태를 선례로 삼아, 우리사회의 인공지능 기술 설계를 특정한 방식과 방향으로 편향시켜 은폐, 구성, 정당화하는 기술신화의 사회적 착근 과정을 분석하고, 이 기술신화의 성립에 의인화가 중요한 역할을 했음을 보이고자 했다. 중요한 기술사회사적 사건으로서 챗봇 이루다는 우리사회가 견지한 인공지능에 대한 암묵적 믿음, 즉 기술신화를 드러내는 전형적인 인공물이었다.

챗봇 이루다의 기술신화는 단지 하나의 행위자나 하나의 국면에서만 완성되는 것이 아니었다. 그리고 그 기술 신화는 복합적인 방식과 단계로 구성됐다. 자유대화형 챗봇을 실현하려 했던 기업의 목표(행위프로그램)는 사실상 '인간다운 인공지능'(의인화) 구상에 의해 은폐됨으로써 역설적으로 인공지능 신화를 완성하고 강화하는 데 기여했다. 달리 말해, 챗봇 검색모델의 기술적 특성(한계성)은 친밀성 같은 젠더 고정관념을 의인화한 이루다의 여성형 페르소나(젠더화된 테크놀로지)와 상호작용하는

성적 대상화에 의해 폭로되면서도 곧바로 이루다의 페르소나와 미성숙한 '어린아이'라는 의인화 담론에 의해 은폐되었다. 개발사는 편향과 혐오 발언 문제와 비판에 직면하여 챗봇 기술 자체를 수정·보완해 문제점을 그저 '기술적으로만' 해결할 문제로, 최신 기술이 갖는 불가피한 부작용 정도로 축소했다. 이 과정에서 문제에 대한 해결책은 '기술적' 결함 보완책이 전부가 됐고, 이루다의 페르소나와 이용자들의 오남용은 기술 외적 문제로 남게 되었으며, 여기에 개보위의 승인이 일종의 공모로 작용했다. 개보위의 조사과정을 비롯한 이루다 사태의 전개 국면들은 인공지능을 문화적 기획의 산물에서(드러나는 문제점에도 불구하고 계속 발전시켜나가야만 하는) 단지 '최신 기술' 그 자체로 번역하고 정화한 셈이다. 결국 챗봇의 기술적 설계(알고리즘)는 문화적 설계(페르소나)와 분리·이원화되고 말았다. 따라서 이에 대한 제도적 대응 역시 공학적·법적 대책과 문화적·사회적·윤리적 대책으로 이원화되고 말았다.

이루다 사건은 단지 챗봇 알고리즘 및 기계학습의 오류나 한계 문제만은 아니었다. 문제는 알고리즘과 기계학습의 기술 공학적 차원을 넘어서서, 이루다가 기술과 문화라는 이종적인 행위소들로 구성된 네트워크 또는 집합체라는 사실에 있었다. 다시 말해, 이루다라는 기술적 존재의 성립과 문제적 원인은 단지 알고리즘 기술만으로 귀인되지 않고, 집합체의 전체적 관계 안에서 설명되어야 했다. 즉 인공지능은 '순수 기술 시스템'이 아니라 기술-문화의 총체적 네트워크에 가깝다는 말이다(Crawford, 2021/2022 참조). 기술을 네트워크로서 이해한다는 것은 기술을 둘러싼 네트워크에 의해 그 기술의 정체성도 변한다는 것을 의미한다(홍성욱, 2010, 141쪽).

결과적으로 이루다 사태는 인공지능에 활용된 정제되지 않은 훈련데이터의 문제뿐만 아니라, 초기 챗봇의 여성형 페르소나로 표상되는 제품 기획과 검색 알고리즘 모델이 갖는 기술적(알고리즘) 특성이 인간 행위자들의 이용과 해석 행위(오남용)를 조건짓고, 또 역으로 인간의 행위가

개발사의 의도를 굴절시키고 번역하는 과정을 통해서 진행된 사건이다. 이것은 우리가 인공지능의 문제를 볼 때, 기술과 문화의 총체적 결합의 양상을 전체적으로 관찰하고 읽어내야 함을 뜻한다. 즉 기술적 설계는 문화적 설계와 분리되지 않는다는 것이다.

인공지능 시장이 이제 형성 단계라고 상정한다면, 이루다 사태는 그 과정에서 불거진 사회적 징후와 같다. 비슷한 인공지능의 사회 이슈가 주기적으로 계속 터져 나올 확률이 크다. 인공지능은 '사회적으로 민감한 기술'이다. 인공지능의 수요는 코로나 팬데믹 이후 '비대면(언택트)' 자동화사회를 요구하면서 점점 늘어나는 추세다. 고독과 우울이 현대인을 짓누르는 현실에서, 챗봇 같은 대화형 인공지능은 이번 사태로 잠시 주춤했지만 그 성장세를 멈추지 않을 것이다. 이루다 사태는 인공지능과 그 설계에 대한 산업계와 우리사회의 기술신화를 드러내고 지능형 기술 체제가 사회적으로 고착화되는 방식을 잘 보여줬다. 이루다 사태는 지능 정보화 기술을 우리 사회 속에 어떻게 안착해야 할 지에 대한 일종의 기술사회사적 '선례(precedent)' 구실을 했다. 물론 이루다가 아직까지는 인공지능에 대한 우리사회의 기술감각 또는 기술문화를 완성하고 지배하는 사례라거나 완전히 안정화된 기술체제를 대표한다고 단언하기는 어려울 것이다. 핀치와 바이커(Pinch & Bijker, 2012)의 주장처럼, 기술의 안정화는 더 이상의 논쟁을 종결짓고 문제의 소멸 상황에 이르겠지만, 이루다는 여전히 진행 중인 프로젝트이고 이와 유사한 AI 기술 서비스의 등장으로 인해 관련된 사회적 논쟁은 계속될 것이기 때문이다. 즉 현재로서 우리 사회의 인공지능 기술신화는 확고한 것이라기보다는 아직까지 열린 구성 과정과 진화 단계에 있다고 보는 편이 맞을 것이다.

개발자의 기술은 이용자의 문화적 실천에 틈입하고, 이용자의 문화적 실천은 기술의 한계를 시험하여 그 실제적 양상을 변화시키며, 정책집행기관은 기술과 문화의 경계를 확정지음으로써 기술과 문화적 실천(사용) 모두에 영향을 끼쳐 우리 사회의 기술신화를 형성해 간다. 그러므

로 개발자와 정책집행기관은 기술이 이용자들의 문화적 실천에 영향을 준다는 차원에서 이를 신중하게 설계·개발·의제화 해야 하고, 문화가 기술적 내용에 대한 개입을 통해 기술설계와 그 긍정적·부정적 효과를 결정할 수 있음을 인식해야 한다. 이것이 최신 기술이 주어진 본성이나 정답이 아니라, 문화적 실천에 의해 변형될 수 있는 것임을 인정하는 태도인 것이다. 이번 이루다 사태의 재발을 방지하기 위해서는 이 사안을 기점으로 사회적으로 민감한 인공지능 등 첨단기술의 사회적 유통과 보급에 앞서, 개발자, 이용자, 정책 집행기관 등 관련 제 주체들이 참여하는 AI 기술의 사회적 영향 평가를 상시화해 사회적 마찰을 줄이는 숙의의 장을 본격적으로 가동시키는 일이 시급하다.

아울러 유사 이루다 사태의 발생을 사전에 차단하기 위해서는 보다 현실주의적인 접근이 필요하다. 일차적으로, 개발사의 해명 과정에서 보듯 개발자 집단 내부에 인공지능 윤리 가이드라인이 부재하다는 점을 알 수 있다. 적어도 인공지능 윤리 내규를 만들고, 개발자들에게 이에 대한 원칙과 가이드라인 교육을 의무화해야 한다. 데이터셋 편향과 어뷰징 여부, 그리고 문화적 의제들에 대해 정보인권전문가와 젠더, 장애, 인권 활동가로부터 시스템 개발 시 자문을 받는 공식 절차를 마련하는 것도 중요해 보인다. 중소 규모 스타트업 사업자들이 기본적으로 기술설계의 문화적 차원, 특히 넓게는 차별과 불평등과 같은 인권 의제에 관한 감수성, 좁게는 AI 윤리 가이드라인을 익힐 채널이 부족하다는 점도 큰 문제다. 물론 개발자 가이드라인만으로는 변화를 이끌기에 부족하다. 때로 가이드라인은 법적 제재를 회피하기 위한 알리바이로 종종 활용되는 까닭이다. 그런 의미에서 인공지능 개발의 윤리나 원칙을 어길 시에 규제하는 법적 근거 마련이 요청된다. 궁극적으로는 이에 대한 사회적 토론의 상대적 부재도 기업이나 개발자들이 기술설계의 문화적 차원을 고려하는데 있어서 약점으로 작용한다. 이에 대한 대책 마련과 활발한 사회적 논의가 시급하다. 이것이 바로 챗봇 이루다를 기술공학-기술문화 하

이브리드의 관점에서 이해함으로써 인공지능 의인화 신화를 극복하는 하나의 방안이 될 것이다.

참고문헌

개인정보보호위원회 (2021. 4. 28). 2021년 제7회 개인정보 보호위원회 속기록. URL: https://www.pipc.go.kr/np/default/minutes.do;jsessionid=oehqZUkob7M6BkxjZmbI3Q40.pips_home_jboss21?op=view&idxId=6883&page=2&mCode=E020010010&fromDt=&toDt=&schCatCd=1&schTypeCd=1&typeCd=1&catCd=1

김경은·강진숙 (2023). 인공지능 (AI) 의 젠더화된 목소리와 주체화 방식에 대한 사례연구: 푸코의 장치와 주체화 사유를 중심으로. 〈한국방송학보〉, 37권 2호, 5-39.

김규희 (2021. 1. 8). 여성 AI까지 성착취…온라인서 '루다 성노예 만드는 법' 공유. 〈여성신문〉. URL: https://www.womennews.co.kr/news/articleView.html?idxno=205906

김민선 (2021. 1. 7). 써본 이들이 먼저 '사람 같다' 칭찬하는 챗봇 '이루다'. 〈ZDNet Korea〉. URL: https://news.naver.com/main/read.naver?mode=LSD&mid=shm&sid1=105&oid=092&aid=0002210002

김수향 (2021 .7. 23) 〔인터뷰〕 과학기술학 연구자 임소연 교수 (1). 〈COMMONS LAB〉 URL: http://commonslab.cc/115/-인터뷰-과학기술학-연구자-임소연-교수-1-

김정민 (2021. 8. 30) 〔팩플〕 '이루다' 그후 반년…"루다는 관계의 불평등 해결할 AI 될 것". 〈중앙일보〉. URL: https://www.joongang.co.kr/article/24119834#home

김종윤 (2020. 12. 1). 오픈도메인 챗봇 '루다' 육아일기: 탄생부터 클로즈베타까지의 기록. 〈DEVIEW(네이버랩스 연례개발자 컨퍼런스) PPT〉. URL: https://tv.naver.com/v/16968268

김현아 (2021. 4. 28). 루다가 내뱉은 혐오발언, AI가 지어낸 게 아니다. 〈이데일리〉. URL: https://www.edaily.co.kr/news/read?newsId=03775286629020712&mediaCodeNo=257

남혜현 (2020. 7. 25). AI 챗봇 '루다'와 랜선 친구가 됐다. 〈바이라인네트워

크〉. URL: https://byline.network/2020/07/24-78

백욱인 (2021). 인공지능 시대의 기계들과 인간들. 〈문화/과학〉, 105호, 27-52.

손희정 (2022). 인공지능과 젠더 테크놀로지: 이루다 1.0 논란을 중심으로. 〈젠더와 문화〉, 15권 2호, 67-94.

이광석 (2021a). 챗봇 '이루다'가 우리 사회에 남긴 문제: 인공지능에 인권 메뉴얼 탑재하기. 〈문화/과학〉, 105호, 138-198.

이광석 (2021b). 〈포스트디지털: 토픽과 지형〉. 파주: 안그라픽스.

이시연 (2018). (인공) 지능은 성별이 없다고? 〈인공지능인문학연구〉, 1권, 77-93.

이정현 (2020). 인공지능 젠더 편향성과 포스트휴먼 주체. 〈AI와 더불어 살기〉(211-235쪽). 서울: 커뮤니케이션북스.

이종임 (2020). AI와 여성 개발자: 기술 산업이 갖는 젠더 불평등.. 〈AI와 더불어 살기〉(237-256쪽). 서울: 커뮤니케이션북스.

이효석 (2021. 1. 8). 'AI 이루다' 개발사 "성희롱 예상했다…심한 게시물 강력 대응". 〈연합뉴스〉. URL: https://www.yna.co.kr/view/AKR20210107153353017

이효석 (2021. 1. 12). 스캐터랩 "이루다, 문장 속 실명 전부 못 걸렀다"…파기는 안해(종합). 〈연합뉴스〉. URL: https://www.yna.co.kr/view/AKR20210112137851017

이희은 (2018). AI는 왜 여성의 목소리인가? 〈한국언론정보학보〉, 90호, 126-153.

임소연 (2021. 3. 5). 여성을 차별하지 않는 인공지능을 만들 수 있을까? 〈한겨레〉. URL: https://www.hani.co.kr/arti/culture/book/985521.html

임종수·신민주·문훈복·윤주미·정태영·이연주·유승현 (2017). AI 로봇 의인화 연구: '알파고'보도의 의미네트워크분석. 〈한국언론학보〉, 61권 4호, 113-143.

임종수·최진호·이혜민 (2020). AI 미디어와 의인화: AI 음성 대화형 에이전트

의 의인화 평가척도 개발 연구. 〈한국언론학보〉, 64권 4호, 436-470.

조행만 (2020. 10. 1). 인간 곁으로 다가온 AI 로봇의 윤리적 설계, '아이 키우는 것처럼 해야'. 〈AI타임즈〉. URL: http://www.aitimes.com/news/articleView.html?idxno=140597

진동수 (2021). 인공지능 챗봇의 성공과 실패에 미치는 요인에 관한 연구. 〈차세대융합기술학회논문지〉, 5권 2호, 168-175.

진보래 (2020). 인공지능은 우리의 친구가 될 수 있을까? 〈AI와 더불어 살기〉(3-41쪽). 서울: 커뮤니케이션북스.

최광민 (2020. 12. 23). 스캐터랩, 세계 최고 수준의 언어능력 보유한 인공지능 '루다' 정식 출시. 〈인공지능신문〉. URL: https://www.aitimes.kr/news/articleView.html?idxno=18758

한애라 (2019). 인공지능과 젠더차별. 〈이화젠더법학〉, 11권 3호, 1-39.

홍성욱 (2010). 인간과 기계에 대한 '발칙한' 생각: ANT의 기술론. 홍성욱 (편), 〈인간·사물·동맹: 행위자네트워크 이론과 테크노사이언스〉(125-154쪽). 서울: 이음.

Blok, A., & Jensen, T. E. (2011). *Bruno Latour: Hybrid thoughts in a hybrid world*. Abingdon, UK: Routledge. 황장진 (역) (2017). 〈처음 읽는 브뤼노 라투르: 하이브리드 세계의 하이브리드 사상〉. 고양: 사월의책.

Bryson, J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues*(pp. 63-74). Amsterdam, Netherlands: John Benjamins Publishing.

Caporael, L. R. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior, 2*(3), 215-234.

Coeckelbergh, M. (2012). Are emotional robots deceptive? *IEEE Transactions on Affective Computing, 3*(4), 388-393.

Costa, P., & Ribas, L. (2019). AI becomes her: Discussing gender and artificial intelligence. *Technoetic Arts: A Journal of Speculative Research, 17*(1-2), 171-193.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. New Haven, CT: Yale University Press. 노승영 (역) (2022). 〈AI 지도책: 세계의 부와 권력을 재편하는 인공지능의 실체〉. 서울: 소소의책.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*(3-4), 177-190.

Epley, N. (2018). A mind like mine: The exceptionally ordinary underpinnings of anthrpomorphism. *Journal of the Association for Consumer Research, 3*(4), 591-598.

Epley, N., Waytz, S., & Cacioppo, J. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition, 26*(2), 143-155.

Feenberg, A. (1999). *Questioning technology*. London, UK: Routledge. 김병윤 (역) (2018). 〈기술을 의심한다: 기술에 대한 철학적 물음〉. 서울: 당대.

Feenberg, A. (2017). *Technosystem: The Social Life of Reason*. Cambridge, MA.: Harvard University Press.

Hartzog, W. (2015). Unfair and deceptive robots. *Maryland Law Review, 74*, 785-829.

Jung, E. H., Waddell, T. F., & Sundar, S. S. (2016). *Feminizing robots: User responses to gender cues on robot body and screen*. Paper presented at the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, San Jose, CA.

Krementsov, N. L., & Todes, D. P. (1991). On metaphors, animals, and us. *Journal of Social Issues, 47*(3), 67-81.

Latour, B. (1992). Where are the missing masses? The sociology of

a few mundane artifacts. In W. E. Bijker & J. Law (Eds), *Shaping technology/building society: Studies in sociotechnical change*(pp. 225-258). Cambridge, MA: The MIT Press.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge, MA: Harvard University Press. 장하원·홍성욱 (역) (2018). 〈판도라의 희망: 과학기술학의 참모습에 관한 에세이〉. 서울: 휴머니스트.

Latour, B. (2010). An attempt at a 'compositionist manifesto'. *New Literary History, 41*(3), 471-490.

Latour, B. (2012). *Cogitamus: Six lettres sur les humanités scientifiques*. Paris, France: La Découverte. 이세진 (역) (2012). 〈브뤼노 라투르의 과학인문학 편지: 인간과 자연, 과학과 정치에 관한 가장 도발적인 생각〉. 고양: 사월의책.

Law, J. (2010). ANT에 대한 노트: 질서 짓기, 전략, 이질성에 대하여. 홍성욱 (편), 〈인간·사물·동맹: 행위자네트워크 이론과 테크노사이언스〉 (37-56쪽). 서울: 이음.

Natale, S. (2021). The ELIZA Effect: Joseph Weizenbaum and the emergence of Chatbots. In S. Natale (Ed.), *Deceitful media: Artificial intelligence and social life after the turing test*(pp. 50-67). New York, NY: Oxford University Press. doi: 10.1093/oso/9780190080365.003.0004

Pinch, T. J., & Bijker, W. E. (2012). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. In W. E. Bijker, T. P. Hughes, & T. Pinch (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology*(pp. 11-44). Cambridge, MA: MIT press.

Ransbotham, S. (2018. 5. 21). Rethink AI objectives: Using AI to create humanlike computers is a shortsighted goal. *MIT*

Sloan Management Review. Retrieved from https://sloanreview.mit.edu/article/rethink-ai-objectives

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Stanford, CA: CSLI Publications.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience, 11*(2), 88-95.

Shneiderman, B. (1989). A nonanthropomorphic style guide: overcoming the humpty-dumpty syndrome, *The Computing Teacher 16* (7) , 1-5.

Stojnić, A. (2015). Digital anthropomorphism: Performers avatars and chat-bots. *Performance Research, 20*(2), 70-77.

Watson, D. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines, 29*(3), 417-440.

Weber, J. (2005). Helpless machines and true loving care givers: A feminist critique of recent trends in human-robot interaction. *Journal of Information, Communication and Ethics in Society, 3*(4), 209-218.

투 고 일 자: 2023년 10월 05일
심 사 일 자: 2023년 11월 01일
게재확정일자: 2023년 11월 23일

Abstract

Questioning Anthropomorphism as an AI Techno-myth

A Case Study of the chatbot 'Iruda'

Hyun Jun Kim
Ph. D. candidate., Seoul National University of Science & Technology
Kwang-Suk Lee
Professor, Seoul National University of Science & Technology

This study considers AI anthropomorphism as a powerful component and mechanism of today's AI technomyth. This paper critically analyzes the chatbot Iruda as a case from a constructivist perspective that reveals the ways in which this technological myth has been created and reproduced. Through the critical analysis of the Iruda case, this study observes that the anthropomorphism of AI tends to reproduce and reinforce the techno-neutralism and tech-fetishism of AI by separating the dual aspects of its technological engineering and technological culture. However, the anthropomorphisation of AI cannot be reduced to the strategic intentions of developers or specific actors, but, is a network effect that involves multiple actors (or actants) within the social debate.
Today, the logic of intelligent information technology has taken on the appearance of independent and autonomous artificial objects in close relation to the myth of 'human-like AI'. Specifically, the chatbot technology used in Iruda, despite being co-constituted by a hybrid network of AI, the technology itself, developers, government regulators, and society (anonymous AI users), paradoxically contributes to the techno-myth of AI sublime by concealing its relational aspects through the logic of "human-like AI"

(anthropomorphism). This study intends to refute the replacement in which the problematic nature of the socially multilayered entanglement of AI is reduced to the problem of solving technological defects or flaws in a narrow sense. For doing that, this study critically explores the anthropomorphic processes of AI that are mobilized to reinforce technological determinism: persona (the first anthropomorphism), sexual objectification (the second anthropomorphism), childlike-ness (the third anthropomorphism), and finally the Personal Information Protection Commissioner's policy intervention into the case of Iruda.

KEYWORDS Human-like AI, chatbot, anthropomorphism, techno-neutralism, technological myth, constructivism, hybrid, ANT