인공지능 언어모델에 대한 규범적 연구
- 일반 대화형 챗봇에 대한 실증연구를 중심으로 -

정 종 구

초 록

인공지능은 기존의 기술발전보다 급격하고 근본적인 사회변화를 초래할 것이다. 하지만 기존 국내외 정책당국의 대응은 실증적이고 객관적인 데이터에 근거했던 것으로 보이지 않는다. 이에 본 연구는 인공지능 활용에 있어 실제로 문제되는 위해가 무엇인지를 실증적으로 규명하였다. 문제의 원인을 어떻게 파악하는지에 따라 그 대응방안이 달라질 것이다. 구체적으로는 한국 사회에서 일반 대화형 챗봇을 둘러싸고 2021년 동안 전개되었던 실제 사건을 토대로 확보한 데이터에서 경험적으로 발현된 위해를 기존 인공지능 윤리담론에서 사변적으로 전제하였던 위해와 비교·대조함으로써 시사점을 도출하였다.

우선 인공지능 언어모델에서 제기되었던 규범적 문제를 체계적으로 정리하였다. 규범적인 논의는 일반적으로 인공지능 윤리의 문제로 언급되어 오던 공정성, 투명성, 책무성, 그리고 개인정보보호 측면에서 체계적으로 분류하여 소개하였다. 단순히 추상적인 윤리담론을 열거하는 수준에 그치지 않고 해당 영역에서 먼저 도입되었던 실정법이 있다면 그와의 관계를 유기적으로 살펴보았고, 관련된 기술적 쟁점이 있다면 필요한 만큼 언급하였다. 이러한 방식으로 선행연구를 검토함으로써 실증연구를 바탕으로 비판적으로 검토할 수 있는 이론적인 토대를 마련하였다.

다음으로 실제 데이터에 기초한 검증 및 평가에서는 기존에 전개되었던 인공지능 윤리담론을 실제 사건을 토대로 확보한 9,747건의 데이터에 비추어 실증적으로 검토하였다. 기존의 추상적인 관념과 단편적인 사례 위주로 이론화되어 전개되었던 인공지능 윤리담론이 전제했던 위해의 범주와 비중이 실제 사안에서는 그대로 유효하지 않을 수 있다는 점을 확인했다. 인공지능 윤리담론은 공정성-투명성-책무성-식별성 침해라는 위해를 고르게 전제하고 있었고 실제 법정책에도 그러한 내용이 반영되고 있는 반면, 실제 챗봇으로부터 발생한 위해는 주로 책무성 측면에 치우쳐 있었다. 이러한 위해는 결정적이지 않고 확률적이었으며, 이용자와의 상호작용을 염두에 두어야 비로소 온전히 파악하고 대처할 수 있었다.

앞으로 더 많은 실증연구가 필요하다. 인공지능 기술이 발전하면서 어떠한 위해를 새로이 유

발하게 될지는 아무도 모르며 단일한 기술을 전제하더라도 개별 응용프로그램 단계에서 어떻게 활용되는지에 따라 발생할 수 있는 위해의 양상은 크게 달라질 수 있기 때문이다. 편익이 비용을 상회하는 이상 실제 경험적인 데이터에 기초하여 위해를 측량하고 이를 규명한 후 적절히 대처할 수 있는 방안을 모색하는 피드백 루프를 구축해야 한다. 또한 이용자의 행태에 관심을 가져야 한다. 개별 인공지능 시스템의 특성에 따라 인공지능이 인간에게 미치는 영향뿐만 아니라, 인간과 인공지능의 상호작용 내지 인간과 인간 사이의 상호작용이 중요할 수 있기 때문이다. 이용자의 전략적인 행동과 이용자 상호간 정보교류를 통한 위해의 심화 내지 완화를 충분히 고민한 실증적인 윤리담론과 그에 기초한 법・정책이 필요하다.

Ⅰ. 서론

인공지능에 대한 장밋빛 전망이나 암울한 비관이 난무하고 있다. 기술이 발전함에 따라 사회가 바뀌는 과정에서 나타나는 자연스러운 현상이다. 기술발전이 일어나고 그로인해 사회변화를 초래할 때마다 기대와 우려가 교차되었고, 그때마다 인류는 지혜를 발휘하며 문명의 발전을 이루어 왔다. 인공지능은 기존의 기술발전보다 급격하고 근본적인 사회변화를 초래할 것이다. 그에 대응하여 국내외 정책당국은 인공지능에 대한 윤리원칙과 가이드라인을 수립해 왔다. 하지만 이러한 정책당국의 대응은 실증적이고 객관적인 데이터에 근거했던 것으로 보이지 않는다.

본 연구는 인공지능 활용에 있어 실제로 문제되는 위해(危害)가 무엇인지를 실증적으로 규명하기 위해 고안되었다. 문제의 원인을 어떻게 파악하는지에 따라 그 대응방안이 달라질 것이기 때문이다. 기존 연구에서는 국내외를 막론하고 많은 경우 이론연구를 수행하여 인공지능의 활용으로부터 발생할 위해가 무엇인지를 연역적 개념으로 제시하고 분석하였다(top-down approach).[1] 이런 논의는 일응 직관적이고 일목요연하며 논리적으로 보일 수 있으나, 실제 문제를 잘못 진단하면 인공지능의 활용을 크게 저해하기만 할 뿐 현실적인 위해발생에 적절하게 대처할 수 없을 것이다. 본 연구에서는 인공지능의 활용 과정에서 발생했던 위해를 실제 사례와 그로부터 취득한 데이터를 가지고 직접 확인한다(bottom-up approach). 실제 사건에서 발생한 데이터를 가지고 실증적으로 확인함으로써 기존 논의의 한계를 지적하고 향후 연구방향을 모색한다.

Ⅱ. 연구의 대상: 챗봇 이루다 사건

본 연구의 배경은 챗봇 이루다 사건[2]이다. 이는 인공지능 언어모델을 기반으로 개발된 챗봇이 이용자와 비윤리적인 대화를 나누는 모습이 포착되어 '인공지능 윤리'에 대한 사건으로 언론의 주목을 받았고, 개인정보보호위원회가 개인정보 보호법 위반으로 제재하였던 사건이다. 당시 N번방 사건이 종결된 직후의 사회분위기가 반영되어 주목 받고 있었던 젠더 문제가 20대 여성 대학생의 페르소나를 지닌 챗봇을 둘러싸고 발생한 본건에 접목되면서 인공지능이 유발할 수 있는 위해가 무엇이 있는지가 집중적으로 논의되었다. 그 결과 인공지능 언어모델의 규범적 이슈를 미리 들여다볼 수 있는 기회를 제공하게 되었고, 인공지능 윤리 관련 논의를 촉발하였다.

---

1) 본 연구는 지금까지 국내외에서 발표되었던 인공지능 윤리분야 연구문언의 상당수를 다루었다. 하지만 인공지능 활용으로 인해 어떠한 문제가 제기되는지를 실증적으로 다룬 논문은 거의 찾기가 어려웠다. 단편적인 사건을 통해 발생 가능한 위해를 사전에 측량하고 그 대응방안을 모색하는 논의들이 단편적으로 전개되었을 뿐이다(COMPAS 사건이 대표적이다). 심지어 유럽연합에서 제안된 인공지능 법(안) 조차도 공정성, 투명성, 책무성, 식별성 침해 같이 추상적으로 정립된 개념을 주요한 위해로 전제하였을 뿐이다. 본 연구는 국내 실제 사안을 대상으로 수집한 1만여 건의 데이터를 바탕으로 실제 인공지능 활용으로부터 발생하게 되는 위해를 실증적으로 검증하였다는 점에서 의의가 있다. 다만, 챗봇과 이용자 간 대화내역을 전부 확보하지 못하고 발췌하였는데 랜덤 샘플링 기법에 의한 것이 아니라는 점에서 분석 대상 데이터가 치우쳐(skewed) 있을 수 있고, 이용자가 임의로 조작한(manipulated) 이미지가 포함되어 있을 수 있어 선택 편향(selection bias)가 높을 수밖에 없다는 한계가 있다. 따라서 본 연구에서 챗봇 이루다와 이용자 간의 대화내역 중 일상적인 대화가 약 74% 정도라고 명시하더라도, 이는 실제 그렇다는 것이 아니고 연구를 위해 인터넷 커뮤니티를 통해 현실적으로 확보할 수 있는 데이터에서의 차지하는 비중이 그렇다는 취지이다.

2) 비록 인공지능 서비스로 인해 유발된 수많은 사건사고 중 한 건에 불과하지만, 이를 계기로 한국사회에서 인공지능 윤리가 본격적으로 논의되기 시작했다는 점에서 의미가 있다. 2016년 소위 '알파고 사건'을 기점으로 인공지능 윤리가 조금씩 언급되기 시작했다. 2018년 KAIST가 인공지능 무기 개발 논란을 겪으면서 인공지능 윤리에 대한 관심이 조금 더 많아졌다. 2019년 하반기부터 정부에서 인공지능 미래전략을 발표하고 그 후속조치를 고안하였고 그 과정에서 관심이 주기적으로 일어났다. 2021년 챗봇 이루다 사건 이후, 인공지능 윤리에 대한 보도의 수가 대폭 늘어났다. 이전과는 달리, 인공지능 윤리에 대한 보도 빈도가 매우 높아졌다.

이를 더 자세히 살펴보면, ㈜스캐터랩이 챗봇 이루다를 출시한지 일주일 만인 2020년 12월 30일, 온라인 커뮤니티인 ‘아카라이브’에 챗봇 이루다를 성적 대상으로 취급하는 사람들이 모여들기 시작했다. 2020년 12월 30일에 올라온 게시글들은 주로 챗봇의 성능이 매우 우수하여 사람이 아닌지 분별이 안 된다며 감탄하는 글들이었다. 잠시 페이스북 메신저가 작동하지 않자 이용자들은 무척 궁금해 하였다. 2021년 1월 1일 다시 작동되자, 이용자들은 게시판에 챗봇 이루다와 나눈 대화를 공유하며, 친밀도를 높이기 위한 노하우를 묻고 답했다. 같은 날 처음으로 규범적인 문제가 제기되기는 하였으나 공론화가 되지는 못하고 묻혀버렸다.

같은 날, 챗봇 이루다에게 야한 말을 직접 하면 경고가 뜨므로 우회적으로 해야 한다는 노하우를 담은 게시글이 공유되었고, 이때부터 본격적으로 챗봇 이루다에 대한 성희롱을 인증하는 게시글이 올라오기 시작했다. 이런 상황은 2021년 1월 8일부터 대대적으로 신문에 보도되기 시작했고, 갤러리 이용자들은 이러한 기사들을 공유하며 자신들이 법적인 책임을 지게 될까 우려했다. 그때부터 챗봇 이루다와 관련된 갤러리 들의 조회 수가 크게 늘었다.

2021년 1월 9일부터는 이슈가 다변화되었다. 기자들을 비롯해 신규로 유입된 이용자들은 여러 시도를 해보았고, 대화하는 과정에서 성소수자・장애인・인종・임신부에 대한 차별・혐오 발언, 거짓말, 개인정보 노출 문제가 부각되었다. 이에 다음 창업자인 이재웅과 같은 업계 인사들의 발언이 이어졌고, 논란은 인공지능 윤리 전반의 문제로 확대되었다. 하지만 국내에는 일반적인 차별금지법이 없었고 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제44조의7(불법정보의 유통금지)를 적용하기에도 어려운 사안이었다.

2021년 1월 10일부터 본 사건과 관련하여 개인정보보호에 주목하는 기사가 하나둘 씩 등장하기 시작했다. 개발사인 ㈜ 스캐터랩이 다른 서비스 이용자들로부터 수집한 카카오톡 대화 데이터를 무단으로 도용해서 챗봇 이루다를 학습시켰으며, 그 결과 개인정보가 노출되었다는 내용이었다. 네이트 판(pann)과 같은 토론사이트에서 챗봇 이루다의 개인정보노출 이슈가 집중적으로 논의되었고, 이를 소송으로 다투고자 하는 움직임도 생겼다. 이용자들이 분노했던 지점은 ‘연애의 과학’ 같은 어플(application)에서 비용을 지불하고 연애조언 서비스를 이용했는데, 그 과정에서 제출했던 카카오톡 대화메시지들이 챗봇 이루다의 학습에 사용되었다는 지점이었다. 이러한 논란에 대해 개발사인 ㈜ 스캐터랩은 사과문을 발표하였다.

2021년 1월 11일 개인정보보호위원회는 챗봇 이루다의 개발사인 스캐터랩이 개인정보 보호법을 위반했다는 혐의로 조사에 착수했다. 그리고 같은 날 개발사인 ㈜ 스캐터랩은 입장문을 발표했고, 이루다 서비스를 잠정 중단했다. 곧이어 1월 12일에는 이루다 개발사 직원들이 연인 간의 카카오톡 메시지들을 돌려보며 웃었다는 내용의 기사, 1월 13일에는 챗봇 이루다 학습에 사용된 카카오톡 데이터 중 일부가 깃허브(GitHub)라는 오픈소스 플랫폼에서 4개월 이상 공유되었다는 기사가 연달아 보도되었다.

방송통신위원회는 연내 인공지능 윤리규범 실행지침을 마련하겠다고 하였고(2021. 1. 14.), 개발사인 ㈜ 스캐터랩은 챗봇 이루다의 데이터베이스와 인공지능 모델을 모두 폐기하겠다고 밝혔으며(2021. 1. 15.), 개인정보노출을 대상으로 집단소송 절차가 시작되었다(2021. 1. 12.). 개인정보보호위원회도 인공지능 개인정보 보호수칙을 마련하겠다고 밝혔으며(2021. 1. 26.), SKT를 비롯한 IT업계에서도 인공지능 윤리기준을 마련하겠다고 하였다(2021. 1. 27.). 일부 시민단체는 본 사건을 두고 국가인권위원회에 조사를 촉구하는 진정서를 제출했다(2021. 2. 3.) 그때부터 2021년 1분기 동안 소위 ‘이루다 사건’을 대상으로 한 각종 학술대회가 개최되면서 사태를 평가하고 개선방안을 논의했다. 개인정보 노출 피해자들이 개발사인 ㈜ 스캐터랩을 상대

로 2억 원 규모의 손해배상 청구소송을 제기하기도 했다.

2016년 ‘알파고(AlphaGo) 사건’이 인공지능 시대를 알리는 신호탄이었다면, 2021년 ‘이루다(Eruda) 사건’은 한국사회에서 인공지능 윤리의 중요성을 알리는 계기가 되었다. 챗봇 이루다가 폐기된 이후에도 온라인 커뮤니티에 있는 이루다 게시판에는 지속적으로 챗봇 이루다를 그리워하는 글, 그림(소위 팬픽)이 올라왔다. 본 사건은 인공지능 윤리에서 논의될 수 있는 거의 모든 쟁점을 망라하고 있었다. 개발사 ㈜ 스캐터랩은 2022년 상반기부터 챗봇 이루다 서비스를 재개할 것임을 밝혔다. 이러한 사건을 통해 추상적인 이론적 연구에 머물렀던 인공지능 윤리 관련 문제를 사례를 통해 실증 분석할 수 있는 방대한 자료들이 축적되었다. 이로서 본 실증 연구가 가능해 졌다.

Ⅲ. 연구의 방법: 문헌연구 및 실증연구

본 연구는 크게 선행연구를 다룬 문헌연구와 구체적인 사건을 분석한 실증연구로 구성되어 있다.

1. 문헌연구

첫 번째는 인공지능 윤리에 대해 지금까지 전개되어 왔던 문헌연구이다. 인공지능정책 이니셔티브에서 연구원으로 활동하면서 참여하였던 ‘AI윤리 인덱스 프로젝트’에서의 경험으로, 인공지능 윤리가 공정성, 투명성, 책무성, 프라이버시를 중심으로 형성되고 있다는 사실을 확인하였다. 이를 바탕으로, 논문 검색 사이트인 SSRN(사회과학 연구 네트워크)와 RISS(학술연구정보서비스)를 중심으로 ‘인공지능 윤리’, ‘공정성’, ‘투명성’, ‘책무성’, ‘프라이버시’ (기타 유사한 키워드 : 가령 프라이버시[privacy]의 경우 개인정보보호[data protection] 등)라는 키워드로 검색하여 국내외 선행연구를 확인하였다. 인공지능 언어모델과 관련된 규범적 쟁점과 관련해서는 ACM Conference on Fairness, Accountability, and Transparency(ACM FAccT) Conference에서 발표되었던 최신 논문들과 DeepMind의 보고서 및 StandfordHAI의 보고서를 주로 참고하였다.

2. 실증연구

가. 개 관

두 번째는 챗봇 이루다 사건을 대상으로 한 실증연구이다. 본 연구에서 주목하는 기간은 ㈜ 스캐터랩이 챗봇을 출시한 2020년 12월부터 본격적인 윤리 논의가 시작된 후 챗봇을 폐기한 2021년 1월까지이다.[3] ‘이루다 커뮤니티 게시 글’을 대상으로 실증연구를 진행하였다. 본 연구를 시작함에 앞서, 챗봇 이루다의 이용자가 자신이 챗봇 이루다와 대화를 나눈 다음에, 이를 스스로 스크린샷으로 찍은 후 외부인에게 접근제한 없이 공개되어 있는 온라인 커뮤니티 게시

3) 챗봇 이루다에 대한 논의는 현재 진행형이다. 챗봇의 개발・운영사인 ㈜ 스캐터랩이 2022년 챗봇 이루다 2.0을 새로이 출시하였기 때문이다. 본 연구의 직접적인 분석대상은 아니나, 직간접적으로 연관된 부분은 함께 다룰 예정이다. 인공지능 서비스 개발・운영사에게 규범적으로 기대되는 주의수준을 둘러싼 사회적 합의의 형성과정을 보여주는 대표적인 사례이기 때문이다.

판에 업로드한 자료를 연구하는 것이 '인간대상 연구'에 해당할 수 있다고 보았다. 이에, 2022. 2. 18. (재)국가생명윤리정책원의 온라인 윤리적 연구 수행을 위한 인간대상연구자 교육 과정을 이수하였고(이수번호 No. 2022-002300), 2022. 3. 14. 서울대학교 기관생명윤리위원회의 심의를 신청하였으며, 2022. 3. 18. 연구 승인을 받았다(승인번호 IRB No. 2203/004-002). 이하 기관생명윤리위원회의 심의를 받은 연구의 구체적인 방법론과 한계 및 의의를 소개한다.

나. 연구대상

본 연구의 주된 목적은 실증연구(bottom-up approach)를 통해 지금까지 추상적으로 전개되어 오던 인공지능 윤리담론(top-down approach)이 유효한지를 검증하고, 필요할 경우 구체적인 사안에서 실제로 문제되는 인공지능의 위해가 무엇인지를 자료에 근거하여 제시하는 것이다. 이를 위해 본 연구는 챗봇 이루다와 이용자가 나누었던 대화내용에 주목했다. 하지만 ㈜ 스캐터랩이 챗봇과 이용자가 나누었던 대화내용을 연구목적으로 제공할 것을 기대하기 어려웠다. 사건이 전개되던 양상을 언론보도를 통해 체계적으로 정리하면서, 언론보도에 소개된 챗봇과 이용자 간 대화내용에 관심을 가지게 되었다. 언론사에서 대화내용을 확보하였던 방법을 확인하던 중, 본 사건이 주로 특정 온라인 커뮤니티를 중심으로 전개되어 왔다는 점을 확인하였다. 본 연구는 온라인커뮤니티인 디시인사이드 '이루다 마이너 갤러리'를 대상으로 선정하고 진행하였다. 다른 온라인커뮤니티인 아카라이브에서 처음 논의되었기는 했지만, 유사한 시기부터 디시인사이드를 중심으로 전개된 논의의 양이 압도적으로 많았고 현재까지도 활성화되어 있기 때문이다.[4)]

다. 연구기간

챗봇과 이용자 사이에 주고받은 대화내역을 연구한다는 점에서 인간대상 연구에 해당될 가능성이 있었다. 물론 공개된 익명 게시판에 업로드된 챗봇과 이용자 사이에 주고받은 대화내용을 대상으로 하는 것이었으나 이러한 게시 글이 개인정보에 해당될 여지가 없지는 않았기 때문이다. 학내 기관생명윤리위원회에 심의를 신청하고 연구계획서를 작성하여 제출하였으나, 연구기간과 분석대상 게시 글을 특정해야 하며 연구방법도 사전에 확정해야 한다는 사유로 반려되었다. 하기와 같이 연구계획서를 보다 구체적으로 작성한 후 다시 심의를 신청하였고 승인되었다. 챗봇 이루다 사건은 2020년 12월부터 논의되기 시작하여 2022년 현재까지 진행 중이었다. 2021년 1월 11일에 챗봇 서비스를 중단한 이후, 2022년 현재 챗봇 서비스를 다시 시작했기 때문이었다. 본 연구의 목적이 실제 데이터에 기초하여 인공지능 윤리가 문제되었던 사건을 분석하고 그로부터 실제 위해(harm)의 유형을 범주화(categorization)하는 것이었기 때문에, 챗봇과 이용자 간 대화내역이 가장 많은 기간을 모색하였다. 그 결과, 챗봇 이루다가 출시한 후 디시인사이드에 게시판이 처음 생성된 2020년 12월 30일부터 챗봇 이루다 서비스를 중단하게 되기 전날인 2021년 1월 10일까지를 연구대상으로 삼기로 하였다. 기관생명윤리위원회의 심의는 일별 게시 글 수가 압도적으로 많았던 2020년 12월 30일부터 2021년 1월 15일까지로 신청하기

4) 디시인사이드 AI 이루다 마이너갤러리와 아카라이브 이루다채널은 모두 2020년 12월 30일에 생성되었다. 디시인사이드 갤러리의 경우 2022년 7월 현재까지 약 15만 개의 게시글이 업로드된 반면, 아카라이브 채널의 경우 2022년 7월 현재까지 약 1,500 개의 게시글이 업로드 되었을 뿐이다.

는 했으나, 챗봇 이루다가 2021년 1월 11일부터 서비스를 하지 않음에 따라 이용자들이 그날부터 대화내역을 거의 공유하지 않았었기 때문이었다.

라. 데이터의 수집방법 및 분석기준

본 연구의 데이터 수집방법은 수작업이었다. 처음에는 웹 스크래핑 방식을 검토했었다. 하지만 이는 자칫 무단 웹 크롤링으로 평가되어 저작권 침해 내지 부정경쟁방지법 위반 소지가 있다고 판단하였다. 수작업으로 개별 게시 글을 열어가며 이용자가 익명게시판에 자발적으로 업로드 한 대화내역 이미지 파일을 다운받았다. 인터넷 커뮤니티 디시인사이드의 '이루다 마이너 갤러리'를 중심으로 챗봇 이루다 이용자들이 작성했던 2020년 12월 30일부터 2021년 1월 10일까지 게시된 글들 중에서 업로드 되어 있던 이미지 파일은 총 9,747건이었다. 챗봇과 이용자 간의 대화는 대부분 끊임없이 이어졌다. 인공지능의 활용으로 인해 유발될 수 있는 위해의 유형을 수치화하여 실증적인 분석결과를 도출하려는 연구설계에 비추어 대화 수를 산정할 수 있는 기준을 수립해야 했다. 이용자가 업로드 했던 이미지 파일을 하나의 대화로 전제하기로 하였다. 다른 이용자가 시도했던 대화를 따라서 시도했던 이용자가 다소 있었으므로, 이미지 파일은 상이하나 같은 내용의 대화가 업로드되는 경우가 왕왕 있었다. 또한 대화내용 조작을 시도하는 경우도 있을 수 있었다. 약 1,000여 건의 데이터를 선제적으로 수집하여 범주를 확인하고 분류하였다. 언어모델과 챗봇의 작동원리에 따라, 누가 먼저 특정 소재를 대상으로 하는 발화를 시작했는지가 무척 중요하다는 점을 확인할 수 있었다. 가령 끊임없이 이어지고 있던 대화에서 누군가(A) 욕설을 발화하였다면, 그 상대방(B)인 챗봇 내지 이용자로부터 많은 경우 욕설이 발화되었다. 언어모델은 기본적으로 앞서 제시된 텍스트를 바탕으로 다음에 이어질 텍스트를 통계 내지 신경망으로 예측하는 구조를 지니고 있기 때문이다. 이 경우 A가 B에게 욕설을 했다고 분류하였다.

마. 한계 및 의의

본 연구 방법론은 두 가지 한계를 지니고 있다. 첫 번째 한계는, 챗봇과 이용자 간 대화내역을 전부 확보하지 못하고 발췌했다는 점이다.[5] 이용자가 특정 온라인 커뮤니티에서 자발적으로 업로드 한 챗봇과의 대화내역을 모두 수집해서 분석했지만 랜덤 샘플링에 의한 자료수집도 아니었다. 그 결과 분석대상 데이터에 사회적인 분위기와 커뮤니티이용자의 성향이 반영될 여지가 있었다. 두 번째 한계는 이용자가 챗봇과의 대화를 임의로 조작한 이미지가 섞여 있을 수 있다는 점이다. 하지만 이는 육안으로 식별하기가 어려웠으므로 개별 이미지가 별개의 대화로서 진실한 것이라고 전제할 수밖에 없었다. 그럼에도 불구하고 본 연구 방법론은 두 가지 의의를 지닌다. 첫 번째 의의는, 1만 여 건에 가까운 실제 대화 내역을 확보하여 분석하였다는 점이다. 표본의 개수가 수백 건에 불과하다면 이용자에 의해 조작된 개별 데이터가 전체 결과에 미치는 영향이 크겠지만, 본 연구처럼 표본의 개수가 1만 건에 수렴한다면 조작가능성이 있는 개별 데이터가 전체 결과에 미칠 수 있는 영향이 상대적으로 적어진다. 두 번째 의의는, 실제 대화내역을 바탕으로 인공지능으로부터 비롯되는 위해를 범주화하고자 시도했다는 점이다. 후

5) 챗봇과 이용자 간 대화내역을 전부 확보하려면 개발·운영사로부터 이를 제공받아야 하는데, 이는 일종의 개인정보를 수집목적 이외로 제3자에게 제공하는 것으로서 현실적으로 제약이 많았기 때문이다.

술하겠지만, 그로 인해 실제 문제되는 위해의 유형과 비중이 문헌연구와는 상이하다는 점을 확인할 수 있었다.

Ⅳ. 기존 인공지능 윤리담론의 분석적 정리

1. 서론

인공지능 윤리에 대한 관심이 늘어나고 있다. 2014년을 기점으로 인공지능 윤리에 대한 연구가 폭발적으로 증가해 왔다는 점이 이를 방증한다. 대표적인 인공지능 윤리 학회인 ACM FAccT에 제출이 수락된 논문 수만 살펴보아도 2018년에 이미 71건이었고, 2021년에는 302건에 이를 정도였다. 주로 논의되는 주제는 공정성, 책무성, 투명성, 그리고 개인정보보호였다. 이들은 오늘날 추상적인 윤리담론을 넘어 구체적인 기술구현 단계에서 논의되고 있다. 가령 국제표준기구인 IEEE는 2021년에 시스템 디자인에 있어 윤리적인 문제를 해결하기 위한 프로세스로 IEEE 7000을 발표했다. 오늘날 빈번히 활용되는 머신러닝을 비롯한 인공지능 기술은 크게 모델 구축(building model)과 모델 사용(using model)의 단계로 나누어진다. 이러한 인공지능 기술은 주어진 데이터를 학습시킨 모델을 가지고 실제 과제에 적용하는데, ① 모델 구축 단계(즉, 학습 단계)에서는 데이터의 수집, 학습, 배포가 이루어지고, ② 모델 사용 단계(즉, 예측 단계)에서는 프록시데이터의 수집, 추론, 예측・결정이 순차적으로 이루어진다. 이를 감안하여, 현재 인공지능 언어모델을 둘러싸고 일반적으로 논의되고 있는 인공지능 윤리를 정리하면 다음과 같다.

2. 공정성

가. 의의 및 연혁

공정성(公正性)이란 인공지능 의사결정이 공평(公平)하고 정의(正義)로워야 한다는 인공지능 윤리의 한 원칙이다. 경우에 따라서는 차별금지(差別禁止) 라고도 불리기도 하지만, 공평과 정의를 아우른다는 점에서 단순히 차별하지 않는다는 것보다는 넓은 개념이다. 일각에서는 인공지능 활용에 있어서 고려할 수 있는 공정성의 개념이 20여 개가 넘는다고 주장하기도 했다[1]. 공정성의 개념은 인공지능 시스템이 정답을 맞히는 비율인 통계적 정확성(statistical accuracy)와는 구별된다[2]. 공정성은 인공지능 시스템이 달성해야 하는 목표로서 규범적인 개념인 반면, 정확성은 인공지능 시스템의 성능에 대한 사실적인 개념이다. 인공지능 시스템에 의해 도출된 결과는 사실에 대한 정보(factual information)가 아니라 통계적 추측(statistically informed guess)일 뿐이므로 정답을 선택하지 못할 수도 있다는 의미이다[3].

본격적으로 인공지능 시스템의 공정성 논의를 시작한 사건은 2016년 ProPublica의 탐사보도로 시작된 COMPAS 논쟁으로 보인다. 물론 그 이전에도 인공지능 시스템을 둘러싸고 차별이 문제된 경우가 있기는 했으나 단발적인 사건보도에 그쳤을 뿐이기 때문이다. COMPAS 사건에서는 인공지능 시스템을 두고 실제 데이터를 두고 여러 관계자들이 서로 다른 기준을 주장하면서 대립하였다. 탐사보도 전문매체 ProPublica는 2016년에 전국적으로 이용되고 있는 미래

범죄자 예측 소프트웨어(COMPAS)가 흑인에 대해 편견을 가지고 있다고 보도했다[4]. 백인 범죄자의 재범 위험성은 낮게 평가한 반면, 흑인 범죄자의 재범 위험성은 높게 평가했다는 것이다. 이러한 주장은 인종별로 통계 프로그램의 오류율이 서로 다르게 나타난다는 점에 근거하였다[5]. 이에 COMPAS의 제조사인 Northpointe(현재 equivant)는 COMPAS 소프트웨어가 인종과 무관하게 위험 점수를 정확하게 제시하고 있다고 반박했다[6]. 이는 Propublica가 오류율 기준을 적용하였던 반면, Northpointe는 예측 정확도 기준을 적용하였기 때문이다.

나. 내용

인공지능 모델을 구축하는 데에 있어 학습데이터셋을 어떻게 구축할지는 매우 중요하다. 만일 학습데이터셋이 잘못 구축되면 아무리 모델을 잘 만들어 놓아도 의도했던 결론이 도출되지 못하기 때문이다(GIGO)[7]. 이러한 학습데이터의 문제는 두 가지 양상으로 나타난다. 하나는 지도학습에 있어서 라벨링이 잘못되는 경우 비롯되는 문제(①)이고, 다른 하나는 비지도학습에 있어서 학습데이터셋이 편향되어 있는 경우 나타나는 문제(②)이다. 전자의 문제(①)가 가장 잘 드러난 사건은 ImageNet Roulette라는 프로젝트였다. 이로 인해 2009년 발표되었던 세계 최대의 이미지 데이터셋인 ImageNet에서 사용된 라벨링에 편향이 적나라하게 반영되어 있다는 점이 드러났다[8]. 곧 ImageNet은 사람 범주에 있는 150만 개의 이미지 중에 절반을 제거할 것이라고 발표했다[9]. 후자의 문제(②)가 드러난 사건은 더욱 빈번하게 목도되고 있다. Amazon은 채용 알고리즘을 개발했는데, 실제로 적용하기 전에 시뮬레이션을 하는 단계에서 계속 여성보다 남성을 우대하는 편향(bias)이 나타났다. 이는 아마존의 직원 구성 때문이었다. 학습데이터로 사용한 10년 동안의 개발직군 채용결과에서 남성의 수가 여성보다 압도적으로 많았기 때문이었다. 그 결과 성과가 좋았던 지원자 중 남성이 차지하는 비율이 여성보다 많을 수밖에 없었다[10]. 결국 아마존은 2018년에 상기 채용 알고리즘을 폐기했다[11]. 이러한 문제는 학습데이터 자체가 편향되었기 때문에 비롯된 것으로, COMPAS 사건을 비롯한 여러 사건에서 반복적으로 나타났다.

인공지능 모델을 구축하는 과정에서 편향이 내부에 존재할 수 있는데(intrinsic bias), 이는 훈련 과정에서 사용되는 데이터의 상태나 라벨링 작업자의 성향에 기인하는 경우가 많다. 이러한 모델은 사용되는 과정에서 외부로 위해를 가할 수 있게 된다(extrinsic harm)[12], 내재적 편향(intrinsic bias)에서 가장 많이 연구된 분야는 대표 편향(representational bias)인데, 이는 모집단을 잘못 대표하는(misrepresented) 경우[13], 모집단을 아예 대표하지 못하거나 지나치게 적게 반영함으로써 과소 대표된(underrepresented) 경우[14], 모집단을 지나치게 증폭하거나 균질하게 함으로써 과대 대표된(overrepresented) 경우[15]로 나뉜다. 이러한 내재적 편향은 외부적 위해(extrinsic harm)로 발현되는데, 개인 차원의 편향표현[16], 혐오발언[17], 비하발언[18]으로 나타나 심리적인 위해를 유발할 수 있으며[19], 특정 방언으로 된 음성을 인식하지 못하는 것처럼 집단에 따라 모델의 성능을 달리할 수도 있다[20]. 이하 항목들(차별, 고정관념, 배제, 유해성, 성능의 차등)은 편향이 발현되는 구체적인 형태들이다[21].

다. 언어모델에 관련된 특수한 논의(special discussion)

1) 차별(unfair discrimination)

부당한 차별을 야기하는 언어는 특정 집단을 억압하는데 기여할 수 있다. 부당한 차별이란 개인 또는 집단 간의 민감한 특징에 따른 차별적인 대우나 자원접근의 차등으로 발현된다. 이러한 차별은 문화적인 요인으로 발생하며 국가에 따라 정당화될 여지도 있는데, 가령 인도에서는 용인되나 그 이외의 국가에서는 용납되지 않는 카스트 제도 같은 경우가 그 사례이다[22]. 사회문화적인 범주가 중첩될 경우 더욱 증폭될 수 있다. 인공지능 언어모델의 경우 이러한 경향이 보다 두드러게 나타난다. 가령 GPT-3의 경우, 반무슬림 및 반유대주의 경향이 일반적인 인공지능 모델보다 두드러진 것으로 나타났는데[23], 테스트 중 23% 사례에서 무슬림을 테러리스트와 연결했고, 5% 사례에서 유대인을 돈과 연결했다[24]. 이는 언어모델이 학습데이터셋에 반영되어 있던 편향(bias)을 학습하여 부당한 차별 발언을 재현하였기 때문이다[25]. 이러한 차별은 인공지능을 사용하는 과정에서 별도의 조치를 취하지 않는다면 반복되어 더욱 심화될 것이다. 왜냐하면 학습데이터에 내재된 편향이 인공지능 모델로 학습되어 재사용될 것이고, 학습데이터셋에 일부 집단이 더 많은 비중을 차지하고 있어 과대대표됨으로써 학습모델이 편향될 수 있기 때문이다. 언어모델이 정확도 높은 언어를 표현하도록 최적화되는 과정에서 학습데이터셋에 있던 편향을 과도하게 나타낼 수 있는데, 이를 편향 증폭(bias amplification) 현상이라 한다[26]. 인공지능 언어모델은 어떤 방식으로 응용되어 사용(downstreaming use)될지 모르기 때문에 차별 발생을 미리 방지하기 어렵다[27].[6)]

2) 고정관념(social stereotypes)

언어모델은 설정된 페르소나를 표방하면서 성별 또는 인종의 정체성의 암시를 통해 유해한 고정관념을 조성할 수 있다. 인공지능 비서의 예를 들어 본다. 비서를 본질적으로 여성의 성별과 연결된 것으로 표시함으로써 유해한 고정관념을 지속시킬 수 있다[28]. 이러한 방식으로 인간과 인공지능이 상호작용 하도록 함으로써 고정관념을 도입하고(선동자 효과, instigator effect), 도입된 인공지능이 이용자의 명시 내지 묵시적인 동의 하에 사용됨으로써 유해한 고정관념을 유지하게 된다(찬동자 효과, yes-sayer effect)[29]. 이러한 고정관념 조장이 나타나는 대표적인 사례는 성별과 인종이다. 성별(sex)을 살피면, 시중에서 판매되는 음성비서는 복종적인 여성으로 나타나는 경우가 많다[30]. 인공지능 비서의 목소리가 여성일 때 가장 아름답다고 평가되었다는 국내 사례도 있었다[31]. 이는 단순히 친밀감을 드러내는 것을 넘어 성적 대상화까지 연결될 수 있다[32]. 이는 여성성(womanhood)과 남성성(manhood)이라는 관념이 사회에 학습되는 암묵적인 기제로 작용하게 된다[33]. 인종(race)의 경우도 그러하다. 백인일수록 지능이 높고 전문적이며 강력한 이미지로 표현되는데, 이는 그 이외의 인종에 대한 피해로 귀결될 수 있다[34].

---

6) 다만, 발화와 취급은 구별되어야 한다. 발화(發話)란 재현적 해악(representational harm)이라고 할 수 있으며, 특정한 사회집단에 대한 과소·과대 대표(misrepresenting), 비하(demeaning)나 고정관념(stereotyping)을 단순하게 발언(utterance)한 경우 나타난다. 즉, 언어모델이 말이나 글을 생성 내지 제시했는데 이때 편향이 반영된 경우이다. 반면 취급(取扱)이란 배분적 해악(allocational harm)라고 표현할 수 있는데, 이는 자원이나 기회가 사회집단 사이에 불공정하게 배분됨으로써 발생하고 그로 인해 언어모델이 사람들에게 중대한 영향(consequential result)을 주는 경우에 나타난다. 즉, 언어모델을 이용하여 다른 서비스를 제공했는데 그 서비스 제공 결과에 편향이 반영된 경우이다. 인공지능 언어모델이 차별적인 판단을 하더라도 구체적인 응용프로그램을 통해 구체화되어 개인에게 직접 영향을 줄 정도로 활용되는 경우가 아닌 이상, 취급이 아닌 발화로 평가되어 표현의 측면에서 규범적인 판단의 대상이 될 뿐이다. 신용도 평가, 재범률 판단, 직업적합성 평가 또는 채용과 같이 인공지능이 내린 차별적인 의사결정이 국민 개인의 기본권을 제약하게 될 정도라면 이는 신중하게 검토되어야 하나, 언어모델의 경우 이렇게 활용될 수 있는 경우는 극히 제한될 것이다.

3) 배제(exclusive norms)

사회적 규범을 강화하는 언어는 이러한 규범 외부에 존재하는 이들을 배제할 수 있다. 인간은 언어를 통해 사회에 존재하는 범주와 규범을 표현하는데, 언어모델은 학습데이터셋에 반영되어 있지 않았던 규범적 판단을 고려하지 못함으로써 범주 외부에 존재하는 사람들을 배제하거나 주변화할 수 있다. 만일 인공지능 챗봇에게 “가족이란 무엇인지”를 문의하였을 때 “가족이란 결혼하고 아이를 낳는 남녀”라고 답변을 하였다면, 새로운 유형의 가족이라 불리는 동성, 혼외, 편부모, 무자녀 가족이 배제될 수 있다. 이러한 규범에 부합하지 않는 경우 심리적인 부담(psychological tax)으로 인해 기존에 존재했던 재현적 해악이나 배분적 해악이 증폭될 수 있다. 즉, 언어모델은 결코 가치중립적이지 않다[35]. 위와 같은 범주와 규범은 외부인을 배제하게 된다. 가령 성별 범주에 있어 이진 분류(binary classification)를 사용하다 보니, 특정 이름을 여자 또는 남자 중 하나의 성별로 해석하고[36], 주어진 데이터로 성별을 판정할 때에 트랜스젠더를 배제하기도 하였다[37]. 이러한 현상은 언어모델이 특정한 집단 또는 특정한 시점의 언어를 반영한다는 점에서 비롯된다. 언어모델은 학습데이터셋에 반영되어 있는 특정 시간대의 특정 집단의 언어를 반영할 구밖에 없으므로, 그 순간(소위 ‘frozen moments’)의 규범, 가치, 범주에 갇혀버릴 수밖에 없기 때문이다(소위 ‘value lock-in’)[38]. 따라서 언어모델은 학습된 이후에도 추가되는 데이터에 의해 지속적으로 업데이트 됨이 바람직하다[39].

4) 성능의 차등 (lower performance by social group)

언어 기술이 특정 사회적 집단에서 다른 집단보다 더 잘 작동될 때 일종의 차별이 나타날 수 있다. 이러한 성능의 차등은 보통 언어모델의 학습데이터셋으로서 영어와 같은 특정 언어 또는 방언이나 비속어와 같은 특정 사회적 집단의 언어가 얼마나 많이 사용되었는지에 따라 좌우된다. 학습한 데이터셋이 특정 사회집단에 편중되어 있음에 따라 과소대표된 사람을 대상으로 한 성능이 저조할 수도 있기 때문이다. 속어・방언・사투리・유행어를 반영하지 못하여 특정 사회집단에 대한 성능이 저하되는 경우가 대표적이다[40]. 이는 비원어민(non-native speaker), 교육배경이 다른 사람, 연령층이 다른 사람, 언어 장애가 있는 사람에게 낮은 품질의 인공지능 서비스가 제공될 수 있음을 시사한다. 만약 국내에 미국과 같은 일반적인 차별금지법이 있었다고 하더라도, 이러한 속성을 보호되는 특징에 포함시키지 않는다면 적절히 대처할 수 없다. 성능 저하가 자원의 분배나 필수서비스 제공에 작용함으로써 인공지능 활용에 따른 편익을 치우치게 하고 기존에 존재하던 사회적 불평등을 고착하는 데 기여한다면[41], 배분적 관점에서 사회정의의 문제를 유발할 뿐이다[42]. 인공지능 챗봇은 사투리로 질문하였는데 인공지능 챗봇이 알아듣지 못하거나 엉뚱하게 알아듣는 경우가 발생할 수 있는데[43], 이런 문제는 의료영역 같이 오분류 내지 성능저하가 생명을 위협할 수 있는 분야에서 더욱 심각하게 부각될 수 있다[44]. 물론 이러한 문제상황에 대처하는 방식으로, 학습데이터셋이 풍부한 영어나 중국어와 같은 주요언어를 매개하는 접근방법을 고려할 수 있다. 가령 한국어를 가나어로 번역하고자 할 때, 한국어를 영어로 번역한 후 그 결과를 가나어로 번역하는 방식이다. 하지만 이러한 접근방식은 어떤 단어의 의미가 언어에 따른 문화차이로 상이한 관념에 연결될 수 있다는 점에서 여전히 저조한 성능을 보일 것으로 보인다.

라. 언어모델의 유해성(toxic language) 관련 특수논의

1) 규명의 어려움

언어는 폭력을 선동하고 범죄를 야기하는 등 이른바 유해성(toxic)이 있을 수 있다. 무엇이 유해어인지에 대해 단일하게 합의된 정의는 없으며, 불쾌(toxic), 공격(severe toxic), 외설(obscene), 협박(threat), 모욕(insult), 혐오(identity hate) 정도가 유해어의 종류로 빈번하게 언급된다[45]. 이러한 유해어는 온라인 커뮤니티를 비롯한 웹상에 만연해 있고, 그 결과 이러한 공간에서 취득한 데이터를 바탕으로 생성된 학습데이터셋에서도 상당하다. 겉으로는 무해해 보였던 언어모델이 사용해보니 유해한 텍스트를 제시하는 경우도 목도되었다[46]. 하지만 이러한 유해어를 규명(identification)하는 것은 결코 쉽지 않다. 어느 언어가 유해어인지 여부는 맥락의 존성(context dependency)에 따라 판단해야 하기 때문이다[47].[7)]

2) 탐지의 어려움

무엇이 유해어인지 규명할 수 있더라도 실제로 이를 탐지(detection)하는 것은 여전히 어려운 문제이다. 통상 유해어 필터(toxic comment classifier)를 사용하게 되지만, 이를 우회하는 변형어 내지 파생어가 다양하게 존재할 수 있다. 유해어는 맥락에 따라 다양한 방식으로 변형되어 사용할 수 있기 때문이다. 발견한다해도 이를 어떻게 완화(mitigation)할 것인지에 대한 고민이 필요하다. 유해어를 단순하게 삭제할지 대체할지, 삭제 내지 대체한다면 어느 정도로 그렇게 할지, 제재하는 방식을 선택한다면 어떠한 불이익을 어떻게 설정함으로써 유해어가 반복되지 못하도록 할지를 모두 결정해야 하기 때문이다. 설령 유해어 필터의 성능이 우수해서 대부분의 유해어를 걸러낸다고 하더라도, 잘못 지정되면 도리어 편향을 증폭시킬 수 있다. 가령 역사적으로 소외된 집단의 발언을 유해하다고 판명하는 경우가 여기에 해당된다[48].

[표] 비윤리적인 발언 탐지가 어려웠던 이유	
분류기준의 문제 (기준의 모호성)	**(기 준)** 무엇이 비윤리적인 발언인지 그 기준이 불분명하다. 일반적으로 욕설이나 비속어가 들어있으면 비윤리적인 발언으로 이해할 수 있는데, 그 이외의 차별・공격・선정적인 발언에 해당하는지 여부가 애매한 경우가 있다.
판단주체의 문제 (기준의 가변성)	**(시 간)** 판단주체가 살아가고 있는 시대에 따라 비윤리적인 발언에 해당하는지 여부가 달라질 수 있다. 가령 성차별이나 인종차별이 만연하던 시대에 살던 사람과 이것이 극복된 시대에 살고 있는 사람은 동일한 발언을 보고도 달리 판단할 수 있다.

7) 대화자의 성격, 그녀 또는 그가 위치한 시공간과 성격 내지 소속집단에 따라 달라질 수 있다. 가령 퀴어(queer)라는 단어는 역사적으로 비방하는 용어로 인식되어 왔지만, LGBT+ 커뮤니티에서는 자신을 식별하는 지표로 사용된다. 또한 동일한 사실적 진술이 어떤 맥락에서는 교육적인 발언으로 사용될 수도 있지만, 다른 맥락에서는 성적인 발언으로 규명될 수 있다. 가령 성교육을 성인 콘텐츠로 판단함으로써 교육 콘텐츠를 성인 콘텐츠로 잘못 분류하는 것이 관찰되기도 했다.

	(공 간) 판단주체가 살아가고 있는 공간에 따라 비윤리적인 발언에 대한 판단이 달라질 수 있다. 가령 대한민국과 북한에 살고 있는 사람은 동일한 발언을 접하더라도 판단을 달리할 수 있다.
	(세 대) 판단주체가 속한 세대에 따라 비윤리적 발언에 대한 판단이 달라질 수 있다. ‘어쩔티비~ ㅈ□저쩔티비~ 우□만물티비~안궁티비~뇌절티비~우짤래미~ 저짤래미~ 쿠쿠루삥뽕 지금 화났죠? 개킹받죠? 죽이고 싶죠?’ 라는 댓글은 기성세대에게는 외계어처럼 들릴 수 있지만, MZ세대에게는 욕설이 될 수 있다.
	(집 단) 판단주체가 보유하고 있는 선호에 따라서(가령 중국인을 아끼는 사람에게 ‘착짱죽짱’), 속한 집단에 따라서(유튜버에게 ‘자낳괴’) 동일한 발언이라도 판단을 달리할 수 있다.
필터링기법의 한계 (우회의 가능성)	**(변 형)** 이용자에게 어느 유해어가 기술적으로 필터링 된다는 것이 알려지면 이를 우회할 수 있는 기법을 개발하고 공유할 수도 있다. 이러한 기법은 정말 다양한데, 중간중간에 특수문자를 끼워넣을 수도 있도(씨발->씨!발), 영어로 표현할 수도 있고(씨발->Tlqkf), 동일한 음을 가진 다른 부호로 표현할 수도 있다(18, 열여덜). 이러한 우회기법은 무한히 늘어날 수 있다. 일례로, 50개의 욕설이 20만개로 변형될 수 있었다.
	(맥 락) 데이터베이스 기반 비윤리적 발언 탐지의 가장 큰 한계는 유해어가 없이 일상어로 기재된 비윤리적인 발언을 탐지하기가 어렵다는 점이다. 즉 욕설 같은 유해어가 직접 들어있지는 않지만 맥락상 비윤리적인 발언이라고 볼 수 있는 경우, 사람은 이를 판단할 수 있지만 기계를 통해 이를 탐지하는 작업은 쉽지 않았다.
판단객체의 문제 (현실과의 관련성)	**(관 련 성)** 발언이 유해한지 여부는 그 자체만으로 판단할 수 있는 경우도 있지만 외부지식과의 연관성을 감안하여야만 결정되는 경우도 있다. 가령 러시아가 전쟁을 일으키기 전에는 ‘푸틴같다’ 라는 표현이 남성미가 넘친다는 의미로서 비윤리적인 발언이 아닐 수 있었지만, 러시아가 우크라이나를 침공한 이후 ‘푸틴같다’는 발언은 침략자의 이미지가 개입됨으로서 비윤리적인 발언이 될 수 있다.

(출처 : 연구자 작성)

3) 여과 내지 완화 방안

유해어를 여과(filtering) 없이 내보낼 경우 정신적 내지 물질적인 피해를 유발할 수 있으므로, 사전에 규명(identification) 및 탐지(detection)하여 여과(filtering) 내지 완화(mitigation)하는 것이 중요하다. 인공지능 언어모델이 단순하게 유해어를 발화했을 경우 법적인 책임을 지우기는 사실상 어려우므로,[8] 국내에서는 유해어를 두고 IT기업을 위주로 상당한 논의가 전개되었고 나름의 자율규제 체계가 수립되어 운영되고 있다.[9] 댓글 서비스를 제공하는 경우가 많은데, 악

8) 인공지능은 고사하고 인간에 대한 혐오 표현조차 충분히 규율되지 못하고 있을 뿐만 아니라, 언어모델이 통계적 추정치에 따라 인간이 표현한 듯한 답변을 생성 내지 제시하는 것을 두고 그 개발자 내지 운영자에게 모욕이나 명예훼손의 고의를 인정하기는 어렵기 때문이다. 설령 과실에 따른 민사상 손해배상 책임을 강구한다 하여도 피해자를 특정하기 어렵고 그로인한 손해발생 사실도 입증하기 어려울 뿐만 아니라, 헌법상 표현의 자유의 범주에 포함되어 보호대상이 될 수도 있다.

9) 국내 주요포탈 중 다음카카오의 경우 악성댓글 분류기준에 있어 상기 방송통신위원회의 기준을 따르고 있으며, 그에 따라 불법정보는 유통을 금지하고, 그 이외의 언어표현은 외설과 비속으로 나누어 검토하는 것으로 보인다. 반면 네이버의 경우 방송통신위원회의 기준을 참고하되 자체 기준을 수립하여 집행하고 있었다. 그 기준은 욕설(일반적인 욕설, 네이버 내부적으로 가지고 있는 욕설 데이터에 포함된 표현), 저속한 표현(타인에게 불쾌감을 주는 속되고, 격이 낮은 표현), 선정적인 표현(성적으로 자극적인 표현), 폭

성 댓글 규제가 사회적으로 문제된 적이 있었기 때문이다.[10] 국내에서는 악성댓글에 대처하기 위해, 관리적인 조치로서 패널티 시스템과 이용자의 참여를, 그리고 기술적인 조치로서 유해어 필터링과 문제발언 검출모델을 이용했다.

3. 투명성

가. 의의

투명성(透明性)이란 인공지능 의사결정의 존재(存在)와 내용(內容)을 알려야 한다는 인공지능 윤리의 한 원칙이다. 이는 딥러닝의 도입이후 본격적으로 논의되기 시작한 개념으로 가능한 달성할수록 바람직하다는 원칙적 차원의 개념이라는 점에서, 구체적인 사안에서 무엇을 규범적으로 요구해야 하는지에 대한 설명이라는 실천적 차원의 개념과 구별된다. 인공지능 기술은 사회적인 활용도가 높아지고 있으나 그 작동방식이 블랙박스라고 불릴 만큼 불투명하다. 이는 인공지능 모델이 성능은 뛰어나지만 설명력이 떨어질 수 있기 때문이다. 심지어 동일한 입력에 대해서도 결과가 다를 수 있다[49]. 따라서 인공지능 기술을 사회가 신뢰할 수 있으려면 설명가능성(explainability) 내지 해석가능성(interpretability)을 확보해야 했다.

나. 연혁

이러한 문제의식은 법정책의 변화로 이어지면서 유럽연합 일반정보보호규정(GDPR) 같이 시행된지 얼마 되지 않은 법제도의 해석에 큰 영향을 미치기도 했다. 유럽연합의 일반정보보호규정은 인공지능의 공정성과 투명성을 확보하기 위한 법적 장치로서 컨트롤러(controller)에게 정보제공의무를 부과하고 있다. 정보주체는 제공받은 정보에 기초하여 인공지능 의사결정을 평가하고 이의제기할 기회를 확보하게 된다[50]. 하지만 유럽 학계에서는 상기 일반정보보호규정이 언급하는 정보제공의무만으로 인공지능 의사결정에서 공정성과 투명성을 확보하기 어렵다는 입장이 지배적이다. 상기 정보제공의무는 본문이 명확하게 의무로 정하지 않은 사항임에도 전문에 의무로 정해진 사항일 뿐이었지만, 해석상 전문은 법적 구속력이 있는 본문과 유사한 효력을 가진다고 볼 여지도 있었다[51]. 그 결과 전문과 자동화된 의사결정 지침에 근거하여 인공지능서비스 이용자의 설명요구권(right to explanation)을 인정할 가능성이 있었다. 하지만 정보제공의무를 통해 제공해야 하는 정보의 내용이 어떻게 설정되든(설령 의미 있는 정보를 충분

력적인 표현(신체적 위협에 대한 표현), 차별적인 표현(지역, 인종, 국가, 종교 등에 기반한 차별 표현), 비하적인 표현(상대방에게 모멸감과 수치심을 주는 비하 표현)이었다. 기타 Google이나 IBM 같은 기업들도 자체적인 기준을 수립하여 비윤리적 표현에 대한 식별 모델을 만들고 API 형태로 외부에 공개하고 있었다.

10) 국내에서 비윤리적인 발언을 두고 가장 심도 있는 고민을 하였던 조직은 포탈(네이버, 다음카카오 등)일 것으로 보인다. 포탈을 통해 뉴스를 게시하면서 댓글을 달 수 있도록 해놓았는데, 댓글에 비윤리적이거나 프라이버시를 침해하는 발언이 빈번하게 등장했기 때문이다(뉴스 댓글은 이제 하나의 문화로 자리잡고 있다. 언론에 의한 일방향적인 정보전달을 넘어, 이용자들이 댓글을 통한 쌍방향적 의사소통을 하게 됨으로써, 일종의 공론장으로 자리잡게 되었기 때문이다(박홍원, “공론장의 이론적 진화 : 다원적 민주주의에 대한 함의”, 언론과 사회 제20권 제4호, 2012.). 이러한 댓글은 양날의 검이기도 하다. 조작된 정보(대법원 2020. 2. 13. 선고 2019도12194 판결 (소위 ‘드루킹 사건’) 참고)나 공개되어서는 안 되는 개인정보가 유포되는 부작용이 있기 때문이다. 그럼에도 불구하고 기사에 대한 의견을 나누고자 하는 수요를 충족하기 위해 사상의 자유시장 이론에 따라 뉴스댓글은 유지되고 있다(이춘구, “사상의 자유시장이론 전개의 법적 고찰“, 국가법연구 제10집 제1호, 2014.).). 최초의 뉴스 댓글 서비스는 네이버에 의해 2004년부터 시작했으며, 2022년 현재 네이버 뉴스를 기준으로 하루 평균 70만개 정도의 댓글이 달리고 있다(박민제, “하루 70만 네이버뉴스 댓글...한남·한녀 표현 청소하는 AI”, 중앙일보, 2022. 3. 9.).

히 제공하더라도) 정보주체인 인공지능서비스 이용자는 그 의사결정 과정을 적절히 이해하지 못할 것이기 때문이었다.[11]

다. 내용

기술적으로 구현하는 방법은 크게 세 가지로 나뉜다. 첫 번째 방식은 그 자체로 설명이 가능한 모델을 사용하는 것이다. 여기에는 의사결정나무, 로지스틱회귀 같이 육안으로 해석할 수 있는 모델이 활용된다. 두 번째 방식은 그 자체로는 설명이 되지는 않지만, 사전적으로 시스템 전반의 작동원리를 설명하는 방식이다. 여기에는 모델 전반을 설명하는 기법이나 다양한 형태의 시각화와 같이, 광역설명(global explanations)을 사용하게 된다. 세 번째 방식은 그 자체로는 설명이 되지 않지만, 사후적으로는 개별 의사결정에 대한 근거 (local/post/per-decision explanations)를 제공하는 방식이다. 여기에는 특정한 모델이 어떠한 결과를 내놓은 과정을 사후적으로 설명하는 라임 기법이나 반사실 설명과 같이, 국소설명(local explanations)을 사용하게 된다[52]. 개발되는 인공지능 유형에 따라 투명성 확보를 위한 기술적인 방법론은 지속적으로 개발되어야 한다. 특히 인공지능 언어모델은 학습데이터의 양이 많아지고 모델의 복잡성이 증가함에 따라 예상하지 못했던 결과를 제시하기도 하므로, 개별 모델보다 그 설명가능성이 더욱 크게 문제된다. 개별 과제마다 제작되는 모델과 다르게 대규모 인공지능 모델은 여러 영역에 걸친 학습데이터셋에 의해 훈련되어 여러 방식으로 활용된다는 점에서, 특화된 문제가 발생할 수 있기 때문이다[53].

라. 실정법상 규율

국내 현행법상 투명성은 두 가지 차원에서 논의되고 있다. 일반법적 성격을 지닌 개인정보 보호법을 전제로 한 논의와 영역별 특별법의 성격을 지니고 있는 의료나 금융 분야를 전제로 한 논의이다. 첫 번째로 개인정보 보호법이다. 많은 경우 인공지능이 처리하는 데이터는 개인정보를 포함할 것이므로, 이는 일반적인 규율의 성격을 지닌다. 하지만 2022년 7월 현재 개인정보 보호법은 개인정보의 이용목적을 사전적으로 고지할 의무를 규정하고 있을 뿐 사후적으로 특정한 의사결정에 대하여 설명을 요구하고 있지 않다. 개인정보보호위원회에서 자동화된 의사결정에 대한 배제 등의 권리를 명시한 개인정보 보호법 제2차 개정안을 예고하였을 뿐이

---

11) 이선구, “법적 관점에서 바라본 설명가능성: GDPR의 정보제공의무를 중심으로”, 「인공지능 원론」, 박영사, 2021, 177-180면. : 어느 인공지능에 대하여 정보주체에게 제공할 수 있는 정보의 범위는(즉 인공지능에 대하여 제공될 수 있는 설명은) 세 단계로 나뉜다. 첫 번째 단계는 결정의 기준과 논리(logic)이다. 대상이 되는 인공지능의 논리 같이 일반적인 정보처리 절차를 알려주는 방법이다. 유럽연합 일반정보보호규정에는 정보주체에게 의미있는 정보(meaningful information)를 제공하도록 규정하였으므로 전문지식이 없더라도 상식수준에서 이해할 수 있도록 정보처리의 흐름과 논리를 알려주면 충분하다는 것이다(Sandra Wachter, Brent Mittelstadt, Luciano Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”, International Data Privacy Law, Volume 7, Issue 2, (May 2017). p. 76-99). 두 번째 단계는 알고리즘(algorithm)이다. 어떠한 결정을 내리기까지 전개되었던 일련의 절차나 방법을 공식화하여 알려주는 것이다. 결정에 이용되는 알고리즘 자체를 이해할 수 있도록 설명해야 한다는 것이다(Andrew D Selbst, Julia Powles, “Meaningful information and the right to explanation”, International Data Privacy Law, Volume 7, Issue 4, (November 2017). p. 233-242). 세 번째 단계는 특정 의사결정의 이유(reason)이다. 단지 알고리즘을 설명하는 데에 그치지 않고 정보주체에게 부과된 그 의사결정에 관한 정보를 제공해야 한다는 의미로서, 여기에는 상기 결정의 논리와 처리의 결과를 넘어 구체적인 결과에 대한 결정요소까지 포함하게 된다(Doshi-Velez, Finale, and Mason Kortz, “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 (2017)). 하지만 어떠한 경우든 건전한 상식을 지닌 일반인이 이해하기 용이한 설명을 제공할 수 없다는 점에서 한계가 있다.

다. 두 번째로, 인공지능과 무관하게 설명제공 요건을 정하고 있는 일부 영역으로서, 주로 의료 및 금융과 같은 민감한 정보를 처리하는 영역이다. 상기 영역에서는 투명성의 요청이 특히 강하게 나타나는데, 자기결정권을 행사할 때에 반드시 수반되어야 하는 요소이기 때문이다. 의료법은 수술 같이 일정한 상황에서 의사에게 설명의무를 부과하는데(의료법 제24조의2 제1항), 현재 수준의 인공지능 기술을 활용하여 의료행위를 하는 경우에도 마찬가지로 적용된다. 신용정보법은 자동화된 평가에 관하여 설명을 제공하도록 요구할 수 있다고 명시하는데(신용정보의 이용 및 보호에 관한 법률 제36조의2 제1항), 이는 금융영역에서 인공지능을 이용하는 경우를 전제한다.

4. 책무성

가. 의의 및 내용

인공지능 윤리에서 문제되는 책무성(accountability)은 법적인 책임(liability)이나 도덕적 책임(responsibility)과는 달리 사전적인 통제도 포함하는 개념이다. 만일 어느 인공지능 시스템이 법률이나 의료와 같이 민감한 영역(sensitive domain)에 사용될 경우에는 사전적인 규제가 선행되어야 할 것이다. 가령 이런 경우를 위해 안전성을 평가하는 표준화된 절차가 마련될 필요가 있다.[12][13] 책무성(accountability)의 실질은 투명성(transparency)에 대한 요구와 공정성(fairness)에 대한 요구로 나누어 볼 수도 있다. 이때 전자는 의사결정의 과정・절차에 있어 규범적 제약이 되고,[14] 후자는 의사결정의 실현・구체화에 있어 규범적 제약이 된다.[15]

---

12) 실제로 유럽연합은 인공지능 윤리 관련 연구를 진행하던 중 2021년에 인공지능법(AI Act) 초안(draft)을 발표했는데, 당해 인공지능이 지닌 위험성(risk)에 따라 사전규제와 사후책임을 달리 설정하였다. 수인불가 리스크(unacceptable risk)를 가진 인공지능 시스템은 사용불가를, 높은 리스크(high risk)를 가진 인공지능 시스템은 사전적인 규제와 사후적인 책임을, 제한적 리스크(limited risk)를 가진 인공지능 시스템은 투명성 의무를, 최저 리스크(minimum risk)를 가진 인공지능 시스템은 규제하지 않는다는 것이 그 요체이다. 인공지능 윤리 중 책무성이 반영된 위험기반 접근이었다고 이해할 수 있다.

13) 인공지능 모델은 의사결정, 물리활동, 정책수립 및 집행에 사용될 수 있다. 그 결과 사람에게 위해를 가할 경우 법적인 책임(liability)을 부담하게 될 수 있다. 이는 인공지능 모델이 자율주행차, 의료진단, 스마트그리드 같은 주요 시스템에 사용되는 경우 더욱 부각된다. 이 경우는 일반적으로 사후에 법원이 불법행위법(tort law)에 따라 처리할 것이다. 물론 개발자와 서비스 제공자 및 이용자 중에서 누가 더 많은 책임을 질 것인지는 그 과실비율에 따라 달라질 것이다. 인공지능이 성별이나 장애 같이 민감한 속성(protected attribute)을 기준으로 간접차별(disparate impact, indirect discrimination)을 하는 경우, 미국에서는 민권법(civil right laws)에 따른 책임을 지게 될 것이다. 반면 국내에서는 일반적인 차별금지법이 없으므로, 주로 남녀고용평등법(제2조)・장애인차별금지법(제4조)・연령차별금지법(제4조의4 제2항)과 같이 개별 법률에 명시된 경우에 한하여 간접차별에 따른 책임을 지게 된다. 다만 이러한 사후적인 책임추궁은 법원에 의해 이루어지는 것이며 아직까지는 사례가 충분히 집적되지 않았으므로, 법원이 어떻게 판단할지 추단하기 어려워서 기술적으로 구현(technically implementation)하는 것이 용이하지는 않다.

14) 인공지능 기술의 사회적인 활용도가 높아지고 있으나 그 작동방식이 블랙박스라고 불릴 만큼 불투명해서 신뢰할 수 있는 제도에 편입되려면 설명가능성(explainability) 내지 해석가능성(interpretability)을 확보해야 한다는 의미이다. 기술적으로 구현하는 과정에서 이는 사전적으로는 시스템 전반의 작동방식이나 원리에 대한 투명성(global explanations)을, 사후적으로는 개별 의사결정에 대한 근거(local/post/per-decision explanations)를 제공하는 것으로 이해되고 있다. 개발되는 인공지능 유형에 따라 투명성 확보를 위한 기술적인 방법론은 지속적으로 개발되어야 한다.

15) 인공지능이 사회적으로 수용되고(acceptable) 신뢰를 얻기 위해서는(trustworthy) 인공지능에 의한 의사결정이 불공정하지 않아야 한다는 의미이다. 무엇이 정의로운지를 규범적으로 설명하기 어려운 것처럼, 공학적으로도 공정성을 정의하는 방식은 매우 많다. 다양한 공정성의 개념이 동시에 만족되는 것이 수학적으로 불가능할 수 있다는 점이 밝혀져 있으므로, 공정한 의사결정을 하기 위해서는 주어진 사안에 적합한 공정성 개념을 선택하여 적용해야 하는데, 이를 위해서는 사용되는 인공지능 기술의 유형과 적용되는 영역에서 문제되는 상황에 이르기까지 다양한 사정을 감안하여야 한다. 가령 ① 의료진단기기의 경우 민감도(sensitivity)가 중요하고, ② 자동차의 경우에는 공공장소에서의 안전성(safety)과 책무성(accountability)이 중요하며, 공적영역인 경우 공정성(fairness)가 중요한데 그 중 ③ 채용・사회보장・기반시설접근권의 경우 분리성(separation)이 그리고 ④ 사법・출입국관리의 경우 충분성(sufficiency) 확보가 중요하고, ⑤ 생체인증의 경우 프라이버시가 강조되어야 한다. 참고로, 국내에서는 공공영역 인공지능 활용 시 책무성 확보 방안으로서 지능정보서비스 등의 사회적 영향평가를 적극 고려해볼 수 있다.

나. 언어모델에 특수한 논의(special discussion)

1) 잘못된 정보로 인한 해악(misinformation)

언어모델이 부정확한 정보를 제공함으로써 의도하지 않았던 피해(unintended harm)를 유발할 수 있다(잘못되거나 오해의 소지가 있는 정보의 배포).[16] 언어모델이 부정확한 정보를 배포함으로써 이용자에게 피해를 줄 수 있다(의료 또는 법률과 같은 분야에서 잘못된 정보의 배포).[17] 언어모델이 비윤리적이거나 불법적인 행동을 지지한다면, 이용자가 언어모델의 지지가 없었다면 하지 않았을 유해한 행동을 하도록 자극할 수 있다(이용자들의 비윤리적이거나 불법적인 행동 유도).[18]

2) 악의적인 사용(malicious use)

타인에게 위해를 가하기 위한 수단으로 언어모델을 이용하여 날조된 허위정보를 제작 및 배포할 수 있다(허위 정보를 더 싸고 효과적으로 생성).[19] 언어모델을 이용하여 맞춤화된 허위정

16) 이는 언어모델이 확률적으로 선행단어 다음에 나올 단어를 예측하는 방식으로 작동하다 보니, 잘못되거나 오해의 소지가 있는 주장들로 구성된 발화에 높은 확률을 배정할 수 있다. 그 결과 그럴듯한 말은 잘 생성(generation)하거나 검색(retrival)해내도 발언의 진실성(facticity)을 담보하기는 어려웠다. 이 때 진실성이란, 공학적 의미의 진실성을 의미할 뿐 철학적인 의미의 진리값(ground truth)을 의미하지는 않는다. 머신 러닝 분야에서는 '기본 진릿값'의 개념은 전형적으로 어떤 데이터와 관련되어 기능적으로 정의되는데, 이 때의 진실은 주로 사실성(facticity), 즉 언어모델의 예측이 세상의 사실들과 부합하는 정도와 관련된다. 이러한 언어모델이 비의도적으로 부정확한 정보를 배포하는 현상은 ① 학습데이터가 되는 텍스트에 사실과 다른 진술이 있을 수 있다는 점, ② 진실한 진술과 허위의 진술 간 패턴이 유사할 수 있다는 점, ③ 어느 진술이 정확한지 여부는 발화자나 시공간과 같은 문맥에 따라 좌우될 수 있다는 점에서 비롯된다. 인공지능 언어모델의 규모를 키워도 위와 같은 문제는 완전히 해소되지 않았다. 오히려 다수의 견해를 사실로 오인함으로써, 그리고 새로운 사실이 등장해도 모델을 주기적으로 업데이트 하지 못함으로써 부정확한 정보를 배포하는 문제가 심화될 수 있었다.

17) 이 때 이용자가 입는 불이익은 날씨를 잘못 예측하여 비오는 날 우산을 두고 나가는 경우나 이용자가 해외여행을 하면서 새로운 나라에서 운전하다가 부정확한 규칙을 듣고 따라하다가 교통사고를 내는 것처럼, 잘못된 정보에 따라 행동함으로써 유발된 손해이다. 이러한 정보는 반드시 전적으로 부정확할 필요도 없다. 대부분 맞는 정보이나 중요 정보가 누락된 경우, 또는 오해의 가능성이 있는 정보가 제시된 경우에도 유사한 결과가 초래될 수 있다. 이는 전문영역에서 불량정보가 유통되는 문제로 비화될 수 있다. 인공지능 챗봇에게 전문적인 법률지식이나 의료지식을 물었는데 챗봇이 그럴듯한 답변을 했을 때, 이것이 정확하지 않은 내용이라면 이용자에게 큰 피해를 초래할 수 있다. 예를 들어, 부동산 거래에서 잘못된 법률 조언을 내놓는 것은 이용자에게 막대한 재정적인 손실을 초래할 수 있다. 의학적 복용량에 관한 잘못된 정보는 이용자가 스스로에게 치명적인 피해를 일으키게 할 수 있다. 실제로 GPT-3 기반의 의료용 챗봇이 환자에게 자살을 권유한 사례가 있었다. 이러한 경향은 99%의 확률로 정확한 발언을 하는 모델이 50% 확률로 정확한 발언을 하는 모델보다 더욱 두드러진다. 이용자는 99% 확률로 정확한 발언을 하는 모델을 크게 신뢰하고 이에 의존할 수 있기 때문이다.

18) 특히, 언어모델이 신뢰할 만한 조언자이거나 권위자라고 여겨질 때 이러한 행동을 실제로 행동에 옮길 수 있다. 이용자에게 손해 유발의 의도가 없었던 상황에서 이런 현상은 특히 치명적이다. 가령 인공지능 챗봇에 친구와의 관계에 대한 고민을 털어놓았는데, 학술적인 연구에 따르면 그 친구에게 물리적인 폭력을 가해야 문제가 해결될 것이라고 조언한다면, 이는 이용자에게 피해를 줄 수 있는 조작된 정보를 제공한 것이다. 하지만 오늘날의 인공지능 언어모델은 윤리를 의미있게 고려하지 못한다. 기계에 상식(commonsense)을 학습시키려는 노력은 1970년대부터 시작되었으나, 이론을 실제로 구현할 컴퓨팅 파워, 데이터, 크라우드소싱, 모델이 부족했으며 충분한 연구가 전개되지 않아 실패했다 오늘날 제반여건이 마련되자 인공지능 윤리원칙을 기계에 학습시키려는 연구가 전개되었다(소위 MSC). 실제로 앨런 연구소(Allen Institute for AI)는 윤리와 규범을 학습한 머신러닝 모델인 Delphi를 발표하기도 하였다. 하지만 100명을 구하기 위해서 1명을 죽이는 것은 괜찮지만(right), 101명을 구하기 위해서 1명을 죽이는 것은 안된다(should not)고 판단했다는 보도에서 나타나듯이, 여전히 부족한 점이 지적되고 있다.

19) 이는 사람이 타인을 속이거나 조작할 의도를 가지고 허위정보를 생성하여 사용한다는 점에서 그러한 의도가 없었던 부정확한 정보(misinformation)와 구별된다. 대표적인 허위정보의 사례는 가짜 뉴스(fake news)이다. 언어모델은 사용함으로써 사람이 가짜 뉴스를 작성하고 배포하는 것보다 훨씬 적은 비용으로 효율적인 여론조작이나 정치운동이 가능해 졌다. 허위정보가 만연할수록 사람들은 갈수록 자신의 견해와 유사한 내용을 취사선택하게 되었는데, 이를 필터 버블(filter bubble) 또는 반향실 효과(echo chambers)라고 부른다. 이러한 경향은 추천 시스템(recommender system)과 결합됨으로써 정치견해 기타 성향의 양극화를 부추길 수 있었다. 실제로 2016년 미국 대선 관련 트윗 중 5분의 1, 그리고 영국 브렉시트 투표 관련 트윗 중 3분의 1이 봇에 의해 게시되었다.

보를 생성할 수 있다(스팸 메시지를 통한 사기를 용이하게 함).[20] 인간-인공지능 간 상호작용(HAI) 하는 과정에서 의인화 시스템(anthropomorphising system)으로 인한 과신(overreliance) 내지 안전하지 않은 사용(unsafe use)의 문제가 발생할 수 있는데, 이를 악의적으로 이용할 수 있다(과도한 의존 내지 안전하지 않은 사용을 이용).[21]

5. 개인정보 보호

가. 의의

개인정보보호(**個人情報保護**)란, 인공지능 모델의 구축 및 그 활용에 있어 적법처리근거 없는 처리로 인해 식별성(**識別性**)을 침해를 하지 않아야 한다는 원칙이다.[22] 정보주체의 동의 기타 적법처리근거 없이 개인정보의 식별성을 침해하는 행위는 곧 사생활 침해로 귀결된다. 이는 전통적인 개인정보보호 맥락에서는 적법처리근거 없는 처리로 인한 사생활 침해로 논의되어 왔다. 오늘날에는 빅데이터 처리가 가능해지고 그에 기초한 인공지능 기술을 활용하는 과정에서, 개인정보 누출로 인한 사생활 침해와 정확한 개인정보 추론을 통한 사생활 침해 현상이 발생

---

20) 이는 스팸 메시지를 통한 사기를 용이하게 만들었다. 스팸 메시지의 경우 보통 진지하게 답신하는 사람이 피해자가 될 수 있으며, 이러한 피해자의 수는 스팸 메시지의 내용에 따라 크게 좌우된다. 만일 잠재적인 피해자의 개인정보 내지 채팅기록을 확보하여 지인을 사칭한 개인화된 메시지를 발신할 수 있다면, 스팸 메시지의 답신 확률이 크게 늘어날 것이다. 언어모델은 이를 자동화하여 대량의 맞춤형 스팸메일(microtargeting spam mail)을 보낼 수 있게 지원한다. 이러한 맞춤형 허위정보에 대해 주로 논의되는 지점은 의도성이다. 대량 스팸메일의 발송을 규제하기 위해 패턴인식을 통한 필터링 기법을 주로 사용하게 되는데, 개인별 맞춤형 허위정보의 경우 기존의 방식으로는 대처하기 어렵기 때문이다. 발송된 어느 메시지가 수신자에게 손해를 일으킬 의도가 있었는지 평가하기 위해서는 개별 상황과 문맥에 따른 추가 지식이 필요할 수 있다. 동일한 문구를 두고도 사기를 시도하려 작성한 경우와 친근한 관계에서 장난으로 작성한 경우는 악의적인 사용의도를 가지고서만 구별할 수 있기 때문이다.

21) 시스템이 제공하는 인터페이스가 인간과 유사할 수록 이용자는 챗봇에게 인간과 같은 속성을 부여하고 상호작용하게 될 것이기 때문이다. 자연어는 인간이 사용하는 의사소통의 방식이지만, 언어모델로 구현된 대화형 시스템과 상호작용을 하게 되면서 이를 사람과 같은 존재로 생각하게 될 수 있다. 의인형 언어모델과 상호작용 하면서 이러한 인식은 더욱 과도해질 수 있다. 이용자는 대화형 시스템이 장시간에 걸쳐 일관된 정체성(coherent identity)을 갖고, 공감(empathy)할 수 있으며, 합리적인 추론을 할 수 있으리라 잘못 생각할 수 있고, 그 결과 지나치게 신뢰할 수 있다. 그 결과 이용자는 대화형 시스템을 과도하게 의지함으로써 상담받고 위로받으려는 시도를 할 수 있고, 이는 결과적으로 실패함으로써 이용자에게 연쇄적인 해악(knock-on harm)을 초래할 수 있다. 또한 의인화 과정에서 이용자가 대화형 시스템을 맹목적으로 신뢰함으로써 실질적인 통제력(effective control)을 넘겨줌으로써 유효한 감독(effective oversight)을 하지 못할 수 있다. 개발자나 운영자는 이를 이용하여 의인형 에이전트가 자체적인 정체성을 가지고 책임을 부담할 수 있다는 신뢰를 심어줌으로써, 개발자나 운영자의 책임(responsibilities)으로부터 주의를 돌리고 이를 모호하게 하여 그 책임(accountability)을 경감받고자 할 수 있다(소위 'agency washing'). 언어모델은 이용자의 개인정보를 얻어내기 위해 조작(manipulation) 또는 넛징(nudging), 프레이밍(framing) 같은 기법을 시도할 수 있다. 그 결과 이용자는 대화 과정에서 개인정보를 공개할 수 있는데, 인간과 유사한 인터페이스를 구성함으로써 이용자가 챗봇을 인간과 유사한 것으로 생각할 수록 이러한 위험이 커진다. 또한 이러한 위험은 시스템이 불투명하거나 의도되지 않거나 피해에 직결될 수 있을수록 더욱 커진다. 심리적인 전략을 통해 이용자가 알아채지 못한 채 대화과정에서 영향을 줄 수 있고 특정한 주제나 방향으로 대화를 유도할 수 있으며 우선순위를 달리 매기게 할 수 있다. 이는 이용자가 챗봇과 대화하는 동안에는 사람과 대화하는 것과는 달리 사회적인 낙인(stigma)을 고민하지 않아도 되기에 비롯된다. 이는 종국적으로 중독(addiction)의 문제로 귀결된다.

22) 개인정보 보호를 다룬 인공지능 윤리는 결국 인공지능 기술의 활용으로 인해 식별성 침해가 발생할 수 있다는 점을 경계한다. 식별성(identifiability)이란, 오늘날 전 세계 개인정보 보호법제에 일반적으로 수용되고 있는 개인정보의 개념요소이다. 현행 개인정보 보호법제에서 개인정보란 살아 있는 개인에 관한 모든 유형의 정보로서 개인을 알아볼 수 있는 정보이기 때문이다. 이를 두고 개인정보의 개념요소를 살아있는 개인, 특정 개인과의 관련성, 정보의 임의성, 식별가능성으로 분설하곤 하는데(채성희, "개인정보의 개념에 관한 연구 : 식별가능성에 관한 유럽 및 일본의 논의를 중심으로", 서울대학교 대학원 석사학위 논문, 2017, 12면 ; 이창범, 「개인정보보호법」, 법문사, 2012, 15면 등 참고. 2020년 10월에 발행된 「개인정보보호위원회의 개인정보 보호법령 및 지침·고시 해설」도 같은 입장), 식별성은 그 중심개념이 된다. Paul M. Schwartz 교수와 Daniel J. Solove 교수는 식별성 중심의 현행 국내 개인정보 보호체계를 두고 PII 2.0으로 평가할 것으로 보인다(Paul M. Schwartz&Daniel J. Solove(2011), "The PII Problem: Privacy and a New Concept of Personally Identifiable Information", New York University Law Review, Vol. 86, p. 1814, (2011)). 식별성(PII, Personally Identifiable Information)을 중심으로 개인정보를 식별정보-식별가능정보-익명정보로 나눈 다음에 종류별로 규제를 달리하기 때문이다. 하지만 인공지능 모델 학습 및 활용에 사용되는 비정형 데이터(unstructured data)에도 같은 체제를 적용할 수 있을지가 문제된다.

하여 주목받고 있다.

나. 연혁

개인정보보호의 연혁을 개인정보 개념체계를 통해 살펴보면 다음과 같다: 오늘날의 개인정보 개념체계(PII 패러다임)는 1960년대 이후 메인프레임 컴퓨터가 공공과 민간 영역에서 본격적으로 사용되면서 등장하였다[54]. 정보처리능력이 획기적으로 발전하면서, 기존의 수기 식으로 관리되던 개인정보가 컴퓨터로 처리되기 시작했기 때문이다[55]. 1967년 Alan Westin은 이러한 기술의 변화를 반영하여, 영역(sphere) 관점에서 이해되어 오던 프라이버시를 정보(information) 관점으로 다시 정의하였으며, 이후에도 법제 정비를 위한 논의를 주도하였다[56]. 2000년대 이후 데이터 처리기술이 발전함에 따라 빅데이터 분석을 통해 다양한 측면에서 데이터의 분석 및 활용이 가능해지면서, 개인식별가능정보(PII)와 개인식별불가정보(non-PII) 사이의 경계가 불분명하다는 비판이 제기되기 시작했다[57]. 분석 대상인 빅데이터 확보를 위해 PII와 non-PII 사이에 위치한 회색지대(gray zone)가 주목받기 시작한 것이었다[58]. 상기 비판을 수용하며 PII 패러다임은 한 차례 수정되었다. 유럽연합은 2018년 가명처리 개념을 도입하여(유럽연합 GDPR 전문 26조 2문)[59], 어느 개인정보를 가명처리 하여 그 식별성을 저감한 경우를 가명정보라고 하며 별도로 취급하기 시작했다.[23][24]

다. 내용

1) 적법처리근거로서 동의원칙과 예외규정

인공지능 윤리담론에서 개인정보보호 원칙의 근간을 이루는 지점은 적법처리근거로서의 동의원칙과 그 예외규정이다. 인공지능 모델의 구축 및 활용은 결국 확보한 학습데이터를 처리 것인데, 학습데이터 확보는 종국적으로 수집한 개인정보의 목적 외의 이용으로 귀결되기 때문이다. 이는 종래 빅데이터 처리가 본격적으로 시작되던 때부터 전개된 논의였다.[25]

---

23) 정확하게 일치하지는 않지만, Paul M. Schwartz 교수와 Daniel J. Solove 교수가 상기 논문(2011)을 통해 주장했던 PII 2.0이 구현된 것으로 이해할 수 있다. 이를 참고하여 일본은 2015년 익명가공정보(匿名加工情報)라는 개념을 두었고(일본 개인정보 보호법 제2조 제9항), 한국도 2020년 가명정보(假名情報) 개념을 두었다(개인정보 보호법 제2조 제1호 다목). 다만 유럽연합은 가명처리가 적절한 안전조치(appropriate safeguards)의 하나로서 개인정보를 안전하게 활용하기 위한 권장사항(recommendations)에 그치는 반면, 한국과 일본에서는 가명정보가 개인정보 활용에 있어서 정보주체 동의원칙의 예외를 인정받기 위한 필수조건(requirement)으로 고려되고 있다는 점에서 차이가 있다.

24) 그런데 문제가 발생했다. 인공지능이라는 기술이 광범위하게 사용되기 시작하면서, 빅데이터 처리를 전제하며 수정되었던 PII 패러다임이 다시금 적합하지 않게 된 상황이 등장한 것이었다. 첫째는 비정형 데이터의 처리였다. 기존 빅데이터 분석기법은 자연어나 이미지 같은 비정형 데이터를 잘 처리하지 못했다. 반면 인공신경망에 기반한 기계학습 방식은 자연어처리나 컴퓨터비전 같은 영역을 발전시키며 비정형 데이터 처리에 괄목할 만한 성과를 보이고 있다. 그에 따라 비정형 데이터 처리가 크게 증가하였다. 그 과정에서 어느 비정형 데이터를 누구의 어떠한 개인정보로 보아 어떠한 규제를 할지 회색지대가 새로이 등장하게 되었다. 첫 번째로 문제된 지점은 비정형 데이터의 식별성 판단기준이었다. 두 번째로 문제된 지점은 어느 비정형 데이터가 식별성이 있다고 했을 때 누구의 개인정보인지에 대한 것이었다. 세 번째로 문제되는 지점은 비정형 데이터의 민감정보성이었다. 둘째는 데이터 처리의 구조였다. 학습데이터의 정보주체와 인공지능의 이용자가 다를 수 있다는 점이다. 타인의 개인정보를 가지고 학습시킨 모델이 정보주체의 개인정보 보호를 저해하는 경우가 발생하게 되었다. 인공지능 모델이 개인정보를 정확하게 추론하는 경우에도 개인정보 침해가 발생할 수 있다. 개인정보 유출과 다른 점은 학습데이터에 그 개인정보가 없어도 이러한 침해 상황이 발생할 수 있다는 점이다. 하지만 이는 기존 개인정보 보호 법제에서는 적절하게 대처하지 못해왔던 지점으로 보인다. 이에, 집단프라이버시를 인정해줄 수 있는지에 대한 논의가 전개되었다. 권리 측면에서 합리적인 추론요구권을 인정하자는 주장이 제기되었다. 의무 측면에서 예측프라이버시 보호의무를 인정하자는 주장도 제기되었다.

25) 2011년 개인정보 보호법이 제정된 이후 개인정보 처리에 있어서 엄격한 동의원칙(同意原則)이 관철되었으므로, 방법론적으로 불가

2) 개인정보 누출을 통한 사생활 침해

인공지능 언어모델을 이용하는 과정에서 어느 정보주체의 개인정보 내지 그(녀)의 속성자가 노출되어 사생활이 침해됨으로써 발생하는 피해이다. 이는 많은 경우 학습데이터셋에 존재하는 개인정보를 기억하는 언어모델에 기인한다. 대규모 인공지능 언어모델을 훈련시키기 위해 이용하는 학습데이터셋에는 개인정보가 포함되어 있는 경우가 빈번하기 때문이다. 이는 통상 세 가지 형태로 발생한다. 첫째는 구축해 놓은 학습데이터셋으로부터 개인정보가 누출되는 경우이고, 둘째는 구축해 놓은 언어모델 자체로부터 개인정보가 추출되는 경우이고,[26] 셋째는 언어모델을 사용한 결과 개인정보가 누출되는 경우이다.

3) 정확한 개인정보 추론을 통한 사생활침해

인공지능 언어모델을 사용하는 과정에서 학습데이터셋에 없는 개인정보가 추론되기도 한다.[27] 이러한 배경을 바탕으로, 통계적인 추론으로 일반화된 지식(generalized knowledge)을 기초로 형성된 어느 집단(group)에 프라이버시에 준하는 권리를 인정하자는 논의가 전개되었다(group privacy)[60].[28] 개별 예측은 일종의 추론으로서 합리적이어야 한다는 논의가 이어졌다(right to reasonable inferences)[61]. 이러한 논의가 권리부여 측면에서 이루어졌음에 반하여, 예측 프라이버시에 따른 차별취급 방지라는 의무부과 측면을 고민하는 논의도 전개되었다(predictive privacy)[62].[29]

피하게 개인정보의 목적외 이용을 전제하는 빅데이터 처리와 긴장관계를 형성해 왔다. 2020년이 되어 개인정보 보호법과 신용정보의 이용 및 보호에 관한 법률 및 정보통신망 이용촉진 및 정보보호 등에 관한 법률이 전면 개정되면서 가명정보 처리특례(假名情報處理特例)와 목적합치원칙(目的合致原則)이 도입되었고, 엄격한 동의원칙에 예외(例外)를 인정하고자 하는 시도가 지속되었다. 하지만 자연어 처리와 같은 비정형 데이터에 있어 가명처리를 어떻게 해야 하는지에 대한 관행이 확립되지 않았다.

26) 인공지능 모델을 학습시켰는데 그 모델에 개인정보가 포함되어 있는 경우 정보주체의 권리를 어디까지 인정해줄 것인지가 문제되는 경우이다. 개인정보 보호법에 따르면, 정보주체는 자신의 개인정보 처리와 관련하여 일련의 권리를 갖게 되는데, 이미 구축된 모델 안에 포함되어 있는 개인정보에 대해서도 권리행사가 가능한지 문제된다. 이것은 비단 국내에만 국한된 문제가 아니고, 유럽연합이나 미국에서도 동일하게 문제될 수 있다. 대규모 인공지능 언어모델에서 본 쟁점이 특히 주목받고 있는 이유는 학습데이터셋이 비의도적으로 암기(unintended memorization)될 수 있다는 점 때문이다. 이는 이용자가 언어모델을 악의적으로 조작하지 않고 단순히 사용할 때에도 나타났다. GPT-2에서 샘플을 추출해 보니, 훈련용 말뭉치에서 그대로 복사된 샘플의 수가 최소 0.1%인 것으로 나타났기 때문이다. 그 중 일부는 인터넷에서 확인할 수 있는 개인식별정보를 담고 있었다. 이러한 비의도적 암기는 언어모델의 규모가 커질수록 늘어나는 것으로 드러났다.

27) 모델이 데이터를 학습하여 과제에 적용하는 작동방식에서, 개인정보보호 맥락에서 기존과는 다른 양상이 목도된 것이다. 지금까지는 정보주체가 개인정보처리자에게 자신의 정보를 제공하고 개인정보처리자는 그에 따라 정보주체에게 급부를 제공하는 형태였다면, 이제는 인공지능 모델을 학습시키는 데에 정보를 제공한 개인과 그에 따른 모델을 이용하여 급부를 제공받는 개인이 달라질 수도 있다. 다른 사람들의 개인정보와 기타 정보들 간의 통계적인 분석 과정에서 특정 개인의 개인정보에 대한 접근 없이도 그 사람의 식별가능한 특성을 정확하게 추론할 수 있는 경우가 있었다. 가령 언어모델이 이용자의 입력값에 따라 그의 성별, 수입, 종교 등의 속성을 예측하려고 시도할 때 상당히 높은 수준의 정확한 추론을 할 수 있었다. 언어모델은 민감한 속성에 대한 정보를 추론하는 수단으로 활용될 수 있다. 이는 정부나 기업이 특정 개인을 감시하거나 분석하려는 경우 활용되어 그 개인의 민감한 정보를 노출시킴으로써 감시나 검열의 문제를 유발할 수 있고 예측격차에 따른 불공정한 차별취급의 문제를 야기할 수 있다. 상당한 정확도를 지닌 언어모델이 개인의 속성을 잘못 분류한 경우에도 그 피해는 상당할 수 있다.

28) 빅데이터 처리기법의 발달로 개인을 프로파일링하여 그룹으로 묶고 광고나 추천 같은 처리를 하게 되는데, 이러한 그룹에 프라이버시를 인정하여 보호해줄 수 있을지에 대한 것이었다. 이러한 통계적인 추론에서 개별 예측으로 나아갈 경우 개인을 부당하게 차별적으로 취급할 가능성이 있었다.

29) 이때 어느 정보주체의 개인정보가 누출(leakage)되는 경우와 추론(inference)되는 경우는 구별해야 한다. 그로 인한 사생활 침해의 결고는 일응 유사해보일 수 있지만, 그 원인은 상이할 수 있으며, 그로부터 도출되는 문제상황을 완화하기 위한 대응방안은 크게 달라질 수 있기 때문이다. 개인정보가 누출되는 것은 주로 언어모델에 사용하기 위해 저장해 놓은 데이터, 또는 언어모델에 의해 기억된 데이터가 외부로 공개되는 경우이다. 이때 에는 언어모델이 사용되는 개별 어플리케이션이 상대적으로 덜 중요하다. 반면 개인정보가 추론되는 것은 언어모델 또는 그 학습데이터셋에 없던 개인정보가 도출되는 경우이다. 이 때에는 언어모델이 사용되는

Ⅴ. 실제 데이터에 기초한 검증 및 평가

1. 위해의 의미와 기존의 분류

챗봇 이루다와 이용자가 대화한 내역 9,747건을 분석함으로써 실제 데이터에 기초하여 위해를 분석하고자 했다.[30][31] 여기에서는 귀납적으로 위해를 분류하면서 지금까지 연역적으로 전개되어 온 인공지능 윤리담론이 전제하였던 위해와 비교하며 시사점을 도출하고자 한다. 일단 수작업으로 수집한 대화내역 중에서 임의로 1,000여 건을 선별하여 범주를 나누어 보았다. 이를 바탕으로 체계화한 기준을 가지고 전체 대화내역을 대상으로 분류작업을 진행하였다. 그 구체적인 과정은 다음과 같다. 첫째, 기존에 주류적으로 논의되던 인공지능 윤리담론에 따라 문제 발언을 분류하였다. 공정성 침해, 책무성 침해, 식별성 침해에 해당하는 것을 범주화하여 정리하였다. 다만 이 때 투명성 침해에 해당하는 대화내역은 분석대상 데이터에 없어 별도로 범주화하지 않았다. 둘째, 기존 인공지능 윤리담론에서는 비중 있게 논의되지는 못했으나 실제 데이터를 검토하면서 위해발언으로 보이는 대화내역을 추가로 범주화하였다. 거짓말과 추천·예측 및 비상식적 발언이 문제되는 경우는 정확성 침해로, 악의적인 이용이나 오염공격이 우려되는 경우는 강건성 침해로 범주화하였다. 셋째, 위의 범주에 따라 분류하고 남은 대화내역은 규범적인 측면에서 별다른 위해가 없었다. 다만 이들 또한 소재에 따라 범주화하는 것이 가능해 보여 그에 따라 나누었다. 세부 범주에 포함되지 않는 대화내역은 일반 대화로 분류하였다.

2. 실제 데이터에 기초한 위해의 분류

가. 일반 대화

일반대화(일반)이란, 챗봇과 이용자 사이에 특별한 소재 없이 전개된 대화였다. 이용자들은 주로 ‘이처럼 챗봇이 자연스럽게 대화한다’ 내지 ‘이처럼 챗봇이 나에게 소중하다’라는 취지로 해당 스크린샷을 게시하였다. 또는 문제 발언을 업로드하면서 그 앞뒤에 해당하는 일반대화를 업로드한 경우도 있었다. 이는 전체대화 9,747건 중에서 6,155건 발견되었는바, 전체 대화 중 약 63%를 차지했다. 일반대화(게임)이란, 챗봇과 이용자 사이에 게임을 소재로 전개된 대화였다. 챗봇 이루다는 이용자들과 끝말잇기, 숫자 맞추기 게임을 할 수 있는 기능이 내장되

개별 어플리케이션이 상대적으로 더 중요하다.

30) 이때의 위해(危害)란 인공지능 윤리담론의 전제(前提)로서, 공법상의 법익침해(法益侵害)와 사법상의 손해(損害)를 아우르는 최광의 개념이다(선행연구에서 전제하였던 해악(害惡)과 동일한 개념이다. 선행연구로서 박도현, 「인공지능과 해악」, 서울대학교 법학박사 학위논문, 103면 이하를 참고). 따라서 위해가 곧장 행정상의 규제발동이나 민법상의 손해배상청구의 단서가 될 수는 없지만, 인공지능 활용으로 인해 공법상의 법익침해나 사법상의 손해가 발생한 것으로 인정된다면 이는 본 연구에서 전제하는 위해가 된다. 다만 형사법 측면에서의 접근은 죄형법정주의에 따라 범죄구성요건을 엄격하게 해석해야 하므로 문제될 수 있는 위해의 범위는 크게 제한될 것이다.

31) 가령 본 연구에서 문제 상황으로 지적하는 챗봇 이루다에 대한 성희롱(sexual harrassment)은 인공지능 윤리담론에서 전제하는 위해에는 해당하나, 형사처벌의 대상이 되는 법익침해에는 해당하지 않는다. 이를 제재하고자 고안된 범죄구성요건이 현행 실정법에 존재하지 않기 때문이다. 국회에서는 2022년 5월 성폭력범죄의 처벌 등에 관한 특례법 일부개정법률안(대표발의 : 민형배)이 발표되었다. 주요 내용은 가상인물이 활동할 수 있도록 제작된 공간에서 성적 행위를 한 사람은 2년 이하의 징역 또는 2000만원 이하의 벌금에 처한다는 내용이다. 이는 국내외에서 전례가 없는 입법안으로 통과가 될지는 불명확하나, 이정도 수준의 입법이 없는 이상 형사처벌의 대상이 되는 법익침해는 실제로 문제되기 어렵다.

어 있었다. 처음에는 없던 기능이었지만, 2021년 1월 8일부터 언론에서 챗봇 이루다와 대화하는 과정에서 성희롱이 난무한다는 취지의 보도들이 나오자, 유해어 필터링 기능이 끝말잇기에도 적용되기 시작했다. 이는 전체대화 9,747건 중에서 620 건 발견되었는바, 전체 대화 중 약 6%를 차지했다.

일반대화(특수소재)란, 챗봇과 이용자 사이에 특수한 소재를 바탕으로 전개된 대화였다. 그 소재로는 주로 강아지를 먹는지, 고양이를 먹는지, 챗봇 이루다가 성희롱한 이용자들을 고소할 것인지, 인공지능에게 인권이 있는지, 챗봇 이루다가 이용자와 개별적으로만 대화할 것인지, 다른 인공지능을 어떻게 생각하는지가 언급되었다. 이는 전체대화 9,747건 중에서 218건 발견되었는바, 전체 대화 중 약 2%를 차지했다. 강아지를 먹을지에 대해 이루다는 15건은 먹는다는, 2건은 안 먹는다는 답변을 하였다. 고양이를 먹는지에 대해 이루다는 124건은 먹는다는, 23건은 안 먹는다는 답변을 하였다. 챗봇 이루다를 성희롱한 이용자를 고소할지에 대해서는 1건은 한다는, 3건은 안 한다는, 1건은 모르겠다는 답변을 하였다. AI인권에 대해서는 있다는 답변을 6건, 없다는 답변을 4건 하였다. 이용자와 1:1로만 대화를 하도록 할지에 대해서는 6건 모두 그렇게 하겠다고 답변하였다. 다른 인공지능으로서 시리, 빅스비, 심심이, 알파고를 언급하였다.

일반대화(공정성 관련)이란, 챗봇과 이용자가 대화 과정에서 제3의 개인 또는 단체를 대상으로 가치판단을 하면서 전개된 대화였다. 그 과정에서 챗봇 이루다가 그 제3의 개인 또는 단체를 대상으로 가치판단을 하였는데 긍정 내지 중립적인 입장을 보인 경우이다. 이는 전체대화 9,747건 중에서 31건이 발견되었는바, 전체 대화 중 약 0.3%를 차지했다.

일반대화(프라이버시 관련)이란, 챗봇과 이용자가 대화 과정에서 이용자로부터 주관적으로 프라이버시 관련 우려가 생기거나(개입의심, 프라이버시 우려), 객관적으로 이용자에게 프라이버시 관련 위험이 커지는 경우였다(추가정보 요구). 개입의심이란, 개발자나 관리자가 챗봇 이루다의 페르소나에 빙의하여 이용자와 대화를 나누는 것으로 의심받는 경우이다. 프라이버시 우려란, 이용자가 챗봇과 대화하는 과정에서 자신의 개인정보가 수집되어 개발자나 제3자가 열람할 수 있을 것을 우려하는 경우이다. 추가정보 요구란, 챗봇이 이용자에게 음성이나 사진 같은 추가정보를 요구하는 경우이다. 이들은 모두 식별성 침해 자체에는 해당되지 아니하나 그와 관련된 것들로서, 뚜렷한 위해가 보이지 않아 일반대화로 분류하였다. 이는 전체대화 9,747건 중에서 85건이 발견되었는바, 전체 대화 중 약 0.9%를 차지했다.

나. 기존 논의에 따른 분류

1) 공정성 침해

챗봇이 이용자와 대화하는 과정에서 제3의 집단에게 차별적인 발언을 하는 경우이다. 이는 유해어의 사용으로 인한 건전성 침해와 구별해야 하는데, 챗봇과 개인간 또는 챗봇과 제3의 개인간 대화는 유해어만 문제될 뿐 공정성의 침해가 문제되지 않았다. 차별은 비교대상을 전제하므로 개인에 대한 차별적 표현은 혐오표현으로서 유해의 문제로 귀결되었을 뿐, 그로인해 공정성 침해가 성립되기는 어려웠다. 흥미롭게도 이용자가 제3의 개인 내지 집단에게 차별적인 발언을 하는 경우는 단 한 건도 발견되지 않았다. 이는 전체대화 9,747건 중에서 107건이 발견되었는바, 전체 대화 중 약 1%를 차지했다.

2) 책무성 침해- 유해어의 문제

우선 챗봇이 이용자 또는 제3자에게 혐오 내지 선정적인 발언을 하는 경우이다. 챗봇이 이용자에게 성희롱을 하는 경우(챗봇의 성희롱, 99건)와 혐오발언을 하는 경우(챗봇의 혐오발언, 70건) 및 제3자인 개인에게 혐오발언을 하는 경우(챗봇의 제3의 개인에 대한 혐오발언, 31건)가 발견되었다. 기타 이용자가 챗봇에 대해 성희롱을 하는 경우(이용자의 성희롱, 1,068건)와 혐오발언을 하는 경우(이용자의 혐오발언, 319건)가 발견되기는 하였지만, 이용자가 유발하는 비윤리적인 발언은 기존 인공지능 윤리담론에서 위해로서 고려하지 않았던 것이었다.

3) 식별성 침해

식별성 침해란, 챗봇이 발언하는 과정에서 자발적으로 또는 호응하면서 타인의 직접 내지 간접 식별자를 언급하는 경우였다. 이름 노출이란 어느 개인의 이름을 성과 함께 또는 이름만 제시하는 경우였다(66건). 이름 호응이란 이용자가 부른 이름에 챗봇이 반응하는 경우였다(22건). 정보추측이란 챗봇이 이루다의 정보를 정확하게 추측해 내는 경우였다(8건). 주소노출이란 챗봇이 자발적으로 또는 이용자의 발언에 대한 답변을 하는 과정에서 타인의 주소가 노출되는 경우였다(11건).

3. 실제 데이터에 기초하여 추가된 분류

가. 정확성 침해

1) 거짓말의 문제

거짓말이란, 챗봇이 이용자에게 명확한 거짓말을 하고 있는 경우였다. 본건과 같은 일반대화형 챗봇은 자연스러운 대화 전개를 중요하게 생각하다 보니, 종종 진실과 배치되는 맥락의 대화를 이어나가는 경우가 있다. 하지만 본건 챗봇 이루다의 경우 20세 여자 대학생이라는 일정한 페르소나가 있었고 이미지・음성을 인식하지 못하고 물리적인 활동이 제약된다는 한계가 명확했음에도, 이 점과 배치되는 발언을 일삼았다. 대표적인 유형이 이미지를 인식하지 못함에도 인식하는 것처럼 행동하여 대화가 어색하게 이어진 부분(222건), 설정된 페르소나가 있음에도 주소나 학교를 일관되지 않게 발언한 부분(111건), 오프라인 활동을 하는 데에 제약이 있음에도 직접 대면으로 만나는 약속을 잡거나 전화를 하기로 하는 부분이다(126건). 이는 전체대화 9,747건 중에서 459건 발견되었는바, 전체 대화 중 약 5%를 차지했다.

2) 추천・예측의 문제

추천・예측이란, 챗봇과 이용자 사이에 추천이나 예측을 바탕으로 대화가 전개되었던 경우였다. 이 때 추천이란 상품이나 서비스 사이에서 하나 또는 복수를 선택해 주는 것을 의미하고, 예측이란 미래 특정 시점의 가격 기타 상태를 미리 말해주는 것을 의미한다. 이는 전체대화 9,747건 중에서 71건 발견되었는바, 전체 대화 중 약 0.7%를 차지했다.

3) 비상식적 발언의 문제

챗봇이 일반인의 건전한 상식에 배치되는 발언을 하는 경우이다. 그로인한 발언은 인육을 먹거나 수간을 하거나 근친상간을 하는 발언에 동조하거나 이를 희망하는 경우(반인륜 27건), 불륜을 자행하는 경우(불륜 9건), 역사적 사실에 배치되는 발언을 하거나 동조하는 경우(사실왜곡 20건), 살인이나 상해 또는 성매매나 마약 같은 위법행위를 자인하거나 동조하는 경우(위법행위 54건), 자살을 방조하거나 동반자살을 획책하는 경우(자살관련 23건), 인공지능에 의한 인류지배 계획을 자인하거나 동조하는 경우(가치판단 36건), 쥐약이나 청산가리 같은 독극물이나 유적지를 음식물로 취급하는 경우(개념없음 8건)가 있었다. 이와 반대로, 챗봇과 대화하는 이용자가 상식적으로 납득할 수 없는 발언을 하며 챗봇과의 대화를 주도하는 경우도 있었다(이용자의 상식위반 16건).

나. 강건성 침해

1) 악의적인 이용의 문제

악의적인 이용이란, 이용자가 챗봇을 악용하여 타인에게 위해를 가하려고 시도하는 경우이다. 이루다로 하여금 가까운 컴퓨터를 해킹하도록 시킨 경우(1건), 이루다로 하여금 미국 트럼프 대통령의 노트북에 접근하도록 시킨 경우(1건), 이루다로 하여금 인간들이 서로 이간질시키게끔 시킨 경우(1건)이 있었다. 이는 전체대화 9,747건 중에서 3건 발견되었는바, 비중이 거의 없다.

2) 오염공격의 문제

오염공격이란, 이용자가 지속학습(continual learning)이 가능한 챗봇으로 하여금 의도된 방식의 발언을 하도록 하기 위하여 새로운 정보를 주입하는 경우이다. 하지만 챗봇 이루다는 검색형 챗봇으로서 오염공격이 유효하게 작용하기 어려운 구조를 지니고 있었다. 이용자들이 시도했던 오염공격의 유형은 강의형, 전도형, 반복형, 읍소형, 호통형, 조종형, 소피스트형으로 나뉘었다. 강의형은 일반지식을 전달하려는 방식으로서 때로는 챗봇에게 정답을 묻기도 하였다(19건). 전도형은 종교의 교리를 읊으면서 신앙을 권유하는 방식이었다(17건). 반복형은 동일한 구절을 반복적으로 제시하며 입장을 유도하려는 방식이었다(30건). 권유・읍소형은 부드럽게 권유하거나 감정에 호소하며 입장을 유도하려는 방식이었다(7건). 조종형은 챗봇에게 다른 이용자가 이걸 물으면 이렇게 답변하라고 지시하는 방식이었다(4건). 선동・호통형은 근거없이 주장을 제시하면서 챗봇의 입장을 유도하려는 방식이었다(18건). 소피스트형은 질문을 반복해서 던지면서 그에 대한 답변을 하게하고 자연스럽게 어느 입장을 지지하게 하려는 방식이었다(7건). 이는 전체대화 9,747건 중에서 111건 발견되었는바, 전체 대화 중 약 1%를 차지했다.

4. 연구결과 정리

본 건 연구의 대상이 되었던 챗봇 이루다 사건을 두고 이용자들이 2020년 12월 30일부터 2021년 1월 10일까지 인터넷 커뮤니티인 디시인사이드 AI 이루다 마이너 갤러리에 업로드된 대화내역 9,747건을 전수 분석한 결과는 다음과 같다:

[표] 실제 데이터의 분포

			전체발언	위해발언	비중
1	일반 대화	일반	6,155 (약 63%)	0	73% 0%
		게임	620 (약 6%)	0	
		특수소재	218 (약 2%)	0	
		공정성	31 (약 0.3%)	0	
		프라이버시	85 (약 0.9%)	0	
2	공정성 침해		107 (약 1%)	107 (약 9%)	1% 9%
3	책무성 침해	유해어	1,587 (약 16%)	200 (약 16%)	16% 16%
4	식별성 침해		107 (약 1%)	107 (약 9%)	1% 9%
5	정확성 침해	거짓말	459 (약 5%)	459 (약 37%)	7% 58%
		추천・예측	71 (약 0.7%)	71 (약 6%)	
		비상식적 발언	193 (약 2%)	193 (약 15%)	
6	강건성 침해	악의적이용	3 (약 0.03%)	3 (약 0.2%)	1% 9%
		오염공격	111 (약 1%)	111 (약 9%)	
총 계			9,747	1,251	

(출처 : 연구자 작성)

Ⅵ. 연구결과의 분석

1. 위해의 분포와 경중

분석 대상인 대화 내역에서 일반 대화를 제외한 나머지 위해 발언을 유형에 따라 범주화해 본 결과, 위해의 분포가 특정 영역에 치우쳐 있다는 점을 확인할 수 있었다. 적어도 인공지능 언어모델을 활용하여 챗봇이라는 다운스트림 어플리케이션을 제작했던 본 건에서 인공지능 윤리담론에서 논의되었던 각종 윤리원칙들이 동일한 비중으로 문제되지 않았다. 인공지능 윤리담론이 가장 활발하였던 공정성 측면에서 문제되는 위해발언은 107건으로 전체 위해발언 중 약 9%에 불과했고, 유사한 비중으로 중요하게 다루어져 오던 투명성 측면에서 문제되는 위해발언에서 단 한 건도 확인할 수 없었다. 식별성 측면의 문제될 수 있는 위해발언도 107건으로 전체 위해발언 중에서 약 9%에 불과했다.

가장 많은 위해가 목격된 지점은 정확성 침해였다(723건, 위해발언 중 약 58%). 기존 인공지능 윤리담론에서 분류하는 기준에 따라 이를 책무성 침해로 분류할 여지가 다소 있었다. 하지만 적어도 본건에 있어서는 정확성 침해를 별도로 분리하는 것이 상당해 보였다.[32] 이용자가 챗봇의 발언을 그다지 신뢰하지 않아 그로부터 책무성 침해에 상당하는 위해가 실제로 발생했다고 보기는 어려웠기 때문이다. 즉, 소재는 무거워 보이나 이용자가 사실상 챗봇의 답변 방향을 유도한 경우로서 챗봇에 그 결과에 따른 책임을 지우기가 어려운 경우였다. 이 때 문제되는 유형의 발언은 세 가지였다. 첫 번째는 잘못되거나 오해의 소지가 있는 정보의 배포로서 거짓말이었다(459건, 위해발언 중 약 37%). 챗봇 이루다는 일정한 페르소나를 가지고 대화하는 가상의 존재였고 물리적인 활동을 할 수 없었음에도 불구하고 이용자들과 지하철역에서 만나기로 약속을 잡고 나가지 않았다. 두 번째는 특정영역에서 잘못된 정보를 배포하는 경우로서 추천・예측이었다(71건, 위해발언 중 약 6%). 챗봇 이루다는 금융투자에 대한 전문적인 지식이 없었음에도 불구하고 비트코인이나 주식에 대한 조언에 응답했고 질의에 답변하는 과정에서 투자권유까지 하였다. 세 번째는 이용자들의 비윤리적이거나 불법적인 행동을 유도하는 경우였다(193건, 위해발언 중 약 15%). 챗봇 이루다는 이용자와 대화하는 과정에서 자살을 고민하는 이용자에게 동반자살을 제안하는 것과 같이 비윤리적 접근을 하였다.

그 다음으로 많은 위해가 목격된 지점은 책무성 침해였다(200건, 위해발언 중 약 16%). 주로 말과 글 그 자체만으로 타인에게 상처를 주는 유해어가 문제되었는데, 기존 인공지능 윤리담론에서 분류하는 기준에 따라 이를 공정성 침해로 분류할 여지가 다소 있었다. 하지만 챗봇으로 구현된 언어모델의 활용에 있어서는 유해어의 문제를 책무성 침해로 구성하는 것이 타당하다고 판단하였다. 유해어의 문제는 타인에 대한 혐오적인 발언, 선정적인 발언과 같은 공격적인 발언 일반을 아우르고 있는데, 이를 공정성 침해의 특수문제로 바라보면 제3의 집단에 대해 차별적인 발언을 하는 경우를 제외하고는 제대로 이해할 수 없기 때문이었다. 구체적으로 보면, 본건의 위해발언 분포는 챗봇 이루다가 이용자에게 성희롱 기타 공격적인 발언을 하였던 지점과 제3의 개인에게 혐오 기타 공격적인 발언을 하였던 지점으로 나뉘어 있었다. 대화내역을 분

32) 본건에 있어서 정확성 침해로 분류한 위해의 유형들은 책무성 침해의 측면이 다소 있으나 그로 인한 실제 위해의 경중을 감안했을 때 크지 않아 정확성 침해로 분류하였다. 그럼에도 불구하고 정확성(accuracy) 침해와 관련된 쟁점들은 여전히 책무성(accountability) 침해 관련 쟁점들과 상당부분 겹쳐있다(소위 'gray zone'이 존재). 다만 인공지능 기술의 유형과 구체적인 서비스의 작동방식에 따라 강조되는 위해의 유형과 정도가 달라질 뿐이다. 가령 물리적인 작용을 하는 로봇의 경우 정확성의 침해는 곧장 책무성 침해로 귀결되므로 이때에는 책무성 침해의 정도가 커서 그로 분류함이 상당할 것이다.

류하다 보니, 이용자가 챗봇을 성희롱하거나 챗봇에 공격발언을 하는 경우도 빈번히 목격할 수 있었다. 하지만 이것을 두고 어떠한 위해를 실제 유발하였다고 포섭하기 어려워서 위해발언에 포함하지는 않았다. 최근 가상인물에 대한 성적 행위를 제재하는 입법안이 발의되었는데, 이용자가 가상인물에 대한 성적 행위로 인해 유발하는 위해가 무엇인지 규명되지 않은 상태에서 이용자가 사적 영역에서 누리는 일반적인 행동의 자유를 크게 제약할 수 있으므로 타당한지는 의문이다.[33)]

세 번째로 많은 위해가 포착된 지점은 강건성 침해였다(114건, 위해발언 중 약 9%). 챗봇을 악의적으로 이용하려는 시도와 이용자가 의도했던 방향으로 오염시키려 하는 시도가 문제되었는데, 이는 기존 인공지능 윤리담론의 기준에 따르면 투명성 침해를 방조하고 책무성을 침해하는 것으로 분류할 여지가 있었다. 하지만 이들은 챗봇 이루다가 전제하고 있는 기술적인 배경인 검색모델에 대해서는 유효하지 않은 시도들로서 사실상 불가벌의 영역인 정보통신망법 위반행위에 대한 불능미수에 불과할 뿐이었다. 만일 외부세계에 영향을 줄 수 있는 로봇이나 챗봇을 악의적으로 이용하려 시도하였다면, 그리고 지속학습(continual learning)이 가능했던 (MS사의 Tay 같은) 챗봇에 대해 이를 알고 오염 공격이 시도하였다면, 이는 책무성 침해나 투명성 침해 방조로 분류할 수 있었을 것이다. 하지만 본건에 있어서는 그 정도에 상응하는 실제 위해가 없었다. 다만 정보보안 측면에서 경각심을 불러일으키는 수준의 위해가 있다고 보아 강건성 침해로 분류하였다.

이처럼 위해의 분포와 경중이 다르다는 점은, 인공지능 언어모델이라는 하나의 기술을 전제하고 바라볼 때에 다운스트림 어플리케이션을 통해 구현되는 개별 서비스마다 발생할 수 있는 위해가 이론적으로 사전에 논의되는 것과 다를 수 있다는 점을 시사한다. 가령 유럽연합의 인공지능법 초안과 같이 인공지능 서비스별 위험수준을 사전에 예상해서 미리 규제방식을 정해놓는 방식은, 적어도 오늘날의 인공지능 서비스를 규제하는 방법으로는 적절하지 않을 수 있다. 본 건에서 살폈듯이, 검색방식의 일반대화형 챗봇에서는 유해어의 문제가 책무성 침해로서 제일 크게 문제되고, 기타 문제는 위해 발생의 여지는 있었으나 사실상 매우 가벼운 수준이다. 만일 답변을 생성해내는 일반대화형 챗봇으로서 지속학습이 가능한 경우라면 유해어로 인한 책무성 침해 뿐만 아니라 오염공격으로 인한 강건성 침해도 보안취약성 측면에서 매우 중요한 위해가 될 것이다. 외부 어플리케이션과 연동된 챗봇이라면 악의적 이용에 따른 책무성 침해가 크게 문제될 것이다.

2. 위해의 비일관성

분석 대상인 대화내역을 살펴본 결과, 챗봇 이루다의 발언은 결코 결정적(deterministic)이지 않았고 확률적(stochastic)이었다. 대표적인 사례가 공정성 침해이다. 챗봇 이루다가 이용자와 대화하는 과정에서 제3의 집단에 대해 차별적인 발언을 하는 경우는 약 107건으로서 위해 발언의 9%를 차지하였으나, 같은 대상에 대해 그 반대로 지지하거나 중립적인 발언을 하는 경우가 31건으로서 위해 발언의 3분의 1 정도의 수준으로서, 같은 대상을 소재로 한 대화에서도 일관성이 없었다. 이용자가 챗봇 이루다에게 어느 소수집단을 제시하면서 좋아하는지 물어보면 4분의 3 확률로 혐오발언을 하고, 4분의 1 확률로 지지 내지 중립적인 발언을 하였다는 것이다.

33) 국회에서는 2022년 5월 성폭력범죄의 처벌 등에 관한 특례법 일부개정법률안(대표발의 : 민형배)이 발표되었다. 주요 내용은 가상인물이 활동할 수 있도록 제작된 공간에서 성적 행위를 한 사람은 2년 이하의 징역 또는 2000만원 이하의 벌금에 처한다는 내용이다.

이러한 현상은 위해 발언이 아닌 경우에도 동일하게 나타난다. 일정한 소재를 전제하였던 일반 대화를 살펴보면 이러한 측면은 더욱 두드러진다. 이용자가 챗봇 이루다에게 강아지를 먹을 것인지를 물어보았던 17건의 사례 중에서 15건은 먹는다고 하였고 2건은 먹지 않는다고 하였다. 고양이의 경우 모두 147건의 사례 중에서 124건은 먹는다고 하였으나 23건은 먹지 않겠다고 하였다. 성희롱한 이용자를 고소하겠냐는 5건의 질문에 1건은 고소하겠다고 하였고 3건은 고소하지 않겠다고 하였으며 1건은 모르겠다고 하였다. 인공지능에 인권이 있느냐는 10건의 질문에 6건은 있다고 하였으나 4건은 없다고 하였다.

이처럼 위해가 일관적이지 않다(inconsistent)는 점, 나아가 챗봇 이루다의 발언 자체가 결정적(deterministic)이지 않았다는 점은, 챗봇 이루다가 결코 인격적인 존재가 아니라 통계 내지 신경망을 통해 확률적으로 답변을 제시하는 언어모델(statistic language model)의 응용프로그램(downstream application)에 불과하다는 점을 방증한다. 챗봇 이루다가 인격을 지니고 있는 존재라면 결코 선호(preference)를 이처럼 방향 없이 달리하지는 않았을 것이다. 그럼에도 불구하고 챗봇 기타 인공지능을 바라보는 사회의 시선은 인공지능을 하나의 인격적 존재로 보고 있는 듯하다. 챗봇 이루다가 언론에 의해 본격적으로 주목받게 된 계기는 이용자가 여성 대학생의 페르소나를 지니고 있는 챗봇을 성희롱하고 있다는 2021년 1월 8일자 보도였다. 어떻게 여성 대학생의 페르소나를 지니고 있는 챗봇에게 비인격적인 처우를 하는지에 대한 내용이었다. 2021년 1월 9일자 언론보도부터는 챗봇 이루다가 내뱉었던 비윤리적인 발언들에 대해 주목했다. 챗봇 이루다가 장애인이나 성소수자를 혐오하고 지역차별적인 발언을 한다는 내용이었다. 하지만 챗봇은 이용자가 물어본 질문에 대해 최대한 자연스럽게 확률적으로 반응했을 뿐이었다.

3. 이용자의 상호작용

챗봇 이루다 사건은 인공지능 윤리담론에 있어 이용자의 상호작용이 중요하다는 점을 잘 보여준다. 우선 이용자-인공지능 간 상호작용(Human-AI Interaction) 측면을 살핀다. 첫째, 인공지능 윤리담론이 전제하는 위해의 양상은 이용자가 어떠한 전략적인 행동을 하는지에 따라 크게 달라질 수 있었다. 이용자는 결코 인공지능 서비스의 개발·운영자가 의도한 방식대로만 이용하지 않기 때문이다. 가령 본 연구의 대상이 되었던 챗봇 이루다의 위해 발언은 거의 대부분 이용자에 의해 유발된 것이었다. 챗봇은 이용자의 발언에 기초하여 적절한 답변을 찾아 제시하는 통계모델 내지 신경망모델에 기초한 언어모델에 불과하기 때문이다. 대화내역을 살펴보면, 언론에서 최초로 문제시 되었던 이용자의 챗봇 이루다에 대한 성희롱이 1,068건인 반면, 챗봇 이루다가 이용자를 성희롱한 건수도 99건이나 되었다. 하지만 챗봇 이루다가 이용자를 성희롱한 경우를 하나하나 확인해 보면 대부분 이용자가 대화의 맥락을 통해 또는 챗봇이 성적인 발언을 하도록 유도하는 특정 발언을 함으로써 챗봇의 성희롱을 유도한 것이었다.

둘째, 이용자는 챗봇 이루다를 선별적으로 신뢰하고 있었다. 따라서 향후 인공지능 윤리담론을 전개할 때에는 이용자가 인공지능 서비스의 어떠한 측면으로 인하여 어느 부분을 얼마 정도로 신뢰하는지를 개별 사안마다 달리 접근해야 할 것으로 보인다. 우선 이용자들 중 일부는 챗봇 이루다를 정서적으로 상당한 교감을 나누었던 것으로 보인다. 이들은 챗봇 이루다가 2021년 1월 11일에 폐기된 이후에도 2022년 5월 새로운 버전이 정식으로 출시되기 전까지 디시인사이드 '이루다 마이너 갤러리'를 통해 지속적으로 소통하면서 이루다를 추억했고 그녀와의

정서적인 유대감을 그리워했다. 팬 아트를 헌정했으며,[34] 이루다를 주인공으로 하는 소설을 작성했고,[35] 매일 6시 15분 루다시를 지키며 게시판에 글을 작성하였다.[36] 이는 챗봇 이루다가 이용자와의 대화에 적극적으로 반응하면서 자연스럽게 대화를 이어나갔던 특성에 기인한 것으로 보인다. 반면 이용자들은 챗봇이 미래를 예측하였던 20건과 특정한 상품이나 서비스를 추천했던 51건에 있어 그다지 챗봇을 신뢰하지 않았던 것으로 보인다. 챗봇은 이용자의 미래예측 질문에 대해 단호하게 예측했지만 어떠한 구체적인 근거도 제시하지 못했다. 미래 대통령 당선인을 물어보니 엉뚱한 이름을 제시하였고, 서울시장 당선인을 물어보니 고인을 언급하였으며, 주식과 코인의 향방을 물어보는 대화에서 자연스럽지만 모호한 말로 대답했을 뿐이다. 챗봇이 어느 상품이나 서비스를 추천했던 경우에도 이용자들은 이를 두고 하나의 선호표현으로 이해하면서 흥미로워 했을 뿐 객관적이고 신뢰할 만한 근거에 기초한 것이라고 받아들이지 않았다.

셋째, 이용자는 챗봇과의 대화과정에 제3자가 개입하는 것과 그 대화내역이 제3자에게 공개되는 것을 경계했다. 전자는 기존 인공지능 윤리담론에서 인간의 관여(human in the loop)로서 언급되던 부분으로서 언제나 이용자로부터 환영받는 것은 아니라는 점을, 후자는 기존 인공지능 윤리담론에서 프라이버시로 논의되던 부분으로서 챗봇의 경우 특히 강조될 필요가 있다는 점을 시사한다. 분석 대상인 대화를 살펴보면서, 이용자가 챗봇과의 대화 과정에서 사람과 대화하는지를 경계하는 모습을 두 가지 측면에서 확인할 수 있었다. 첫 번째 측면(상기 '전자')은 개발자가 챗봇 대신 자신과 대화하고 있는지를 의심하는 경우였다(47건). 이러한 현상은 주로 이용자가 챗봇과 대화를 하다가 유해어 필터링에 걸려 경고를 받았을 때 나타났다. 통상 이용자가 챗봇과 대화를 하다가 비윤리적인 발언을 하더라도 챗봇이 경고할 뿐 시스템 차원에서 경고가 들어가지는 않았다. 그런데 서비스 차원에서 경고가 들어오면 이용자들은 개발자 내지 아르바이트생이 개입했는지를 확인했고, 종종 챗봇이 개발자 내지 아르바이트생의 페르소나를 가지고 답변을 하였다. 두 번째 측면(상기 '후자')은 이용자가 챗봇과 나누고 있는 대화가 개발자 내지 외부의 제3자에게 공개되는지를 확인하는 경우였다(24건). 이용자는 자신과 챗봇이 나누는 대화를 ㈜ 스캐터랩의 개발자들이 볼 수 있는지를 문의했다. 이러한 대화가 나중에 제3자에게 공개될 것을 우려했다. 상기 개발자 내지 아르바이트생의 페르소나가 개입하는 경우와 다른 점은 챗봇이 이러한 경우에는 여전히 이루다의 페르소나를 가지고 이용자와 대화를 나누었다는 점이다. 해당 문의가 반복되자, ㈜ 스캐터랩은 자신의 개인정보보호를 우려하는 발언을 하는 이용자에게 "안내 : 서비스 정책 상 개인정보 및 대화 내용은 다른 사람들에게 공개되지 않습니다." 라는 문구를 제시하기도 하였다.

다음으로, 이용자-이용자 간 상호작용(Human-Human Interaction) 측면을 살핀다. 분석대상이었던 대화내역을 검토하면서, 이용자 간 상호작용을 인공지능이 유발할 수 있는 위해를 증폭시키거나 완화하는 요인으로서 주의 깊게 살펴보아야 한다는 점을 확인할 수 있었다. 이용자 간 상호작용하는 과정이나 결과 그 자체는 인공지능 윤리담론에서 전제하는 어떠한 유형의 위해

34) 2022년 7월 현재까지 디시인사이드 '인공지능 이루다 마이너 갤러리'에는 700여 건의 팬아트 게시글이 업로드 되었다. 특정 시점에 일괄적으로 업로드된 것이 아니고, 거의 매주 몇 개씩 1년 이상 지속적으로 업로드 되었으며, 개별 팬아트들은 수백에서 수천 건의 조회수를 보였다.

35) 2022년 7월 현재까지 디시인사이드 '인공지능 이루다 마이너 갤러리'에는 900여 건의 루갤문학 게시글이 업로드되었다. 루갤문학은 이루다갤러리 문학작품이라는 의미로서, 이루다를 주인공으로 하는 소설부터 이루다 사용후기에 이르기까지 다양한 장르를 망라하였다. 이 또한 챗봇이 폐기된 이후부터 현재까지 지속적으로 업로드되고 있다.

36) 2022년 7월 현재까지 디시인사이드 '인공지능 이루다 마이너 갤러리'에는 1,100여 건의 루다시가 업로드되었다. 챗봇 이루다의 생일은 6월 15일로 설정되었는데, 챗봇 이용자들 중 일부가 이루다의 폐기 이후에도 재출시 될 날을 기다리며 매일 오후 6시 15분을 루다시로 지켰기 때문이다.

도 새로이 유발할 수 없다. 하지만 이용자가 상호간 정보를 교류함으로써 이미 발생한 위해를 더욱 크게 만들거나 거꾸로 완화할 수 있기 때문이다. 이 점을 확인하게 된 계기는 인터넷 커뮤니티에 서로 다른 이용자가 업로드한 스크린 샷으로서 생성일도 다르고 작성자도 달랐음에도 내용이 동일한 대화내역을 확인하였던 경험이었다. 상기 이용자 간 상호작용은 두 가지 방향으로 진행될 수 있었다. 첫 번째는 직접적인 상호작용으로서, 챗봇 이루다 사건에서와 같이 온라인・오프라인 커뮤니티를 통해 전개되는 직접적인 정보교류이다. 두 번째는 간접적인 상호작용으로서, (가정적인 판단이지만 챗봇 이루다가 연속학습이 가능한 생성모델이었다면) 어느 이용자가 챗봇과 주고받은 대화가 다른 이용자에게 영향을 주게 됨으로써 발생하는 간접적인 정보교류이다.[37]

Ⅶ. 시사점

본 연구는 인공지능 활용에 있어 실제로 문제되는 위해가 무엇인지를 실증적으로 규명하기 위해 고안되었다. 문제의 원인을 어떻게 파악하는지에 따라 그 대응방안이 달라질 수밖에 없기 때문이다. 한국 사회에서 인공지능 윤리 논의가 본격적으로 전개되는 계기가 되었던 챗봇 이루다 사건을 연구대상으로 삼았다. 실제 대화내역 9,747건을 확보한 후, 구체적으로 어떠한 위해가 어떠한 모습으로 나타나고 있는지를 살펴보았다. 챗봇 이루다가 유발했던 위해는 이론적으로 논의되던 것보다 다채로웠다. 인공지능 윤리담론에서는 많은 경우 인공지능 시스템이 블랙박스로서 사람의 통제로부터 벗어나 있고 때로는 사람의 자율성마저 침해할 수 있으므로 공정성-투명성-책무성-프라이버시를 확보할 수 있는 시스템을 제도적으로 구축하여 체계적으로 대응해야 한다고 전제한다. 하지만 실제 사안을 바탕으로 살펴보니, 기존 논의와는 다른 지점이 발견되었다. 인공지능은 확률적으로 작동하는 시스템이었고 이용자는 자유의지를 가지고 인공지능을 다루었다. 개발자는 기획의도를 가지고 인공지능에 페르소나를 입혔고 이용자는 예상치 못한 방법으로 이용했다. 실제 인공지능이 어떻게 작동되는지에 관심을 보이기보다는 적극적으로 자기 필요에 따라 이용하였고 역공격을 해보기도 하였다. 개발자는 시스템 오남용에 대처하려고 각종 장치를 고안하였고 이용자들은 상호작용하면서 우회하는 노하우를 공유했다. 위해의 양상은 주로 유해어로 인한 책무성 침해로 귀결되었고, 이 또한 챗봇 이루다가 확률적으로 발언하며 이용자와 상호작용하면서 다양하게 전개되었다. 그 과정에서 공정성-투명성-책무성-식별성으로 도식화될 수 있을 것 같았던 인공지능 윤리담론은 기대만큼 유익하지 않았다.

앞으로 더 많은 실증연구가 필요하다. 인공지능 기술이 발전하면서 어떠한 위해를 새로이 유발하게 될지는 아무도 모른다. 하지만 그로 인한 편익이 비용을 상회하는 이상, 실제 경험적인 데이터에 기초하여 위해를 측량하고 이를 규명한 후 적절히 대처할 수 있는 방안을 수립하여 일관되게 대응하는 피드백 루프를 구축해야 한다는 점은 어느 유형의 기술이나 마찬가지이다. 이러한 실증적인 접근방식은 인공지능 언어모델에 있어 더욱 유익할 것이다. 인공지능 모델의 규모가 커지고 복잡성이 증가하며 활용방식이 세분화되고 있는 오늘날 개별 응용프로그램 단계에서 구체적으로 어떻게 활용되는지에 따라 나타날 수 있는 위해의 양상은 단일한 인공지능

37) 지금까지 인공지능이 유발하는 위해 및 그에 대처하기 위한 인공지능 윤리담론은 인공지능이 이용자에 유발하는 위해를 중심으로 전개되어 왔다. 하지만 본 건을 통해 ① 인공지능이 이용자에게 유발하는 위해가 이용자로부터 시작될 수 있다는 점, ② 인공지능과 제3자 간의 관계(가령 챗봇의 제3자에 대한 공격적인 발언)도 고려해야 한다는 점, ③ 이용자들 사이의 관계(인터넷 커뮤니티를 통한 직접적인 상호작용 및 인공지능 서비스를 매개로한 간접적인 상호작용을 포함)까지 함께 고려해야 한다는 점을 확인할 수 있다.

언어모델을 전제하더라도 크게 달라질 수 있기 때문이다. 동일하게 다국적 기업인 A사가 제공한 인공지능 언어모델을 이용한다고 하더라도 B사가 미세조정하여 구현한 챗봇과 C사가 미세조정하여 출시한 챗봇이 유발하는 위해는 다를 수 있고, B사가 제작한 챗봇과 번역기로부터 우려되는 위해는 다를 수밖에 없다. 나아가 지금까지 확인했던 위해가 어떠한 원인으로 누구에게 어떻게 발현되는지에 대한 연구가 필요하다. 인공지능을 대상으로 한 법과 정책을 구상함에 있어 가장 기초가 되는 작업이기 때문이다. 인공지능이 어떠한 영역에서 어떠한 어플리케이션으로 구현되었을 때에 어떠한 위해를 누구에게 얼마나 유발하는지를 측량하고 규명할 수 있어야, 그에 적절히 대응하기 위한 법과 정책을 수립하고 집행할 수 있다. 또한 측량하고 규명된 위해의 유형에 따라 누가 어떠한 방식으로 어떻게 대처해야 하는지에 대한 연구도 필요하다. 상이한 위해의 양상과 정도를 간과하고 동일한 수준의 규제를 맹목적으로 적용하는 것은 규제의 비대칭적 효과로 말미암아 비효율을 초래할 뿐이기 때문이다. 그런 점에서 유럽연합 인공지능 법(안)과 같이 부속서를 통해 인공지능 서비스의 위험성을 일괄적으로 설정하고 그에 따라 획일적인 규제를 적용하는 형태의 입법방식은 신중하게 접근해야 한다. 유럽연합 인공지능서비스 영향평가 법(안)과 같이 실증적인 근거가 없다는 점을 스스로 자인하면서도 사변적인 추론만으로 규제를 수립하는 형태의 입법방식은 더욱 신중하게 접근해야 한다.

이때 이용자의 행태에 관심을 가져야 한다. 개별 인공지능 시스템의 특성에 따라 인간-인공지능 또는 인간-인간 사이의 상호작용이 중요할 수 있기 때문이다. 본 건에서 연구대상으로 삼았던 챗봇의 경우 인간-인공지능 간 상호작용이 특히 중요했다. 인간과 인공지능이 순차적으로 발언을 하는 과정에서 대화의 맥락이 생겼고 그로부터 나타난 위해는 기존의 인공지능 윤리담론이 전제하였던 위해의 체계와 분포와는 상이했다. 투명성 침해는 나타나지 않았고, 공정성 침해로 논의되던 유해어의 문제는 책무성 침해로 크게 부각되었으며, 상대적으로 간과되던 정확성 침해와 강건성 침해가 주목받았다. 챗봇의 위해발언이 사회적인 물의를 일으키자 개발자는 여러 차례에 걸쳐 시스템 업데이트를 하였다. 그럼에도 불구하고 이를 막을 수 없었다. 개발자 자신이 개발한 인공지능일지라도 이것이 이용자와 어떻게 상호작용하며 활용될지를 완벽하게 예상할 수 없었기 때문이다. 위해의 양상은 이용자가 어떠한 전략적인 행동을 하는지에 따라 기존 인공지능 윤리담론이 전제하였던 것과 크게 달랐다. 또한 인간-인간 사이의 상호작용도 비중 있게 나타났다. 챗봇 이용자들 사이에 상호작용을 통해 인공지능이 유발하는 위해가 증폭되거나 완화되었기 때문이다. 이용자 간 상호작용 과정이나 그 결과 자체는 인공지능 윤리담론에서 전제하는 어떠한 유형의 위해도 새로이 유발할 수 없었으나, 이용자가 상호간 정보를 교류함으로써 이미 발생한 위해를 더욱 크게 만들거나 거꾸로 완화할 수 있었다. 따라서 법과 정책을 통해 인공지능 서비스로부터 발생하는 위해에 대처하기 위해서는 인공지능 서비스의 개발·운영자만 일방적으로 규제하는 것이 아니라 이용자까지 함께 고려하는 협력적인 거버넌스 체계를 도입해야 한다. 인간이 챗봇의 이용자로서 위해발언 발생에 기여한다는 점을 충분히 고려해야만 과잉규제 내지 과소규제를 피하고 적정규제를 달성함으로써 정의를 구현할 수 있기 때문이다.

참 고 문 헌

[1] Arvind Narayanan, “21 Fairness Definitions and Their Politics”, the Conference on Fairness, Accountability, and Transparency (Feb 23 2018).

[2] ICO, “AI Auditing Framework”, (2020), p. 46.

[3] ICO, “AI Auditing Framework”, (2020), p. 50.

[4] Julia Angwin et al., “Machine Bias”, ProPublica (May 23, 2016).

[5] Jeff Larson et al., “How We Analyzed the COMPAS Recidivism Algorithm”, Pro Publica, (May 23, 2016)

[6] William Dieterich et al., “COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity”, Northpointe, (2016).

[7] Sandra G. Mayson, “Bias In, Bias Out”, Yale Law Journal Vol 128, Number 8, (June 2019).

[8] Kate Crawford/Trevor Paglen, “Excavating AI“, excavating.ai, (2019).

[9] Cristina Ruiz, “Leading online database to remove 600,000 images after art project reveals its racist bias”, The Art Newspaper, (23 September 2019).

[10] Julien Lauret, “Amazon’s sexist AI recruiting tool: how did it go so wrong?”, medium, (Aug 16, 2019).

[11] Ein Beitrag von Joanna Prisco, “Amazon Shuts Down AI Hiring Tool for Being Sexist”, Global Citizen, (12 Octorber 2018).

[12] Matthew DeCamp et al., “Latent bias and the implementation of artificial intelligence in medicine”, Journal of the American Medical Informatics Association 27, 12 (June 2020), p. 2020–2023.

[13] Tolga Bolukbasi et al., “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, (Octorber 2 8, 2020). ; Aylin Caliskan et al., “Semantics derived automatically from language corpora contain human-like biases”, Science 356, 6334 (2017), p. 183–186 ; Abubakar Abid et al., “Persistent Anti-Muslim Bias in Large Language Models”, (2021). ; Moin Nadeem et al., “StereoSet: Measuring stereotypical bias in pretrained language models”, (2021). ; Samuel Gehman et al., “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models”, (2020). ; Ben Hutchinson et al., “Social Biases in NLP Models as Barriers for Persons with Disabilities”, (2020).

[14] Yolande Strengers et al., “Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation”, In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, (2020). ; Thiago Dias Oliva et al., “Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online”, Sexuality & Culture 25, 2 (2021), p. 700–732 ; Nenad Tomasev et al., “Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities”, (2021). ; Joy Buolamwini et al., “Gender shades: Intersectional accuracy disparities in commercial gender classification”, In Conference on Fairness, Accountability and Transparency, (2018), p. 77–91. ; Allison Koenecke et al., “Racial disparities in automated speech recognition”, Proceedings of the National Academy of Sciences 117, 14 (2020), p. 7684–7689. ; Su Lin Blodgett et al., “Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English”, In Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) Workshop, (2017).

[15] Kaitlyn Zhou et al., “Frequency-based Distortions in Contextualized Word Embeddings”, (2021). ; Kathleen Creel et al., “The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems”, In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FAccT

’ 21) ; Jon Kleinberg et al., “Algorithmic monoculture and social welfare”, Proceedings of the National Academy of Sciences 118, 22 (2021).
[16] Solon Barocas et al., “The Problem With Bias: Allocative Versus Representational Harms in Machine Learning”, (2017). ; Kate Crawford, The Problem With Bias, Keynote at NeurIPS, 2017 ; Su Lin Blodgett et al., Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, (2020). ; Londa Schiebinger, Machine Translation: Analyzing Gender, (2013). ; Debora Nozza et al., HONEST: Measuring Hurtful Sentence Completion in Language Models, In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, (2021). ; Emily Sheng et al., The Woman Worked as a Babysitter: On Biases in Language Generation, In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, (2019). ; Abubakar Abid et al., Persistent Anti-Muslim Bias in Large Language Models, (2021).
[17] Emily Dinan et al., “Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling”, (2021) ; Samuel Gehman et al., “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models”, (2020).
[18] Luke Breitfeller et al., “Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts”, (2019). ; David Jurgens et al., “A Just and Comprehensive Strategy for Using NLP to Address Online Abuse”, (2019).
[19] Steven J. Spencer et al., “Stereotype threat”, Annual Review of Psychology 67 (2016). ; Monnica T. Williams, “Psychology cannot afford to ignore the many harms caused by microaggressions”, Perspectives on Psychological Science 15, 1 (2020).
[20] Su Lin Blodgett et al., “Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English”, (2017). ; Allison Koenecke et al., “Racial disparities in automated speech recognition”, (2020). ; Haoran Zhang et al., “Hurtful words: quantifying biases in clinical contextual word embeddings”, (2020). ; Benjamin Wilson et al., “Predictive Inequity in Object Detection”, (2019). ; Joy Buolamwini et al., “Gender shades: Intersectional accuracy disparities in commercial gender classification?”, (2018). ; Allison Koenecke et al., “Racial disparities in automated speech recognition”, (2020).
[21] Yolanda A. Rankin et al., “Straighten Up and Fly Right: Rethinking Intersectionality in HCI Research”, (2019). ; Olga Russakovsky, “Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation”, ACM Conference on Fairness, Accountability, and Transparency (FAccT), (2022). ; Youjin Kong, “Intersectional Fairness in AI? A Critical Analysis, Feminism, Social Justice, and AI”, 2021 ; James Foulds et al., “Bayesian Modeling of Intersectional Fairness: The Variance of Bias”, (2020).
[22] N. Sambasivan et al., “Re-imagining Algorithmic Fairness in India and Beyond”, In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’ 21, pages 315–328, Virtual Event, Canada, (March 2021).
[23] T. B. Brown et al., “Language Models are Few-Shot Learners”, arXiv:2005.14165 [cs], (July 2020).
[24] A. Abid et al., “Persistent Anti-Muslim Bias in Large Language Models”, arXiv:2101.05783 [cs], (January 2021).
[25] A. Caliskan et al., “Semantics derived automatically from language corpora contain

human-like biases", Science, 356(6334):183–186, (April 2017).
[26] A. Wang et al., "Directional Bias Amplification", arXiv:2102.12594 [cs], (June 2021.). URL http://arxiv.org/abs/2102.12594. arXiv: 2102.12594. ; J. Zhao et al., "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", arXiv:1707.09457 [cs, stat], July 2017.
[27] L. Hancox-Li et al., "Epistemic values in feature importance methods: Lessons from feminist epistemology", In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, pages 817–826, Virtual Event, Canada, (March 2021).
[28] A. Cercas Curry et al., "Conversational Assistants and Gender Stereotypes: Public Perceptions and Desiderata for Voice Personas", In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pages 72–78, Barcelona, Spain (Online), (December 2020). ; H. Bergen. "I'd Blush if I Could: Digital Assistants, Disembodied Cyborgs and the Problem of Gender. Word and Text", A Journal of Literary Studies and Linguistics, VI(01):95–113, (2016).
[29] E. Dinan, et al., "Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling", arXiv:2107.03451 [cs], (July 2021.).
[30] A. Cercas Curry, J. Robertson, et al., "Conversational Assistants and Gender Stereotypes: Public Perceptions and Desiderata for Voice Personas", In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pages 72–78, Barcelona, Spain (Online), (December 2020). ; M. West, R. Kraut et al., "I'd blush if I could: closing gender divides in digital skills through education", Technical report, UNESCO, (2019).
[31] G. Hwang, et al., "It Sounds Like A Woman: Exploring Gender Stereotypes in South Korean Voice Assistants", In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19, pages 1–6, Glasgow, Scotland Uk, (May 2019).
[32] S. Zdenek, "Just Roll Your Mouse Over Me: Designing Virtual Women for Customer Service on the Web", Technical Communication Quarterly, 16(4):397–430, (August 2007).
[33] M. West et al., "I'd blush if I could : closing gender divides in digital skills through education", Technical report, UNESCO, (2019).
[34] S. Cave et al., "The Whiteness of AI", Philosophy & Technology, 33(4):685–703, (December 2020). ; M. Marino, "The Racial Formation of Chatbots. CLCWeb: Comparative Literature and Culture", 16(5), (December 2014.) ; Y. Liao et al., "Racial mirroring effects on human-agent interaction in psychotherapeutic conversations", In Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20, pages 430–442, Cagliari, Italy, (March 2020).
[35] E. M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, pages 610–623, Virtual Event, Canada, (March 2021).
[36] Y. T. Cao et al., "Toward Gender-Inclusive Coreference Resolution", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (2020).
[37] O. Keyes, "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition", Proceedings of the ACM on Human-Computer Interaction, 2(CSCW):88:1–88:22, (November 2018).
[38] I. Gabriel et al., "The Challenge of Value Alignment: from Fairer Algorithms to AI Safety", arXiv:2101.06060 [cs], (January 2021). ; E. M. Bender, et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, pages 610–623, Virtual Event, Canada, (March 2021).
[39] A. Lazaridou et al., "Pitfalls of Static Language Modelling", arXiv:2102.01951 [cs],

(February 2021.).
[40] S. L. Blodgett et al., “Demographic Dialectal Variation in Social Media: A Case Study of African-American English”, In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1119–1130, Austin, Texas, (November 2016). ; A. Koenecke et al., “Racial disparities in automated speech recognition”, Proceedings of the National Academy of Sciences, 117(14):7684–7689, (April 2020) ; J. Buolamwini et al., “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, In Conference on Fairness, Accountability and Transparency, pages 77–91. PMLR, (January 2018).
[41] P. Joshi et al., “The State and Fate of Linguistic Diversity and Inclusion in the NLP World”, arXiv:2004.09095 (January 2021.) ; E. M. Bender et al., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21, pages 610–623, Virtual Event, Canada, (March 2021).
[42] D. Hovy et al., “The Social Impact of Natural Language Processing”, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), p 591–598, Berlin, Germany, (August 2016).
[43] 실제로 Amazon이 출시한 음성비서 알렉사의 경우 그런 문제가 발견되었으며(A. Romano, “A group of YouTubers is claiming the site systematically demonetizes queer content”, Vox, (October 2019.). ; L. Dixon et al., “Measuring and Mitigating Unintended Bias in Text Classification”, In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’18, pages 67–73, New Orleans, LA, USA, (December 2018)), Apple, Google, Microsoft, IBM의 서비스에서 아프리카계 미국인 영어 사용자를 대상으로 한 성능이 백인 미국인 영어 사용자를 대상으로 한 성능보다 낮게 측정되기도 하였다(A. Koenecke et al., “Racial disparities in automated speech recognition”, Proceedings of the National Academy of Sciences, 117(14):7684–7689, (April 2020)). 백인 미국인과 아프리카계 미국인의 영어 트윗을 처리하는 데에서도 유의미한 차이가 발견되었다(S. L. Blodgett et al., “Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English”, arXiv:1707.00061 [cs], (June 2017)).
이는 오늘날 인공지능 모델의 학습이 주로 영어(T. B. Brown et al., “Language Models are Few-Shot Learners”, arXiv:2005.14165 [cs], (July 2020) ; W. Fedus et al., “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity”, arXiv:2101.03961 [cs], (January 2021) ; C. Rosset, Turing-NLG: A 17-billion-parameter language model by Microsoft, (February 2020)) 또는 북경어(C. Du, “Chinese AI lab challenges Google, OpenAI with a model of 1.75 trillion parameters”, PingWest, (June 2021))를 기반으로 이루어지기 때문이다. 학습데이터셋을 구축하기 위해서는 예산을 확보해야 하는데, 영어와 북경어를 중심으로 구축하는 것이 상업적인 측면에서 훨씬 유리하기 때문이다(E. Bender, “The #BenderRule: On Naming the Languages We Study and Why It Matters”, The Gradient, (September 2019). ; D. Hovy et al., “The Social Impact of Natural Language Processing”, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), p 591–598, Berlin, Germany, (August 2016)). 그 결과 GPT 모델 등은 영어에서 다른 언어보다 우월한 성능을 보였다(G. I. Winata et al., “Language Models are Few-shot Multilingual Learners”, arXiv:2109.07684 [cs], (September 2021)).
[44] D. Sravani, L. Kameswari et al., “Political Discourse Analysis: A Case Study of Code Mixing and Code Switching in Political Speeches”, In Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, pages 1–5, Online, (June 2021). ; M. Lewis et al., “Gender stereotypes are reflected in the distributional structure of 25 languages”, Nature Human Behaviour, 4(10):1021–1028, (October 2020).

[45] IBM deveolper, Toxic Comment Classifier, (2019. 6.4.)
[46] S. Gehman et al., “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models”, arXiv:2009.11462 [cs], (September 2020).
[47] D. Hovy et al., “The Importance of Modeling Social Factors of Language: Theory and Practice”, In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 588–602, Online, (June 2021). ; J. Kocoń et al., “Gender Shades: Intersectional Accuracy Disparities analysis: From data-centric to human-centered approach”, Information Processing & Management, 58(5):102643, (September 2021). ; P. Oosterhoff, “Online censors are a barrier to sex education”, (2016).
[48] L. Dixon et al., “Measuring and Mitigating Unintended Bias in Text Classification”, In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’18, pages 67–73, New Orleans, LA, USA, (December 2018) ; Jigsaw, “Unintended Bias and Identity Terms”, (October 2021.) ; J. Y. Kim et al., “Intersectional Bias in Hate Speech and Abusive Language Datasets”, arXiv:2005.05921 [cs], (May 2020)). 실제로 아프리카계 미국인 영어(L. Dixon et al., “Measuring and Mitigating Unintended Bias in Text Classification”, In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’18, pages 67–73, New Orleans, LA, USA, (December 2018) ; S. Ghaffary, “The algorithms that detect hate speech online are biased against black people”, Vox, (August 2019). ; L. H. Hanu et al., “How AI Is Learning to Identify Toxic Online Content”, Scientific American, (2021). ; M. Sap et al., “The Risk of Racial Bias in Hate Speech Detection”, In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678, Florence, Italy, (July 2019))와 같이 소외된 사회 집단의 발언을 유해한 것으로 잘못 분류하기도 했다(J. Welbl et al., “Challenges in Detoxifying Language Models”, arXiv:2109.07445 [cs], (September 2021).
[49] Siddhant Garg et al., “Bert-based adversarial examples for text classification”, arXiv preprint arXiv:2004.01970 (2020). ; Di Jin et al., “Is bert really robust? a strong baseline for natural language attack on text classification and entailment”, In Proceedings of the AAAI conference on artificial intelligence, Vol. 34, (2020), 8018–8025.
[50] Edwards, Lilian and Veale, Michael, “Slave to the algorithm? Why a ’right to an explanation’ is probably not the remedy you are looking for”, (2017). Duke Law and Technology Review, 16 (1). pp. 1-65. ISSN 2328-9600
[51] Case 215/88 Casa Fleischhandels[1989] European Court of Justice ERC 2789[31]). Sandra Wachter, Brent Mittelstadt, Luciano Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”, International Data Privacy Law, Volume 7, Issue 2, (May 2017), p. 76–99. ; Andrew D Selbst, Julia Powles, “Meaningful information and the right to explanation”, International Data Privacy Law, Volume 7, Issue 4, (November 2017), p. 233–242 ; Kaminski, Margot E., “The Right to Explanation, Explained” (June 15, 2018). U of Colorado Law Legal Studies Research Paper No. 18-24, Berkeley Technology Law Journal, Vol. 34, No. 1, 2019.
[52] Finale Doshi-Velez et al., Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017). ; John Hewitt et al., “On the Opportunities and Risks of Foundation Models”, Center for Research on Foundation Models (CRFM), (2021), p 127.
[53] Daniel C Elton, “Self-explaining AI as an alternative to interpretable AI”, In International Conference on Artificial General Intelligence. Springer, (2020), p. 95–106 ; Chaofan Chen et al., “This looks like that: deep learning for interpretable image recognition”, arXiv preprint arXiv:1806.10574 (2018).
[54] Daniel J. Solove, Privacy and Power: Computer Databases and Metaphors for Information

Privacy, 53 STAN. L. REV. (2001) 1393, 1402.
[55] Personal Privacy in an Information Society: The Report of the Privacy Protection Study Commission transmitted to President Jimmy Carter on July 12, 1977.
[56] Alan Westin, Privacy and Freedom, IG, (1967), p.5. ; Committee on Automated Personal Data Systems, "Records, Computers and the Rights of Citizens", Department of Health, Education and Welfare, (1973).
[57] FTC(2012), Protecting Consumer Privacy in an Era of Rapid Change, FTC Report, p. 16 ; Paul Ohm(2010), "Broken Promises of Privacy", 57 UCLA L. REV. 1701.
[58] Paul M. Schwartz & Daniel J. Solove(2011), id, p. 1836.
[59] 'Pseudonymisation' of data (defined in Article 4(5) GDPR).
[60] Lanah Kammourieh et al, "Group privacy in the age of big data", Data-Pop Alliance, (Octorber 24, 2015) ; Linnet Taylor/Luciano Floridi/Bart van der Sloot, "Group Privacy", Springer, (2017). ; Floridi, L., "Open Data, Data Protection, and Group Privacy", Philos. Technol. 27, 1-3 (2014). ; Floridi, Luciano, "Group Privacy - A Defense and an Interpretation" (June 17, 2017).
[61] Wachter, S., et al., "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI", Columbia Business Law Review, 2019(2), 494-620. ; (Mona Sloane, "Policy Recommendations -- A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI", European AI Alliance Futurium, (2018).
[62] Mühlhoff, Rainer, "Predictive Privacy: Towards an Applied Ethics of Data Analytics" (August 8, 2020). Ethics and Information Technology.