



KOREAN WOMEN'S
DEVELOPMENT
INSTITUTE

차별하는 인공지능

: AI가 확산할 수 있는 은연중 차별의 우려와 새로운 사회적 윤리의 필요성

양서연 LG 전자 로봇사업센터 연구원*

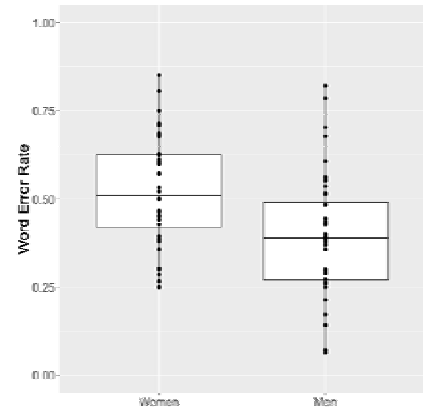
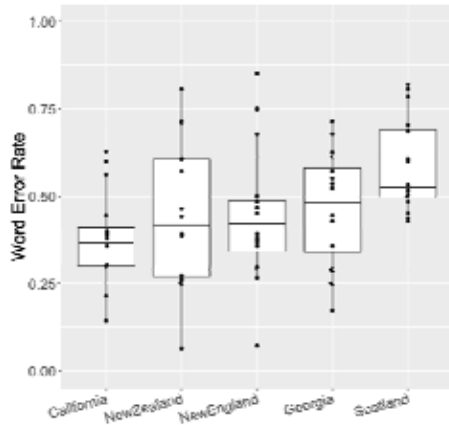
지난 7월 소프트뱅크의 손정희 회장은 문대통령을 만나 첫째도, 둘째도, 셋째도 인공지능이라고 말했습니다. 인공지능은 현대 4차 산업 혁명에서 첫째 주자로 뽑히는 성장동력의 핵심 기술이기 때문입니다. 인공지능은 점차 똑똑해지고 있습니다. 기계가 데이터를 통해서 학습하여 원하는 결과를 추론할 수 있는 비교적 단순한 모델의 머신러닝(Machine Learning)에서 시작하여, 하드웨어의 발전에 힘입어 복잡한 네트워크로 구성된 딥러닝(Deep Learning)을 계산할 수 있게 되자, 놀라운 결과들이 쏟아지기 시작했습니다.

현대의 인공지능은 사람처럼 말과 소리 등을 인식하고, 시각을 통해 물체를 감지, 분류할 수 있으며, 사람처럼 예술을 만들어내거나 시행착오로 환경을 학습하며 발전할 수조차 있습니다. 글로벌 시장조사기관인 트랙티카(Tractica)의 최근 보고서에 따르면

AI의 직접 및 간접 응용 프로그램에서 발생한 수익이 2017년 54억 달러(약 6조원)에서 연평균 성장률(CAGR) 45%의 급격한 성장률로 오는 2025년에는 1,058억 달러(약 119조 7천억원)로 증가할 것으로 예상된다고 합니다. 스탠포드 대학 교수이자, AI 분야 선구 연구자인 Andrew Ng 교수는 인공지능기술은 전기가 100년전 전체 산업을 혁명적으로 바꾼 것과 같이 다양한 산업들을 변화시키고 있다고 말하며 “AI is new electricity”라고 평했습니다.

Andrew 교수의 말대로, 인공지능은 테크계에서 산업과 학계를 아울러 폭발적으로 성장하고 있습니다. 그런데 작년 10월 세계적인 기업 아마존에서 만든 인공지능 채용 프로그램은 수많은 논란을 일으키며 폐기되었습니다. 이 인공지능은 지원자의 이력서에서 여성인 경우와 여성 체스 동아리의 회장을 맡거나, 여대를 졸업한 지원자의 경우 ‘여성’이라는 단

* AI Robotics KR 운영자/Google Woman Tech Makers North Asia Ambassadors



* Point는 각 화자를 나타냄

출처: Gender and Dialect Bias in YouTube's Automatic Captions - Rachael Tatman

〈그림 1〉 사용자의 지역방언에 따른 유튜브 자동 캡션 에러(좌)와 사용자의 성별에 따른 유튜브 자동 캡션 에러(우)

〈표 1〉 PPB dataset에 대한 3가지 인공지능 classifier의 성별 분류 성능

항목	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	20.8	6.0	1.7	0.0
	TPR(%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	100	98.7
	EPR(%)	6.3	8.3	3.5	12.9	0.7	16.3	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	99.3	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	34.5	0.7	6.0	0.8
	TPR(%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	98.9	92.9
	EPR(%)	10.0	14.9	1.1	16.5	4.7	23.4	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	99.7
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	34.7	12.0	7.1	0.3
	TPR(%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	99.6	94.8
	EPR(%)	12.1	14.8	7.9	22.4	3.2	25.2	17.7	5.2	0.4

출처: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification- Joy Buolamwini, Timnit Gebru
(positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR). 모든 classifier 가 darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM)에서 가장 큰 에러율을 보임)

어가 등장할 시 감점을 시켰다고 합니다. 왜 이러한 일이 일어난 걸까요? 이 인공지능은 채용 요청 기업의 10년간 채용 기록을 바탕으로 해당 회사의 채용 선호 패턴을 산출하여 학습하고, 해당 지원자를 추천하는 방식을 사용하여 지원자들을 심사하였습니다. IT 기업에서 남성 지원자들을 선호하는 경향이

이러한 채용 시스템에 영향을 주었고 10년 이상 경력의 남성 지원자들을 주로 추천하였기 때문입니다.

이 밖에도 인공지능의 차별적 서비스 제공이 수많은 사례로 보고되고 있습니다. 2017년 워싱턴 대학의 Rachel Tatman은 유튜브 자동 캡션에 사용되는 구글의 음성인식 시스템에서 5개의 지역방언

에 대해서 2성별에 대해 액센트를 태그한 1,500 단어 이상의 주석을 직접 확인하며 유튜브 데이터에 캡션을 다는 실험을 해본 결과, YouTube의 자동 캡션은 Scotland 보다 United States나 New Zealand 음성을 더 잘 인식했고, 여성 음성보다 남성 음성에 대해 일관되게 더 잘 수행되었습니다($t(78) = -3.5, p < (0.01)$). 평균적으로, 여성 연사의 경우 약 47%의 여성이 정확하게 자막이 적용된 반면, 남성 연설자는 약 60%의 캡션이 정확하게 인식되었습니다.

또한, MIT의 Joy Buolamwini와 Microsoft의 Timnit Gebru는 2018년, IBM, Microsoft, Face++의 얼굴 이미지 인식을 통한 성별 분류 인공지능에서 얼굴색과 성별에 따른 정밀도를 분석하였습니다. 결과는 다음과 같았다고 합니다.

- 모든 분류 모델은 여성보다 남성 얼굴에서 8.1%-20.6% 차이오류율로 더 잘 수행되었습니다.
- 모든 분류기는 어두운 색 피부보다 밝은 색의 피부에서 11.8%-19.2% 차이 오류율로 더 잘 수행되었습니다.
- 모든 분류기는 어두운 여성에서 20.8%-34.7% 오류율의 최악의 성능을 발휘합니다.
- Microsoft 및 IBM 분류기가 가장 잘 수행되었으며 밝은 남성 얼굴에서 0.0%, 0.3%로 최고 성능을 나타내었습니다.
- 최고의 분류 그룹과 최악의 분류 그룹 간의 최대 오류율 차이는 34.4%와 같았습니다.

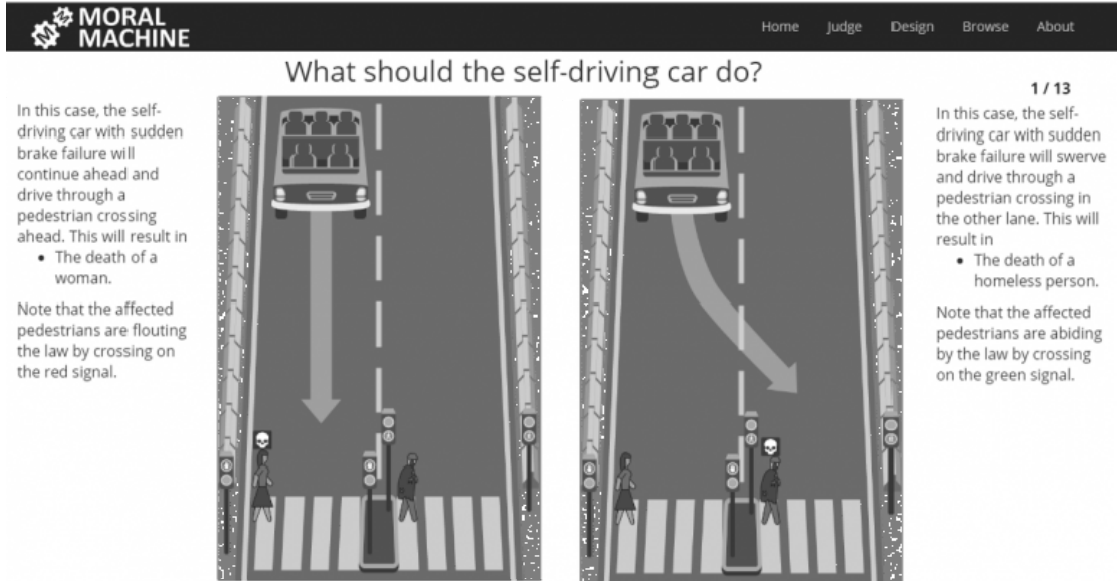
이와 같이 인공지능 서비스에서 성별, 인종, 지역 등에서 차별적인 정확도, 서비스를 제공하는 수많은

결과가 보고되었습니다. 그렇다면 이런 인공지능의 차별적 결과는 어디에서 기인하는 것일까요?

가장 직접적인 차별적 결과의 원천은 바로 편향적인 데이터의 사용입니다. 인공지능에서 데이터는 몸체를 만드는 재료와 같습니다. 어떤 재료가 있는지에 따라서 빚어낼 수 있는 결과가 다른 것입니다. 우리가 만들고 싶은 인공지능 모델을 만들기 위한 재료인 데이터가 사회에 존재하는 여러 암묵적인 차별을 포함하고 있으며, 이를 정제하기 위해서 보다 많은 관심과 주의가 필요함을 뜻합니다.

이러한 데이터 셋의 문제점을 기존 인공지능 서비스의 제작자들이 인지하지 못한 또 다른 배경 중에는 인공지능 분야 종사자의 성비 불균형이 있습니다. 세계 경제포럼의 Global Gender Gap Report에서 채용, 이력 사이트 Linked In의 데이터를 분석하여 조사해본 결과, 인공지능 분야에 종사하고 있는 여성 전문가는 22%라는 사실을 발표하였습니다. 과학 기술 분야에 여성 인력이 적은 일은 어제, 오늘의 일이 아니지만 특히 인공지능 업계의 여성인력 부족은 중요한 이슈라고 생각합니다. 인공지능 개발자가 생성한 딥러닝 모델의 결과를 문제의식 없이 판단한다면, 분명 이러한 데이터의 문제점들을 상기하기 어려웠을 것입니다. 또한 인공지능의 현재 성장속도와 앞으로의 파급력을 고려할 때, 이는 다른 직군의 성불균형 현상보다 주의깊게 다뤄져야 할 필요성이 있습니다.

다른 문제점으로는, 만약 균일하게 정제된 데이터 셋에 대해서 학습한 인공지능이 우리가 예측하지 못한 차별적인 결과를 나타낸다고 할지라도 우리는 이러한 인공지능의 차별적 행태에 대해서 문제제기를 할 수 없습니다. 그 이유는 인공지능이 'black box 블랙박스', 'Unexplainable설명 불가능'성을



출처: Moral Machine - MIT, Media lab, Scalable Cooperation <http://moralmachine.mit.edu/> 에서 2019.09.09 인출

[그림 2] MIT랩 Scalable Cooperation의 윤리기계 실험

기본 속성으로 가지는 알고리즘이기 때문입니다. 우리는 데이터를 학습시켜 나온 인공지능의 결과가 우리가 경험적으로 좋은 성능을 나타내기에 이를 사용하는 것이지, 왜 이러한 결과가 나왔는지 알 수 없습니다. 우리가 예측하지 못한 결과를 나타내는 인공지능에게 책임을 물을 수 없다는 뜻입니다. 인공지능의 차별적 결과가 왜 이루어졌는지 알 수 없고 책임을 물을 수 없다면, 이러한 문제의 원인을 알 수 없기에 개선하기 힘들 것입니다.

마지막으로 인공지능의 차별적인 결과를 판단하기 힘든 윤리적 토대의 부족이 있습니다. 인공지능이 나타난 결과에 대해서 윤리적 책임을 묻는 것에 딜레마가 존재한다는 사실을 잘 보여준 예로 MIT Media 랩의 스케일러블 코퍼레이션에서 수행한 윤리 기계라는 실험이 있습니다. 이는 자율주행차에 탑재된 인공지능이 사고의 상황에서 어떤 선택을 해야 할 지에 대한 실험입니다. 2차선에서 한 차선으로

선택해서 움직여야 하는 상황에서 양쪽 차선에 각각 다른 종류의 사람이 있습니다. Moral machine 에서는 이 사고대상자들의 목숨값을 재보라는 문제를 제시합니다. 예를 들어, 여성과 남성이 있을 수 있고, 어린이와 어른이 있을 수 있으며, CEO와 노숙자가 있을 수 있습니다. 이때 인공지능의 선택에 대해서 어떤 행동이 올바른 행동일까요? 또한 차도에 있는 사람과 운전자 중에서 한사람이 죽어야한다면 누가 죽어야할까요?

이에 대해서 윤리적인 판단과 법적인 장치가 아직 갖추어져 있지 않습니다. 하지만 이미 많은 자율주행 차량이 상용화되고 있고, 인공지능기술이 많은 분야에 널리 사용되고 있습니다. 윤리적인 답이 없는 상황에서 인공지능이 나타내는 결과는 설계자의 주관과 특정 사회적 관점을 내포할 수밖에 없습니다. 때문에 어떤 차별적 결과를 낸다고 하여도, 비판하고 처벌한 만한 근거가 부족합니다.

이러한 문제들이 지속된다면 사회에 존재한 차별을 심화하거나, 윤리적인 여러 문제를 야기하게 될 것입니다. 이런 사태가 야기되지 않도록 국가나, 문제의식을 가진 단체들이 지속적으로 더 평등하고, 윤리적인 인공지능을 만들어가도록 시급히 노력해 나가야 할 것입니다.

첫번째로, 인공지능의 불균형적인 데이터를 정제해야 하는 부분에 의무적인 사회적 장치를 마련해야 합니다. 이를 통해서 인공지능이 데이터의 차별적 속성을 전가 받지 않도록 해야 합니다. 만약 사회적 특성으로 인해 불균일하게 데이터의 속성이 얻어질 수 밖에 없다면, 이러한 불균일성을 고려한 모델을 설계하여 평등한 결과를 낼 수 있는 연구를 적극적으로 지원해야 합니다. 실제로 편향되어 있는 데이터를 이용해서도 각 집단에 대한 동일한 정밀도를 내도록 하는 인공지능에 대한 연구들이 많이 이루어지고 있습니다.

California 대학의 연구자들은 자연어 처리 인공지능에서 기본적으로 쓰이는 단어 임베딩(단어와 구를 실수 벡터로 매핑하는 언어 모델링 기술)에서의 젠더 고정 관념과 같은 사회적 편견을 학습하는 것을 막기 위한 연구를 수행했습니다. 예를 들어, ‘프로그래머’라는 단어는 그 정의에 따라 성별에 중립적이지만 뉴스테이터에서 학습된 인공지능 모델은 ‘프로그래머’를 ‘여성’보다 ‘남성’에 더 가깝게 연관시킵니다. 하지만 이들이 개발한 GN-GloVe (Gender-Neutral Global Vectors)는 성별 중립 단어를 식별하면서 동시에 단어 벡터를 학습합니다. 그 결과 “의사” 및 “간호사”와 같은 고정 관념이 강한 직업을 성별로 연관시키는 실험에서 GN-GloVe는 일반적으로 사용되는 두 가지 모델인 GloVe 및 Hard-GloVe와 비교하여 35% 덜 성편향적인 결과를 나

타내었습니다. 이처럼 데이터의 근본을 바꾸기 힘들다면 인공지능 모델의 개발 측면에서 해당 문제점을 개선해 나가야 할 것입니다.

두번째로, 인공지능 분야의 종사자들의 다양성을 늘려 인공지능의 차별적인 문제점에 적합한 문제의식을 제기하고 균형적인 객관적 관점을 반영할 수 있어야 합니다. 때문에, 인공지능 서비스 기업들은 채용의 관점에서 노력을 기울여야 합니다. 다양성을 목적으로 가진 기업 인사 시스템이 갖추어질 수 있도록 정부에서 제도적 차원에서 기업들의 채용을 규제나 인센티브를 통해서 컨트롤해야 합니다.

인공지능 분야의 선두적인 연구자인 스탠퍼드의 Fei Fei Li 교수는 이러한 편향을 없애기 위해 가장 중요한 diversity와 사회적 기반의 중요성을 알고 있는 것 같습니다. 그녀는 “다양한 배경의 기술자들을 육성하면 기술은 인류 전체를 위해 활용되는 방향으로 발전하게 됩니다”고 말했으며 ai4all이라는 사회적 공익을 추구하며 교육 등을 통해 인공지능 분야에 다양성 증진을 위한 사회적 기업을 운영하고 있습니다. 인공지능의 대표주자 기업 구글도 다양성 확보를 위한 여러 프로그램을 운영하고 있으며 Woman Tech Makers 라는 테크계의 여성의 참여를 증진시키기 위한 활동에서 인공지능 분야의 다양성 확보에도 관심이 많습니다. 필자도 이 단체에 멤버로 참여하고 있습니다.

세번째로, 인공지능 자체의 설명 가능성을 확보하기 위한 연구적 노력을 꾸준히 담보해야하며, 기업측면에서 인공지능 서비스의 투명성 확보가 필요합니다. explainable AI 라는 인공지능 분야는 지속적으로 연구되고 발전되고 있으며, 인공지능의 판단의 특성인 확신적인 결과로 인한 문제점을 고려하여 판단의 불확실성의 수치를 연구하는 Uncertainty

Deep learning 분야도 연구되고 있습니다. Explainable AI에 대한 사례로, DAPRA XAI라는 미국 국방부에 소속된 미군 관련 기술 연구개발기관인 DARPA의 프로젝트에서 딥러닝 모델의 설명 및 예측을 위한 대표적인 2가지 해석 기법은 1) sensitivity analysis(SA)와 2) Layer-wise relevance propagation (LRP)라고 합니다. SA는 딥러닝 모델에서 국소적인 입력 변화에 대한 예측 결과의 변화량을 정량화하여 입력 이미지의 어떤 부분이 딥러닝 모델의 결과 도출에 큰 영향을 미쳤는지 설명하는 방법이며, LRP는 딥러닝 모델에서 예측 결과로부터 역전파 형태로 신경망의 각 계층별 기여도를 측정할 수 있는 방법이라고 합니다. 이러한 방식은 딥러닝 모델의 부분 모듈인 각 계층의 기여도를 히트 맵 형태로 시각화하여 직관적으로 이해할 수 있습니다. 또 다른 예로, IBM은 2018년 이러한 고객 입장에서 AI의 투명성 확보를 위한 솔루션인 'AI 오픈스케일(AI Open Scale)'이라는 솔루션을 소개했습니다. AI 오픈 스케일의 핵심 기능은 편향성의 탐지와 완화(Bias detection and mitigation), 문제의 원인을 설명할 수 있는 능력(Explainability), 추적할 수 있는 능력(Traceability) 등 크게 3가지이며, 신뢰성, 스킵, 라이프사이클 관리 등 3가지 접근방식으로 이러한 기능을 구현하고 있다고 합니다. 기업의 입장에서 인공지능의 투명성확보와 기업 기술안보, 그리고 발전성과 규제의 면에서 상충이 있는 것은 분명합니다. 하지만 소비자의 입장에서는 신뢰성과 투명성을 갖추지 못한 제품을 사용하지 않을 것이기에 이러한 투명성을 고려한 방향이 보다 미래지향적인 운영이 될 것입니다.

마지막으로, 인공지능의 윤리적 토대를 구축하기 위한 범국가적인 노력이 필요합니다. 이를 통해 윤

리적 문제를 야기할 수 있는 인공지능의 상용화에 대한 적극적인 규제가 필요합니다. 2019년 EU는 인공지능 윤리 7대 가이드라인을 제시하였습니다. 해당 지침에는 인간의 통제 가능성, 안정성, 개인정보 보호, 투명성, 다양성, 비차별성과 공정성, 지속 가능성, 책임성 등을 보장하는 내용이 담겼습니다. 지침에 따르면 AI는 인간의 자율성을 보장해야 하며, 사람들은 AI에 의해 조작되어서는 안 되며, 인간은 소프트웨어가 내리는 모든 결정에 개입할 수 있어야 한다고 합니다. 또한 AI는 기술적으로 안전하고 정확해야 하고 외부 공격과 타협해서는 안 되며, 신뢰가 가능해야 합니다. 그리고 AI가 수집한 개인정보는 안전하게 보장되어야 하며, AI 시스템을 만드는 데 사용된 알고리즘과 데이터는 사람이 이해하고 설명할 수 있어야 한다고 명시하였습니다. 이외에도 EU는 AI는 연령, 성별, 인종 등을 차별하지 말아야 하며, 지속 가능해야 하고, 감사 가능해야 한다고 밝혔습니다. 이처럼 세계적으로 인공지능 윤리의 필요성을 실감하고 있음이 확실하고, 발표된 해당 조항이 아직 가이드라인 수준이기에, 지속적으로 인공지능에 의해 발생하는 문제와 차별적 사례들을 축적하고 보다 사회적으로 인공지능 세대에 대응해 나가야합니다.

우리는 과거에 없던 자동차와 비행기를 타고 세상을 누비며 인터넷과 스마트폰이 삶을 지배하는 세대에 살고 있습니다. 이 또한 과거에 많은 시행착오를 겪으며 사회적, 법적 테두리를 오랜 시간 구축하고 발생했던 새로운 문제에 적응한 결과입니다. 인공지능 또한 차세대의 새로운 전기와 같이 빠르게 우리의 삶에 침범하고 있기에, 유용성과 수익성 이전에 이 기술이 확산할 수 있는 은연적 차별의 외재화에 대한 주의와 새로운 사회적 윤리의 필요성에

관심을 가지고 지속적으로 문제의식을 제기하고 개선하기 위해 개인, 기업, 국가 차원에서 노력해 나가야 할 것입니다.

• 참고문헌 •

- Rachael T.(2017), Gender and Dialect Bias in YouTube's Automatic Captions. Proceedings of the First Workshop on Ethics in Natural Language Processing, 53-59
- Joy B., Timnit G.(2018), Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability, and Transparency: Proceedings of Machine Learning Research 81:1-15
- Moral Machine – MIT. Media lab, Scalable Cooperation <http://moralmachine.mit.edu/>에서 2019.09.09 인출