# 인공지능 편향의 식별과 조정 기술

김효은
한밭대학교 인문교양학부 교수 | 인공지능과 가치연구소 소장
hyoekim26@gmail.com

앞으로 인공지능을 제작하거나 연계, 활용하는 기업은 인공지능 편향 및 투명성과 관련된 국내외 정책의 영향을 받게 된다. 2018년에서 2022년 사이 과학기술정보통신부에서는 AI 윤리 가이드라인을 제정(관계부처합동, 2021)하고 체크리스트를 공개했으며, 지난 6월 14일에는 국가기술표준원에서 「인공지능윤리 국가표준(KS)」을 제정하기도 했다. KS표준에서는 인공지능 제품 및 서비스 개발 시 고려해야할 윤리적 항목과 점검방안이 제시됐다. 이는 한국이나 특정 기관의 단일한 행보가 아니며, 세계 산업 및 경제에 영향을 미칠 세계적인 행보에 걸맞은 단계이기도 하다. 같은 날 유럽 의회에서는 윤리적으로 문제가 되는 AI 시스템 유형 등에 대한 규제 내용을 포함한 AI 법률 협상안이 최종 제시됐다.[1]

주된 내용은 성별, 인종, 국가, 장애여부 등 민감한 보호 속성으로 규정된 특성을 변수로 사용하는 생체 인식 분류 시스템이나 재범률 예측과 같은 AI 시스템을 금지하는 것이다. 더 나아가, 요즘 많은 AI 관련 기업이 지향하는 범용 AI(General Purpose AI)의 활용에 있어 투명성 조건을 준수하고 저작권이 있는 데이터 정보를 공개해달라 요구했다. 이는 사실 새로운 것은 아니다. 2018년 5월 발효돼 국내에서도 적용되는 GDPR(General Data Protection Regulation, 일반정보보호규정)의 연속선상에서 제시되는 정책이다. GDPR은 정보주체가 프로파일링에 대한 반대 권리, 자동화된 의사결정의 대상이 되지 않을 권리, 그리고 설명을 요청할 권리가 있음을 규정하고 있다. 이에 따라 최근 출시된 OpenAI의 GPT 또한 GDPR의 필터링 대상이 됐으며, 지속적으로 개발 중인 대규모 언어모델들을 활용한 서비스 또한 국내외적으로 AI 법률의 영향을 받게 된다.

이 글에서는 세계 여러 기업과 국가기관에서 내세우는 소위 '책임있는' 인공지능의 핵심이슈인 '인공지능 편향의 인식과 완화'를 다룬다. 주로 인공지능 구성 과정의 기술적 측면에서 인식과 완화를 논의하면서 국내외에서 새롭게 요구될 인공지능윤리 영향 평가에 대처할 기초적인 인공지능 편향완화 방안과 남은 문제들을 함께 고민하려 한다.

## 인공지능 편향의 특성

현재 인공지능은 1950년대에서부터 계속돼 온 규칙 기반의 자동 시스템에서 '자율지능 시스템(Autonomous Intelligent System)'으로 발전했다. 여기서 '자율'의 개념에 대해 인간의 자유의지나 자율성과 유사한 의미로 해석하지 않도록 주의해야 한다. '자율지능시스템'에서

1 유럽 의회 Press Release, 2023년 6월 14일

'자율'은 외부 개입 없이 "복잡한 환경에서 복잡한 임무를 수행하기 위해, 스스로 인식하고, 계획하고, 학습하고, 진단하고, 제어하고, 중재하고, 협업하는 등 다양한 지능적 기능들을 가지는 시스템"으로 정의되며 시스템의 기능상 여러 수준으로 구분된다.

이 특성이 규칙기반 자동 시스템에서는 제기되지 않았던 '편향'이나 '투명성' 문제를 제기하게 만든다. 필자는 정보윤리의 문제와 AI 윤리 문제를 다른 유형으로 구분한다. 정보보안이나 프라이버시와 관련된 문제는 컴퓨터가 규칙기반의 자동시스템이든 현재 발전된 자율시스템이든 관련 없이 제기되는 문제다. 반면 새로이 생긴 윤리적 사안은 자율 시스템으로서의 딥러닝 기반 인공지능이 가진 특징인 '블랙박스' 문제 때문에 촉발됐으며, '편향' 문제를 인지하고 조정하기 위해 투명성을 확보할 필요가 있다. 인공지능의 최종 출력은 컴퓨터 자체가 하는 일이라기보다는 인공지능이 학습하는 데이터와 기계학습과정에서 생기는 블랙박스, 그리고 이 과정에 자동으로 데이터를 제공하는 우리 인간, 컴퓨터 알고리즘을 조정하는 인간이 개입된 결과이다. 따라서 인공지능에 개입되는 '편향'이란 인공지능이 기계학습을 할 때 사용되는 데이터를 선택, 수집, 분류, 사용할 때 그리고 알고리즘을 구성할 때 사회적으로 공평하지 않은 기준이 개입되는 것을 의미한다.

기계학습에 사용되는 데이터는 인간이 만들어낸 데이터이며 자연스럽게 인간이 가지는 편향이 내재해 있을 수밖에 없다. 예컨대 확증 편향(Confirmation Bias)(Gale, M. et al., 2002)이란 심리학에서 기존의 관점을 반영하는 방식으로 데이터를 선택하고 분석하는 편향을 의미하는데 이런 편향이 기계학습 과정에서 다른 편향과 결합해 증폭될 수밖에 없다. 이러한 증폭 현상이 종종 언급되는 '필터버블 현상'(Pariser, 2011)이다. 예컨대 검색엔진을 사용할 때 우리가 선택한 정보와 유사한 정보가 더 많이 보이거나 관련된 광고가 내 컴퓨터 검색 창에서 무수히 많이 보인다. 이러한 현상은 상품의 고객이 될 대상에게 맞춤형 쇼핑정보를 제공하는 마케팅 방법으로 사용되는 반면, 소비자에게는 유용함과 동시에 선택의 폭이 줄어드는 부작용 또한 존재한다.

또, 타인종효과(Cross-Race Effect)(Beaupré MG, et al., 2006) 라는 사회심리학 현상 또한 기계학습 과정을 거친 인공지능에서 보일 수 있다. '타인종효과'란 인간경험이 특정 인종에 더 친숙해 타인종의 인간적 특성들을 알아차리지 못하는 현상으로 인종에 제한되지 않고 자신에게 익숙하지 않은 대상을 잘 지각하지 못하는 현상이다. 기계학습에 사용되는 얼굴 데이터들은 인터넷상에서도 백인 얼굴에 대한 데이터가 흑인 얼굴에 대한 데이터보다 훨씬 많이 존재하고, 이러한 데이터에 기반한 기계학습으로 만들어진 인공지능은 상대적으로 소수 인종의 얼굴을 잘 인식하지 못하는 경향이 있을 수밖에 없다.(Martineau, 2019) 이는 단순히 얼굴인식에만 그치지 않고 스포츠 게임에서의 승패를 가르는 중요한 의사결정에서도 나타날 수 있다. 볼과 스트라이크를 판정하는 야구심판의 예를 살펴보면 관련 인공지능 시스템 구성 시 인간 야구 심판의 데이터를 사용할 수밖에 없는데, 이때 편향이 존재한다. 스트라이크 및 볼의 판정은 일견 투수의 개인 특성과는 무관하다고 생각할 수 있다. 그러나 야구에서 스트라이크와 볼을 판별하는 경우 본인의 판단에 영향을 미치는 요인 등에 대해 그 원인이나 인지적 배경을 자발적으로 필터링하기는 어렵다. 편향이 개입될 수 있는 지점은 스트라이크와 볼 간의 경계선(Borderline) 사례로 판정자가 가진 편향에 따라 판정이 달라질 수 있다. 파슨스(Parsons, A., et al., 2011)는 MLB(Major League Baseball)의 야구 심판이 인종과 민족이 같은 투수에게 더 관대한 판결을 내리는 편향이 있다고 분석했다. 이러한 경향이 사실로 밝혀진 상황에서, 인간의 부정확한 판단을 보조하고자 인공지능 야구 심판을 만들 때 사용할 데이터는 인간심판이 내려왔던 무수한 판정들이다. 이 데이터를 토대로 인공지능 야구심판을 만든다면 인종편향적 볼-스트라이크 판정이 그대로 재현 내지 증폭될 수밖에 없다.(김효은 2022)

## 인공지능 구성 단계들에서 개입되는 편향들

편향은 기본적으로 데이터 자체에 내재하지만 인공지능을 구성하는 단계들 거의 모두에서 발생가능하다. 이런 점에서 서비스 및 상품이 만들어진 후의 문제를 다루는 컴퓨터 및 정보윤리와 차이가 발생한다. 인공지능의 구성 단계는 ▲ 데이터 수집 단계, ▲ 라벨링과 같은 데이터 전처리 단계, ▲ 기계학습 단계, ▲ 모델의 채택 단계 등으로 나눌 수 있으며, 각 단계마다 편향이 개입되는 내용은 구체적으로 다음과 같다.

**[그림 1] 인공지능의 구성 단계**


* 출처: 김효은 2021의 그림 차용

### 데이터 수집 단계에서의 편향

인공지능 구축의 첫째, 둘째 단계인 데이터 수집과 전처리는 인공지능 구성 과정에서 80%를 차지한다고 할 정도로 큰 비중이며 사람이 대부분 개입하므로 사람의 편향이 그대로 전이 및 반영된다. 따라서 차후 단계들에서 사후 조치를 취하는 방법보다 기계학습의 원재료를 다루는 단계에서 조정하는 것이 편향완화의 효과가 높다고 알려져 있다. 이제는 많이 알려진 구글 포토서비스에서 흑인의 얼굴을 고릴라라고 인식한 2015년의 사례는 흑인 얼굴 이미지 수의 부족으로 발생했고, 이러한 사례에는 인종뿐만 아니라 소수자 성(Gender), 나이, 장애여부, 가족 형태, 국가, 민족, 군 복무 여부 등 다양한 보호 속성(Protected Attributes) 범주가 포함된다. 보호 속성이 적용돼 인공지능의 인식에서 정확도 저하와 편향을 일으키는 대상은 명시적으로 보이는 얼굴뿐만 아니라 국가마다 다르게 표현되는 모든 대상, 예컨대 가구 인식 인공지능(Vries et al., 2019)에서도 관찰된다.

### 데이터 라벨링 과정에서의 편향

데이터 라벨링에서의 편향은 부적절한 도메인 지식으로 잘못 라벨링하거나, 라벨된 데이터들에서 라벨 종류 간의 양이 비대칭적일 때 발생할 수 있다. 라벨링(Labelling)은 이미지, 영상, 텍스트 등 수집한 원데이터를 인공지능이 학습할 수 있도록 목적에 맞게 분류해 주석을 다는 작업이다. 텍스트의 경우 글의 내용이나 인공지능 활용 목적에 따라 고난도의 개념 이해 및 미묘한 맥락 파악이 필수적이기에 특정 분야에 대한 지식, 즉 도메인 지식이 요구된다. 예를 들어 20~30대 연령층에서 주로 사용하는 신조어가 포함된 글을 기계가 이해 가능하도록 라벨링해야 한다. 언어처리를 하는 인공지능을 구성할 때 글의 특성에 따라 맥락을 잘 이해할 수 있는 사람들이 라벨링의 일을 담당해야 한다.

적절한 도메인 지식을 갖추지 않은 사람이 라벨링을 하게 되면 데이터 편향으로 인해 최종 구축된 인공지능 시스템의 의사결정 문제가 생길 가능성이 있고 그 정확도 또한 낮아진다. 또, 여러 유형의 라벨 간에 비율 차이가 크면 기계학습에서 균형 잡힌 학습이 불가능해 최종 구축되는 인공지능 시스템의 특정 목적을 달성하기 어려워진다. 즉, "인공지능 시스템은 기술뿐만 아니라 사용자들, 영향받는 이들, 배포된 환경 모두를 포함한다."(Microsoft, 2021)



### 모델링 단계에서의 편향

인공지능 구축에 필요한 기계학습을 위해서는 예측 정확도 및 성능을 높이기 위해 변수를 튜닝하는 작업을 진행해야 한다. 모델링 단계에서의 편향은 ▲ 인공지능이 예측하고자 하는 목표와 관련성이 없거나 적은 변수가 포함돼 있을 경우, ▲ 중요도가 상대적으로 낮은 변수에 높은 가중치가 주어진 경우, ▲ 데이터셋에 인종, 젠더, 나이 등에서 불균형한 데이터가 있는 경우 발생할 수 있다. 전처리 단계에서 언급한 '보호 속성'들이 특정 속성과 통계적 종속성을 가지는 경우 편향이 생긴다. 종종 언급되는 미국의 재범예측 인공지능 컴파스(Compas)는 백인이나 부자보다는 흑인이나 빈민촌 거주자들, 그리고 남성의 재범률을 더 높게 판정해 변수 및 가중치 설정에 편향이 있는 것으로 분석됐던 사례이다.(Tan, et al., 2017) 이 같은 사례는 인공지능 면접과 같은 여타의 다양한 평가 시스템에서도 나타날 수 있다.

## 스무 가지 유형의 계산 가능한 인공지능 공정성과 한계

인공지능 구성의 어떤 단계에서든 편향이 있다고 인식하고 이를 완화 및 조정하려면 그 근거로서 공정성 기준을 필요로 한다. 그런데 공정성이란 그 맥락과 의미가 다양하다. 심리적, 윤리적 차원에서 등가교환적 정의문제인 '형평(Equity)'은 사실 차원에서 거론되는 '평등(Equality)'으로 다 설명할 수 없다. 또 분배의 공정성만 보더라도 동등하게 분배하는 '객관적 평등', 기여도에 따른 분배인 '상대적 평등', 개인의 필요에 따르는 '주관적 평등', 비용에 따른 '서열적 평등', 그리고 '기회의 평등'으로 구분된다.(Eckhoff, 1974) 이 개념들은 '필요'라는 상황맥락(Deutsch, 1975)과 균형을 맞추는 것 또한 고려해야 한다.

이렇듯 다양한 공정성의 맥락을 기준으로 삼아 인공지능 속 편향성을 판단하려면 통계적으로 계산이 가능해야 할 것이다. 이를 위해 최근 등장한 컴퓨터 기반 기술분야가 기계학습과 관련된 '공정한 인공지능'이다. '공정한 인공지능'의 기술적 의미는 인공지능 모델의 최종 판단결과가 인종, 성별과 같은 특정 보호 특성들에 종속변수가 되지 않도록 무관하게 제시되는 기술이다. 대표적 예로 열두 종류의 통계지표를 활용 및 조합해 작성한 20가지의 공정성 유형이 정의되고 있다.

[표 1] 통계적 공정성의 유형

| 유형 | 종류 | 수학적 의미 |
|---|---|---|
| 예측 기반 | 집단 공정성 | 집단별 긍정적 예측값을 할당받을 확률이 동일 |
| | 조건부 통계적 동등성 | 특정 데이터 속성을 통제했을 경우 그룹별로 긍정적 예측값을 할당받을 확률이 동일 |
| 예측 및 실제 결과 기반 | 예측적 동등성/결과적 동등성 | 긍정적 예측값의 비율이 집단 간에 실제로 동일 |
| | 위양성율 (False Positive Error Rate) 균형 | 위양성 예측값을 할당받을 확률 |
| | 위음성율 (False Negative Error Rate) 균형 | 위음성 예측값을 할당받을 확률 |
| | 동등 확률 | 예측값 기반 양성예측도(PPV, Positive Predictive Value)와 음성예측도(NPV, Negative Predictive Value)의 동등 |
| | 조건부 사용 정확도 동등성 | 예측값 기반 양성예측도(PPV, Positive Predictive Value)와 음성예측도(NPV, Negative Predictive Value)의 동등 |
| | 전체 정확도 동등성 | 위양성(False Positive)과 위음성(False Negative)의 비율이 집단 간 동일 |
| | 대우 동등성 | 위양성(False Positive)과 위음성(False Negative)의 비율이 집단 간 동일 |
| 예측 확률 및 실제 결과 기반 | 테스트 공정성 (조건빈도) | 예측된 확률 점수에 대해 보호집단과 비보호집단의 피험자가 실제 양성일 확률이 동일할 때 |
| | Well-Calibration | 예측된 확률점수에 대해 보호집단과 비보호집단의 피험자가 양성에 실제로 속할 확률이 같아야 할 뿐만 아니라 예측된 확률점수와도 같을 때 |
| | 양성 집단에 대한 균형 | 보호 그룹과 비보호 그룹의 양성 클래스를 구성하는 피험자가 동일한 평균 예측 확률 점수 S를 갖는 경우 |
| | 음성 집단에 대한 균형 | 보호집단과 비보호집단 모두에서 음성인 피험자는 평균 예측 확률 점수가 동일해야 함 |
| 유사성 기반 | 인과적 차이 | 정확히 동일한 속성을 가진 두 주제에 대해 동일한 분류를 생성할 때 |
| | 블라인드(Unaware)를 통한 공정성 | 의사 결정 과정에서 민감한 속성이 명시적으로 사용되지 않을 때 |
| | 인식을 통한 공정성 | 유사한 개인이 유사한 분류를 가질 때 |
| 인과 추리 | 반사실적 공정성 | 예측된 결과가 보호된 속성의 자손변수에 의존하지 않는 경우 |
| | 미해결된 차별없음 | 보호된 속성에서 예측된 결과까지의 경로가 존재하지 않는 경우 |
| | 대리차별금지 | 보호된 속성에서 대리 변수에 의해 차단되는 예측된 결과까지의 경로가 없는 경우 |
| | 공정한 추론 | 인과 관계 그래프의 경로를 정당 혹은 부당한 것으로 분류 |

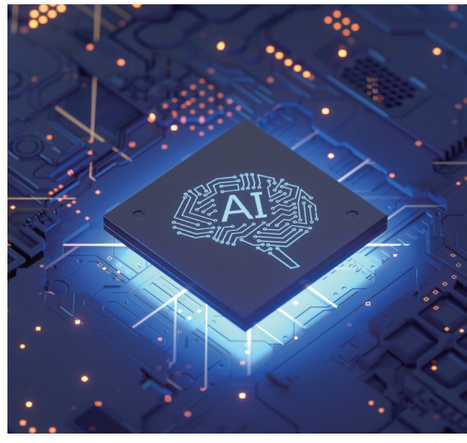
* 출처: 김효은(2021)에서 Verma et al., 2018을 표로 정리한 것을 재인용



이렇게 많은 의미의 공정성이 제기되는 배경은 [표 1] 예측 기반 유형에 '집단 공정성'이라고 표현되는 통계적 균등성(Statistical Parity)이 표준적 기준이 될 수 없기 때문이다. 예컨대 A 집단과 B집단은 샘플에 있어 양성으로 예측될 확률이 애초부터 동일하지 않을 수 있다. 그렇다면 이 차이는 두 집단에 있어서 위양성(실제는 음성인데 양성으로 예상한 경우)과 진양성(예상과 결과가 모두 양성인 경우) 비율의 차이를 만든다. 이 경우 두 집단의 통계적 균등성 거리를 측정하면, 결과적 균등함 여부를 알 수는 있으나 과정상 공정함이 고려되지 않는다. 따라서 통계적 균등성 수치 자체로 공정함을 판단하기 어려우며, 때때로 통계적 균등성이 오히려 알고리즘의 정확도를 낮추는 결과(Menon, 2018)를 가져온다.

통계적 균등성이 유일한 공정성 측정치가 될 수 없기에 베르마와 루빈(2018)은 통계적 지표들을 활용 및 조합해 [표 1]과 같이 다양한 공정성 정의를 정리했다. 다만 이러한 공정성의 다양한 정의를 동시에 충족하기는 어려우며 서로 상충하는 특징이 있다. 특정 상황에 하나의 공정성 계산을 적용한다 해도, 절차적 공정성 등을 함께 고려해 최종 공정성 여부를 판단해야 한다. 그리고 상황에 따른 맥락들을 고려하면 더 많은 종류의 공정성 개념이 구성될 수 있다. 또한, 각각의 스무 가지 공정성 기준들은 자체적으로 더 잘 사용되는 맥락이 있는 등 각기 장단점이 존재하며, 각 기준에 대한 연구자들의 평가 또한 상이하다. 즉, 완벽하게 이상적인 공정성이라는 유일한 기준은 존재하지 않는다. 그러므로 해당 상황의 맥락에서 고려해야 할 부분이 정량적, 통계적으로 계산 가능하지 않은 경우에는 정성적 분석을 필수적으로 병행하고, 이를 통해 해당 인공지능 및 데이터의 편향 여부를 결정해야 한다.

## 인공지능 편향완화 기술

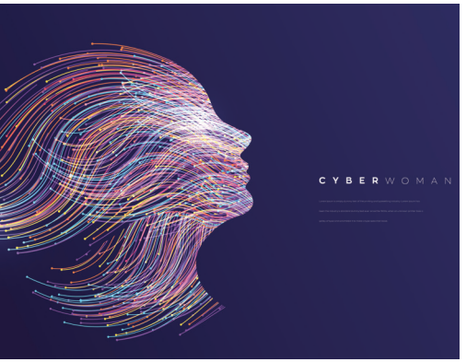



앞서 살펴보았듯 편향은 인공지능 구성 단계(전처리 단계, 처리 중 단계, 처리 후 단계)마다 내재할 수 있다. 편향의 완화방법 또한 각 단계마다 적용할 수 있는 여러 방법이 존재하고, 지속적으로 개발 중에 있다. 특히 전처리 단계에서 적용하는 완화방법이 효율적인데, 예컨대 맥락 속에서 의미를 파악해야 하는 텍스트를 벡터로 처리하는 과정에서 선제적으로 편향을 인지하고 조정한다면 기계학습을 거쳐 증폭되는 것을 방지할 수 있다.

## 전처리(Pre-Processing) 단계에서의 편향완화

탈편향(Debias) 알고리즘이라고 칭해지는 전처리 단계에서의 편향완화는 임베딩구조에서 조정하는 방법과 반사실적 인과를 만들어내 조정하는 방법이 있다. 임베딩 구조에서 편향을 조정하는 방법을 적절히 이해하려면 텍스트 데이터의 임베딩 구조를 알 필요가 있다. 자연언어를 디지털로 처리할 때는 워드임베딩(Word Embedding)의 방법을 활용해 벡터로 표현한다. 이때 텍스트 데이터에 편향이 내재돼 있다면, 이후 기계학습으로 데이터를 훈련시켜 함수를 파악하는 과정에서 편향이 증폭된다. 따라서 이후 단계보다는 전처리 단계에서 편향을 완화하는 작업이 효율적이다. 워드임베딩의 한 방법인 워드투벡(Word2Vec)은 인공신경망 기법을 활용해 어휘를 일정한 길이의 벡터로 표시하는 방법으로 가장 많이 사용되므로 대표적 예시로 활용할 수 있다. 워드투벡 방법의 기본 가정은 '분포가설(the Distributional Hypothesis)'로 표현되는데"한 단어의 의미는 그것이 동반하는 것들을 통해 알 수 있다."(You shall know a word by the company it keeps.)(Firth, 1957)는 의미이다. 이 방법은 저차원에서 단어의 의미를 여러 차원 공간에 분산해 표현하기 때문에 단어 간 유사도를 맥락의 의미 손실 없이 기계 안에서 빠르게 계산할 수 있게 한다. 유사 의미를 가진 단어들 주변에는 유사한 의미를 가진 어휘들이 있는 구조이다. 예를 들어 '코로나 감염병'이란 단어는 '백신', '방역', '선별진료소' 등의 단어들과 함께 출현하는 횟수가 다른 단어들보다 많다.

워드투벡 등 맥락을 반영하도록 만들어진 텍스트 데이터 안에 편향이 있다면, 편향과 관련된 단어들은 벡터 구조 안에서 서로 가까운 거리로 표현된다. 많이 인용되는 사례를 보면, 벡터 구조 안에서 '비서'는 '남성'보다는 '여성'이나 '여자'에 한층 더 가까운 거리에 위치해있다. 이러한 거리는 인공지능이 텍스트를 출력할 때 반영되므로 편향을 그대로 내보이게 된다.




이런 구조에서 일견 가장 간단해 보이는 편향완화방법은 언어벡터 구조상에서 보호 속성과 특정 속성 간의 연결, 예컨대 '비서'와 '여성' 사이의 연결을 끊는 것으로 보일 수 있다. 하지만 텍스트의 경우, 데이터 구조상 여러 개념 벡터들이 복잡하게 상호 관련돼 있어서 편향된 관계속성을 완전히 삭제하지 못한다.(Barocas et al., 2017) 예컨대, '여성' 벡터와 '비서' 벡터는 직접적으로 관련 없어 보이는 거주지나 주민번호 벡터와도 텍스트 데이터 구조 안에서 연결돼 있다. 그래서 '여성-비서' 연결고리를 삭제해도 여성과 관련되는 주민번호 특성이나 거주지 등의 연관정보는 여전히 '여성'을 간접적으로 '대리'해서 나타내게 된다.(따라서 '대리 속성(Proxy Attributes)'이라고 일컬어짐) 만약 이러한 대리속성을 포함하는 모델을 기계학습에서 활용할 경우, 편향은 여전히 발생할 수 있다. 대리속성이 내부에 남아 있고 이로 인한 편향이 발생하기 때문에 이를 '대리 차별(Proxy Discrimination)'이라 칭한다.(Prince., et al., 2019) 이러한 현상은 널리 알려진 재범 예측률 인공지능에서 유색인종뿐만 아니라 '남성'과 '재범률 높음'과의 연결성이 높다고 보는 것에도 마찬가지로 적용된다. 이러한 언어모델 구조 자체의 특성 때문에 편향 제거보다는 '완화'가 현실적 차선책이다.

다른 방법으로는 임베딩 구조에서 중립적인 단어에 해당하는 벡터가 특정 단어와 편향적인 거리를 갖지 않도록 동등화(Equalize)하는 방법이 있다. 예를 들어 '비서'라는 중립적 단어를 '여성'과 '남성'이라는 보호 속성 관련 단어 모두에 동일한 거리를 가지게끔 조정하는 것이다.(Bolukbasi et. al., 2016) 이 방법의 부작용은 특정 목적의 시스템에서는 중요한 구분까지 제거할 수 있다는 점이다. 즉 공정성을 확보하는 방향이 정확도와 상충(Accuracy-Fairness Tradeoff)하는 경우이다. 특정 목적의 시스템에서는 '아버지'의 의미가 가지지 않는 의미를 '어머니' 표현이 가질 수 있는데, 이 경우 '어머니'에 가중치를 더 부여해 모델을 구성할 수 있다. 이런 모델이 편향을 가진다고 보기보다는 특정 목적의 시스템에 맞게 설계됐다고 볼 수 있다.

또 다른 방법은 반사실적 텍스트나 영상 데이터를 활용하는 것이다. 이 방식의 편향완화 방법은 보호/민감한 속성이 가상적 상황에서 바뀌더라도 시스템의 의사결정은 동일하도록 조정해 편향이 작동되지 않도록 한다.(Chiappa et al., 2019) 그럼에도 데이터 하위범주 내에서 데이터의 양적 불균형이 발생해 편향이 생길 수 있다. 특정 집단의 데이터가 부족하면 기계학습에서 주요하게 파악하는 분류의 정확도가 낮아지고 결과적으로 의도치 않은 간접 차별이 나타날 수밖에 없기 때문이다. 이런 양적 비율의 불균형 문제는 특정 집단 혹은 관련 집단들의 샘플 크기를 늘려 해결할 수 있다. 따라서 편향완화는 인공지능 구성 과정의 각 단계에서 질적, 양적 차원 모두를 검토해 진행해야 한다.

## 처리 중(In-Processing) 및 처리 후(Post-Processing) 단계에서의 편향완화

모델링 단계의 편향완화를 위해 많이 알려진 방법은 적대적 학습의 방법을 사용하는 것이다. 적대적 학습방법은 '적대적 공격(Adversarial Attack)', 즉 시스템이 싫어할만한

데이터를 만들어내는 방법으로 데이터에 인위적 조작을 가해 인공지능 모델의 성능을 높이는 방법이다. 사람이 인지할 수 없을 정도의 작은 교란(Perturbation) 샘플을 만들어내 인공지능의 판단을 흐리게 만들고 그를 극복하면서 기계학습의 분류 성능을 높이고 모델을 최적화할 수 있다. 보호 속성의 소수자 집단에 해당하는 흑인종이나 연장자 연령의 집단 등에 나타난 편향이 학습데이터에 있다고 가정할 때, 이를 토대로 기계학습을 진행하면 이 편향을 그대로 반영할 수밖에 없다. 반면 적대적 학습방법을 사용하면 보호 속성 집단에 대한 변수, 예측 변수, 그리고 적대적 공격자 모두를 동시에 학습하는데, 이때 적대적 공격자는 보호 속성인 흑인 집단의 주민번호뿐만 아니라 대체 특성인 우편번호까지 모델링하기 시작한다. 그리고 이러한 적대적 공격자의 시도능력을 최소화하는 프레임워크가 적용된다.(Zhang et al., 2018) 이 방법은 다양한 유형의 공정성에 적용할 수 있다는 장점을 가진다고 평가받는다. 이외에도 구글, 링크드인(Linkedin) 등은 오픈소스 공정성 도구를 제공해 편향완화를 보조한다.

앞서 전처리나 처리 중 단계에서 데이터 수집, 라벨링 등에서 발생하는 편향을 모두 검토하고 조정한다 해도 비의도적으로 생겨나는 편향이 있을 수 있다. 이 경우 데이터 처리 단계나 기계학습 과정에서 인지가 불가능해 영향을 주지 않는 것처럼 보이지만, 출력 결과 단계에서 이해관계자에게 영향을 주게 된다. 예측 이후 단계에서 편향을 완화하는 방법은 기계학습에 사용된 데이터와 분류기는 변화시키지 않고 결과의 임곗값을 선택한 공정성 기준에 맞춰 조정하는 방식이다.

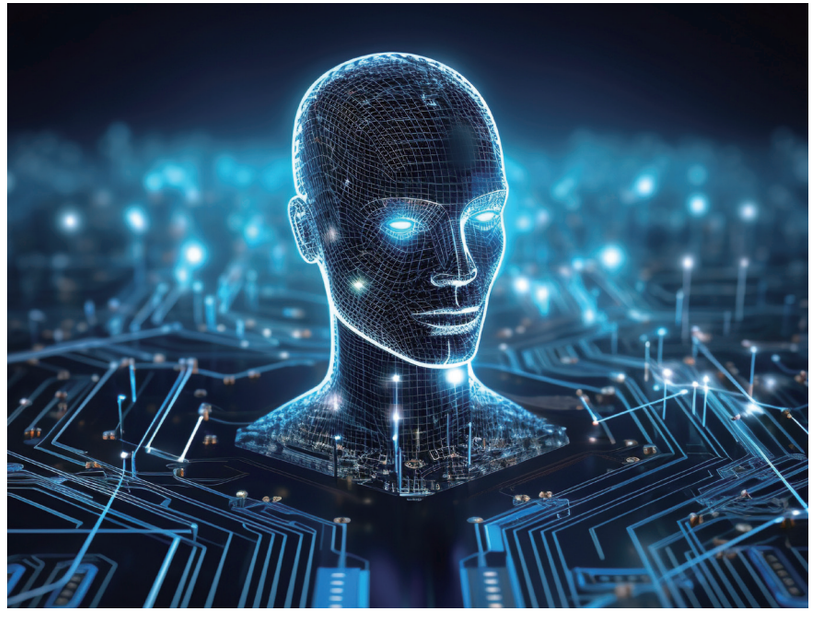
예컨대, 전세자금 대출 대상자를 결정하는 모델을 만든다고 가정하면, 일반적으로 이미 경제력을 구축해온 중년층을 선호하는 모델이 될 수 있다. 이러한 편향을 완화하고자 후처리 기술을 사용하면, 만들어진 분류 모델을 그대로 유지하면서 대출 대상자 전체 수락률이 모든 세대에 공평하도록 결과를 조정할 수 있다. 조정하는 방식으로는 해당 작업의 특정 목표에 따라 공정성의 정의 중 적절한 것을 선택하고, 관련된 집단 공정성이나 개인 공정성의 지표에 따라 조정한다.(Lohia et al., 2019) 이러한 후처리 방법은 전처리나 처리 중에 편향을 완화하는 방법처럼 굳이 모델 안을 들여다볼 필요가 없다는 편리성이 있다. 또한 인공지능의 블랙박스 특성으로 인해 그 원인을 정확히 추정하기 어렵다는 난점 또한 후처리 단계에서는 문제가 되지 않는다. 이외에도 다양한 방법들은 아래의 표와 같으며, 공정성 기준 간의 충돌 문제, 편향완화와 시스템정확도의 조화 문제 등을 해결하기 위한 방법들(Barocas et al., 2017)이 개발되고 있다.

**[표 2] 편향완화 방법들**

| 데이터 차원 | 완화 방법 | 종류 | 방법 |
|---|---|---|---|
| 시스템 | 훈련 데이터 편향완화 | 최적화된 전처리 (Optimized Pre-Processing) | 훈련 데이터의 특징과 라벨을 수정 |
| | | 가중치 재할당 (Reweighing) | 훈련 데이터의 가중치를 수정 |
| | | 이질적인 영향 제거 (Disparate Impact Remover) | 그룹 공정성을 개선하기 위해 특성값을 편집 |
| | | 공정한 표상 학습 (Learning Fair Representations) | 보호 속성에 대한 정보를 난독화해 공정한 표현을 학습 |
| | 분류기의 편향완화 | 적대적 편향성 제거 (Adversarial Debiasing) | 적대적 기술을 사용해 예측에서 정확도를 최대화하고 보호 속성의 증거를 감소시킴 |
| | | 편견 제거자 (Prejudice Remover) | 학습 목표에 차별을 인식하는 정규화 항을 추가 |
| | | 메타 공정 분류기 (Meta Fair Classifier) | 공정성 메트릭을 입력의 일부로 사용하고 해당 메트릭에 최적화된 분류기를 적용 |
| | 예측에서의 편향완화 | 거부 옵션 분류 (Reject Option Classification) | 분류기에서 예측을 변경 |
| | | 보정된 균등 배당률 후처리 (Calibrated Equalized Odds Post-Processing) | 공정한 출력 레이블로 이어지는 보정된 분류기 점수 출력을 최적화 |
| | | 균등 배당률 후처리 (Equalized Odds Post-Processing) | 최적화 체계를 사용해 예측된 레이블을 수정 |
| 집단 | 특권/ 비특권 집단 간 차이 | 통계적 parity 차이 (Statistical Parity Difference) | 특권집단에 비해 비특권 집단이 받은 유리한 결과의 비율차이 |
| | | 평등한 기회의 차이 (Equal Opportunity Difference) | 비특권:특권 집단의 진양성 비율의 차이 |
| | | 평균 배당률 차이 (Average Odds Difference) | 비특권: 특권 집단 간의 위양성 비율(위양성/음성)과 진양성 긍정 비율(진양성/양성)의 평균 차이 |
| | | 이질적 영향 (Disparate Impact) | 특권 집단의 비특권 집단 대비 유리한 결과의 비율 |
| | 데이터셋 | 유클리드 거리 (Euclidean Distance) | 두 데이터셋의 샘플 간의 평균 유클리드 거리 |
| | | 마할라노비스 거리 (Mahalanobis Distance) | 두 데이터셋의 샘플 사이의 평균 Mahalanobis 거리 |
| | | 맨해튼 거리 (Manhattan Distance) | 두 데이터셋의 샘플 사이의 평균 맨해튼 거리 |
| 개인 | 개인 | 부품색인 (Theil Index) | 개인에 대한 혜택 할당의 불평등 측정 |

## 맺음말

인공지능 편향을 인지하고 완화하고자 할 때 인공지능 편향이 나타난 사례를 아는 것은 도움이 될 수 있지만, 편향을 인식하고 완화할 방향을 생각해 내기에는 불충분하다. 인공지능에서 주요한 비중을 차지하는 데이터 전처리와 기계학습을 이해하고, 그 과정에서 편향이 개입되는 지점을 인지하는 것이 중요하다. 이를 파악하는데 중점이 되는 것은 인간 사회차원에서 벌어지는 인공지능 사용 후의 결과 및 영향이 아니다. 앞서 인공지능을 만드는 데 사용하게 되는 원료인 데이터가 어떻게 처리되는지, 기계학습이 대략 어떤 과정을 거치는지를 기초적인 차원이나마 아는 것이 필요하다. 이러한 인식과 교육은 개발자에게만 요구되는 것이 아니라 한 회사 및 기관의 의사결정자, 대표자, 그리고 사용자에게도 필요하다. 일반인들이 인공지능의 설계, 제작에 참여하지 않더라도, 인공지능 기반의 시스템이 내리는 의사결정에 크게 영향을 받게 된다. 따라서 인공지능의 구성과정을 최소한이나마 이해하고 편향이 개입되는 방식과 조정 가능성을 인지해, 개인 차원에서는 권리를 지키고 사회적 차원에서는 지속가능성을 상호 도모할 필요가 있다.





▧ 참고문헌


- 관계부처합동 (2021). 신뢰할 수 있는 인공지능 실현전략, 5월 13일. 과학기술정보통신부 인공지능기반정책과. https://www.korea.kr/common/download.do?fileId=195009613&tblKey=GMN
- 김효은 (2022), 머신러닝포키즈를 활용한 데이터 편향 인식 학습: AI야구심판 사례, 정보교육학회 논문지 26:4: 273-284.
- 김효은 (2021). "인공지능 편향식별의 공정성 기준과 완화, 한국심리학회지: 일반, 40: 4, 459-485.
- Barocas, S., Andrew D. Selbst, (2016). Big Data's Disparate Impact, California Law Review, 104:671, 679-680.
- Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. Nips tutorial, 1. https://arxiv.org/ct?url=https%3A%2F%2Fdx.doi.org%2F10.1007%2F978-3-030-43883-8_7&v=31e44ca0
- Beaupré MG, Hess U (2006). An ingroup advantage for confidence in emotion recognition judgments: the moderating effect of familiarity with the expressions of outgroup members. Personality & Social Psychology Bulletin. 32(1): 16-26.
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29, 4349-4357.
- Chiappa, S. (2019). Path-specific counterfactual fairness. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(1),7801-7808.
- Dave, P. (2018). Fearful of bias, Google blocks gender-based pronouns from new AI tool. Reuters, November, 27. https://www.reuters.com/article/us-alphabet-google-ai-gender-idUSKCN1NW0EF
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis, vol. Special Volume of the Philological Society, 1-32.
- http://cs.brown.edu/courses/csci2952d/readings/lecture1-firth.pdf
- Gale, Maggie; Ball, Linden J. (2002). Does positivity bias explain patterns of performance on Wason's 2-4-6 task?, Gray, Wayne D.; Schunn, Christian D., Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society, Routledge, p.340. https://eprints.lancs.ac.uk/id/eprint/11136
- Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., & Beutel, A. (2019, January). Counterfactual fairness in text classification through robustness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 219-226.
- Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2019, May). Bias mitigation post-processing for individual and group fairness. Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp) (pp. 2847-2851). IEEE.
- Menon, A. and Williamson, R. (2018). The cost of fairness in binary classification. In Conference on Fairness, Accountability and Transparency. 107-118.
- Microsoft (2021). Transparency note and use cases for Custom Neural Voice. https://docs.microsoft.com/en-us/legal/cognitive-services/speech-service/custom-neural-voice/transparency-note-custom-neural-voice
- Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.
- Parsons, A., et al., 2011. "Strike Three: Discrimination, Incentives, and Evaluation." American Economic Review, 101 (4): 1410-35
- Prince, A. E., & Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. Iowa L. Rev., 105, 1257.
- Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2017). Detecting bias in black-box models using transparent model distillation. arXiv preprint, https://arXiv:1710.06169
- Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In 2018 ieee/acm international workshop on software fairness (fairware) (pp. 1-7). IEEE.
- Vries et al., (2019). Does Object Recognition Work for Everyone?, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, 52-59.
- MEPs ready to negotiate first-ever rules for safe and transparent AI, Press Releases PLENARY
- SESSION IMCO LIBE 14-06-2023 - 12:52, European Parliament https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai?fbclid=IwAR1SLZX9K_zUwIsJRYcRx-6FmqB8Zt7oBjoVIXOxu4wE4EEhlCSCthqu8ml