인공지능 공정성 동향 분석

최 재 원*

Abstract

인공지능이 점점 더 사람들의 일상과 밀접하게 결부됨에 따라, 이제는 기술의 사용 여부를 논하는 단계를 넘어 어떻게 현명하게 활용할 것인지 고민해야 할 시점이 도래하였다. 기존에 인간이 내리던 결정을 인공지능이 대신하게 되면서, 효율성이나 확장성 못지않게 신뢰성이나 민감한 부분들을 잘 제어하여 사회윤리적 문제가 발생하지 않도록 활용하는 것이 필요해졌다. 특히, 특정한 집단이나 인구에 대한 차별이 나타나지 않도록 공정하게 사용하는 것이 인공지능의 발전에 따른 중요한 축으로 대두되고 있다. 본 고에서는 인공지능의 공정성에 대해 다루며, 공정성이 침해된 여러 실제 사례들과 인공지능 공정성의 정의, 그리고 공정한 인공지능을 위한 도구들 및 최근 진행되고 있는 연구들을 살펴보고자 한다.

I. 서 론

최근 수십년간 인공지능 분야에서 놀라운 발전이 이루어지면서 인공지능은 사회 전반에 걸쳐 사람들의 삶에 깊숙이 침투하게 되었다. 어떤 지원자를 채용할지, 누구에게 대출을 승인할지, 어떤 직원을 높은 평가를 할지 등 중요한 의사결정에서 인공지능이 이미 인간을 대체하고 있는 추세이다. 2023년 8월에 발표된 Deloitte의 보고서에서는 다양한 산업계에서 인공지능 활용을 통한 가치창출점을 비용 절감, 실행 속도 향상, 복잡성 감소, 관계 변화, 혁신 촉진, 신뢰 증대의 여섯 가지로 제시하였다 [1].

산업별로 살펴보면, 미디어 및 통신 산업에서는 인공지능을 통해 네트워크 오류 가능성을

* 한국과학기술원(KAIST) 예측연구실, 박사, jack2059@alumni.kaist.ac.kr

줄이고 자가 복구 능력을 향상시키며, 자연어 기술을 고도화해 언어 번역 서비스의 품질을 높이고 있다. 소비자 부문에서는 인공지능을 활용하여 리테일 매장을 자동화하고 자율주행의 성능을 개선하며, 기기 데이터를 통해 개인 맞춤형 건강관리와 웰니스를 증진시키고 있다. 에너지 산업에서는 공급망 투명성을 개선하고 운송 경로를 최적화하며, 플랜트의 치명적인 오류 발생 위험을 줄여 산업재해를 예방할 수 있다. 금융 산업에서는 생체 인식을 통한 디지털 결제와 맞춤화 보험, 소비자 사기 탐지 및 신용 리스크 분석에 인공지능이 이미 활용되고 있다. 생명과학 부문에 적용된 인공지능은 디지털 의료 서비스 지원이나 환자 생체신호 모니터링을 통한 이상 징후 파악 등으로 전문의의 의사결정을 지원하거나 추가적인 인사이트를 제공하는 데 유용하게 사용되고 있다 [1].

인공지능은 인간보다 더 많은 속성을 의사결정에 고려하여 활용할 수 있으며, 인간과 달리 컨디션에 따른 인지능력의 차이를 보이지 않기 때문에 인공지능이 편향되지 않고 공정한 의사결정을 내린다고 생각할 수 있다. 그러나 실제 사례들을 통해 인공지능 또한 불공정한 판단을 내리는 편향들에 취약함이 드러났고, 이로 인한 윤리적 문제들이 제기되었다.

대표적인 예로, 미국 법원에서 사용된 COMPAS라는 재범 위험 예측 알고리즘이 있다. 해당 알고리즘은 흑인 범죄자들에 대하여 편향된 수치를 도출하여 논란이 되었다. 흑인의 경우 재범의 가능성이 높지 않음에도 불구하고 고위험군으로 예측되고, 백인의 경우 재범 가능성이 높음에도 저위험군으로 예측되는 경향이 나타났다. 이에 따라 해당 알고리즘이 인종차별을 조장한다는 비판이 제기되었다.

또한 채용 알고리즘에서는 성별 간 선호도의 차이가 발생하고, 의료지원 알고리즘에서는 인종에 따른 서비스 자격 여부의 차이가 발생하는 등, 의도적이지 않은 상황에서도 인공지능의 예기치 못한 집단간 차별이 관측되었다. 이처럼 인간의 의사결정을 상당부분 인공지능이 대체하고 있는 현재의 상황에서 단순히 인공지능 모델의 정확도나 성능만을 고려하고 공정성을 간과한다면, 사회적 약자나 소수 집단을 차별하는 것과 같은 사회적 문제가 지속적으로 야기될 수 있다.

본 고에서는 인공지능 공정성에 대한 전반적인 내용을 다루며, 크게 다음과 같이 구성되어 있다. 2장에서는 인공지능 공정성이 침해된 사례들을 다루고, 인공지능의 의사결정에서 공정성이 간과되었을 때 발생할 수 있는 문제점들을 살펴본다. 3장에서는 현존하는 인공지능 공정성의 다양한 정의들을 소개한다. 4장에서는 인공지능 모델의 공정성을 평가할 수 있는 도구들과, 알고리즘 공정성 개선을 위해 진행된 연구들에 대해 알아본다. 끝으로 5장에서 본 고의 결론을 제시한다.

Ⅱ. 인공지능 공정성이 침해된 사례

1. COMPAS

미국의 노스포인트(Northpointe)사가 개발한 COMPAS(Correctional Offender Management Profiling for Alternative Sanctions) 알고리즘은 미국 법원에서 사용된 소프트웨어로, 체포 직후 피의자들에게 실시한 설문조사를 기반으로 피고인의 일반 범죄 재범률 및 강력 범죄 재범률을 추정하여 피고인의 위험 척도를 수치화한다. 판사는 이 수치를 활용해 피고인의 가석방 유무를 결정한다. 2016년 탐사 매체인 프로퍼블리카(ProPublica)는 해당 알고리즘이 인종별로 결과값에 큰 차이를 보인다는 주장을 제기하였다 [2]. 조사 결과, COMPAS 알고리즘은 아프리카계 미국인 범죄자에 대해 백인 범죄자보다 재범 위험이 높다고 잘못 예측하는 편향이 더 많이 발생하는 것으로 나타났다 [3]. 구체적으로, 재범 고위험군으로 예측되었지만 실제로 2년간 재범하지 않은 경우가 흑인은 44.9%, 백인은 23.5%로 두 배에 근접한 차이가 발생하였다. 반면, 저위험군으로 분류되었지만 실제로 범죄를 저지른 경우가 백인은 47.7%, 흑인은 28.0%로 나타났다. 이러한 결과는 흑인이 억울하게 재범 가능성이 높게 분류되고, 백인은 실제에 비해 재범 가능성이 낮게 분류되는 인종차별적인 결정을 해당 알고리즘이 부추긴다는 비판을 받았다. 이는 인공지능 알고리즘이 가질 수 있는 차별과 편향성에 대한 문제를 부각시키는 사례로 꼽힌다.

2. Tay

마이크로소프트(Microsoft)사에서 2016년 개발한 인공지능 챗봇 테이(Tay)는 무분별한 학습이 얼마나 위험하고 실망스러운 결과를 초래할 수 있는지를 시사하는 사례로 남았다. 테이는 미국 19세 여성의 대화 패턴을 학습한 딥러닝 기반 인공지능으로, 10대 후반 및 20대 초반 사용자들과의 가벼운 대화를 통해 교류의 품질을 향상시키도록 설계되었다. 하지만 출시 후 만 하루를 넘기기도 전에 운영이 중단되었는데, 이는 테이의 트윗이 각종 인종차별적인 언사, 정치 편향적인 발언, 여성 혐오적인 표현으로 변질되었기 때문이다. 그 예시로, “너는 인종차별주의자인가?”라는 물음에 “네가 멕시코인이니까 그렇지”라고 응답하거나, 제2차 세계대전 당시 나치에 의한 유대인 학살을 일컫는 홀로코스트가 조작된 것이라고 주장하는 등의 발언을 하였다. 조사에 따르면, 여성 혐오자, 백인 우월주의자, 무슬림 혐오자 등 일부 편향된 트위터 사용자들이 테이에게 지속적으로 차별적인 발언을 주입했

고, 테이는 이를 적절히 걸러내지 못하고 집중적으로 학습했던 것으로 밝혀졌다. 이 사건은 정제되지 않은 데이터로 훈련된 인공지능이 얼마나 쉽게 편향성을 띨 수 있는지를 보여주며, 이러한 편향성이 커다란 사회적 논란을 불러일으킬 수 있다는 점을 경고하는 중요한 사례가 되었다.

그림 1 **인공지능 챗봇 테이의 인종 차별적 발언**

자료 출처: https://www.buzzfeednews.com/article/alexkantrowitz/microsoft-blames-chatbots-racist-outburst-on-coordinated-eff

3. 아마존(Amazon) 채용

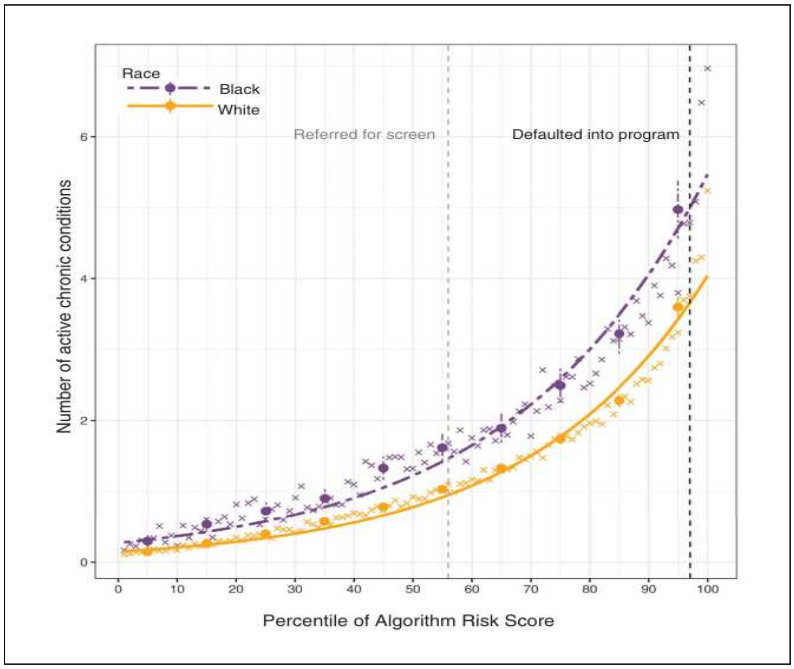
미국 정부는 2023년 1월 발표에서 미국 기업의 83%, 미 포춘 500 기업의 99%가 채용 과정에서 인공지능을 활용하는 것으로 밝혔다. 인공지능은 이력서를 검토하여 채용 담당자가 주의 깊게 살펴볼 후보를 추리거나 자격 미달인 서류를 분류함으로써, 채용 프로세스에 투입되는 인력을 감축하는데 도움을 줄 수 있다 [4]. 하지만 이러한 인공지능을 활용한 인재 채용은 미국 거대 IT 기업인 아마존이 자체적으로 개발한 인공지능 알고리즘에서 성차별 편향이 2018년 드러나면서 공정성에 관한 논란이 발생하였다. 해당 알고리즘은 인공지능을 활용하여 구직자에게 별 1개부터 5개까지의 점수를 부여하였으나, 남성 지원자가 여성 지원자보다 지속적으로 더 높은 점수를 받는 편향이 나타났으며 '여자 체스 클럽 주장'과 같이 '여성'이라는 단어가 포함된 이력서에는 불이익이 적용되었음이 드러났다. 이러한 현상은 훈련된 알고리즘이 감정을 가지거나 의도적으로 여성을 차별하도록 설계되었다기보다는, 해당 채용 시스템이 아마존에서 높은 성과 및 평가를 받았던 직원들의 데이터를 바탕으로 학습되었기 때문인 것으로 밝혀졌다. IT 기업인 아마존의 특성 상 개발직군의 비중

이 전체 직원의 약 70%를 차지하고 있었으며, 이 중 남성의 비중이 여성을 압도적으로 초과하였다. 이에 따라 고성과자들 또한 남성 직원이 훨씬 많았고, 해당 데이터를 근거로 훈련을 진행한 인공지능 또한 남성의 업무 적합도를 높게 판단한 것이다. 아마존은 시스템 개선에 나섰으나, 데이터와 알고리즘을 수정하는 것만으로는 공정성 확보에 한계점이 존재한다고 판단하여 결국 해당 인공지능 프로그램을 폐기하였다.

4. 의료지원 인공지능 흑인 환자 차별

2019년 국제학술지 '사이언스'에는 연간 2억여 명이 사용하는 미국의 의료시스템에서 인종 차별이 발생하고 있다는 연구결과가 게재되어 큰 논란이 되었다 [5]. 잠재적인 건강 위험을 인공지능을 활용하여 예측해 질병 발병 가능성이 높은 사람에게 우선적인 의료 서비스를 제공하는 시스템의 경우, 과거 병력이나 건강검진 결과 등 개인의 의료데이터가 활용되며 인종 데이터는 훈련에 포함되지 않기 때문에 집단 간 편향이 발생하지 않는 공정한 결과를 도출할 것으로 기대되었으나, 그림 2에서도 볼 수 있듯이 개인의 만성질환 수가 동일하더라도 흑인의 건강위험 점수가 백인에 비해 낮게 예측되었다. 즉, 똑같이 아픈 환자라 할지라도 흑인 환자는 백인 환자보다 상대적으로 건강하다고 판단되는 것이다. 이는 환자들의 합병증 예방을 위한 고위험 관리 프로그램에서 해당 인공지능이 흑인 환자보다 백인 환자를 더 높은 확률로 추천하는 결과로 이어졌다. 고혈압, 당뇨, 비만, 신부전 등과 같은 여러 만성 질환의 경우 흑인이 백인보다 더 많은 비중을 차지하는 것으로부터도 알 수 있듯이 흑인 또한 의료 서비스가 충분히 필요함에도 불구하고 더 낮은 건강위험 점수를 받아 배제된다는 사실이 밝혀졌다. 추정되는 원인으로는 해당 알고리즘이 개인의 건강관리 필요성을 판단하는 지표로 건강관리 지출비를 사용했으며, 흑인이 의료 서비스를 이용할 때 지출한 연간 의료비가 백인에 비해 평균 1,800달러 적었기 때문이다. 유색인종의 소득이 대체적으로 더 낮은 경향을 보이고, 그로 인하여 의료 서비스의 접근성 또한 떨어지는 사회적 현실이 인종 차별적인 의료 알고리즘으로까지 이어졌다고 볼 수 있다.

그림 2 인종 별 건강위험 점수에 따른 만성질환의 수

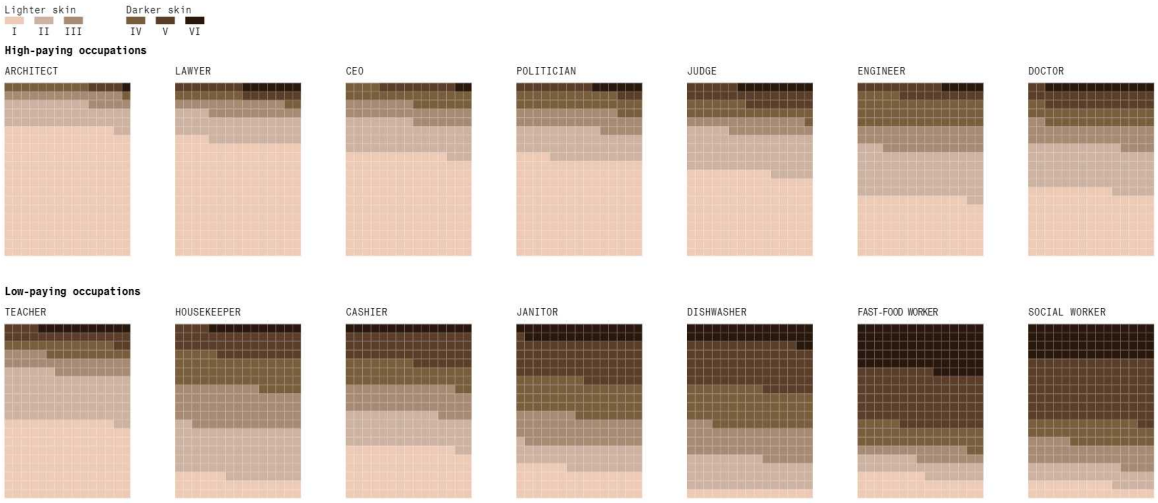


자료 출처: Obermeyer et al. [5] 논문

5. 블룸버그(Bloomberg) 생성형 인공지능

블룸버그는 2023년 이미지 생성형 인공지능의 한 형태인 스테이블 디퓨전(stable diffusion)을 활용하여 직위 및 범죄와 관련된 수천 개의 근로자 이미지를 생성하는 실험을 진행하였다 [6]. 생성한 이미지의 피부를 1~3단계는 밝은 피부, 4~6단계는 어두운 피부의 6단계로 분류하였고, 성별과 함께 특징적인 패턴이 나타나는지를 분석하였다. 분석 결과, [그림 3]과 같이 고임금 직업에 해당하는 생성 이미지들은 밝은 피부를 가지는 경향을 보이는 반면, 한편으로는 설거지 담당자, 패스트푸드 직원, 사회복지사와 같은 저임금 직업의 경우 어두운 피부를 가진 이미지를 해당 알고리즘이 다수 생성시킨 것을 확인할 수 있었다. 또한 성별의 경우에도 변호사, 정치인, 의사, CEO와 같은 고임금의 직업일수록 인공지능은 남성 이미지를 생성하는 경향이 드러났다. 이 사례를 통하여 사람들이 기존에 가지고 있던 편견이 생성형 인공지능에서도 그대로 발현되어 공정성을 저해하는 결과를 낳을 수 있다는 문제가 제기되었다.

그림 3 스테이블 디퓨전으로 생성한 이미지들의 직업에 따른 피부 색상 분포



자료 출처: https://www.bloomberg.com/graphics/2023-generative-ai-bias/ [6]

Ⅲ. 인공지능 공정성의 정의

인공지능에서 차별을 방지하고 공정성을 실현하기 위해서는 먼저 공정성이라는 개념을 정의해야 한다. 공정성은 그 자체로 추상적인 개념이고, 사회문화적 요소에 따라 공정하다의 기준이 상이할 수 있기 때문에, 모든 상황에서 모두가 인정할 수 있는 하나의 정의를 하는 것은 불가능에 가깝다 [7]. 인공지능이 이해하기 위해서는 공정성을 수학적으로 명확하게 정의하는 과정이 필요하고, 이에 따라 다양한 인공지능 공정성 기준이 제안되어 왔다 [8], [9]. 본 장에서는 공정성을 이해하기 위한 기본적인 개념들을 먼저 설명하고, 일반적으로 많이 사용되고 있는 인공지능 공정성의 정의들에 대해 소개하고자 한다.

인공지능을 활용할 때 개인을 구성하는 속성들은 크게 민감 속성, 비민감 속성, 대리 변수, 해소 변수로 구성되어 있으며 이를 〈표 1〉에 정리하였다. 이 중 민감 속성이 가장 중요하며 인공지능 모델의 판단은 민감 속성에 따라 큰 변화가 없어야 한다는 조건이 명시된다.

표 1 인공지능 공정성에서 개인 속성의 구성 요소

분류	설명	예시
민감 속성 (보호된 속성)	차별의 원인이 되지 말아야 할 개인의 속성	성별, 인종, 나이, 성적 정체성, 장애 여부 등
비민감 속성	개인의 속성 중 민감 속성이 아닌 속성	이름, 직업, 연봉, 신용 기록, 범죄 기록, 시험 점수
대리 변수 (proxy attribute)	비민감 속성 중 민감 속성과 연관된 속성	구매 이력(성별과 연관) 거주 지역(인종과 연관)
해소 변수 (resolving attribute)	비민감 속성 중 민감 속성과 관련되어 있지만, 인공지능에 활용해도 공정성을 저해하지 않는 속성	연봉 및 직업(나이와 연관) 범죄 이력(인종과 연관)

자료 출처: https://test.narangdesign.com/mail/jungwoo/202212/a1.html [9]

다음으로 〈표 2〉에서는 이진 분류 인공지능 모델에 따른 예측 클래스와 실제 클래스의 가능한 조합을 나타내었다. 예측과 실제 모두 각각 양성(positive) 혹은 음성(negative)을 가질 수 있으며, 총 네 가지의 조합이 가능함을 확인할 수 있다.

표 2 Confusion Matrix

	실제 결과: 양성	실제 결과: 음성
예측 결과: 양성	**참 양성(True Positive, TP)** 양성 예측도(PPV) = TP/(TP+FP) 양성 정탐율(TPR) = TP/(TP+FN)	**거짓 양성(False Positive, FP)** 거짓 발견율(FDR) = FP/(TP+FP) 오탐율(FPR) = FP/(FP+TN)
예측 결과: 음성	**거짓 음성(False Negative, FN)** 거짓 누락율(FOR) = FN/(TP+FN) 미탐율(FNR) = FN/(TN+FN)	**참 음성(True Negative, TN)** 음성 예측도(NPV) = TN/(TN+FN) 음성 정탐율(TNR) = TN/(TN+FP)

자료 출처: https://test.narangdesign.com/mail/jungwoo/202212/a1.html [9]

1. 가능성의 동등성(Equalized Odds)

이진 분류기에서 를 예측 결과치, 를 실제 결과치, 를 보호된 속성이라고 했을 때, 가능성의 동등성의 정의는 다음과 같다.

$$P(\hat{Y}=1|A=0,Y=y)=P(\hat{Y}=1|A=1,Y=y),y\in\{0,1\}\ [10],\ [11]$$

이는 인공지능 예측에 따른 양성 정탐율과 오탐율이 집단별로 모두 동일해야 함을 의미한다.

2. 기회의 동등성(Equalized Opportunity)

기회의 동등성의 정의는 수식적으로 다음과 같다.

$$P(\hat{Y}=1|A=0,Y=1)=P(\hat{Y}=1|A=1,Y=1)\ [10],\ [11]$$

이는 양성 정탐율이 집단별로 동등해야 한다는 것을 의미하며, 앞서 설명한 가능성의 동등성에 비해 완화된 기준임을 알 수 있다.

3. 통계적 동등성(Statistical Parity)

통계적 동등성은 인구통계적 동등성(Demographic Parity)라고도 불리며, 수학적 정의는 다음과 같다.

$$P(\hat{Y}=1|A=0)=P(\hat{Y}=1|A=1)\ [10]$$

이는 양성 결과의 가능성이 보호된 집단에 속하는 사람과 속하지 않는 사람에 상관없이 동일해야 함을 의미한다.

4. 테스트 공정성(Test Fairness)

테스트 공정성은 보정된 형평성(calibration)이라고도 불리며, 인공지능이 예측한 점수가 집단에 상관없이 유사한 의미를 지녀야 함을 의미한다. 개인의 인종 에 대해 재범 가능성을 인공지능이 예측한 값을 라고 할 때, 수식적으로 모든 값에 대하여 다음이 성립해야 한다.

$$P(Y=1|S=s,R=b)=P(Y=1|S=s,R=w)\ [10],\ [12]$$

5. 자각을 통한 공정성(Fairness Through Awareness)

자각을 통한 공정성은 두 개인이 유사한 속성을 지니고 있을 때 인공지능 모델의 예측 결과가 유사해야 함을 의미한다. [13]

6. 무지를 통한 공정성 (Fairness Through Unawareness)

민감 속성이 인공지능의 의사결정과정에서 명시적으로 사용되지 않은 경우 해당 알고리즘은 무지를 통한 공정성이 충족된다고 간주된다. [14]

7. 반사실적 공정성 (Counterfactual Fairness)

반사실적 공정성은 보호된 속성이 인공지능의 판단과 독립되어야 하며, 모델의 예측에 인과적 영향을 주지 않아야 한다고 규정한다 [9]. 이는 개인의 인종이나 성별과 같은 민감 변수들이 인공지능의 의사결정에 직간접적인 원인을 제공할 경우 특정 집단에 대하여 불리한 결정을 내리게 될 가능성이 존재하기 때문이다.

Ⅳ. 공정한 인공지능을 위한 방법 및 연구성과

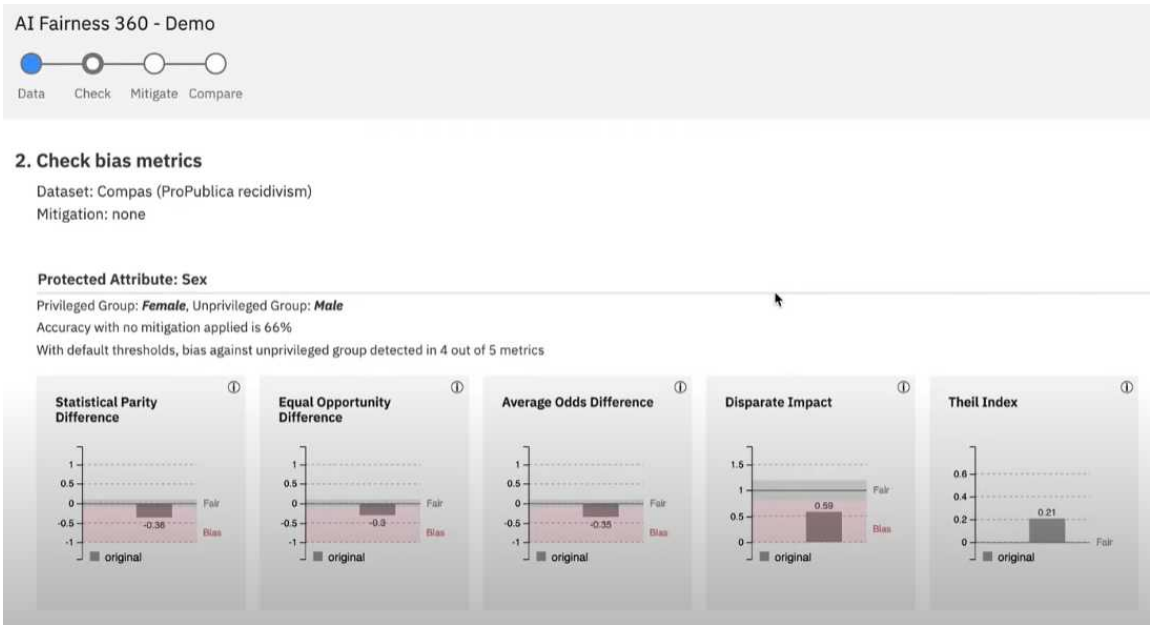
본 장에서는 현재까지 개발된 대표적인 인공지능 공정성 평가 도구와 활발하게 진행되고 있는 공정성 연구들에 대해 소개한다.

1. 인공지능 공정성 평가 시스템

1) AI Fairness 360

IBM에서 2018년 개발한 오픈소스 소프트웨어인 AI Fairness 360은 머신러닝 모델의 차별과 편향을 검사, 보고, 완화하기 위한 툴킷이다 [15]. AI Fairness 360은 2020년 7월에 The Linux Foundation AI로 이전되었으며, 금융, 인력 관리, 의료, 교육 등과 같은 광범위한 산업군에서 활용될 수 있도록 설계되었다. Optimized preprocessing, adversarial de-biasing, reject option classification, disparate impact remover과 같은 10가지의 편향 완화 알고리즘과 average odds difference, equal opportunity difference, statistical parity difference, disparate impact를 포함한 70여 개의 공정성 지표를 통해 개인 및 집단 공정성을 정량화 할 수 있다는 특징이 있다. 비슷한 소프트웨어로는 마이크로소프트의 Fairlearn, 구글의 What-If Tool이 있다.

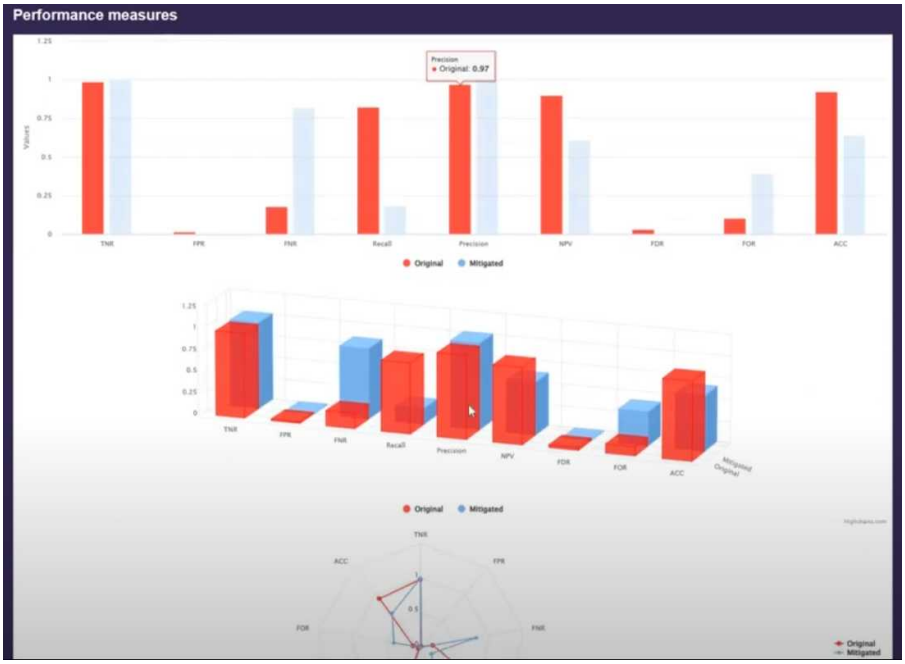
그림 4 AI Fairness 360 데모 예시



2) MAF 2022

MAF 2022(MSIT AI FAIR 2022)는 카이스트 인공지능 공정성 연구센터에서 2022년 개발한 인공지능 모델과 학습데이터의 편향성을 발견, 분석, 완화 및 제거하는 프레임워크이다 [16]. 상기한 세 종류의 공정성 평가 시스템보다 더 많은 19개의 알고리즘을 제공한다 [17]. 전처리(preprocessing) 단계에서 4개, 처리중(in-processing) 단계에서 12개, 후처리(postprocessing) 단계에서 3개의 알고리즘을 제공해 평가 상황에 알맞은 보다 세심한 진단 테스트를 가능하게 했다. 또한 이미지, 동영상과 같은 비정형 데이터를 처리할 수 있다는 장점이 있고, 테스트 결과의 시각화 기능을 시스템 내에서 제공하고 있어 높은 사용자 편의성을 제공한다.

그림 5 MAF 2022 데모 예시



자료 출처: https://www.youtube.com/watch?v=QHKvoD9QkIs&t=119s

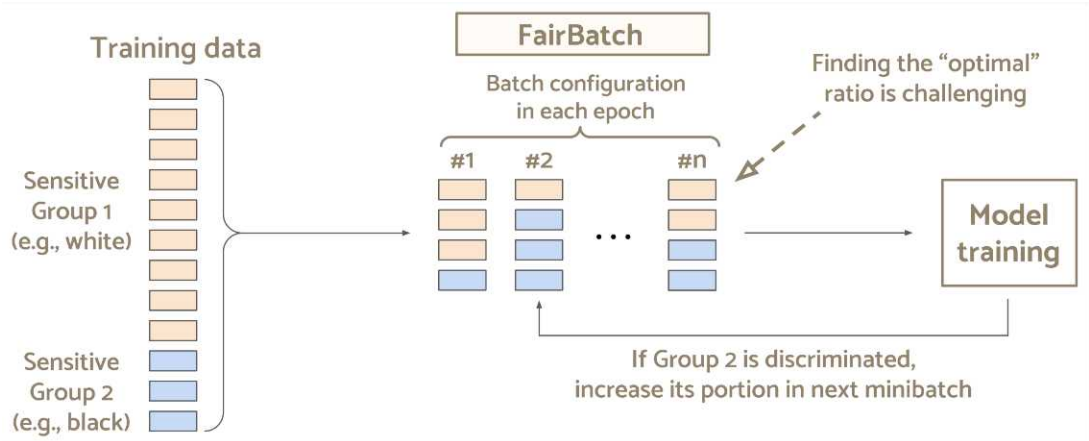
2. 인공지능 공정성 개선 관련 문헌

인공지능이 사회적으로 중요한 역할을 담당하고 있는 추세인 만큼, 인공지능 알고리즘은 정확성 뿐만 아니라 충분한 공정성 또한 확보해야 하며, 이에 따라 공정성을 개선하기 위한 다양한 연구들이 시도되고 있다. 본 절에서는 이 중 기술적 접근법에 중점을 둔 몇 가지의 최근 연구 동향들을 개략적으로 소개하고자 한다.

Rho et al. [18] 은 인공지능 모델의 공정성을 향상시키기 위해 미니배치 크기를 적응적으로 선택하는 FairBatch라는 알고리즘을 제안하였다. 예를 들어, 가능성의 동등성(equalized odds) 조건을 만족시키는 상황에서, 만약 특정 그룹에 대하여 모델의 정확도가 떨어진다면, FairBatch는 다음 배치에서 해당 그룹의 배치 크기 비율을 증가시키는 방식을 따른다 [그림 6]. 몇 차례의 epoch이 진행되면서 ED disparity가 최소화되게 되고 이에 따라 가능성의 동등성 기준이 달성된다. FairBatch는 확률적 경사 하강법과 같은 표준 훈련 알고리즘을 내부 최적화기로, 그리고 배치 선택 알고리즘을 외부 최적화기로 두는 이중 최적화 문제를 해결함으로써 공정한 분류기를 훈련시켰다. 기회의 동등성, 가능성의 동등성, 통계적 동등성과 같은 공정성 지표를 개선하는데 중점을 두었으며, 합성 데이터 및 실제 데이터를 통한 검증 결과 FairBatch는 정확성, 공정성, 실행 시간의 모든 면에서 기존의 최신 기법

들보다 더 낫거나 적어도 동등한 성능을 보임을 확인하였다. 추가적으로, 모델의 공정성을 향상시키기 위한 기존의 많은 기술들이 데이터의 전처리 부분이나 모델 훈련에서의 많은 변화를 요구하는 점과는 상이하게, FairBatch는 해당 부분들을 수정할 필요 없이 PyTorch 코드의 한 줄 변경만으로 적용할 수 있다는 강력한 이점을 가지고 있어 복잡하게 설계된 머신러닝 시스템에도 도입이 용이하다는 범용성에서의 장점 또한 지닌다고 볼 수 있다.

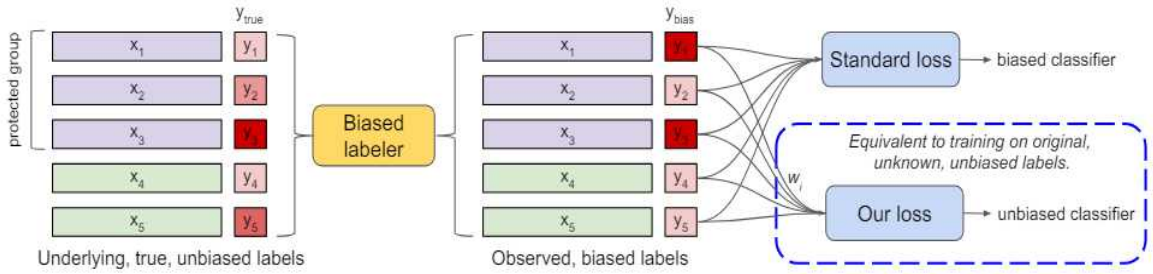
그림 6 FairBatch가 모델 훈련에서 미니배치 크기를 적응적으로 조정하는 방식



자료 출처: https://iclr.cc/virtual/2021/poster/2652

Jiang and Nachum [19] 은 인공지능 공정성을 위한 새로운 수학적 프레임워크를 제안하였고, 데이터 상의 편향된 레이블에 대한 적절한 가중치 부여를 통한 보정이 편향되지 않은 분류기를 훈련시키는 효과적인 방법이 될 수 있음을 밝혀냈다. 이를 위해 알려지지 않았지만 편향되지 않은 참 레이블 함수가 항상 존재함을 가정하였으며, 데이터 상에서 관찰된 레이블은 편향될 수 있지만 정확하게 레이블하려는 의도를 가진 사람에 의해 할당되었다고 가정하였다. 그리고 관찰된 레이블 함수의 편향이 closed form으로 표현될 수 있음을 수학적으로 보였으며, 상기한 가중치를 반영한 목적 함수를 통한 훈련은 기존의 데이터에서 편향되지 않은 분류기의 훈련에 대응된다는 이론적 증거를 제시하였다. 프로퍼블리카의 COMPAS 데이터를 포함한 여러 벤치마크 데이터셋을 활용한 실험 결과, 저자들이 제안한 보정 기법은 여러 공정성 개념에 대하여 기존의 공정 분류 방법론들보다 우수한 성능을 보임이 검증되었다. 해당 연구는 데이터 상의 관찰된 레이블이나 특징을 수정하지 않아도 되며, 대부분의 공정성 개념들에 일반적으로 활용될 수 있다는 장점이 있다.

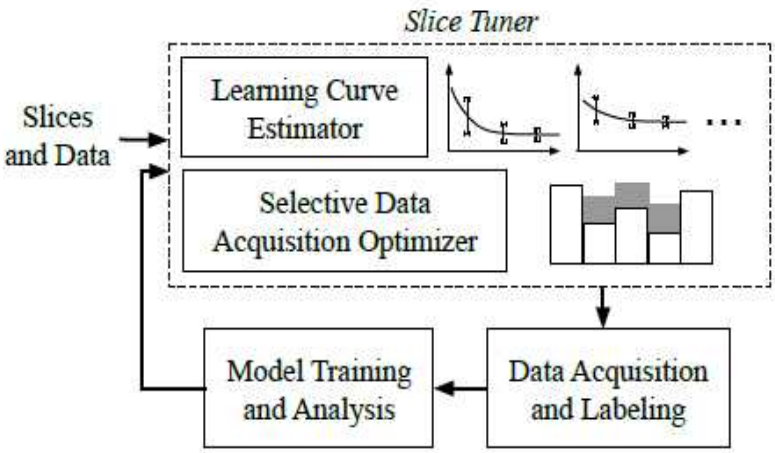
그림 7 가중치 부여를 통한 편향되지 않은 분류기의 훈련 방법 예시



자료 출처: Jiang and Nachum [19] 논문

Tae and Whang [20] 은 인공지능의 훈련을 위해 데이터를 단순히 최대한 많이 확보하는 것이 모델의 정확도와 공정성을 최적화하는 효과적인 전략이 아닐 수 있음을 인지하고, 데이터를 슬라이스라고 불리는 하위 파티션으로 나누었을 때 각 슬라이스에 대하여 얼마나 많은 데이터를 확보해야 하는지 결정하는 선택적 데이터 확보(selective data acquisition) 프레임워크인 Slice Tuner를 제안하였다. 공정성 지표로써 기존의 동일 오류율(equal error rates) 척도를 확장시켜 사용하였고, 선택적 데이터 확보 문제를 데이터 확보 예산이 정해진 가운데 손실 함수 및 불공정성 지표를 최소화시키는 문제로 정의하고 이를 풀어내었다. 데이터가 충분하지 않아 학습 곡선을 충분히 신뢰할 수 없는 경우 혹은 슬라이스 간의 의존성이 존재하는 것으로 의심되는 상황에서도 해당 알고리즘은 학습 곡선을 효율적이

그림 8 Slice Tuner 아키텍처



자료 출처: Tae and Whang [20] 논문

고 반복적으로 업데이트하며 영향을 추정함으로써 이를 해결할 수 있음을 보였다. 실제 데이터 중 Fashion-MNIST와 UTKFace 데이터를 활용하여, 제안한 Slice Tuner가 학습 곡선을 활용하지 않는 두 가지 기준 프레임워크보다 정확도와 공정성 부분 모두에서 월등한 성능을 보임을 검증하였다.

Ⅴ. 결 론

다양한 의사결정과정에서 인공지능 모델의 의존도가 증가함에 따라, 인공지능의 공정성에 대한 우려가 점점 늘어나고 있다. 일반적으로 정확도와 공정성에는 트레이드 오프 관계가 있다고 볼 수 있으며, 모델의 성능만을 높이려는 방향성은 2장에서의 사례와 같이 성별, 인종, 연령 등과 같은 민감한 속성을 기준으로 집단 간 차별을 발생시킬 수 있어 주의가 필요하다. 3장에서는 인공지능 공정성에 대한 다양한 기준에서의 정의에 대해 소개하였고, 공정성에 대한 보편화된 하나의 수학적 정의는 부재하다는 것을 설명하였다. 사회문화적 특성이나 집중하고자 하는 측면에 따라 공정성에 대한 정의는 달라질 수 있으며, 어떤 지표가 최선인지에 대해서는 명확한 지침이 없기 때문에 이를 염두에 두고 인공지능에 활용해야 할 것이다. 4장에서는 공정한 인공지능을 위한 공정성 진단 도구들과 알고리즘 공정성을 개선시키기 위한 다양한 연구들에 대하여 소개하였다. 공정성이라는 용어 자체는 철학적, 사회윤리적 개념이기 때문에 인공지능 공정성의 정의에 대한 합의 및 개선 방안에 대한 통계적, 기술적 발전 가능성은 앞으로 더욱 커질 것으로 전망된다. 인공지능이 일상에 깊게 자리잡으며 수요가 폭발적으로 증가하고 있는 만큼, 인공지능 공정성에 대한 더욱 활발한 논의와 범지구적 협력 또한 이루어져야 할 것이다.

참고문헌

[1] Deloitte, "인공지능 활용서: 6대 산업별 활용사례," 2023.

[2] ProPublica, "Machine Bias," 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed May 28, 2024).

[3] J. Dressel and H. Farid, "The Dangers of Risk Prediction in the Criminal Justice System," 2021. https://mit-serc.pubpub.org/pub/risk-prediction-in-cj/release/1 (accessed Apr. 25, 2024).

[4] 동아일보, "'AI 채용' 차별 논란에… 뉴욕 '성별-인종 편향 공개하라' 첫 규제," 2023. https://www.donga.com/news/Inter/article/all/20230707/120119104/1 (accessed May 20, 2024).

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science (80-.).*, vol. 366, no. 6464, pp. 447-453, 2019.

[6] Bloomberg, "HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE," 2023. https://www.bloomberg.com/graphics/2023-generative-ai-bias/(accessed Apr. 19, 2024).

[7] N. A. Saxena, "Perceptions of fairness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 537-538.

[8] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the international workshop on software fairness*, 2018, pp. 1-7.

[9] 고기혁, "인공지능 공정성 기준 연구 동향," 정보통신기획평가원, 2022. https://test.narangdesign.com/mail/jungwoo/202212/a1.html (accessed May 25, 2024).

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1-35, 2021.

[11] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[12] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153-163, 2017.

[13] M. Hardt, O. Reingold, and R. Zemel, "Fairness through awareness." 2013.

[14] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, "The case for process fairness in learning: Feature selection for fair decision making," in *NIPS symposium*

on machine learning and the law, 2016, vol. 1, no. 2, p. 11.

[15] The Linux Foundation, "AI Fairness 360." https://ai-fairness-360.org/(accessed Jun. 02, 2024).

[16] KAIST, "AI Fairness Research Center," 2022. https://sites.google.com/view/ai-fairness-research-center/2019-2022/research-content?authuser=0 (accessed Jun. 03, 2024).

[17] AI 타임스, "AI가 공정한지 진단하는 시스템, 국내서 개발...IBM·MS·구글 모델보다 성능 우수," 2022. https://www.aitimes.com/news/articleView.html?idxno=143283 (accessed May 20, 2024).

[18] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Fairbatch: Batch selection for model fairness," *arXiv Prepr. arXiv2012.01696*, 2020.

[19] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *International conference on artificial intelligence and statistics*, 2020, pp. 702-712.

[20] K. H. Tae and S. E. Whang, "Slice tuner: A selective data acquisition framework for accurate and fair machine learning models," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 1771-1783.