한국심리학회지: 일반
Korean Journal of Psychology: General
2021, Vol. 40, No. 4, 459-485
http://dx.doi.org/10.22257/kjp.2021.12.40.4.459

인공지능 편향식별의 공정성 기준과 완화

김 효 은†

국립한밭대학교 인문교양학부

인공지능 편향은 사회적 영향과 거버넌스의 문제일 뿐만 아니라 인공지능 시스템의 강건성 문제이기도 하다. 부호처리 패러다임의 인공지능에서는 제기되지 않던 인공지능 편향 문제는 컴퓨터가 인공신경망 기반의 자율지능시스템 단계가 되면서 시스템 구축 절차 각각에서 개입된다. 이 논문의 목적은 인공지능의 구성 절차에서 개입되는 편향의 양상, 편향 판단의 공정성 기준들, 편향완화 방법을 탐색하는 것이다. 공정성의 다양한 유형들은 동시에 만족되기 어렵고 인공지능의 적용 분야 및 맥락에 따라 상이한 기준과 요소의 결합이 필요하다. 학습 데이터, 분류자, 예측 내용의 편향을 완화하는 방법 또한 편향을 완전히 차단하는 것은 아니며 편향완화와 정확도 간의 조화를 모색해야 한다. 인공지능 감사를 통해 알고리즘에 무제한으로 접근하여 편향을 식별해낸다 하더라도 해당 알고리즘의 편향 여부를 단정하기는 어렵다. 편향완화 기술은 단순히 편향을 제거하는 단계를 넘어서서 편향완화와 시스템의 강건성을 동시에 확보하는 과제, 그리고 다양한 공정성 유형들을 조정하는 과제를 해결하는 단계로 나아가고 있다. 결론적으로, 이러한 특성들은 인공지능 편향을 인지하고 해결책을 모색하는 과제가 개념적 차원의 사안 인지를 넘어서 시스템 이해에 기반한 편향 인식 및 조정 차원에서 모색되어야 함을 암시한다.

주요어 : 기계학습, 편향, 공정성, 데이터, 알고리즘

† 교신저자: 김효은, 국립한밭대학교 인문교양학부, (08827) 대전시 유성구 동서대로 125
Tel: 042-821-1738, E-mail: hyoekim26@hanbat.ac.kr

인공지능윤리 이슈들에서 '편향'의 위치

현재 한국을 포함한 전 세계에서 진행되는 인공지능윤리 관련 기술 및 정책의 방향은 단순히 인공지능 윤리적 사안에 대한 주목과 이해를 넘어서 있다. 개념적, 이론적 차원에서는 2016년부터 최근까지 인공지능윤리 가이드라인 및 지침이 각 국가의 정부 및 기관들에서 발표되어 왔다. 최근에는 실천적 차원에서 정책의 적용이 이루어지고 있다. 인공지능 활용 기관 및 개인들이 인공지능기술 활용시 점검하여 실행할 인공지능윤리 목록들이 정부 및 기관 주도로 만들어지고 있다(관계부처합동, 2021). 산업계에서도 일부 대형 플랫폼 회사들이 알고리즘 편향 및 설명 가능한 인공지능 관련 점검 도구들을 개발하여 이미 사용 중이다. 이는 인공지능윤리 이슈들이 전 세계적으로 사회적 문제를 발생시킴과 동시에 산업과 비즈니스 차원에서도 비생산적 결과를 산출하였기 때문에 그 해결책으로 관련 이슈를 기술로 구현해야 할 수 밖에 없었기 때문이다.

이러한 맥락에서, 인공지능윤리 이슈들을 인문학에서 다루어지는 윤리이론과의 접점으로 해석하는 접근도 있으나 인공지능을 활용하는 로봇의 움직임을 제어할 때의 규칙 적용 방식들 중 하나로 윤리이론을 적용하는 제한적 해석이 현실적이다. 특히 이 논문에서 다루어질 인공지능 편향(bias)은 복합적 사안이다. 상위개념적 차원에서는 인간의 도덕심리의 측면으로 다루어질 수 있지만 인공지능의 차원에서는 시스템 안의 변수들로 표상되는 요소들을 사회의 기준에 어울리도록 성능을 최적화하는 사안이며 이 과정에서 데이터를 만들어내는 인간과 기계간의 상호작용, 그리고 이를 위해 필요한 이해관계자들 간 조정과 관련된 거버넌스 사안이기도 하다. 이 논문에서 다룰 주제인 편향(bias)은 인공지능윤리의 여러 이슈들 중 하나로 간주되지만 인공지능 기술과 한 몸으로 다뤄져야 할 토대적 주제이다. 투명성(transparancy) 이나 설명가능성(explainability) 이슈는 인공지능 구성 후의 문제들인 반면, 편향은 그에 앞서 갖추어야 할 출발점으로 먼저 언급되어야 할 대상이다. 이 논문에서는 개념적 측면보다는 실제적 측면에 더 방점을 둔다. 먼저 인공지능 편향의 특성 및 인간 편향과의 차이를 설명하고, 인공지능 구성 단계에 따른 편향의 개입양상과 오픈소스로 접근하여 사용되고 있는 편향완화 기술들을 소개함으로써 인공지능윤리의 기본 이슈인 편향완화에 접근할 수 있는 방안 그리고 이를 위한 교육의 방향을 모색할 것이다.

인공지능 편향의 특성

편향은 편견과 어떻게 다른가

'편향(bias)'이란 '편견(prejudice)'과 다소 다른 의미를 지닌다. '편견'은 인간 개개인의 국한된 경험 때문에 가지게 되는 심리적 고정관념을 의미한다. 인간은 제한된 감각능력과 시공간적 경험을 가지기 때문에 대상을 타인과 동일한 각도나 방식으로 볼 수 없고 이 점이 개인의 주관적 생각과 개성을 만들어낸다. 어느 인간도 이러한 편견에서 자유로울 수 없다. 인공지능이 사회적으로 문제가 되는 의사결정을 내리게 되는 경우 그 의사결정의 원인은 '편견' 때문이 아니다. 편견을 가지려면 '주관성'이 있어야 하는데 이를 어떻게 정의하는가에 따라 논의의 결론은 달라질 수 있다. 그러

나 현재 '주관성'이란 것이 어떤 것인지, 적용 범위가 어디까지인지에 대한 단일한 정의가 없다는 점 그리고 인간 중심으로 정의가 내려지므로, 인공지능은 주관적 관점을 가진다고 하기 어렵다. 주관성 개념을 넓게 해석하여 인공지능에 적용한다고 할지라도, 즉 인공지능이 독자적인 의사결정을 내리고 이에 기반하여 자신만의 편견을 가진다고 해도 이는 윤리적 문제가 되지는 않는다. 예컨대 노인에게 특화된 세부 목적을 가진 인공지능로봇이나 개인맞춤형 인공지능은 모든 유형의 내용이 아니라 제한되고 필요한 내용을 중점적으로 하여 구성되므로 일종의 주관성이나 편견을 나름대로 가진다고 비유적으로 이야기할 수 있지만, 그 목적의 특수성 때문에 윤리적으로 문제가 되지 않는다.

'편견'이 주로 심리적 차원에서 가지게 되는 상태인 반면 '편향'은 통계 분석에서 데이터 모집단 및 분석 과정에서 발생한다. 비유컨대 만약 한 사람의 뇌가 어떤 사고로 인해 어릴 때부터 손상되어 사람들을 책상으로 인식하는 비정상적인 상태라면 그 사람은 사람들에 대해 편견을 가진 것이 아니라 책상으로 생각하는 '편향'을 가지고 있는 것이다. 인공지능을 구성하는 과정의 한 절차인 기계학습은 데이터 안의 비선형적인 패턴을 인식하기 위해 통계테크닉이 활용되는 절차로, 사용하는 모델이 복잡해질 수록 편향은 작아지고, 분산은 커진다. 시스템에서의 오류를 최소화하는 방법은 편향을 작게 하는 것이 아니라 편향과 분산의 합을 최소화하는 것이므로 편향이 전혀 없을 수는 없다. 또, 인간이 직접 인터넷 등에 입력하거나 전자기기를 사용함으로써 생겨나는 데이터들 자체가 인간의 편향을 그대로 반영하고 있고 이 데이터들은 인공지능을 만드는 산업체들에서 활용하므로 이를 사용하는 일반인들은 편향의 영향을 그대로 혹은 증폭되어 받게 된다.

인공지능 편향에 자율지능시스템의 의미가 왜 핵심인가?

현재 인공지능은 과거의 규칙 기반의 자동시스템으로부터 일종의 자율시스템으로 발전했다. 이 특성이 기존에 없던 편향, 블랙박스 문제 등의 윤리적 측면을 인공지능에 내재하게 한다. 이는 기존 컴퓨터 윤리, 기술윤리에서 흔히 언급되어왔던 프라이버시 등의 윤리 문제와는 다소 다른 차원의 맥락이라는 점을 주지할 필요가 있다. '인공지능' 대신 사용되는 '자율지능시스템(autonomous Intelligent system)'이라는 더 구체적 표현에서 알 수 있듯, 인공지능은 더 이상 단순한 기계가 아니라 인간의 일부 '의사결정'을 대신하는 복잡한 시스템이다. '자율지능시스템'에서 '자율'은 "복잡한 환경에서 복잡한 임무를 수행하기 위해, 스스로 인식하고, 계획하고, 학습하고, 진단하고, 제어하고, 중재하고, 협업하는 등 다양한 지능적 기능들을 가지는 시스템"으로 정의된다.'자율'이란 표현은 외부 개입 없이 스스로 학습하는 능력을 가리킨다. 그런데 여기서 '자율'개념은 인간에게 부여하는 자율성과 동등한 의미가 아니다. 인간에게 적용하는 자율성 개념과 기계에 적용하는 자율성 개념은 다르다는 점을 주의해야 한다. 그렇지 않으면 무인자동차를 때로 '자율주행차'로 칭할 때, 그리고 인공지능을'자율시스템'으로 언급할 때 자칫 인간의 자유의지나 자율성과 유사한 심적 상태가 기계에도 있다는 엉뚱한 가정을 하게 된다. 물론, 인공지능이 인간과 같은

자유의지나 자율성을 지닌다고 할 수 있는지에 대한 물음이 있으나 이는 미래의 인공지능 기술수준인 '범용인공지능'(AGI: artificial general intelligence)이나 인간수준을 넘어선 지능 수준인 '초인공지능'(ASI: artificial super intelligence) 수준에서 논의될 사안이며 현재 개발 및 사용되고 있는 영역별 인공지능인 '좁은 인공지능'(ANI: artificial narrow intelligence) (Joshi, 2019)의 사안은 아니며 그와 구분되어야 한다. 자율지능시스템으로서의 인공지능을 이 논문에서 논할 때는 사변적 가설의 논의보다는 후자의 좁은 인공지능에 국한하여 현 시대의 인공지능 수준과 관련되는 편향을 다룬다.

'자율' 시스템으로서의 인공지능 학습방식이 인공지능 편향과 어떤 관련이 있는가를 보기 위해 인공지능의 간략한 역사를 볼 필요가 있다. '인공지능'이라는 용어는 튜링의 초기 컴퓨터와 계산기계의 지능이라는 아이디어가 제시된 후 1956년 여름 미국 다트머스(Dartmouth) 대학에서 열린 학술회의에서 존 매카시(John McCarthy), 마빈 민스키(Marvin Minsky), 클로드 섀넌(Claude Shannon) 등의 학자들이 모여 컴퓨터의 추론과 탐색에 대해 대화하면서 지능의 요소들을 형식화하는 연구를 하고자 하는 취지로 만들어졌다. 지능을 기계적 계산과정으로 설명할 수 있다는 아이디어는 계산을 수행하기 위한 '기호'와 이 기호들을 계산할'규칙'을 통해 가능하기 때문에 초기 인공지능은 규칙기반(Rule-based AI)로 발전하였다. 대표적 예가 미국 퀴즈쇼에서 광범한 지식들로부터 빠르게 해답을 찾아내어 일등을 거머쥔 인공지능 '왓슨(watson)'이다.

반면, 현재 개발 및 활용되는 인공지능은 '규칙기반' 뿐만 아니라 뇌신경망을 모형으로 발전된 '인공신경망(artificial neural network)' 기반의 인공지능도 사용하며, 과거에 해결할 수 없었던 컴퓨터 파워와 충분히 많은 양의 데이터에 힘입어 발전하고 있다. 기존의 규칙기반 인공지능은 규칙과 입력값이 주어지면 그에 해당하는 출력값을 내어놓는 일종의 자동시스템이다. 이와 반대로 신경망 기반 인공지능은 자료들이 주어지면 자료들 안의 패턴이나 규칙을 찾아내는 방식이다. 이러한 기계학습을 통해 규칙기반 인공지능에서는 난제였던 패턴인식 등이 가능하게 되어 기계가 사물의 특징을 파악할 수 있게 되었다. 인공신경망 학습 유형은 학습에 대한 정답이 있고 없음을 기준으로 다시 지도 학습(supervised learning) 기법과 비지도 학습(unsupervised learning) 기법으로 나뉜다. 기법의 차이는 학습에 대한 정답이 있고 없음에 있다.

규칙기반의 인공지능이 주류일 때에는 컴퓨터 사용과 관련한 윤리 쟁점들, 예컨대 정보보안, 개인정보, 프라이버시 이슈들이 있었다. 반면, 현재의 인공지능윤리 이슈들은 딥러닝 기반의 인공지능 시스템 구성에서 제기되는 새로운 차원의 사안들이다. 정보보안이나 프라이버시와 관련한 문제들은 컴퓨터가 규칙기반의 자동시스템이든 현재 발전된 자율시스템이든 관련없이 제기되는 문제들이다. 반면 자율 시스템으로서의 인공지능의 내부 특징인 '블랙박스'문제 때문에 생기는 윤리적 사안들이 있다. 기계학습을 통해 인공지능이 구성될 때 인간은 최소한의 제한 규칙만 제공할 뿐 정해진 입력-출력의 법칙이 존재하지 않는다. 정해진 알고리즘에 따라 움직이는 기존 컴퓨팅 방법을 넘어서는 학습 능력을 갖추기 위해 기계학습의 여러 방법이 사용된다. 기계학습의 일종인 심층학습(deep-learning)에서 인간은

적은 수의 코드만 설계한다. 심층 신경망에서 학습을 통해 만들어지는 층과 연결망은 인간이 모두 입력한 것이 아니다. 그런데 어떻게 특정 인종이나 성, 계층 등에 대한 선호나 편향이 개입될 수 있을까. 심층학습 프로그램의 학습 방법은 뇌신경세포인 뉴런이 학습하는 방식과 유사한 점이 있다. 연결망은 아래층의 출력이 위층의 입력으로 이어지는 방식으로 구성되어 있다. 심층학습은 받아들인 데이터에서 패턴을 찾아내 연결망 사이의 연결 강도를 변화시키면서 차원을 높여 학습해가는 방식이다.

자율시스템의 이러한 학습 과정은 인간이 예측하지 못하는 의사 결정으로 이어진다. 이런 방식으로 학습하고 훈련한 과정을 거쳐 나온 최종 출력에 대해서는 심층학습 설계자조차도 어떻게 그러한 최종 판단이 나왔는지 모르게 된다. 이렇게 인공신경망에서 이루어진 학습 과정의 내부를 알 수 없다는 의미에서 인공지능은‘블랙박스(black box)’로 불린다. 이때 블랙박스라는 표현은 자동차의 블랙박스와 달리, 안의 상황을 들여다볼 수 없다는 의미로 사용된다. 이러한 블랙박스 상황 때문에 사람의 생명과 인권을 다루는 인공지능 의사나 인공지능 변호사의 위험한 의사 결정을 방지하기 위해 인공지능이 내린 의사 결정의 근거를 알 필요가 있다. 인공지능 구축에 필요한 데이터와 알고리즘은 기계 안에서 작동하기 때문에 일견 인간의 기호나 정치적 견해 등과 관련 없는 것으로 오인하기 쉽다. 그래서 인공지능 판사나 인공지능 의료 진단기가 어떤 진단을 내릴 때 인간보다 더 객관적이고 더 믿을 만한 것으로 착각할 수 있으나, 현재는 인공지능을 활용한 시스템에 사용되는 데이터나 알고리즘을 인간이 만들어내고 조정할 수 있다는 사실을 알아가고 있다. 인공지능의 최종 출력은 컴퓨터 자체가 하는 일이라기보다는 인공지능이 학습하는 데이터와 기계학습 과정에서 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.

인간의 편향과 인공지능의 편향

인공지능이 자율적 학습을 할 경우 야기되는 윤리적 문제들이 모두 인간이 의도한 것이라고 보기는 어렵다. 인간이 고의로 인공지능을 악용하는 사례도 있겠지만, 이 경우는 인간이 자율적이고 의식적으로 행동한 것이므로 기존의 법률 체제 안에서 법적 조치를 취하거나 윤리적 평가를 내리기가 비교적 용이하다. 법적 판결이나 윤리적 평가는 행위 당사자가 자유의지와 의도를 가진다는 가정을 평가의 전제 조건으로 삼기 때문이다. 반면, 인공지능에 필요한 최소한의 알고리즘을 설계할 때 의도적이지는 않아도 설계자의 문화적, 개인적 편향이 변수나 가중치 설정에 반영되며, 기계학습을 위한 데이터 선정에도 데이터를 만들어낸 일반인들의 편향이 개입될 수 있다. 인공지능으로 데이터를 학습하고 이를 의사결정에 활용할 경우 상술한 블랙박스 상황 때문에 문제 상황의 원인을 찾아내거나 그에 책임을 지우기가 어려워진다. 최종 출력단계에서 곧바로 문제가 발견될 수도 있지만, 기계학습 과정이나 인간의 알고리즘 설계, 데이터 수집

및 전처리를 하는 단계에서 발생한 것 혹은 이들 중 일부 요소들의 결합에서 그 문제의 원인을 찾을 수도 있다.

그렇다면 인공지능 구축 단계들 중 데이터 수집, 선택, 전처리, 알고리즘 설계 단계에서 의식적, 무의식적으로 인간의 편향이 개입되었다고 가정하자. 이 때 인간의 편향과 최종적으로 구축된 인공지능에서 나타나는 편향은 유사할까? 기계학습을 위해 데이터를 수집하고 훈련데이터를 선택, 데이터를 전처리하는 과정, 그리고 주요 변수의 중요도를 조정하는 작업은 인간이 수행하므로 인간심리에서 발생하는 종류의 편향들이 기계학습을 통해 만들어진 인공지능에 내재할 수 있다. 예컨대 확증편향(confirmation bias) (Gale, M. et al., 2002)이 기계학습을 통해 나온 결과물에 반영될 수 있다(김인식 외, 2021). 확증편향은 기존의 관점을 반영하는 방식으로 데이터를 선택하고 분석하는 데에서 나온 편향을 의미한다. 이러한 편향이 개입된 데이터를 사용하여 기계학습 할 경우 그 편향이 다른 편향과 결합하여 증폭되는 결과를 낳는데, 예컨대 검색 엔진이 검색자의 견해를 강화하는 효과를 낳는다는 '필터버블 가설'(Pariser, 2011)과 관련된다. '표본 편향', '선택 편향'은 유사한 종류의 편향으로 분석을 위해 데이터를 선별하거나 대표성이 없는 샘플을 부주의하게 선택할 때 발생하는데, 기계학습에 사용될 훈련데이터(training data)를 고를 때 데이터 소스, 샘플의 대표성 여부에 따라 인간 편향에서와 마찬가지로 선택 편향이 개입될 수 있다.

사회심리학에서 논의되는 타인종효과(cross-race effect)(Beaupré MG, et al., 2006) 또한 기계학습을 통해 만들어지는 인공지능에서 보여질 수 있다. '타인종효과'란 인간경험이 특정 인종에 더 친숙하여 타인종의 인간적 특성들을 알아차리지 못하는 현상으로 인간이 가진 편향이다. 그러나 인간 경험이 특정 인종에 더 친숙해져 지각에 영향을 미치는 것과 마찬가지로, 기계학습에서 특정 인종의 특성 데이터를 주로 기계학습하게 되면 최종 결과에서도 마찬가지의 타인종효과가 나타난다. 예컨대 기계학습에 사용되는 훈련 데이터에서 소수 인종 집단에 속하는 사람 얼굴을 혼동할 가능성이 더 높다는 점(Martineau, 2019)이 발견되었다. 이러한 기계의 편향은 기계학습에서 사용하는 데이터가 인간의 편향을 반영하기 때문이다. 실제로 구글 이미지를 검색해보면 백인의 얼굴 이미지가 흑인 얼굴 이미지보다 양적으로 월등하게 많다. 이는 인공지능 시스템을 만들 때 흑인 얼굴에 대한 가중치를 적게 설정한 것과 마찬가지의 결과를 초래하며 결국 사회에서 사용될 인공지능 시스템의 정확도나 성능을 저하시키게 된다.

데이터 편향의 이러한 특성은 통계에서의 역설인 '심슨 역설'(Simpson's paradox)에서처럼 데이터 전체를 단위로 하여 얻어지는 결론만 관찰하고 더 작은 단위로 하여 얻어지는 변수들을 신경쓰지 않을 때 오류가 발생하는 상황과 유사하다. 심슨의 역설로 잘 알려진 사례는 대학원 입학에서 성비 사례이다. 1973년 초 UC 버클리 대학의 대학원 입학 시 지원자 수 대비 입학률을 보면, 12763명의 지원자가 101개 학과 및 학과간 전공 중 하나에 지원하였는데, 지원한 4,321명의 여성 중 대략 35%가 합격한 반면 지원한 8,442명의 남성 중 44%가 합격했다. 상위 6개 학과의 입학 결정을 보아도 남성의 합격률은 약 44%인 반면, 여성의 경우 약 30%에 불과해 여성에 대한 차별이 의심되었다. 그러나 이는 전체 단위에

서의 평균만 고려하였기에 내려진 결론이며, 표준 통계적 유의성 검정에 따르면 기본 합격률에 차이가 없는 경우였고, 학과별로 나누어 살펴보면 실제로 여성의 지원자 수 대비 입학률이 더 높음을 볼 수 있던 사례이다. 데이터를 관찰하는 표본을 어떻게 선택하는가에 따라 분석결과가 크게 달라진다. 심리학의 조사방법론에서 주요 변수를 고려하지 못하면 잘못된 분석결과를 산출하듯, 기계학습에 사용되는 훈련데이터 역시 해당 도메인에 대한 지식이 충분하지 않거나 직관을 가지고 있지 않으면 데이터를 잘못 분류할 수 있다. 이렇게 인공지능을 구축하는 과정에서 인간이 만든 데이터들을 사용하기 때문에 인간의 편향이 기계의 편향에 그대로 반영되거나 증폭될 수 있다.

다만, 기계는 인간과 달리 기억의 한계가 없으므로 기억과 관련되는 한 인공지능 편향은 없다고 생각할 수도 있다. 그러나 인공지능의 학습과정에서 인간과 마찬가지로 망각이 존재한다. 실제로 인공신경망에서는 새로운 것을 학습하는 과정에서 이전에 학습한 내용들에 대한 가중치가 바뀌어 낮아지면서 인공지능의 성능이 떨어지는데 이를 '파괴적 망각(catastrophic forgetting/interference)'이라고 한다. 이 점을 극복하여 생산성을 높이는 것이 하나의 과제(Kirkpatrick, et al., 2017)로 여겨져 왔다. 다른 한편으로는 거꾸로 이러한 망각 메커니즘을 이용하여 개인정보가 포함된 데이터 등 편향이 개입될 수 있는 내용들에 가중치를 적게 할당함으로써 과거의 정보가 망각될 수 있는 '의도적 망각'(Timm et al., 2018) 기술이 있으며 이를 활용하면 인공지능 시스템 안에 있지만 명시적으로 드러나지 않아 편향이 증폭될 수 있는 확률을 낮출 수 있다.

인공지능 구성 절차단계에서의 편향

인공지능 편향은 인공지능을 구성하는 절차의 단계마다 발생할 수 있다. 이를 위해 간단히 인공지능의 구축 절차를 살펴보자. 인공지능은 기계학습을 통해 모델을 훈련하여 예측 시스템을 구축하게 되는데, 기계학습은 주어진 데이터를 바탕으로 모델의 변수(parameters)를 조정하여 최적의 변수를 찾아가는 과정이다. 이때 변수가 많을수록 학습이 어렵고 과적합(overfitting) 등의 문제가 있어서 예측값과 실제 정답값 사이의 차이 혹은 오차(목적함수)를 최소화하는 변수를 찾는다. 예컨대 숟가락과 젓가락을 구분할 수 있는 인공지능의 분류작업은 선형분류로 가능한 반면, 이세돌 9단을 이긴 알파고의 경우 바둑의 다음 수를 어디에 둘 것인지 등의 결정은 비선형분류가 필요하다. 후자의 경우 사례들로부터 학습하는 인공신경망에 뿌리를 둔 딥러닝(deep learning) 즉 심층학습에서 최적화된 해를 반복적으로 찾는 방식으로 인공지능이 구성된다. 심층학습은 과거에는 다른 영역보다는 시지각에서 주로 우수한 성능을 보였고(김청택, 2019) 현재 딥러닝 알고리즘이 발전하여 수많은 영역에서 눈에 띄게 발전하고 있다.

인공지능의 구성단계는 그림 1과 같으며, 데이터수집 단계, 라벨링과 같은 데이터 전처

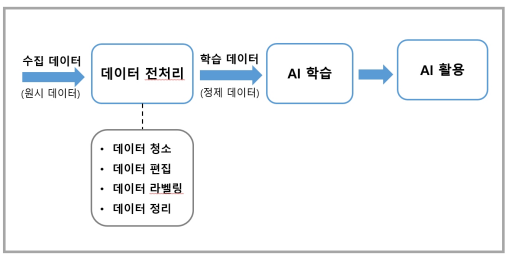

그림 1. 인공지능의 구성 단계

리 단계, 기계학습 단계, 이후 모델의 채택 등의 각 단계마다 편향이 개입되는 내용은 구체적으로 다음과 같다.

데이터 수집 단계에서의 편향

인공지능 구축 절차 중 첫 단계인 데이터 수집과 선정은 전체 단계들 중 편향과 관련하여 가장 큰 영향을 준다고 할 수 있다. 데이터 관련 단계는 다시 데이터 청소(cleaning), 데이터 편집, 데이터 라벨링 등으로 나뉘며 이 단계 각각마다 편향이 발생할 수 있다. 학습 데이터를 전처리하는 작업은 인공지능 구성의 전체 과정에서 80퍼센트를 차지한다고 할 정도로 큰 비중을 차지하며 사람이 대부분 개입하므로 사람의 편향이 그대로 전이 및 반영된다. 따라서 차후 단계들에서 변수와 가중치를 조정하거나 시스템이 구축된 후 사후 조치를 취하는 방법보다 기계학습의 원재료를 다루는 단계에서 조정하는 것이 근본적인 편향완화의 효과가 높을 수 있다. 이는 서베이나 여론조사에서도 유사하게 나타날 수 있는 문제로 선택편향과 관련된다. 서베이에서는 표본추출(Sampling)이 중요한데, 인구의 특정 연령이나 특정 집단에만 편중된 표본 추출을 하면 그 조사 결과는 편향성이 있다. 이는 고의적일 수도 있지만 자신에게 익숙한 상황을 자연스럽게 받아들여 이와 관련한 데이터를 더 많이 수집하는 경향과 관련이 있다. 이와 마찬가지의 편향이 기계학습에 사용될 데이터 수집에서도 나타난다.

기계학습을 통해 완성된 구글 포토 서비스는 사람들의 얼굴을 자동 인식해 이름이나 분류를 붙이는 기능이 있는데 이와 관련된 편향 이슈가 발생했다. 사건은 2015년 구글 포토 서비스를 이용하는 흑인 프로그래머 앨신이 흑인 여성 친구와 함께 찍은 사진을 인터넷에 올리면서 발생했다. 공개된 사진에는 앨신의 여성 친구 얼굴 아래에 사람 이름이 아닌 '고릴라'라는 이름이 달렸다. 트위터를 통해 퍼져 논란이 되자 구글은 '프로그램 오류'라고 사과했고 프로그램 개선을 약속했다. 그러나 당시 구글은 근본적 개선보다는 고릴라로 태그하는 부분만 수정했다고 알려졌다.

이 사건이 프로그램 오류 때문에 발생했을까? 사람들의 얼굴을 학습하여 적절한 태그를 붙이는 연습을 하는 기계학습이 어떤 과정을 거쳐 고릴라라는 잘못된 태그를 붙일 수 있는지 생각해볼 수 있다. 기계학습이 가능하려면 먼저 훈련데이터가 있어야 한다. 어떤 사람들의 얼굴을 기계학습의 훈련 데이터로 삼았는지 검토할 수 있다. 사람 얼굴 데이터는 인터넷에서 이미지 파일들을 많이 발견할 수 있다. 구글 검색사이트에서'Americans' 혹은 '미국인'을 입력해서 이미지 결과를 보면 대다수가 백인 이미지이다. 구글 포토서비스를 만들 때 의도적으로 백인 얼굴 위주로 얼굴인식 인공지능을 학습시킨 것은 아니지만 우리가 가진 사람 얼굴 데이터의 대다수가 백인이므로 흑인 얼굴에 대한 패턴인식의 정확도가 잘 달성되기 어렵다. 데이터 분석에서 흔히 거론되는 'garbage in, garbage out'의 경우이다. 이미지 데이터들을 수집하고 그중 학습시킬 이미지들을 분류할 때 흑인의 얼굴 이미지 수가 절대적으로 부족하다면 충분한 수의 얼굴 이미지를 추가하여 기계가 이미지를 충분히 학습하여 패턴을 잘 인식하도록 조정하는 작업이 필요하다. 구글의 고릴라 태그 사건과 유사한 사례로 니콘 카메라 얼굴 인식 시스템이 아시아인 피사체에 대해'눈을 감았다'고 인식

한 사례가 있다.

이러한 사례는 보호집단(Protected Group)으로 간주되는 소수 인종뿐만 아니라 소수자 성(gender), 나이, 문화에서의 소수자 등도 포함된다. 2019년 연구(Vries et al., 2019) 에서는 가구와 같은 대상 인식 서비스를 통해 54개 국가를 비교하여 평가했다. 54개국에서 가정용품을 인식하는 이미지 인식 시스템을 비교 평가한 결과, 저소득 국가의 가정용품 이미지의 정확도가 낮게 나타났고 국가 간 정확도에 큰 차이를 보였다. 인종 간 데이터 차이나 물품 이미지의 정확도 차이는 해당 종류의 데이터의 양과 비율을 보충하면 해결될 것처럼 보인다. 그러나 주어진 시간과 재원을 고려하여 대표적 샘플들을 주로 고려할 때, 샘플의 '대표성' 또한 논의의 대상이 된다. 의도된 인종 차별은 아니지만 인공지능이 학습하는 데이터는 그 양에 있어서 특정 집단, 특정 속성에 치중될 가능성이 높다. 이 편향은 인공지능이 어서 가지게 되는 편향이라기보다는 인간 사회가 이미 가지고 있는 데이터의 불균형이 반영된 결과이다. 만약 데이터를 수집하고 기술 구축 및 구현을 담당하는 팀에 다양성이 부족할수록 부정적인 결과의 가능성이 커진다. 예를 들어, 특정 집단과 관련한 데이터가 의도치 않게 배제되면, 데이터가 모델 학습에 사용되는 방식 때문에 특정 집단이 상당히 불리한 위치에 처하게 될 뿐만 아니라 궁극적으로 데이터 품질에 영향을 주어 인공지능의 성능이나 정확도가 떨어지게 된다.

이미지 및 영상 데이터뿐만 아니라 언어 데이터에서도 유사한 결과가 나타난다. 언어 데이터의 경우 보통 기존에 사람들이 사용해왔던 글들을 가공한 말뭉치(corpus)를 통해 기계학습한 인공지능 시스템인 번역이나 글 생성 시스템에서 볼 수 있다. 인공지능을 만들기 위해 활용하는 기존의 글이나 책에서 남성과 여성을 나타내는 대명사, 명사, 형용사의 비율에서 남성 쪽 단어의 비율이 현저히 높게 나타났으며, 직업 유형이 언급되는 글에서도 남성이나 여성에 연결되는 비율이 현저한 차이가 있었다(Prates, et al., 2020). 온라인 번역에서는 중립적인 대명사가 예컨대 "그는 의사다, 그녀는 간호사다"처럼 특정 젠더로 번역되는 경향(Barocas, et al., 2017)이 나타났다. 주지할 점은 인공지능윤리 이슈가 현재 중요한 반면 데이터 종류에서의 불균형이 항상 부정적 의미의 편향을 함의하는 것은 아니라는 점이다. 인공지능이 특정 영역 혹은 특정 회사 혹은 사람에게 맞추어 만들어지는 '맞춤형'의 경우 목적에 따라 특정 종류의 데이터를 더 많이 수집 및 활용하는 것이 필요할 수 있다.

데이터 라벨링 과정에서의 편향

데이터 라벨링에서의 편향은 부적절한 도메인 지식으로 잘못 라벨링될 때, 그리고 기계학습시 사용될 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 분류하여 주석을 다는 작업이다. 먼저, 부적절한 도메인 지식으로 잘못 라벨링될 때 예컨대 이미지 인식 인공지능의 목표가 고양이 이미지를 고양이로 인식하게 하는 것이라면 한 이미지가 고양이인지 고양이와 비슷하게 생긴 동물인지를 인식해서 고양이는 '고양이'로 정확하게 라벨링이 되어야 한다.

특히 텍스트의 경우는 글의 내용이나 인공

지능 활용 목적에 따라서는 고난도의 개념 이해가 필요할 수 있어서 특정 분야에 대한 지식, 즉 도메인 지식이 요구된다. 예를 들어 20-30대 연령층에서만 사용하는 신조어가 포함된 글을 기계가 이해가능하도록 라벨링해야 한다. 이 경우 신조어의 의미를 파악하지 못하면 적절한 라벨링이 이루어질 수 없고 텍스트 정보가 기계에 인식될 수 없어 인공지능이 해당 글에 대한 적절한 이해를 할 수 없다. 더구나, 텍스트의 경우 동음이의어나 맥락에 따라 동일한 표현도 다르게 해석이 되기 때문에 특정 텍스트가 어떤 맥락에서 어떤 의미로 파악되는지를 세밀하게 알고 이해할 수 있어야 데이터 라벨링과 정리가 가능해진다. 만약 20대가 사용하는 신조어의 의미를 50대 연령의 라벨러가 정확히 이해하지 못한 채로 글에 라벨을 붙일 경우 해당 텍스트는 20대 연령층에 사용되기에 부적합한 텍스트 데이터가 된다. 언어처리를 하는 인공지능을 구성할 때 글의 특성에 따라 해당 글의 맥락을 잘 이해할 수 있는 사람들이 라벨링의 일을 담당해야 한다. 또, 특정 목적 예컨대 소비자에 맞는 책을 추천하는 인공지능시스템의 경우, 소비자가 과거에 고른 책 제목이나 내용의 키워드와 밀접하게 연관되어 있는 표현들을 분류하는 방식으로 라벨링이 이루어진다. 이러한 작업은 데이터에 포함된 의미를 이해하고 이를 책 추천이라는 목적에 맞게 어떻게 연결해야 하는지의 태그(tag) 기준이 따로 마련되어 라벨링은 이러한 세부 목적에 맞게 수행된다. 이러한 도메인 지식의 중요성은 거의 모든 영역에서 강조될 수 있다. 적절한 도메인 지식을 갖추지 않은 사람이 라벨링을 하게 되면 데이터 편향으로 인한 최종 구축된 인공지능 시스템의 의사결정의 문제점뿐만 아니라 정확도 또한 낮아진다. 또, 여러 유형의 라벨들 간에 비율 차이가 크면 기계학습에서 균형잡힌 학습이 불가능하여 최종 구축되는 인공지능 시스템의 특정 목적을 달성하기 어려워진다.

데이터 처리 과정에서의 도메인 지식의 중요성을 가장 극명하게 보여주는 영역은 의료이다. 의료사진에 나타난 종양이 폐암인지를 식별하는 인공지능을 구성하기 위해서는 폐암인 사진과 아닌 사진을 학습시켜야 한다. 이미지에 라벨을 붙이는 작업은 단순 반복적이어서 교육만 받으면 쉽게 할 수 있다고 생각할 수 있다. 그런데 폐암인 사진을 구분하여 태그를 붙이는 것은 의학 훈련을 받지 않은 일반인은 할 수 없고 의사가 할 수 밖에 없다. 인공지능 설계 전반에 걸쳐 도메인 지식의 중요성은 인공지능이 단순한 공학적 시스템이 아니라 인문, 사회 등 거의 모든 영역의 명시적, 암묵적 지식이 필수적이라는 점을 보여준다. 이 점을 최근 마이크로소프트에서는 다음과 같이 표현한다. “인공지능 시스템은 기술 뿐만 아니라 사용자들, 영향받는 이들, 배포된 환경 모두를 포함한다”(Microsoft, 2021). 편향이 인공지능 시스템 구성의 여러 절차들 각각에 내재하는 만큼, 이에 개입하는 인간 또한 인공지능 시스템의 구성요소에 속한다.

모델링 단계에서의 편향

인공지능 구축에 필요한 기계학습을 위해서는 라벨드데이터(labelled dataset)가 필요할 뿐만 아니라 예측 정확도 및 성능을 높이기 위해 변수를 튜닝하는 작업을 진행하게 된다. 모델링 단계에서의 편향은 인공지능이 예측하고자 하는 목표와 관련성이 없거나 적은 변수가 포함되어 있을 경우, 중요도가 상대적으로

낮은 변수에 높은 가중치가 주어진 경우, 그리고 데이터셋에 인종, 젠더, 나이 등에서 불균형한 데이터가 있는 경우 발생할 수 있다. 차별받는 소수자 보호를 위해 미국 법률은 어떤 사람을 인종, 젠더, 나이, 결혼 여부, 시민권 소지 여부, 임산부, 인종, 종교, 성적 지향성 등의 특성들을 이유로 차별하는 경우 위법으로 간주하며, 이러한 특성들을 '보호속성'(protected attributes)라고 한다. 모델링 단계에서는 이러한 보호 속성들이 특정 속성과 통계적 종속성을 가지는 경우 편향이 생긴다.

먼저 변수 및 가중치 설정에 편향이 있는 것으로 밝혀졌던 사례로는 알고리즘 편향으로 종종 언급되는 사례인 미국의 재범예측 인공지능 컴파스(Compas)가 있다. 이 시스템은 현재 미국 법원에서 판결의 근거자료로서는 아니지만 참고자료로 사용되고 있다. 컴파스는 위스콘신주에 거주하는 에릭 루미스가 '재범율'이 높다고 판단했고 이를 참고하여 법원에서는 루미스가 2013년 총격 사건에 사용된 차량을 운전한 것에 대해 징역 6년을 선고했다. 에릭 루미스는 인공지능 컴파스가 내린 높은 재범율의 판단기준을 문의했으나 답변이 없었고 회사도 어떤 요소에 어떤 배점이 내려지는지에 대해 공개하지 않았다. 그러나 기계학습 연구자들이 모형 추출과 비교 접근법을 사용하여 범죄위험도를 컴파스와 유사하게 흉내내어 검사해본 결과, 컴파스는 범죄 이력이 있는 이들 중 백인이나 부자보다는 흑인이나 빈민촌 거주자들의 재범율을 더 높게 판정한다(Tan, et al., 2017)고 조사되었다. 또, 모방 모델을 만들어 비교 분석한 결과, 컴파스는 피고인의 나이, 성별, 이전 유죄 판결 건수, 이전 교도소 체류 기간처럼 직접적으로 범죄와 관련된 개인정보 뿐만 아니라, 범죄와 직접적으로 관련이 없는 성격, 가족, 대인관계, 라이프스타일에 대한 정보 또한 사용했다(Skeem, et al., 2016)는 점이 밝혀졌다. 나이나 성별 그리고 이전 유죄 판결 건수는 이미 일부 인종에 편중되어 있기에 사실을 반영하는 데이터라고 할 수 있지만 이 데이터를 사용하여 인공지능을 구축할 경우 컴퓨터는 자동으로 특정 인종에 편향되게 판단할 가능성이 높다. 또한 범죄와 직접 관련이 없는 요소인 성격이나 가족 등의 요인에 가중치를 어떤 비율로 할당했는지에 따라 재범률이 잘못 계산될 수 있다.

이 같은 사례는 평가 시스템에서도 나타난다. 최근 국내 회사들에서 도입하고 있는 인공지능 영상면접 인공지능 구축에 있어 기계학습을 위해 수집하는 데이터는 그 회사에서 성공적인 실적을 올린 이사급 직원들의 데이터이다. 이사급 직원들의 연령은 50-60대로, 면접후보군인 20대의 특성을 반영하지 못한다. 게다가 표정이나 태도 분석을 포함하는 경우 맥락에 따라 같은 표정이 다른 감정을 표현할 가능성도 있어서 적절한 사전면접이 되기 어렵다. 미국에서 교사평가를 하는 데 활용했던 교육부가가치평가시스템(EVAAS) 프로그램은 그 기준과 가중치를 공개하지 않았고 시민권 침해 우려로 결국 소프트웨어 사용이 중단되었다(Sample, 2017). 또, 비의도적이지만 불가피하게 어느 특정 집단이나 사람에 편향되는 결과를 낳을 수 있다. 알고리즘 설계에서 결과를 나열할 때 가장 일반적으로 사용하는 기준은 가나다 순서, 알파벳 순서다. 이 방식으로 순서를 나열하는 것은 편의상 많이 사용된다. 이 기준을 따르면 호텔 검색 사이트에서는 'ㄱ'으로 시작하는 호텔들이 소비자에게 더 우선적으로 노출될 수밖에 없다. 이

는 불가피한 편향이지만 순서를 번갈아가며 보이는 방식의 완화방안이 필요하다.

마지막으로, 보호 속성(protected characteristic)과 특정 속성이 통계적 종속성을 가지는 경우에 편향이 나타난다. 예컨대 '여성'은 검색이나 말뭉치에서 '비서'와 종종 함께 나타난다. 이러한 편향적 연결관계를 완화하기 위해 말뭉치 상에서 두 속성 간의 연결고리를 삭제하는 방식을 가장 간단한 해결책으로 고려할 수 있다. 최근 구글은 얼굴인식 프로그램에서 '남성'인지 '여성'인지를 구분하는 라벨을 제거하는 방법(Dave, 2018)으로 편향완화의 노력을 보였다. 하지만 텍스트의 경우 데이터 내부의 요인들이 상호관련되어 있어 삭제방법은 오히려 부작용이 있다. 예컨대, '여성'은 '비서'와 관련없어 보이는 다른 속성과도 연관되어 있기 때문에 여성-비서 연결고리를 삭제해도 유사한 편향이 여전히 남는다. '여성'과 관련되는 주민번호나 거주지 등의 연관정보는 여전히 '여성'을 간접적으로, 여성을 '대리'해서 나타내기 때문에 (따라서 '대리 속성(proxy attributes)'이라고 일컬어짐) 만약 기계학습에서 이러한 대리속성을 포함하는 모델을 활용할 경우 젠더 관련 편향이 여전히 발생할 수 있다. 대리속성이 내부에 남아 있고 이로 인한 편향이 발생하기 때문에 이는 '대리 차별'(proxy discrimination)(Prince, et al., 2019)로 불린다. 따라서, 편향은 해당 보호 속성과 명시적으로 보여지는 속성과 연관되어 있다는 점만 원인으로 보기 어렵고, 암묵적으로 연관되어 있는 대리속성들 또한 포함하므로 보호 속성-특정속성 간의 연결 삭제는 편향의 근본적 해결책은 아니다(Barocas et al., 2017). 그렇다면 명백하게 드러나는 요인이나 보호 속성-중립속성 간 연결을 삭제하는 방법 외에도 대리 특징과의 암묵적 연결을 약화하는 방법도 사용할 수 있다.

편향인지 및 완화 기준으로서의 공정성

편향인지의 시도들

상술한 각 단계별 편향의 가능성을 인지하더라도 편향이 어느 지점에서 발생하며 어떻게 바로잡을지 파악하기란 어렵다. 최근 편향을 식별하고 어느 정도 위험한지 평가하는 방법이 개발되고 있다. 코넬대학교 연구진과 산업체가 공동으로 진행한 연구(Tan et al. 2017)에서는 편향된 알고리즘의 작동 방식을 밝히기 위해 두 가지 방법을 사용했다. 먼저 블랙박스 알고리즘을 모방해서 모델을 만들고 초기 데이터를 기반으로 그 블랙박스가 어느 정도 위험한지 점수를 낸다. 그리고, 초기 데이터가 아닌 다른 데이터를 이용하여 학습 이후의 모델을 만들어 비교하고 최종 결과에서 어떤 변수가 중요한지 추정했다. 인공지능으로 대출을 결정하는 한 회사의 인공지능을 이 방식으로 분석한 결과, 대출 결정의 기준에 대출의 동기는 없었고 중요한 변수를 무시하는 것으로 분석되었다.

이러한 기술적 해결책은 인공지능을 설계 및 제작하거나 활용하는 이들이 사용하기에는 효율적이지 않을 수 있고 비용이 든다. 또, 인공지능 시스템의 사용자인 일반인들은 예컨대 인공지능으로 만들어진 면접평가 시스템이나 주식추천 시스템 등을 사용할 때 시스템의 의사결정에 대한 문의나 문제제기를 하기 어렵다. 2016년 4월 14일 유럽연합에서 발표되어 국내에도 채택한 일반데이터보호규

정(GDPR, General Data Protection Regulation) (https://www.eugdpr.org)에서는 알고리즘 편향이나 통계절차상의 차별방지조항 등 인공지능의 윤리적 쟁점들을 포함하고 있다. 그 예가 정보이동권(right to portability), 설명을 요구할 권리(right to explanation), 잊힐 권리(right to be forgotten)이다. 이러한 권리에 기반하여 일반인도 인공지능에 어떤 편향이 의심될 경우 문의하고 이의를 제기할 수 있는 법적 근거가 마련되었다(김재완, 2019). 이에 선대응하기 위해 일부 기업들은 자체적으로 편향을 감지 및 완화할 수 있는 시스템을 마련하고 (완화 기술 절 참조), 일부 기업은 자체적으로 인공지능 시스템의 편향을 감사(audit) 받기도 하였다.

이렇게 인공지능에 내재한 편향은 불공정함을 야기할 뿐만 아니라 인공지능산업의 발달 또한 지연시킨다. 이러한 부정적 영향을 선(先)방지하려는 대표적 시도는 게임을 통해 성격유형을 판단하여 특정 기업에서 필요한 직원 유형과 비교평가하는 인공지능기반 평가도구 제작 및 판매사인 미국의 파이메트릭스(Pymetrics)사(https://www.pymetrics.ai/)의 사례이다. 파이메트릭스 사는 외부 대학에 자사의 도구가 편향을 가지는지 여부에 대한 감사를 의뢰하였고 문제가 없음의 결과를 통지받아(Wilson et al., 2021) 자사의 인공지능 시스템이 편향이 없음을 밝히고 있다. 이 알고리즘 감사 보고서에서는 파이메트릭스 사에서 문서, 데이터, 소스 코드를 외부 비공개 계약으로 제공하고 감사관은 모델링을 분석하기 위해 데이터를 합성하기도 하여 분석하였는데 그 기준은 미국의 차별방지 관련 법적 기준인 '5분의 4 규칙'(the four-fifth rule)이며 이에 기반하여 위해영향분석을 수행했다. 보고서에는 법적 기준을 충족했음을 밝히고 있다.

미국의 균등고용기회위원회(Equal Employment Opportunity Commission, (EEOC))에서 1978년 발표된 '5분의 4 규칙'이란 인종, 성별, 문화등과 같은 보호 속성(protected characteristic) 에서 특정 집단이 가장 높은 비율을 차지하는 집단의 5분의 4(또는 80%)에 미치지 못할 경우, 이러한 회사의 고용 절차가 해당 집단에 불리한 영향을 끼치는 증거로 간주하는 규정이다. 보호 속성에는 인종, 성별 외에 연령, 종교, 장애, 임신상태, 퇴역군인 등이 포함된다. 이 규칙은 보호 속성을 가진 집단이 타 집단에 대해 최소한의 비율적 균등성을 충족하도록 함으로써 공정성을 확보하고자 하는 것이다.

이러한 규칙에 근거한 편향의 인지와 감사가 편향방지와 투명성의 확보를 위해서 도움이 되는 것은 사실이다. 그러나 부작용도 예상할 수 있다. 파이메트릭스 사례에서처럼 감사의 대상목록은 감사관 쪽이 아니라 자사에서 요청한 항목 특히 5분의 4 규칙에만 국한되며 감사관 측과의 이해충돌여부는 알 수 없다. 더 중요한 요소는 아래에 살펴볼 것처럼, 인구통계학적 편향 여부의 평가에서 문제가 없다고 해서 공정성을 담보할 수 없다는 점이다. 인구통계학적 공정성은 수많은 공정성 유형 중 한 종류에 불과하기 때문에, 해당 사안의 맥락, 즉 기업의 직무특징이나 유형에 맞는 공정성 계산 방법인지를 고민할 필요가 있다.

비형식적, 형식적 차원의 공정성

인공지능 편향을 완화하려는 시도는 그 근거로서 공정성 기준을 필요로 한다. 그런데 사회, 철학 등 인문학적 의미에서 뿐만 아니

라 통계 및 기술적 견지에서도 다양한 공정성의 의미가 있다. 스무 가지가 넘는 의미로 분류되기도 하지만 불완전하다(Narayanan, 2018)고 할 정도로 상황과 맥락이 다양하다. 공정성의 다양한 의미들을 크게 비형식적 의미와 형식적 의미로 구분해볼 수 있다. 먼저, 비형식적 의미를 살펴보면, 심리적, 윤리적 차원에서 등가교환적 정의문제인 '형평(equity)'과 사실 차원에서 거론되는 '평등(equality)'은 그 의미가 다르다. 평등한 배분이 형평한 배분의 바탕이 될 수는 있어도 평등한 배분이 곧 형평 배분을 의미하는 것은 아니다. 또한 공정한 분배라는 개념에는 동등하게 분배하는 '객관적 평등', 기여도에 따른 분배인 '상대적 평등', 개인의 필요에 따르는 '주관적 평등', 비용에 따른 '서열적 평등', 그리고 '기회의 평등'의 원칙(Eckhoff, 1974)이 있다. 이 중 분배적 정의는 평등과 필요 간의 균형을 맞추는 문제(Deutsch, 1975)이기도 하기 때문에 원칙 외에 '필요'라는 상황맥락을 검토하는 작업이 각 사례의 공정성을 검토할 때 필요하다.

형식적 기준으로는 최근 부상한 기계학습과 관련한 '공정한 인공지능'기준을 꼽을 수 있다. '공정한 인공지능'의 기본 개념은 인공지능 모델의 최종 판단결과가 인종, 성별과 같은 특정 보호 특성들에 종속변수가 되지 않도록 무관하게 제시되는 기술을 의미한다. 그런데 예상할 수 있는 것처럼 비율적 균등성 만으로는 공정성을 온전히 확보하기에 불충분하다. 첫째 이유는 사안들은 언제나 복합 요인으로 구성되기 때문에 (표 1 참조) 다른 종류의 공정성 또한 개입된다는 점이다. 둘째 이유는 단순요인으로 구성되는 매우 단순한 사안이라 할지라도 편향을 교정한 알고리즘으로 만들어진 인공지능으로 내려진 결정의 차후 결과가 이상적이지 않거나, 업무의 우수성과 같은 목표가 확보되지 않을 가능성이 있다. 따라서 공정성의 기본 원칙과 함께 이해관계의 다양한 상황에 따라 조정이 필요하므로 다양한 유형의 공정성을 고려할 수밖에 없다.

통계적 공정성의 유형

다양한 의미의 공정성이 제기되는 배경은 통계적 균등성(statistical parity)이 기본적인 공정성 개념이지만 표준적 기준이 될 수 없기 때문이다. 만약 영향을 미치는 어떤 매개변수가 있다면 통계적 균등성 거리를 계산하여 측정할 수 있고 그 정도를 수량화하여 조정할 수 있다. 예컨대 특권 집단이 비특권 집단에 비해 받은 유리한 결과가 어느 정도 비율의 차이가 나는지를 계산할 수 있다. 그런데 통계적 균등성의 계산을 통해 얻어지는 동등함은 결과적 균등함이며, 과정이나 그 전 단계에서의 공정함은 보장할 수 없다. 이 경우 인공지능에 구현될 때 결과적으로 알고리즘 정확도를 낮출 수 있기 때문에 결과적으로 적절한 공정성을 확보하지 못하게 한다. 예컨대 A집단과 B집단은 샘플에 있어서 애초부터 양성으로 예측될 확률이 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. 그래서 이 경우 결과적으로 균등함을 보여줄 뿐 공정함이라고 평가하긴 어렵다. 이러한 이유로, 통계적 균등성은 알고리즘 정확도를 오히려 낮추는 결과(Menon, 2018)를 가져온다.

통계적 균등성에 대한 대안으로, 형식적 의미에서의 기계학습과 관련한 '공정한 인공지

표 1. 공정성의 유형

유형	종류	수학적 의미
예측 기반	집단 공정성	집단별 긍정적 예측값을 할당받을 확률이 동일
	조건부 통계적 동등성	특정 데이터 속성을 통제했을 경우 그룹 별로 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	예측 값 기반 양성예측도(PPV, Positive Predictive Value)와 음성예측도(NPV, Negative Predictive Value)
	조건부 사용 정확도 동등성	예측 값 기반 양성예측도(PPV, Positive Predictive Value)와 음성예측도(NPV, Negative Predictive Value)
	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 동일할때
	잘 보정됨 (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 차별	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
	대리차별 금지	보호된 속성에서 대리 변수에 의해 차단되는 예측된 결과까지의 경로가 없는 경우
	공정한 추론	인과 관계 그래프의 경로를 정당 혹은 부당한 것으로 분류

능'은 여러 통계적 지표들을 활용 및 조합한 다양한 공정성 정의를 사용한다. 기계학습에서 편향 인지 및 완화를 위해 활용할 통계 지표(metric)들은 기존의 통계적 균등성을 포함한 베르마와 루빈(2018)이 제시하는 다음의 열 두 가지이다. 1. 진양성(True Positive): 예상 및 결과가 모두 양성인 경우, 2. 위양성(False Positive): 실제는 음성인데 양성으로 예상한 경우, 3. 위음성(False Negative): 실제는 양성인데 음성으로 예상한 경우, 4. 진음성(True Negative): 예상 및 결과가 모두 음성인 경우, 5. 양성예측값(Positive Predictive Value): 예측된 모든 양성 사례 중 양성으로 올바로 예측된 양성 사례의 비율, 6. 위발견율 False discovery rate): 예측된 모든 양성 사례 중 양성 클래스로 잘못 예측된 음성 사례의 비율, 7. 오탈락률(False Omission Rate): 예측된 모든 음성 사례 중 음성으로 잘못 예측된 양성 사례의 비율, 8. 음성예측값(Negative Predictive Value): 예측된 모든 음성 사례 중에서 음성 클래스에 올바로 예측된 음성 사례의 비율, 9. 진양성율(Treu Positive Rate): 모든 실제 양성 사례 중에서 양성으로 정확하게 예측된 양성 사례의 비율, 10. 위양성율(False Positive Rate): 모든 실제 음성 사례 중에서 양성으로 잘못 예측된 음성 사례의 비율, 11. 위음성율(False Negative Rate): 모든 실제 양성 사례 중 음성으로 잘못 예측된 양성 사례의 비율, 12. 진음성율(True Negative Rate): 모든 실제 음성사례 중 음성으로 올바르게 예측된 음성사례의 비율.

이 열두 종류의 통계지표를 활용 및 조합하여 표 1과 같이 20가지의 공정성 유형(Verma, et. al., 2018)이 정의되고 있다.

공정성의 이러한 다양한 정의는 동시에 충족하기 어려우며 상충하는 경우가 많다. 또, 실제로 한 종류의 공정성의 의미를 특정 상황에 적용할지라도 비형식적 의미의 공정성 종류, 절차적 공정성 등을 함께 고려해야 하며, 적용영역, 상황 및 맥락에 따른 다양한 필요들과 함께 고려할 경우 더 많은 종류의 공정성 개념이 구성될 수 있다.

공정성 정의에 기반한 AI편향 인식 및 조정도구

구성한 모델에 편향이 있는지를 검사할 수 있는 도구는 오픈소스 형태로 다양하게 개발되어 있다. 구글의 WIT(What-If Tool), ML 페어니스 짐(Fairness Gym), IBM의 AI 360 페어니스(Fairness), 애퀴타스(Aequitas), 페어런(FairLearn)이 있다. 구글의 'What-If Tool'은 이용자들이 기계학습 과정에서 직접 다섯 유형의 공정성 도구를 적용해 볼 수 있도록 만든 도구다. 알고리즘 공정성 제약 조건을 테스트하고, 추론 결과를 시각화하고, 데이터 포인트를 편집해 피처(feature) 변경에 따라 모델 성능이 어떻게 달라지는지를 볼 수 있게 한다. 기계학습 전 데이터에 반영된 인간의 편향을 식별하는 방법과 편향이 반영된 모델의 성능과 예측을 평가할 수 있도록 하고 있다.

IBM의 편향성 완화 툴킷인 'AIF 360(AI Fairness 360 Open Source Toolkit)'은 편향성 평가 메트릭스를 제공하기 위해 개발된 오픈소스로 전처리 과정에서 가중치 조정을 수행하고 편향을 수치화해 제공한다. 70개의 알고리즘 공정성 지표와 10개의 편향보정 알고리즘을 제시하고 있으며, 콤파스, 신용평가알고리즘(automated credit-scoring algorithms), 성인 소득 분류 등 기존에 존재하는 데이터를 기반으로 편향을 예측해 볼 수 있게 한다. 마이크로

소프트의 '페어런(Fairlearn)'은 평가 대시보드를 통해 모델의 예측이 다른 집단에 어떤 영향을 주는지 평가하기 위한 도구를 제공하고 공정성 및 성능 메트릭스를 사용하여 여러 모델을 시각화해 비교할 수 있게 한다. 이 중 구글의 도구를 간략히 소개하면, 구글의 what-if 도구는 아래처럼 다섯 가지의 공정성 유형을 제공(https://pair-code.github.io/what-if-tool/ai-fairness.html) 하면서 성별을 대표적 예로 설명하고 있다. 사용자는 본인이 테스트해보고 싶은 데이터셋을 업로드하고 특정한 유형의 공정성을 선택하여 그 기준에 따라 편향을 완화시킨 모델을 검토해볼 수 있다. 다섯 가지 공정성 유형은 다음과 같으며 앞서 소개한 20가지 유형의 공정성 중 일부를 채택한다.

1) Group unaware (집단 블라인드) 공정성을 확보하기 위해 대상자의 성별을 모르게 하는 것이다. 이 방법을 통해서 순전히 목표 가치만으로 선택하여 어느 한쪽 성별만 최종적으로 선택되더라도 이것이 공정하다고 해야 한다. 이를 위해서는 데이터셋에서 성별과 젠더-대체 정보를 제외하여 시스템을 구성하고 예측해야 한다.

2) group thresholds (집단 임계값): 첫 번째 유형의 기준만으로는 부족할 수 있다. 모델구성에 사용된 수집데이터 자체가 과거의 현상을 반영하기 때문에 특정 보호 속성의 가치가 더 낮게 반영될 수밖에 없다. 예컨대, 여성의 경력은 출산 등으로 중단될 가능성이 남성보다 높으므로 이는 기계학습에서 불리하게 작용한다. 따라서 이러한 경우 남성에 대한 신뢰 수준보다 여성에 대한 신뢰 수준을 낮추는 방식으로 남성 집단과 여성 집단에 대한 신뢰 임계값을 독립적으로 조정할 수 있다.

3) demographic parity (인구통계학적 동등성): 데이터셋에서의 보호 속성의 비율은 결과값에 반영이 되어야 한다. 예컨대 만약 특정 성별의 데이터가 30% 였다면 결과값에서 그 성별의 비율도 30%이어야 한다.

4) equal opportunity (평등한 기회): 1,2,3번 기준은 비적격자에게 잘못된 대출과 같은 의사결정을 내릴 수 있는 위험이 있다. 대안은 성공적인 결과값이 산출될 것으로 여겨지는 이들이라면 남성이든 여성이든 동일한 비율로 대출승인 결정이 내려지는 경우이다.

5) equal accuracy (동일한 정확도): 4번의 기준에서 더 나아가 대출 대상자로 선정된 이들과 아닌 이들의 비율 또한 각 젠더 간에 동일해야 한다. 4번의 기준은 여전히 대출상환을 하지 않는 이들이 나올 가능성이 있으며, 대출승인 및 거부 기록에서 잘못 결정되었던 비율이 두 성별에서 동일하게 조정되어야 한다.

시카고 대학교에서 만든 공정성 검토 도구인 '애퀴타스(Aequitas)' (http://www.datasciencepublicpolicy.org/projects/aequitas/)는 기계학습 개발자, 분석자, 정책결정권자들이 기계학습 모델에 있어서 차별과 편향을 탐지할 수 있도록 만든 오픈 소스 도구로 웹 사이트에서 편향성 탐지가 필요한 데이터를 업로드하고 편향 문제가 있어서는 안 될 집단을 선택한 뒤 툴킷이 제공하는 공정성 메트릭스(젠더, 연령, 인종 등)를 선택하면 해당 데이터가 갖고 있는 편향성에 대해 보고서를 발행해 준다.(Saleiro et al., 2018) 미국의 주정부가 연구자들과 함께 자체 개발한 공정성 체크 도구들도 있다. 국내의 경우, 과기부에서 인공지능을 적용하는 기업 및 연구소를 위한 인공지능윤리 체크리스트를 구성하는 작업을 진행 중(관계부처합동, 2021)이며 플랫폼 기업들은 인공지능윤리 헌장을 발표하며 방향을 모색하고

있다.

인공지능 편향완화 기술

앞서 살펴보았듯 편향은 인공지능 구성 단계마다 내재할 수 있다. 편향의 완화방법 또한 각 단계마다 적용될 수 있다. 상술한 편향 완화의 방법들은 인간이 수동방식으로 직접 개입하여 조정이 가능한 단계들이다. 반면 편향을 완화하는 방법 중 디지털 구조에서 조정하는 기술이 있다. 이 기술이 적용되는 단계들은 기계학습 전 전처리, 처리 과정 중, 그리고 결과산출 후 단계들에서이며 각 단계들에 대표적인 기술들은 '탈편향 (debias)'기술로 임베딩구조와 반사실적 공정성을 이용하기, 처리 과정 중에 분류자의 편향을 완화하거나 적대적 방법을 사용하기, 그리고 결과가 나온 후 편향을 완화하는 방법이 있다.

전처리(pre-processing) 단계에서의 편향완화

기계학습 전 단계인 데이터 처리 단계에서 해당 편향성을 제거하는 방식을 탈편향(debias) 알고리즘이라고 부른다. 최근 여러 가지 방법이 개발되고 있지만 대표적으로 임베딩구조에서 조정하는 방법과 반사실적 인과를 만들어 내어 조정하는 방법을 소개한다. 기계학습에 사용되는 데이터들에서 편향성이 확인될 경우 이 부분을 조정하는데 예컨대 특정 인종이나 성별이 가지는 직업과 관련한 편향성의 경우 데이터 처리 단계에서 조정하는 것이 보다 효과적이다. 이 단계에서 처리하는 것이 중요한 이유는 데이터의 특성에 있다. 영상과 텍스트 중 특히 맥락 속에서 의미를 파악해야 하는 텍스트가 컴퓨터에서 이해될 수 있도록 벡터로 처리되는 과정에서 아래와 같이 그 특징을 잘 볼 수 있다.

임베딩구조에서의 편향완화

자연언어, 특히 일상어나 의견표명 그리고 인문학 분야의 내용은 관련한 핵심어를 사용하지 않고도 다양한 동의어, 다의어, 축약어, 신조어 등을 사용하여 표현되며, 애매모호한 표현을 사용하기도 하기 때문에 관련 개념들을 전체 맥락 속에서 얼마나 적절하게 이해하는가가 중요하다. 이를 위해 자연언어를 디지털로 처리할 때 워드임베딩(word embedding)의 방법을 활용하여 벡터로 표현한다. 워드임베딩 안에서 처리되는 텍스트 데이터에 편향이 내재한다면, 추후 기계학습으로 데이터를 훈련시켜 함수를 파악하는 절차에서 그 편향이 증폭된다. 따라서 이후 단계에서 보다는 워드임베딩의 구조 안에서 편향을 완화하는 작업이 상대적으로 효율적이다. 워드임베딩의 한 방법인 워드투벡(word2Vec)은 인공신경망 기법을 활용하여 어휘를 일정한 길이의 벡터로 표시하는 방법으로 가장 많이 사용되므로 여기서는 이 벡터들의 구조 안에서 편향이 어떻게 표현되는지를 보겠다.

워드투벡 방법의 기본 가정은 "한 단어의 의미는 그것이 동반하는 것들을 통해 알 수 있다."(You shall know a word by the company it keeps.) (Firth, 1957)는 '분포가설(the Distributional Hypothesis)'이다. 이 방법은 저차원에서 단어의 의미를 여러 차원 공간에 분산하여 표현하기 때문에 단어 간 유사도를 맥락의 의미 손실 없이 기계 안에서 빠르게 계산할 수 있게 한다. 유사 의미를 가진 단어들 주변에는 유

사한 의미를 가진 어휘들이 있는 구조이다. 예컨대, '코로나 감염병'이라는 단어는 '무섭다', '아프다', '마스크', '방역', '백신'등의 단어들과 출현하는 횟수가 다른 단어들보다 많다.

텍스트 데이터 안에 편향이 있다면, 워드투벡의 이러한 분산표현의 특성 때문에 편향 관련단어들이 벡터 구조 안에서 가까운 거리로 표현된다. 예컨대 벡터 구조 안에서 '비서'는 '남성'보다는 '여성'이나 '여자'에 한층 더 가까운 거리에 위치해있는 것이 일반적이다. 이렇게 벡터구조 안에 내재한 편향을 완화하려면 이 구조의 특성을 활용해야 한다 (Bolukbasi et al., 2016). 기하학적 구조를 활용하여 편향을 완화하는 방법은 임베딩 구조에서 중립적 단어에 해당하는 벡터가 특정 단어에 편향적인 거리를 갖지 않도록 동등화(equalize)하는 방법이다. 예컨대 '비서'라는 중립적 단어를 '여성'과 '남성'이라는 보호 속성 관련단어 모두에 동일한 거리를 가지게끔 조정(Bolukbasi et. al., 2016)하는 것이다.

이 방법의 부작용은 특정 목적의 시스템에서는 중요할 수 있는 구분까지 제거할 수 있다는 점이다. 즉, 공정성을 확보하는 방향이 정확도와 상충관계 (accuracy-fairness tradeoff)가 되는 경우이다. 특정 목적의 시스템에서는 '아버지'의 의미가 가지지 않는 의미를 '어머니' 표현이 가질 수 있는데, 이 경우 '어머니'에 가중치를 더 부여하여 모델을 구성할 수 있다. 이 때 이렇게 구성된 모델이 편향을 가진다고 보기 보다는 특정 목적의 시스템에 맞는 설계라고 볼 수 있다.

이러한 편향완화 방법을 통해서 만들어진 인공지능은 데이터 전처리 단계의 질적 측면에서 보호 속성과 중립적 속성 간에 독립성을 확보할 수 있다. 그럼에도 불구하고 데이터의 양적 측면에서의 편향의 요소는 여전히 내재할 수 있다. 모델이 만들어지기 전 기계학습을 위해 수집된 데이터의 양에 있어서 특정 속성이 상대적으로 많은 비중을 차지할 수 있다. 예컨대 면접에 사용될 인공지능을 구성할 때 학습 데이터는 사회에서 높은 성과를 올리는 50대 남성의 특성들이 현실적으로 대다수이기에 데이터도 해당 데이터의 양이 더 많이 확보될 수밖에 없다. 따라서 임베딩 구조 안에서 '우수 후보자'로부터 '남성'과 '여성' 양자 간의 기하학적 거리를 동등하게 위치시키는 방식으로 조정했다 할지라도 50대 남성들의 특성들이 기계학습에 더 잘 학습될 수밖에 없다. 특정 집단의 데이터가 부족하면 결과적으로 기계학습에서 주요하게 파악하는 분류의 정확도가 낮아지고 최종적으로는 의도하지 않더라도 불가피하게 간접 차별이 나타날 수밖에 없다. 양적 비율의 불균형에 있어서는 특정 집단 혹은 관련 집단들의 샘플 크기를 늘리는 방식으로 문제를 해결할 수 있다. 따라서 편향완화는 인공지능 구성과정의 각 단계에서 질적, 양적 차원 모두 검토대상이다.

반사실적 인과를 이용한 편향완화

반사실적 텍스트 및 영상 데이터를 이용한 편향완화 방법은 보호/민감한 속성이 가상적 상황에서 바뀌더라도 시스템의 의사결정은 동일하도록 조정하여 편향이 작동되지 않도록 하는 것이다(Chiappa et al., 2019). 반사실적 (counterfactual) 상황이란 현재의 사실적 상황을 'X --> Y'로 기술했을 때 이와 반대되는 상황을 가정하는 것이다. 예컨대 "우리나라 대통령은 남자이다"라는 텍스트를 가정했을 때,

반사실적 텍스트는 “우리나라 대통령은 여자이다”이다. 보호 속성과 관련하여 소수자에 대한 데이터는 보통 양적인 측면에서 크게 부족한 경우가 많으며 이에 데이터를 보충할 필요가 있을 때 이러한 반사실적 텍스트를 기존의 데이터셋에 추가하는 방법을 선택할 수 있다. 예컨대 “영수는 아름다운 런던에 산다.”, “영수는 서울 시내에 산다”, “영수는 공동묘지 근처에 산다” 등 무수한 반사실 텍스트를 만들어낼 수 있는데 이렇게 다양한 반사실적 텍스트 샘플이 만들어지고 나면, 영수가 어느 지역에 살든 지역에 따른 차별된 내용이 생성되지 않도록 조정된다. 수많은 반사실적 텍스트들을 만들어 훈련데이터로 사용하는 방식으로 해당 텍스트들의 모델이 편향을 만들어내지 않도록 조정할 수 있다. 그런데 이러한 방법의 난점은 동일한 예측값을 가지도록 하는 반사실 상황을 산출하는 것이 바람직한 맥락과 그렇지 않은 맥락을 분명하게 구분해야(Garg et al., 2019)만 이 방법을 정확히 활용가능하다는 점이다. 현재는 이러한 난점을 해결하여 편향완화와 인공지능 시스템의 강건성을 동시에 확보할 수 있는 방법들이 연구되고 있다.

처리 중(in-processing) 단계에서의 편향완화

모델링 단계의 편향완화를 위해 많이 알려진 방법은 적대적 학습의 방법을 사용하여 편향을 제거하는 것이다. 적대적 학습방법은 ‘적대적 공격(adversarial attack)’즉 시스템이 싫어할만한 데이터를 만들어내는 방법으로, 데이터에 인위적 조작을 가하여 인공지능 모델의 성능을 높이는 방법이다. 사람이 인지할 수 없을 정도의 작은 교란(perturbation) 샘플을 만들어내어 인공지능의 판단을 흐리게 만들고 그를 극복하는 과정을 통해 기계학습에서의 분류 성능을 높이고 모델을 최적화할 수 있다.

예컨대 보호 속성의 소수자 집단에 해당하는 흑인종이나 연장자 연령의 집단과 관련한 편향이 학습데이터에 있다고 가정하자. 그러면 이를 토대로 기계학습을 하게 되며, 이 편향을 그대로 반영할 수밖에 없다. 적대적 학습방법을 사용하면, 보호 속성 집단에 대한 변수, 예측 변수, 그리고 적대적 공격자 모두를 동시에 학습하는데 이 때 적대적 공격자는 보호 속성 예컨대 흑인 집단의 주민번호 뿐만 아니라 대체 특성인 우편번호까지도 모델링하려고 시도하게 된다. 그리고 이러한 적대적 공격자의 시도능력을 최소화하는 프레임워크가 적용된다(Zhang et al., 2018). 이 방법은 다양한 유형의 공정성에 적용할 수 있다는 장점을 가진다고 평가받고 있다.

또 다른 방법은 최근 구글이 발표한 프레임워크로 기존 공정성 지표의 난점을 극복하면서 시스템의 안정성 또한 확보하는 방법이다. 상술한 공정성의 지표들 중 예컨대 ‘기회 균등(Equality Of Opportunity)’은 기계학습에 사용되는 데이터를 범주별로 분류할 때 해당 데이터에 속하지 않은 집단에 대한 편향이 자연스럽게 만들어지는데, 이 때 편향측정에 기회균등 지표를 사용하여 상이한 집단 간의 잘못 표상된 양성률의 차이를 최소화한다. 이는 앞서 소개한 IBM 에서도 사용하는 지표이다. 다만 단점은 기록되지 않은 어떤 정보가 있을 경우 정보 조정 자체가 어려워서 데이터셋에서 집단 간 균형을 이루기가 어렵다. 구글은 이러한 난점을 해결해서 시스템의 강건성도 강화하면서도 편향완화도 한층 개선하는 ‘민

표 2. 편향완화 방법들

차원	완화 방법	종류	방법
시스템	훈련 데이터 편향 완화	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 특성 값을 편집
		공정한 표상 학습 (learning fair representations)	보호 속성에 대한 정보를 난독화하여 공정한 표현을 학습
	분류기의 편향 완화	적대적 편향성 제거 (adversarial debiasing)	적대적 기술을 사용하여 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 편향 완화	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 사용하여 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 진양성긍정 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

디프(MinDiff)'라는 프레임워크를 발표(https://ai.googleblog.com/2020/11/mitigating-unfair-bias-in-ml-models.html)했다. 이 방법은 유해하다고 간주한 텍스트를 식별·제거하는 새로운 학습 방식으로 데이터 샘플이 매우 적은 집단에 대한 데이터까지도 최적화하는 방식으로 편향완화를 돕는다. 구체적으로, 예측과 인구통계학적 그룹 간 상관관계를 최소화하여 편향을 완화하고, 분포가 다를 경우 평균과 예측의 변화량을 동일하게 미세 조정하여 모델 정확도도 확보한다(Prost et al., 2019).

링크드인(LinkedIn)은 'LiFT'라는 오픈소스 공정성 도구(https://github.com/linkedin/LiFT) 를 제공한다. 이 도구는 머신러닝 모델의 학습 도중에도 배치되어 사용될 수 있으며, 학습 데이터셋의 편향성에 대한 점수를 매기고, 모델의 공정성을 평가함과 동시에 학습 모델의 서브 그룹들에 대한 성능 차이를 탐지할 수 있다.

후처리(post-processing) 단계에서의 편향완화

앞서 전처리나 처리 중 단계에서 데이터수집이나 라벨링 등에서의 가능한 편향들을 모두 검토하고 조정한다 해도 비의도적으로 생겨나는 편향이 있을 수 있다. 이는 일종의 간접적 편향으로 법적 용어로는 '결과적으로 차별이 된다'는 의미로 '차등적 영향'(disparate impact) 또는 '간접차별'(Zliobaite, 2015)로 표현된다. 인간이 인지하지 못하는 부분에서 상호연관되는 특성들이 오랜동안 사용되면 결과적으로 차별을 더 강화하게 된다(Barocas et al., 2016). 이 경우 데이터 처리 단계나 기계학습 과정에서는 인지가 불가능하여 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 임계값을 선택한 공정성 기준에 맞추어 조정하는 방식이다. 예컨대, 전세자금 대출 대상자를 결정하기 위해 모델을 만든다고 가정하면, 일반적으로 이미 경제력을 구축해온 중년층을 선호하는 모델이 될 수 있다. 이러한 편향을 완화하고자 후처리 기술을 사용하면, 만들어진 분류 모델을 그대로 유지하지만 대출 대상자 전체 수락률이 중년층 뿐만 아니라 청년층, 장년층에게도 공평하도록 결과를 조정할 수 있다. 조정에 있어서는 해당 작업의 특정 목표에 따라 공정성의 정의 중 적절한 것을 선택하고, 관련된 집단 공정성이나 개인 공정성의 지표에 따라(Lohia et al., 2019) 조정한다. 이러한 후처리 방법은 전처리나 처리 중에 편향을 완화하는 방법처럼 굳이 모델 안을 들여다 볼 필요가 없다는 편리성이 있다. 또한 인공지능의 블랙박스 특성으로 인해 그 원인을 정확히 추정하기 어렵다는 난점 또한 후처리 단계에서는 문제가 되지 않는다.

대표적으로 소개한 편향완화 방법들을 포함한 다양한 방법들(Barocas et al., 2017)은 표 2와 같으며, 상술한 공정성 기준들 간의 충돌 문제, 편향완화와 시스템 정확도와의 조화 문제 등을 해결하기 위한 방법들이 개발되고 있다.

결 론

인공지능 편향 이슈가 기존의 부호처리 방식의 컴퓨터 발전 단계가 아니라 딥러닝이라

는 정보 처리방식이 만연하게 된 현재에 등장한 배경은 편향이 인공지능 시스템의 설계 및 구성에 내재해 있음을 암시한다. 그런 의미에서, 인공지능 편향을 인지하고 완화하고자 할 때 인공지능에서 주요한 비중을 차지하는 데이터 전처리와 기계학습을 이해하고 그 과정에서 편향이 개입되는 지점들을 인지하는 것이 중요하다. 인공지능이 사회의 전 영역에서 활용되고 또한 역으로 우리의 판단과 행동에 영향을 주고 있음을 고려하면, 인공지능 편향에 대한 감수성이 필요하다.

인공지능 편향은 인간이 만들어 기계 안에 내재하게 되며 어느 정도 조정가능한 만큼 이에 대한 교육은 기존의 개념중심의 학습이 아닌 경험 및 체험 (hands-on) 방식의 학습이 필수적으로 포함되어야 인공지능 개발자뿐만 아니라 사용자로서도 편향에 의한 부작용의 피해를 방지할 수 있다. 인공지능 편향의 시스템적 이해를 위해서는 상술한 인공지능 구성 절차의 일부라도 체험해보는 것이 도움이 된다. 현재 어린이도 경험가능한 기계학습 플랫폼이 다양하게 존재하며 수준을 다양하게 조정할 수 있다. 이를 통해 데이터의 종류를 달리해보고, 변수를 변화시켜 봄으로써 비교적 간단하게 기계학습의 원리와 편향의 발생하는 지점을 직접 이해할 수 있다(김효은, 2020).

이러한 학습방식은 비단 인공지능 산업 관련 개발자뿐만 아니라 인공지능 기반의 제품을 사용하는 일반인들에게도 유용하다. 일반인들은 인공지능의 설계, 제작에는 참여하지 않지만 인공지능 기반의 시스템이 내리는 의사결정에 영향을 크게 받는다. 따라서 인공지능의 구성과정과 편향이 개입되는 지점, 조정가능하다는 사실과 이것이 사회의 거버넌스와 연관되어 있다는 점을 인지하면 세계적으로 적용되고 있는 인공지능윤리 관련 권리를 정확히 행사하여 시민권을 지킬 수 있다. 또 사회로 나가기 전의 젊은이들은 인공지능을 활용하는 모든 영역에서 보다 나은 활용을 할 수 있다.

인공지능 편향에 대한 상술한 적절한 인지를 위해서는 인공지능 공정성이나 인공지능 편향완화 기술에 대한 막연한 기대보다는 인공지능 시스템 구성 절차와 거버넌스 과정에서 공정성의 요소를 고려하는 것이 필요하다. 본문에서 소개된 공정성의 정의들이나 편향완화 알고리즘들은 다양하지만 상충하는 사례들도 있으며, 어느 하나가 대표적이거나 완벽한 해결책이 될 수 없다.

그 이유는 다양한 사례들이 한 유형의 공정성이나 편향완화 알고리즘만으로 해결되기 어려운 복잡한 맥락을 가지고 있다는 점에 그치지 않는다. 여러 유형의 공정성을 아우르는 알고리즘이 개발된다고 할지라도 많은 경우에 공정성 즉 편향완화가 이루어질수록 모델의 정확도가 낮아지기 때문이다(Friedler et al., 2019). 상기에 소개한 몇몇 사례들처럼 편향을 완화하면서도 시스템의 강건성 또한 확보하는 새로운 기술들이 연구되고 있는 반면, 양자를 모두 만족시키기란 쉽지 않다. 그렇다면 자연스런 대안은 편향과 정확도 간에 '균형'을 찾는 것이다. 이 균형을 설정하기 위해서는 인공지능 기술의 발전만으로는 불가능하며, 사회의 이해관계자들 간의 논의와 합의를 필요로 한다. 기술과 사회적 거버넌스, 그리고 이 과정에서 인간의 심리적 편향의 검토와 완화의 과정은 새로운 공정성 개념들과 편향완화 기술들을 만들어 내게 될 것이다.

참고문헌

관계부처합동 (2021). 신뢰할 수 있는 인공지능 실현전략, 5월 13일. 과학기술정보통신부 인공지능기반정책과. Retrieved from https://www.korea.kr/common/download.do?fileId=195009613&tblKey=GMN

김인식 외 (2021). 유튜브 알고리즘과 확증편향, 한국컴퓨터교육학회 학술발표대회논문집 25. Retrieved from https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002555225

김재완 (2019). EU 일반정보보호규정(GDPR)의 알고리즘 자동화 의사결정에 대한 통제로써 설명을 요구할 권리에 대한 쟁점 분석과 전망. 민주법학, 69, 277-298. doi:10.15756/dls.2019..69.277

김청택 (2019). 빅데이터를 이용한 심리학 연구 방법. 한국심리학회지: 일반, 38(4), 519-548. doi:10.22257/kjp.2019.12.38.4.519

김효은 (2020). 공학적 방법을 결합한 인공지능윤리 학습. 윤리연구, 1(129), 133-153. doi:10.15801/je.1.129.202006.133

Barocas, S., Andrew D. Selbst, (2016). Big Data's Disparate Impact, *California Law Review,* 104, 671-732. doi:10.2139/ssrn.2477899

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1. doi:10.1007/978-3-030-43883-8_7

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. doi:10.1177/0146167205277097

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. doi:10.5555/3157382.3157584

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1),7801-7808. doi:10.1609/aaai.v33i01.33017801

Dave, P. (2018). Fearful of bias, Google blocks gender-based pronouns from new AI tool. *Reuters*, November, *27*. Retrieved from https://www.reuters.com/article/us-alphabet-google-ai-gender-idUSKCN1NW0EF

Firth, R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis,* vol. *Special Volume of the Philological Society*, 1 – 32. Retrieved from http://cs.brown.edu/courses/csci2952d/readings/lecture1-firth.pdf

Friedler, A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019, January). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, 329-338. Retrieved from https://arxiv.org/abs/1802.04422

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. Retrieved from https://eprints.lancs.ac.uk/id/eprint/11136

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, H., & Beutel, A. (2019). .Januaryerfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 219-226. doi: 10.1145/3306618.3317950

Josh, N. (2019, June 19) 7 Types of Artificial Intelligence, *Fobes Media* LLC. Retrieved from https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificialintelligence/#145fe100233e

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13), 3521-3526. doi:10.1073/pnas.1611835114

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)*, 2847-2851. IEEE. Retrieved from https://www.ibm.com/downloads/cas/WM4MWDOE

Menon, A. and Williamson, R. (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, 107-118. Retrieved from https://arxiv.org/abs/1705.09055

Microsoft (2021). Transparency note and use cases for Custom Neural Voice. Retrieved from https://docs.microsoft.com/en-us/legal/cognitive-services/speech-service/custom-neural-voice/transparency-note-custom-neural-voice

Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp.*, 1170, New York, USA. Retrieved from https://fairmlbook.org/tutorial2.html

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK. Retrieved from https://dl.acm.org/doi/10.5555/2029079

Prates,M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10), 6363-6381. Retrieved from https://arxiv.org/abs/1809.02208

Prince, A. E., & Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.,* 105, 1257. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3347959

Prost, F., Qian, H., Chen, Q., Chi, E. H., Chen, J., & Beutel, A. (2019). Toward a better trade-off between performance and fairness with kernel-based distribution matching. arXiv preprint. Retrieved from https://*arXiv*:1910.11779.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A.,& Ghani, R. (2018).

Aequitas: A bias and fairness audit toolkit. arXiv preprint, *arXiv.* Retrieved from https://*arXiv*:1811.05577.

Sample, I. (2017, Nov. 5). Computer says no: why making AIs fair, accountable and transparent is crucial. *The Guardian,* 5, 1-15. Retrieved from https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, *54*(4), 680-712. doi:10.1111/1745-9125.12123

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2017). Detecting bias in black-box models using transparent model distillation. arXiv preprint, Retrieved from https://*arXiv*:1710.06169

Timm, J., Staab, S., Siebers, M., Schon, C., Schmid, U., Sauerwald, K., & Beierle, C. (2018, September). Intentional forgetting in artificial intelligence systems: Perspectives and challenges. In *Joint German/Austrian Conference on Artificial Intelligence* (Künstliche Intelligenz), 357-365. Springer, Cham. doi:10.1007/978-3-030-00111-7_30

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware),* 1-7. IEEE. doi:10.1145/3194770.3194776

Vries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does object recognition work for everyone?. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,* 52-59. Retrieved from https://arxiv.org/abs/1906.02659

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., & Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666-677. doi:10.1145/3442188.3445928

Zhang, H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society,* 335-340. Retrieved from https://arxiv.org/abs/1801.07593

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. arXiv preprint. Retrieved from https://arxiv.org/abs/1511.00148

1차원고접수 : 2021. 10. 27.
2차원고접수 : 2021. 12. 14.
최종게재결정 : 2021. 12. 14.

한국심리학회지: 일반
Korean Journal of Psychology: General
2021, Vol. 40, No. 4, 459-485
http://dx.doi.org/10.22257/kjp.2021.12.40.4.459

Fairness Criteria and Mitigation of AI Bias

Hyo-eun Kim

Department of Humanities, Hanbat National University

AI bias is not only an issue of humanities and social impact and governance, but also of systemic robustness. The algorithm bias has the characteristic of being intervened in the system construction process as the computer becomes an artificial neural network-based autonomous intelligence system. The objective of this paper is to deal with the aspects of bias that are involved in each stage of artificial intelligence, the fairness criterion for the judgment of bias, and the bias mitigation methods. Different types of fairness are difficult to satisfy simultaneously and require different combinations of criteria and factors depending on the field and context of AI application. Each method for mitigating the bias of training data, classifiers, and prediction alone do not completely block the bias, and a balance between bias mitigation and accuracy should be sought. Even if bias is identified through unlimited access to the algorithm through AI auditing, it is difficult to determine whether the algorithm is biased. The bias mitigation technology goes beyond simply removing the bias, and is moving toward solving the problem of both reducing the bias and securing the robustness of the system, and adjusting the various types of fairness. In conclusion, these characteristics imply that policies and education that recognize AI biases and seek solutions should be explored in terms of bias recognition and coordination based on system understanding beyond recognizing issues at the conceptual level.

Key words : machine learning, bias, fairness, data, algorithm