creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

법학박사 학위논문

인공지능 공정성 심사 기준에 관한 연구

Assessing Fairness in Artificial Intelligence: Contextual Justification for Normative Criteria

2023 년 8 월

서울대학교 대학원

법학과

김병필

인공지능 공정성 심사 기준에 관한 연구

지도교수 박상철

이 논문을 법학박사 학위논문으로 제출함

2023 년 5 월

서울대학교 대학원

법학과

김병필

김병필의 박사 학위논문을 인준함

2023 년 7 월

위 원 장 허 성 욱 (인)

부위원장 임 용 (인)

위 원 박 상 철 (인)

위 원 박 혜 진 (인)

위 원 이 두 갑 (인)

국 문 초 록

인공지능의 활용 사례가 크게 증가하고 있다. 인지모형 인공지능은 종래 컴퓨터로 구현하기 어려운 것으로 여겨져 왔던 이미지, 음성, 텍스트 인식 작업을 인간 수준으로 수행해 내기 시작했다. 판별모형 인공지능은 인간의 의사결정을 보조하거나 의사결정을 자동화하는 데 널리 활용되고 있다. 나아가 생성모형 인공지능을 통해 인간이 인공지능과 자연스럽게 소통하거나 인공지능이 예술적 창작물을 생성해 낼 수 있게 되었다. 이러한 활용 사례의 증가와 함께 인공지능 모형이 학습 데이터에 반영된 기존 사회적 편향을 반영하여 차별적 영향을 초래할 위험에 대한 우려도 점차 커지고 있다.

이러한 법적, 윤리적 위험을 막기 위한 주요한 방안으로 인공지능 공정성 감사가 주요하게 고려되고 있다. 공정성 감사는 차별금지법제, 개인정보 보호법제, 인공지능 규제 법제 등에 있어 인공지능에 대한 주요한 규제 수단으로 여겨지고 있다. 이를 위해서는 공정성의 개념을 수학적으로 정의한 통계적 기준을 설정하고, 감사 대상 인공지능이 해당 기준을 만족하는지 평가하는 방법론이 마련될 필요가 있다. 이에 인공지능 공정성 연구자들은 감사에 활용할 수 있는 다양한 공정성 지표를 조작적으로 정의하는 방안을 제시해 왔다. 그러나, 특히 분류모형에 관한 여러 공정성 지표를 동시에 충족시킬 수 없다는 불가능성 정리가 증명되면서, 공정성 감사를 위한 공정성 지표의 선택 문제가 중요하게 부상하고 있다. 이에, 본 연구는 분류모형과 생성모형을 구분하여, 인공지능의 특정한 이용 사례의 구체적 맥락(context) 하에서 어떠한 공정성 지표를 선택하는 것이 규범적으로 정당화될 수 있을지 모색한다.

우선, 분류모형에 대한 공정성 심사는 그 모형의 예측 결과가 역사적으로 차별받아 온 사회적 집단에 대해 차별적 처우를 하거나 차별적 결과를 초래하는지에 관한 판단에 초점을 둔다. 그런데 이제껏 발전해 온 차별심사에 관한 법적 판

단 기준은 인공지능에 의한 차별 심사에 활용하는 데 있어 여러 어려움을 초래하고 있고, 기술적으로 정의된 공정성 지표는 기존의 법 체계와 조화되지 못하는 문제가 발생하고 있다. 이러한 난점을 해소하기 위해 본 연구는 분류모형에 대해 시뮬레이션을 활용하는 새로운 방법론을 적용한다. 즉, 다양한 시나리오에 걸쳐 가상의 차별적 학습 데이터를 생성하여 분류모형을 학습시킨 다음, 해당 모형이 어떠한 공정성 심사기준을 어떠한 방식으로 위반하는지 조사함으로써, 해당 시나리오에 적합한 공정성 기준을 도출한다. 본 연구의 시뮬레이션 결과는 인지모형 인공지능의 경우 위양성율과 위음성율의 동등성을 일차적 평가 지표로 고려하는 것이 타당하지만, 판별모형 인공지능의 경우 양성 예측도와 음성 예측도의 동등성(즉, 이른바 '캘리브레이션')이 일응 적합한 공정성 지표일 수 있다는 점을 시사한다. 나아가, 본 연구는 이제까지 진행되어 온 공정성 지표에 관한 논의의 한계를 지적하고, 대안적 공정성 심사의 방향을 제시한다.

다른 한편, 생성모형 인공지능에 대한 공정성 심사는 분류모형과는 다른 양상을 띤다. 분류모형은 의사결정의 보조 또는 자동화 등에 활용되어 분배적 위해를 초래할 가능성이 있는 것에 비해, 생성모형의 산출물은 특정 의사결정에 영향을 끼치는 것이 아니라 사회 전반에 걸쳐 문화적이고 확산적인 표상적 위해를 초래할 위험이 있기 때문이다. 더욱이 표상적 위해는 사회적 편향과 고정관념의 확산 및 강화, 유독한 표현의 생성과 유포 등 다양한 차원의 문제를 야기할 수 있는데, 이러한 경우 그 피해자 특정이나 손해 산정이 어려우므로 전통적 사적 구제수단을 활용하여 통제되기 어렵고, 기존의 행정규제 체계 내에서 규율하기에 어려운 문제가 있다. 그러므로 생성모형에 대해서는 그 개발자 또는 이용자에 의한 사전적 자기 감사가 중요하게 요구되고 있다. 이에 본 연구는 생성모형에 대한 공정성 심사를 위한 구체적 지표와 평가 방법을 편향, 협의의 공정성, 유독성의 차원으로 구분하여 검토한다. 나아가 생성모형이 초래할 수 있는 법적, 윤리적 위험성에 대한 적절한 법적 규율을 위해서는 국내에서 활용되는 대규모 언어

모형 및 이미지 생성모형 등에 대한 법적, 윤리적 위험성을 다면적으로 평가할 수 있는 벤치마크 데이터 및 평가 지표가 마련될 필요가 있음을 지적한다.

주요어 : 인공지능 공정성, 인공지능 차별, 공정성 지표, 알고리즘 공정성, 불가능성 정리
학　번 : 2021-33081

목 차

제 1 장 서 론 1

제 1 절 연구의 배경과 목적 1

1. 인공지능 공정성 감사 1

2. 국제 기술 표준의 제정 3

3. 규제 해석 및 집행 기준 마련의 필요성 4

4. 본 연구의 범위 및 목적 7

제 2 절 연구 방법론 7

1. 문헌조사 7

2. 시뮬레이션 방법론 10

제 3 절 본 연구의 구성 13

제 2 장 인공지능의 개념과 유형 16

제 1 절 인공지능의 개념 16

1. 인공지능 개념에 관한 일반적 이해 16

2. 인공일반지능의 문제 19

3. 본 연구의 정의 22

제 2 절 인공지능의 유형화 26

1. 분류모형의 유형화 26

2. 생성모형의 유형화 37

제 3 절 인공지능에 의한 공정성 침해 위험성 46

1. 공정성의 개념 46

2. 인공지능에 의한 공정성 침해 양상 60
제4절 소결 70
제3장 인공지능 공정성 보장을 위한 법적 규제 72
제1절 개관 72
제2절 차별금지법제 74
1. 인공지능에 의한 직접차별 77
2. 인공지능에 의한 간접차별 82
3. 직접차별 기준과 간접차별 기준의 충돌 86
제3절 개인정보 보호법제 98
1. 자동화된 의사결정에 대한 규제 98
2. GDPR 상의 자동화된 의사결정 시스템 규율 100
3. GDPR 상의 설명요구권 논쟁 103
제4절 인공지능 규제법(안) 111
1. EU AI 법안 112
2. 미국 121
3. 국내 입법 동향 127
제5절 소결 129
제4장 분류모형 인공지능의 공정성 지표 130
제1절 개관 130
제2절 입력변수에 대한 공정성 기준 132
1. 의의 132
2. 한계 136
3. 대안적 시도 및 평가 141

제3절 출력변수에 대한 공정성 기준 143

1. 결과 동등성 143
2. 오류율 동등성과 예측도 동등성 154

제4절 인공지능 공정성 지표 간의 충돌과 불가능성 정리 175

1. COMPAS 사례 175
2. 인공지능 공정성 지표에 대한 불가능성 정리 178
3. 인공지능 공정성 지표에 대한 선택 논쟁 180
4. 소결 182

제5장 분류모형 인공지능의 공정성 지표 선택 183

제1절 시뮬레이션 방법론 183

1. 목표변수의 확률분포 183
2. 머신러닝 모형 187

제2절 인지모형 시나리오 189

1. 시뮬레이션 방법 189
2. 시뮬레이션 결과 191
3. 장기적 개선 방향 - 인공지능의 성능 개선 197

제3절 판별모형 시나리오 200

1. 시뮬레이션 방법 200
2. 시나리오 2-1 - 집단간 잠재적 능력의 확률분포가 다른 경우 202
3. 시나리오 2-2 - 동일 잠재적 확률분포, 특성의 편향 209
4. 확률적 목표변수 예측 과제의 인식론적 한계 212

제4절 분류모형 공정성 지표의 법적 함의 214

1. 인공지능 개발 단계에서의 공정성 지표 활용 215

2. 인공지능 도입 평가 단계에서의 공정성 지표 활용 217
3. 사후적 사법 심사 단계에서의 공정성 지표의 활용 219
4. 소결 226
제5절 기존 공정성 지표의 한계와 대안적 방안 227
1. 기존 공정성 지표의 한계 227
2. 대안적 공정성 감사 방안 234
제6장 생성모형 인공지능의 공정성 지표 241
제1절 개관 241
1. 생성모형의 편향과 좁은 의미의 공정성 243
2. 생성모형의 유독성 245
제2절 생성모형에 관한 공정성 지표 및 평가 방법 248
1. 편향 평가 249
2. 좁은 의미의 공정성 평가 256
3. 유독성 평가 263
제3절 생성모형의 공정성 보장을 위한 기술적 조치 267
1. 학습 데이터 구축 단계 268
2. 생성모형 학습 단계 269
3. 생성모형 이용 단계 274
제4절 생성모형의 공정성 규제를 위한 법적 과제 278
1. 표현 생성 목적 생성모형의 공정성 보장을 위한 법적 규율 279
2. 기반모형의 공정성 보장을 위한 법적 규율 288
제7장 결론 299
참 고 문 헌 303

별 지 - 시뮬레이션의 상세 322

Abstract 329

표 목 차

표 1 인지모형과 판별모형의 구분 71

표 2 자동화된 의사결정 시스템에 대한 설명의 유형 105

표 3 이진 분류의 혼동 행렬과 통계적 정확성 지표 156

표 4 인지모형 시나리오의 시뮬레이션 결과 191

표 5 집단간 위음성율 차이 (시나리오 2-1) 204

표 6 차별 시정 조치 후 False Discovery Rate (시나리오 2-1) 207

표 7 집단간 위음성 오류율 차이 (시나리오 2-2) 210

표 8 생성모형의 공정성 지표와 평가 방법 예시 249

표 9 생성모형에 의한 허위조작정보 생성, 유포 위험성 증가 가능성 281

그 림 목 차

그림 1 집단간 능력치의 확률분포 가정 11

그림 2 생성된 가상의 학습 데이터 12

그림 3 시뮬레이션 상의 머신러닝 과정 13

그림 4 가상의 대출심사 학습 데이터 28

그림 5 학습된 대출심사 판별모형의 결정경계선 예시 29
그림 6 정형 데이터 생성모형 39
그림 7 적대적 생성신경망의 구조 44
그림 8 확산 모형의 학습 및 생성 단계 45
그림 9 GPT-2 에 의해 생성된 편향된 텍스트 66
그림 10 재범 예측을 위한 입력변수에 대한 설문 결과 135
그림 11 대용변수의 유형 138
그림 12 버클리 대학 입시 사례의 인과성 도표 148
그림 13 집단 캘리브레이션이 충족되지 못하는 사례 171
그림 14 집단 캘리브레이션이 충족되는 사례 172
그림 15 COMPAS 소프트웨어에 대한 프로퍼블리카의 감사 결과 176
그림 16 COMPAS 소프트웨어에 대한 노스포인트 사의 감사 결과 177
그림 17 목표변수가 퇴행분포인 경우 분류모형의 적용 과정 184
그림 18 확률적 목표변수의 절단정규분포 186
그림 19 목표변수가 정규분포인 경우 분류모형의 적용 과정 187
그림 20 로지스틱 회귀 모형 예시 사례 188
그림 21 베이스라인 시나리오의 학습 데이터 분포 190
그림 22 인공지능 성능 개선을 통한 공정성 개선 가능성 (unawareness 기준 미적용시) 198
그림 23 인공지능 성능 개선을 통한 공정성 개선 가능성 (unawareness 기준 적용시) 199
그림 24 집단간 원래 능력의 분포가 다른 사례의 가정 201
그림 25 집단간 능력의 확률분포에 차이가 있는 경우의 예측 결과 202
그림 26 머신러닝 모형에 의해 예측된 능력 값의 분포 203
그림 27 집단 간 캘리브레이션이 충족되는 결과 (시나리오 2-1) 204
그림 28 오류율 불균등이 발생하는 원인 (시나리오 2-1) 206

그림 29 관측된 특성변수가 편향된 경우 머신러닝 모형의 결과 (시나리오 2-2) 210
그림 30 캘리브레이션 기준을 위반하는 결과 (시나리오 2-2) 211
그림 31 집단 간 예측의 신뢰구간 차이가 큰 경우 237
그림 32 직업별 여성 비율과 이미지 검색 결과 비교 252
그림 33 자연어 처리에서의 공정성과 편향 상충관계 263
그림 34 ChatGPT 에 대한 추가적 학습 단계 273
그림 35 대규모 언어모형의 창발성 291
그림 36 프롬프트-기반 인공지능 도입 전략 292

제1장 서 론

제 1 절 연구의 배경과 목적

1. 인공지능 공정성 감사

빅데이터에 기반하여 학습된 인공지능과 그에 기반한 자동화된 의사결정 시스템의 활용이 증가하면서, 여러 영역에서 인공지능의 불공정성과 차별에 대한 문제제기가 이루어지고 있다. 또한, 이에 대응하기 위해 기존 규제가 적용되거나 새로운 규제가 마련되고 있기도 하다. 이러한 규제 준수나 새로운 규제 마련에 있어 중요한 방안으로 인공지능에 대한 감사(audit)가 주목받고 있다.[1] 인공지능에 대한 감사는 인공지능 개발 및 도입 과정에서의 사전적 감사 형태를 취할 수도 있고, 이미 도입된 인공지능에 대한 사후적 감사의 형태를 취할 수도 있다. 다른 한편, 인공지능을 제공하거나 이용하는 사업자에 의한 내부 감사 방식과 규제 기관이나 중립적 전문가에 의한 외부 감사가 있을 수 있다.[2]

어떠한 감사 방식에 의하든지 간에 인공지능에 대한 감사를 위해서는 특정한 평가 기준을 설정하고, 대상 인공지능의 공정성을 '측정(measure)'하여 해당 기준을 충족시키는지 평가해야 한다. 이때 '공정성(fairness)'의 개념을 어떻게 측

---

1 Inioluwa Deborah Raji, et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing", *Proceedings of the 2020 conference on fairness, accountability, and transparency*, (2020).

2 Adriano Soares Koshiyama, et al., "Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms", *Technical Report,* (2021).

정할 수 있는지에 대한 문제가 제기된다. '공정성'은 법학, 철학, 윤리학, 사회학, 경제학 등 여러 학문의 관점에서 다양한 방식으로 정의될 수 있고, 누구든지 동의할 수 있는 한 가지 정의를 채택하는 것은 극히 어려운 과제가 된다. 이러한 어려움에도 불구하고, 최근 다수의 인공지능 공정성 연구자들은 가능한 여러 '공정성' 개념 중 특정 징표를 계량하여 표시하는 방법을 발전시켜 왔다. 이하에서는 인공지능의 '공정성'의 개념을 계량할 수 있도록 정의한 지표를 인공지능에 관한 '공정성 지표(fairness metrics)'라 칭한다.[3]

인공지능에 관한 공정성 지표는 '공정성' 개념에 대한 일종의 '조작적 정의(operational definition)'라 할 수 있다. '조작적 정의'란 어떠한 개념이 어떻게 측정되어야 하는지를 정확하게 구체화하는 것을 지칭한다.[4] 예를 들어 사회경제적 지위(socioeconomic status)에 대한 '조작적 정의'는 해당 개념을 소득 수준과 교육 수준을 더한 값으로 정의하는 것이다. 이처럼 공정성 개념을 조작적으로 정의함으로써 인공지능의 개발 또는 도입 과정에서 편향을 예방 또는 제거하거나, 도입된 인공지능에 대한 공정성 감사를 위하여 사용할 수 있다.

이처럼 조작적으로 정의된 인공지능 공정성 지표를 유용하게 활용할 수 있다고 하더라도, 공정성 개념의 본래 의미를 온전히 포착하지 못할 위험이 있는 것은 사실이다.[5] 이에 각각의 인공지능 공정성 지표가 어떠한 공정성 개념에 부합하는지를 판단하여, 각각의 지표를 더욱 적절하게 정의하고 활용하기 위한 노력이 필요한 상황이다.

---

3 Deborah Hellman, "Measuring algorithmic fairness", *Virginia Law Review* 106(4), (2020).

4 얼 바비 저, 고성호 외 옮김, *사회조사방법론* (제13판), 센게이지러닝, (2015), 제5장

5 조작적 정의에 대한 철학적 논의와 그 한계에 대한 논의로는 Hasok Chang, "Operationalism", *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.) 참조.

2. 국제 기술 표준의 제정

나아가 여러 국제 표준화 기구는 인공지능에 관한 공정성 지표를 기술 표준으로 제정하고자 노력하고 있다. 그 중에서도 국제 전기 전자 기술자 협회(Institute of Electrical and Electronics Engineers, 이하 "IEEE")는 인공지능 윤리에 대한 국제표준 수립에 대한 논의를 선도하고 있다. IEEE 산하 'AI 및 자율시스템의 윤리적 고려사항에 대한 IEEE 글로벌 이니셔티브'는 2019년 3월 '윤리 기반 설계(Ethically Aligned Design): 자율지능시스템에서의 인간 안녕을 우선시하기 위한 비전' 보고서를 발간하였다.[6] 위 보고서는 '자율지능시스템(autonomous and intelligent systems; A/IS)'에 대해 제기될 수 있는 윤리적 문제를 포괄적으로 검토하고 있는데, 자율지능시스템에 의한 차별 가능성과 관련하여 "자율지능시스템은 시스템 성능의 잠재적 편향으로 인해 특정 사회적 또는 인구통계적 집단에 대해 불리한 영향을 끼치는지에 대해 신중하게 평가되어야 한다. 해당 평가 절차에서는 그와 같은 편향을 진단하고 바로잡는 데 있어 잠재적으로 불이익을 얻는 집단의 구성원을 포용해야 한다"고 권고한다. 나아가, IEEE는 그 후속 작업으로 '윤리 기반 자율지능시스템(ethically aligned autonomous and intelligent system)'에 관한 다수의 기술 표준을 마련하기 위한 작업을 진행해 오고 있다.[7]

또한, IT 분야의 주요한 국제 기술 표준화 기구인 국제표준화기구(International Organization for Standardization, 이하 "ISO")와 국제전기기술위원회(International Electrotechnical Commission, 이하 "IEC") 역시 인공지능 신뢰성에 관한 기술 표준을 준비하고 있다. ISO/IEC의 제1합동기술위원회(JTC 1)는 2018. 4. SC 42(Subcommittee 42)를 신설하여 인공지능 기술의 표준

---

6 Institute of Electrical and Electronics Engineers (IEEE), *Ethically Aligned Design*, (2019).

7 위의 글.

화를 전담하도록 하고 있는데,[8] 그 산하 인공지능의 신뢰성에 관한 WG3 작업반이 설립되어 있다.[9] WG3 작업반은 2020. 5. "인공지능 신뢰성 개요에 관한 기술보고서(ISO/IEC TR 24028:2020)"를 발간하였고, 그 이외에도 인공지능에 관한 위험 관리 표준, 인공지능에 대한 윤리적, 사회적 우려에 관한 기술보고서, 인공지능 시스템이나 인공지능의 조력을 받는 의사결정에 있어서의 편향(bias)에 관한 표준, 신경망의 견고성(robustness) 평가를 위한 표준을 준비하고 있다. 현재 개발 중인 기술 표준 중에는 인공지능 공정성 지표를 정의하고 이를 적용하는 방법을 다루는 것들이 포함되어 있다.[10]

3. 규제 해석 및 집행 기준 마련의 필요성

향후 인공지능 공정성 지표에 대한 기술 표준이 마련되면, 이는 인공지능 기술과 관련된 여러 규제의 해석과 집행에 있어서도 중대한 영향을 끼칠 것으로 전망된다. 유럽연합의 AI Act 법안(이하 "**EU AI 법안**")[11]의 경우, 사업자가 동법상의 의무를 준수하는지 여부를 평가하기 위하여 기술 표준을 활용하는 방향으로 접근하고 있다. 현재 논의 중인 EU AI 법안은 인공지능에 관한 기술 표준을 준수할 경우 동법상의 의무사항을 이행한 것으로 추정하는 조항을 두고 있다.[12]

---

8 신성필·하수욱·이강찬, "ISO/IEC JTC 1/SC 42 (인공지능) 국제 표준화 동향", *TTA저널* 제187호, 한국정보통신기술협회, (2020), 16 – 21면.

9 강상욱, "인공지능 기술의 안전성 확보 동향", *주간기술동향* 제1978호, 정보통신기획평가원, (2020), 2 – 11면.

10 관련된 IEEE 표준으로는 IEEE P7003 Algorithmic Bias Considerations; ISO/IEC 표준으로는 ISO/IEC AWI TR 24368 Artificial intelligence — Overview of ethical and societal concerns, ISO/IEC 24027 Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making 등이 있다

11 법안의 정식 명칭은 "Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS"이다. EU AI Act에 관하여는 본 연구 제3장 제4절 참조.

12 본 연구 제3장 제4절 참조.

나아가 EU 집행위원회 산하 AI Watch가 2021년 발표한 보고서는 EU AI 법안 상 요구되는 의무사항과 국제 표준화 기구의 기술 표준을 대응시키고 있다.[13]

나아가 유럽연합의 일반 개인정보 보호 규정(General Data Protection Regulation, 이하 "GDPR")[14] 등 개인정보 보호법제에서 요구되는 개인정보보호 영향평가의 과정에서도 인공지능 공정성 지표가 고려될 필요가 있다.[15] 예컨대, 영국 개인정보 보호기구인 Information Commissioner's Office (이하 "ICO")는 인공지능 및 개인정보 보호를 위한 가이드라인 초안에서 여러 공정성 지표를 간략히 소개하면서, 이를 "인공지능 개발자들이 자신의 시스템에서 발생할 수 있는 잠재적 차별을 측정하고 완화하기 위해 *취할 수 있고 취해야 하는(can and should)* 몇 가지 기본적 접근법이 있다"(역자 강조)고 언급하였다.[16] 위 ICO 인공지능 감사 가이드라인 초안은 인공지능 편향의 원인을 ① 학습 데이터 자체의 불균형과 ② 학습 데이터에 반영된 기존의 편향으로 구분한 다음, ① 학습 데이터 자체에 불균형이 있는 경우 과소/과대 대표되는 집단에 대해 학습 데이터상의 균형을 유지하는 기법을 적용할 수 있고, ② 학습 데이터가 종전의 차별을 반영하는 경우 해당 데이터를 수정하거나, 학습 과정을 수정하거나, 해당 모형을 학습 후 수정할 수 있다고 설명한다.

---

13 Stefano Nativi and Sarah De Nigris, *AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework*, EUR 30772 EN, Publications Office of the European Union, (2021).

14 정식 명칭은 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)이다.

15 GDPR 제35조 참조.

16 ICO, "ICO consultation on the draft AI auditing framework guidance for organisations.", (2020. 2. 19.) https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-ai-auditing-framework-guidance-for-organisations/. (2023. 5. 26. 접속 확인) (이하 "**ICO AI 감사 가이드라인 초안**").

다른 한편, 금융 감독기관이 감독 대상 금융기관에게 공정성 지표를 어떻게 적용할 수 있는지를 보여주는 모범 사례를 제시하기도 한다. 대표적으로 싱가포르 통화청(Monetary Authority of Singapore, MAS)은 2019. 11. 인공지능의 책임성 있는 도입을 위한 프레임워크를 구축하는 계획을 발표하고, 금융기관이 인공지능 윤리원칙에 부합하도록 인공지능을 도입할 수 있는 오프소스 도구를 개발해 오고 있다.[17] 싱가포르 통화청은 2021. 1. 1단계 과정을 마무리하고, 고객 마케팅과 신용점수 산정 분야에 있어 인공지능 공정성 기준을 적용한 사례 보고서를 발표하였다.[18]

그러나, 실제로 인공지능 공정성 감사에서 공정성 지표를 적용하고자 할 때에는 여러 어려움에 직면하게 되는데, 그 중 하나는 공정성 지표가 매우 다양하다는 점이다. 예를 들어, 인공지능 공정성 연구자 Arvind Narayanan은 이제까지 제안된 주요 인공지능 공정성 지표가 21개에 이른다고 소개한다.[19] 이처럼 다양한 지표 중 과연 어떠한 맥락에서 어떠한 지표를 사용해야 하는지에 대한 기준은 아직 마련되어 있지 못하다. 더욱이 알고리즘 공정성 지표에 대한 불가능성 정리(impossibility theorem)로 인해 여러 지표를 동시에 충족하는 것은 불가능

---

17 Monetary Authority of Singapore (MAS), *Fairness Metrics to Aid Responsible AI Adoption in Financial Services*, (2020), https://www.mas.gov.sg/news/media-releases/2020/fairness-metrics-to-aid-responsible-ai-adoption-in-financial-services/. (2023. 5. 26. 접속 확인).

18 MAS, "Veritas Initiative Addresses Implementation Challenges in the Responsible Use of Artificial Intelligence and Data Analytics", Media Releases, (2021. 1. 6.) https://www.mas.gov.sg/news/media-releases/2021/veritas-initiative-addresses-implementation-challenges (2023. 5. 26. 접속 확인)

19 Arvind Narayanan, "21 fairness definitions and their politics", *Conference on Fairness, Accountability, and Transparency* (2018. 2. 23). 한편 Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, online textbook, https://fairmlbook.org/ (2023. 5. 26. 접속 확인) Chapter 3에 소개된 공정성 지표는 모두 19가지이다.

하다는 사실이 증명되었다.[20] 따라서 여러 지표 중 하나를 선택하는 일은 더욱 어려운 문제가 되고 있다.

4. 본 연구의 범위 및 목적

이상과 같은 배경 하에서 본 연구는 인공지능을 개발 또는 도입하는 조직이 공정성 감사에서 활용할 수 있는 인공지능 공정성 지표를 검토하고, 각 지표가 어떠한 법적 함의를 갖는지 해명하고자 한다. 특히 인공지능에 의한 차별 사안에 대해 현재 적용되고 있는 규제 내용에 비추어, 어떠한 맥락에서 각각의 공정성 지표를 적용하는 것이 규범적으로 타당한지를 밝히고자 한다. 다만, 본 연구는 개별 사업자나 기관이 활용할 수 있는 공정성 지표를 중심으로 다루고, 인공지능의 시장 경쟁에 대한 효과나 거시적인 분배 효과 측면에서의 공정성이나 인공지능의 도입과 활용에 있어서의 절차적 공정성 문제는 직접적인 연구 범위에 포함하지 않는다.

제2절 연구 방법론

1. 문헌조사

2010년대 이후 기업과 정부 부문에 걸쳐 인공지능의 활용이 확산되자, 여러 활용 분야에 걸쳐 인공지능의 공정성에 대한 문제제기가 이루어져 왔다. 이러한

---

20 인공지능 공정성 기준 간의 불가능성 정리에 관해서는 본 연구 제4장 제4절 참조.

문제제기가 이루어진 활용 영역은 대출,[21] 채용,[22] 온라인 광고,[23] 형사사법절차,[24] 이민자 구금,[25] 아동학대 스크리닝,[26] 복지 대상자 선별[27] 등 다양하다. 이러한 문제제기에 따라 인공지능의 공정성을 보장하기 위한 다양한 기술적 방법론이 제시되어 왔다.[28] 이러한 흐름 속에서 인공지능 공정성에 관한 학제간 연구를 위한

---

21 Moritz Hardt, Eric Price, and Nati Srebro, "Equality of opportunity in supervised learning", *Advances in neural information processing systems* 29, (2016); Lydia T. Liu et al., "Delayed impact of fair machine learning", *International Conference on Machine Learning*, (2018); Andreas Fuster, et al., "Predictably unequal? The effects of machine learning on credit markets", *The Journal of Finance* 77(1), (2022).

22 Claire Cain Miller, "Can an algorithm hire better than a human?", *New York Times*, (2015. 6. 26.); Lily Hu and Yiling Chen, "A short-term intervention for long-term fairness in the labor market", *Proceedings of the 2018 World Wide Web Conference*, (2018).

23 Latanya Sweeney, "Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising", *Queue* 11(3), (2013).

24 Julia Angwin et al., "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks", *ProPublica*, (2016. 5. 23.); William Dieterich, Christina Mendoza, & Tim Brennan, "COMPAS risk scales: Demonstrating accuracy equity and predictive parity", *Northpointe Inc,* (2016); Jeff Larson, et al., "How we analyzed the COMPAS recidivism algorithm", *ProPublica,* (2016).

25 Robert Koulish, "Immigration detention in the risk classification assessment era", *Connecticut Public Interest Law Journal* 16 (2017).

26 Rhema Vaithianathan, et al., "Children in the public benefit system at risk of maltreatment: Identification via predictive modeling", *American journal of preventive medicine* 45(3), (2013); Alexandra Chouldechova, et al., "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions", *Conference on Fairness, Accountability and Transparency*, (2018); Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St. Martin's Press, (2018).

27 Eubanks(2018) (각주 26).

28 Shira Mitchell, et al., "Algorithmic fairness: Choices, assumptions, and definitions", *Annual Review of Statistics and Its Application* 8, (2021).

국제 학술대회인 ACM FAccT(Conference on Fairness, Accountability, and Transparency)가 2018년부터 개최되어 왔다.[29]

본 연구는 이상과 같은 연구들에 있어 ① 컴퓨터 공학적 접근, ② 법제도적 접근, ③ 법이론적 접근이라는 세 가지 흐름을 종합적으로 조사한다. 우선 컴퓨터 공학적 접근은 인공지능 공정성을 수학적으로 정의하고, 인공지능 학습 과정에서 공정성 지표의 제약조건을 만족시킬 수 있는 방법을 탐구한다.[30] 법제도적 접근은 대체로 인공지능 공정성 문제에 대한 현재의 법적 규율 현황을 조사하고, 현 제도의 한계를 지적하거나 대안을 모색하는 시도라고 평가할 수 있다.[31] 마지막으로 법이론적 접근은 현재 논의되고 있는 인공지능 공정성 기준이 이제까지 법철학 혹은 정치철학에서 논의되어 온 '공정성' 개념에 비추어 어떻게 평가되고, 정당화 또는 거부될 수 있는지를 다룬다.[32]

---

29 https://facctconference.org/ (2023. 5. 26. 접속 확인).

30 본 연구에서는 최근의 연구를 종합한 서베이(survey) 문헌으로 Ninareh Mehrabi, et al. "A survey on bias and fairness in machine learning", *ACM Computing Surveys (CSUR)* 54(6), (2021)을 주요하게 참조하였다.

31 Solon Barocas and Andrew D. Selbst, "Big data's disparate impact", *California law Review* 104(3), (2016)은 이러한 접근법에 있어 선도적인 시도로 평가된다. 이후 다양한 연구가 이루어져 왔는데, 예컨대 금융 분야에 있어서는 Matthew Adam Bruckner, "The promise and perils of algorithmic lenders' use of big data", *Chicago-Kent Law Review* 93, (2018); Mikella Hurley and Julius Adebayo "Credit scoring in the era of big data", *Yale Journal of Law & Technology* 18, (2016); 노동 분야에 있어서는 Pauline Kim, "Data-driven discrimination at work", *William & Mary Law Review* 58, (2016); Charles A. Sullivan, "Employing AI", *Villanova Law Review* 63, (2018) 등이 있다.

32 최근 이러한 접근을 취한 연구가 여럿 발표되고 있다. Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. "Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics", *AI and Ethics* 1(4), (2021); Brian Hedden, "On statistical criteria of algorithmic fairness", *Philosophy and Public Affairs* 49(2), (2021); Sune Holm, "The Fairness in Algorithmic Fairness", *Res Publica,* (2022); Corinna Hertweck, Christoph Heitz, and Michele Loi. "On the moral justification of statistical parity", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (2021).

본 연구는 위와 같이 다양한 학제에 걸친 인공지능 공정성 기준에 대한 상이한 접근법을 종합하여 적절한 법적 규율 체계를 수립하는데 기여하고자 한다.

2. 시뮬레이션 방법론

본 연구는 분류모형 인공지능에 있어 시뮬레이션 작업을 통해 어떠한 맥락에서 어떠한 공정성 지표를 선택하는 것이 규범적으로 정당화될 수 있을지 모색한다. 즉, 다양한 시나리오에 걸쳐 가상의 차별적 학습 데이터를 생성하여 머신러닝 알고리즘을 학습시킨 다음, 해당 알고리즘이 어떠한 공정성 지표를 어떠한 방식으로 충족하거나 위반하는지 조사함으로써, 해당 시나리오에 적합한 공정성 지표를 도출하는 것이다.

본 연구는 분류모형 인공지능 $f_{classfier}$를 입력변수 X로부터 목표변수 Y를 예측하는 일종의 함수로 전제한다.

$$f_{classfier}\ X \longmapsto \hat{Y}$$

여기서 X는 머신러닝 알고리즘에 의한 예측에 활용되는 특성(feature)이고, $\hat{Y}$는 머신러닝 알고리즘이 예측한 출력변수이다. 본 연구에서는 논의의 단순화를 위해 Y가 0 또는 1의 값만을 갖는 이진(binary) 변수인 경우로 한정한다.

본 시뮬레이션은 인공지능을 통한 과제 수행 과정을 아래와 같이 추상화한다.

(1) 우선 시뮬레이션을 위해 실제 사회에서의 각 개인이 갖는 어떠한 확률분포를 가정한다. 예컨대 확률변수 p가 개인의 어떠한 능력(업무 수행 능력, 변제 능력 등)을 표상한다고 하자. 나아가 A 집단은 과거의 역사적 차별로 인해 p값

이 낮게 분포되어 있고, B 집단은 높게 분포되어 있다고 하자.[33] 이러한 상황은 아래와 같이 표시할 수 있다.

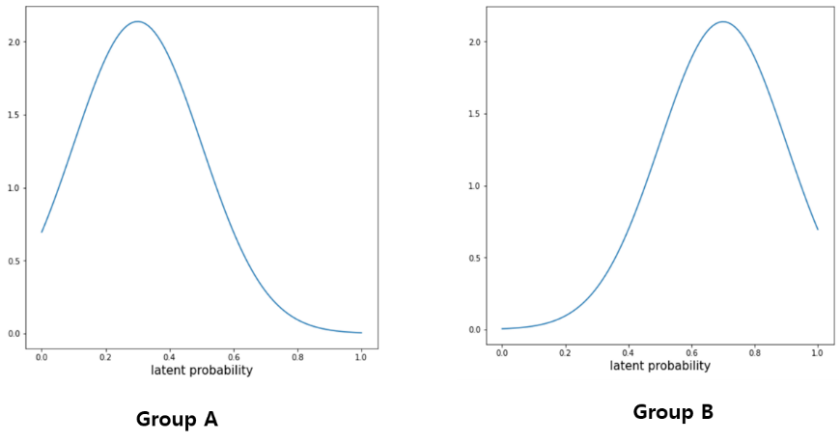


그림 1 집단간 능력치의 확률분포 가정

(2) 이처럼 가정된 p값으로부터 실제 능력에 따른 차이를 반영한 가상의 관측 가능한 변수를 생성한다. 이는 머신러닝 과정에서 '특성 추출(feature extraction)'에 해당하는 것으로 이해될 수 있다. 본 연구에서는 단순화를 위해 머신러닝 알고리즘이 X_0와 X_1이라는 2가지 특성을 관측할 수 있는 상황을 전제한다. 이때 X_0와 X_1 변수는 실제 p 값을 약간의 노이즈와 함께 반영할 뿐 특성 추출 과정 자체에서는 어느 집단에 대해서도 편향되어 있지 않은 시나리오와,[34] X_0와 X_1 변수가 집단에 따라 편향되어 생성되는 시나리오를 구분하여 고려한다.

(3) 학습 데이터 생성에 있어서는 p 값의 확률에 비례하여 목표변수 Y 값을 결정한다. 만약 p 값이 0.5일 경우 50%의 확률로 합격(1) 또는 불합격(0)으로

---

33 여기서는 모형을 만들기 위해 각 개인마다 p 값이 존재하는 것으로 가정하는 것일 뿐, 고정된 p 값이 존재하지 않을 수도 있다.

34 이러한 가정은 현실에 부합하지 않을 수 있다. 가령 개인의 지적 능력을 IQ 시험을 통해 측정하는 것은 사회문화적 편향을 반영할 수 있다. IQ 결정론에 대한 비판으로는 스티브 제이 굴드 지음, 김동광 번역, *인간에 대한 오해*, 사회평론, (2003) 참조.

결과가 정해진다. 이처럼 {(X_0, X_1), Y} 쌍으로 구성된 것이 학습 데이터가 된다. 위 가상의 사례에서 생성된 학습 데이터를 표시하면 아래 그림과 같다. A 집단이 과거 역사적 차별의 대상이 되어 온 것으로 가정했으므로, 학습 데이터 상으로 집단 A는 데이터상 불합격자(붉은 X 마크)가 더 많은데 비해, 집단 B는 합격자(파란 O 마크)가 더 많은 것을 확인할 수 있다. 한편, 만약 머신러닝 모형이 각 개인이 어떠한 집단 소속인지 알 수 있는 경우, 더미(dummy) 변수인 X_2 변수를 두어 집단 소속을 입력한다. 이 경우 학습 데이터는 {(X_0, X_1, X_2), Y} 쌍이 된다.

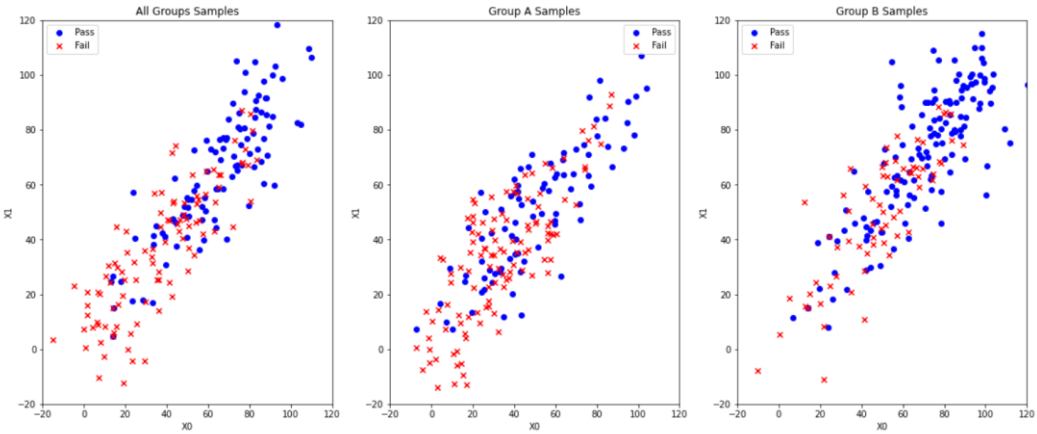


그림 2 생성된 가상의 학습 데이터

(4) 머신러닝 과정은 관찰된 특성을 이용하여 원래의 확률분포를 추정한다. 예컨대, 위 사례에서 머신러닝 모형은 X_0와 X_1이라는 관측된 특성을 이용하여, p의 분포를 추정한다. 그런데, 머신러닝 모형은 진정한 p값을 관측할 수 없고, 데이터 집합에 기록된 결과변수 $Y(0/1)$만을 관측할 뿐이다. 따라서 머신러닝 과정은 관측된 결과변수를 바탕으로 $\hat{p}$의 분포를 추론해 내는 역할을 한다. 이후 분류모형에 대해 검증 및 평가 데이터 집합을 활용하여 학습에 사용되지 않은 데이터에 대해 $\hat{p}$이 Y를 잘 예측하는지 여부를 검증한다.

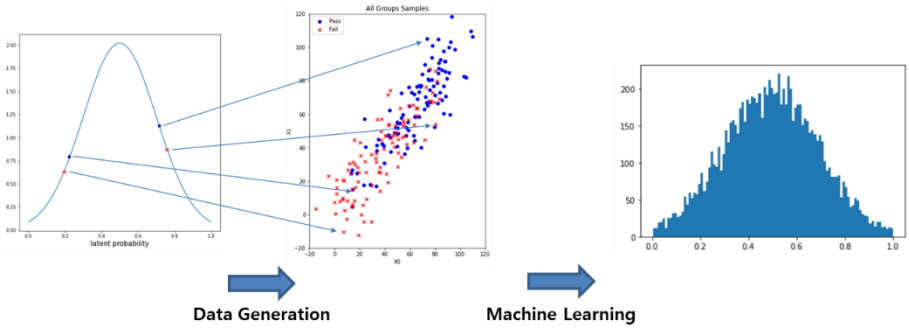


그림 3 시뮬레이션 상의 머신러닝 과정

(5) 마지막으로 알고리즘에 의한 의사결정 과정은 $\hat{p}$이 일정한 문턱 값(threshold) 이상인 사람을 합격시키는 것이라고 가정하는데, 본 시뮬레이션에서는 문턱 값으로 0.5를 사용하였다.

본 연구는 위와 같은 시뮬레이션을 통해 각각의 인공지능 공정성 지표가 어떠한 상황에서 발생하는 차별을 포착해 내는지를 살펴보고, 이를 통해 각 상황에 적합한 공정성 지표가 무엇인지 검토한다.

제3절 본 연구의 구성

본 연구의 구성은 다음과 같다. 우선, 제2장은 본 연구의 전제를 소개한다. 즉, 본 연구에서 사용되는 인공지능의 개념을 정의하고, 선행 연구를 기반으로 이를 유형화한다. 특히 본 연구는 인공지능을 크게 분류모형(classification model)과 생성모형(generative model)으로 구분하고, 분류모형을 다시 인지모형(cognitive model)과 판별모형(discriminative model)으로 구분한다. 나아가 후속 논의의 전제로서, 인공지능이 어떠한 방식으로 공정성을 침해할 위험이 있는지 개관한다.

제3장은 인공지능 공정성 보장을 위한 여러 실정법상 규제 및 현재 논의 중인 입법안들을 검토한다. 인공지능이 초래할 수 있는 위해를 사회적으로 규율하는 데 있어 여러 규제 형태가 가능하고, 현재 다양한 정책적 선택지를 고려한 입법활동이 활발히 전개되고 있다. 특히, 인공지능의 공정성 문제는 여러 법제에 의해, (부분적으로 중첩되어) 규율이 이루어지는 양상을 보이고 있는데, 본 연구는 이러한 규제(안)을 그 입법 목적 및 규율 방식을 기준으로 (1) 차별금지법제, (2) 개인정보 보호법제, (3) 인공지능 규제법제로 구분하여 살핀다.

제4장 및 제5장은 분류모형의 공정성 지표를 검토한다. 제4장은 이제까지 제안된 인공지능 공정성 지표 중 주요하게 고려되고 있는 것들을 조사한다. 본 연구는 선행 연구 및 정부 가이드라인, 보고서 등에 사용된 기준을 참조로 하여 인공지능 공정성 지표를 유형화하여 제시한다. 우선 (1) 입력변수에 대한 심사 기준과 (2) 출력변수에 대한 심사 기준으로 구분한다. 나아가 출력변수에 대해 적용할 수 있는 공정성 지표를 ① 결과 동등성(outcome parity), ② 오류율 동등성(error parity), ③ 예측도 동등성(predictive value parity)으로 구분하여 각각의 정의를 제시하고 그 의의와 한계를 살핀다. 다음으로 제5장은 시뮬레이션 결과를 통해 제4장에서 제시된 인공지능 공정성 지표가 어떠한 이용 사례의 맥락에서 정당화될 수 있는지 조사한다. 본 시뮬레이션에서는 인지모형과 판별모형을 구분하여 살피는데, 이러한 구분을 통해 각 유형마다 적용되어야 할 공정성 지표가 달라질 수 있다는 점을 보일 것이다. 나아가 시뮬레이션 결과를 바탕으로 기존 지표가 갖는 한계를 확인하고, 그에 대한 대안적 인공지능 감사 방안을 제시한다.

제6장은 생성모형에 관한 공정성 지표를 검토한다. 우선 선행 연구에서 활용되어 온 생성모형의 공정성 개념은 (1) 편향 (2) (좁은 의미의 공정성) (3) 유독성의 차원으로 구분될 수 있고, 특히 그중 편향에 대한 평가는 ① 대표성 평가, ② 고정관념화 평가, ③ 폄훼 평가로 나눌 수 있다는 점을 보인다. 나아가 현재 생성모형의 공정성 보장을 위해 활용되고 있는 기술적 조치를 학습 데이터 구축,

생성모형 학습 및 생성모형 이용의 각 단계로 나누어 조사한다. 마지막으로 이상의 공정성 평가 기준과 공정성 보장을 위한 기술적 조치를 고려하여 생성모형의 공정성에 대한 적절한 법적 규율 방안을 모색한다. 특히 최근 진행되고 있는 이른바 기반모형(foundation model)으로의 패러다임 전환과 프롬프트-기반 인공지능 도입 전략의 확산이 이제껏 논의되어 온 인공지능 규제 방안에 어떠한 영향을 끼치는지 살핀다.

제2장 인공지능의 개념과 유형

제 1 절 인공지능의 개념

1. 인공지능 개념에 관한 일반적 이해

일상적인 용례에서 인공지능은 흔히 '인간처럼 사고하는 기계'로 여겨진다.[35] 하지만 이러한 이해는 인공지능에 관한 논의에 있어 적절한 기준으로 삼기 어렵다. 최근의 인공지능은 인간의 신경망 구조로부터 영감을 받아 구현된 것이기는 하지만,[36] 그와는 구분되는 통계적 기법에 따라 구현되고 있으므로 인간의 '사고' 과정을 모방한 것이라고 보기 어렵다.[37] 또한 인공지능은 인간이 수행하기 어려

---

35 Harry Surden, "Artificial Intelligence and Law: An Overview", *Georgia State University Law Review* 35, (2019).

36 Warren McCulloch and Walter Pitts, "A logical calculus of the ideas immanent in nervous activity", The bulletin of mathematical biophysics 5, (1943). 위 연구는 생물학적 뉴런의 구조와 기능에 기반한 계산 모형을 제시함으로써 이후 인공신경망 연구의 기초를 마련했다고 평가된다.

37 특히, 현재 광범위화게 활용되고 있는 트랜스포머(Transformer) 모형은 생물학적 뉴런 구조에 영감을 받은 것이 아니라, 종래의 자연어 처리 및 머신러닝 분야의 수학적 발전에 기반한 것으로 평가될 수 있다. 트랜스포머 모형을 제시한 Ashish Vaswani, et al., "Attention is all you need", *Advances in neural information processing systems* 30, (2017) 연구는 멀티-헤드 셀프 어텐션 매커니즘(multi-head self-attention mechanism) 및 위치 인코딩(positional encoding) 기법을 사용하는데, 이는 종래 생물학적 뉴런 구조에 영감을 받은 순환신경망(recurrent neural network)이나 합성곱 신경망(convolutional neural network) 구조와는 구별되는 것이다. 이처럼 초기 인공신경망 연구는 생물학적 신경망 구조에 큰 영감을 받았지만, 최근의 인공신경망 발전 동향은 이러한 영향에서 벗어나 수학적 개념이나 새로운 혁신적 구조를 제시하는 데 중점을 두고 있다고 평가할 수 있다.

운 작업에 대해 우수한 능력을 보이거나 그 반대인 경우도 적지 않다.[38] 따라서, 이처럼 '인간성'을 기준으로 인공지능의 개념을 이해하는 것이 적절하지 못하다.[39]

그 대신 대부분의 연구자들은 인공지능의 개념을 정의하기 위해 '합리성(rationality)' 개념을 사용한다. 이는 최적화할 수 있는 기준과 가용한 자원이 주어진 경우, 어떠한 목적을 달성하기 위한 최선의 행동을 선택할 수 있는 능력을 가리킨다.[40] 이러한 접근에 따르면 인공지능은 결국 어떠한 소프트웨어 및/또는 하드웨어로 구현된 기계-기반 시스템으로서 합리성을 구현하고 있는 것이라고 볼 수 있다. 그러나 이러한 합리성에 기반한 인공지능 개념 이해는 현재 컴퓨터가 사용되는 대부분의 활용 사례를 인공지능의 개념으로 포섭하는 문제를 초래할 수 있다. 예를 들어 2018년 뉴욕시는 자동화된 의사결정 시스템의 태스크포스를 수립하였는데, 이는 학계, 산업계, 시민단체 및 정부 기관의 전문가로 구성되어 시정부에 자동화된 의사결정 시스템에 관한 권고를 제공하는 것을 목표

---

38 현재 발전하고 있는 인공지능은 인간 수준의 지능을 구현하는 것이 아니라 동물의 능력을 모사하는 것에 불과하다는 견해도 제기되는데, 이에 관하여는 Adnan Darwiche, "Human-level intelligence or animal-like abilities?", *Communications of the ACM* 61(10), (2018) 참조.

39 Stuart Russel and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th edition, Pearson, (2020)은 인공지능은 어떠한 행위주체(agent)가 갖는 특성을 지칭하는 것으로서, (1) 그것의 사고 과정에 관한 것인지, 아니면 행동 방식에 관한 것인지, (2) 인간의 능력을 기준으로 하는지, 아니면 합리성을 갖는지를 기준으로 하여, 인공지능을 다음 4가지로 구분할 수 있다고 설명한다. 즉, ① 인간적으로 행위하는 것(acting humanly), ② 인간적으로 사고하는 것(thinking humanly), ③ 합리적으로 사고하는 것(thinking rationally) ④ 합리적으로 행위하는 것(acting rationally)으로 구분해 볼 수 있다. 이러한 구분을 기초로 이들은 '인간성' 대신 '합리성'을 기준으로 하고, '사고 과정' 대신 '행동'을 기준으로 삼아, 인공지능이란 주어진 상황에서 "최상의 결과(불확실성이 존재할 경우 기대할 수 있는 최상의 결과)를 내도록 행동"하는 합리적 행위주체(rational agent)라 정의 내린다.

40 High-Level Expert Group on Artificial Intelligence (AI HLEG), *A definition of Artificial Intelligence: main capabilities and scientific disciplines*, (2019. 4. 8).

로 하였다.[41] 그런데 이들은 엑셀 스프레드시트를 사용하는 것이 "자동화된 의사결정 시스템"에 해당하는 지와 같은 문제를 논의하면서 첫 수 개월 동안 태스크포스의 업무 대상의 범위가 어떻게 되는지에 관해 합의를 이루지 못하였다고 한다.[42] 더욱이 시 정부는 위 태스크포스가 "새로운 인공지능" 기술을 사용하는 시스템을 다루어야 하는 것이지, 이제껏 사용되어 온 알고리즘 시스템에 대한 재검토를 하는 기구가 아니라는 입장을 유지하였다. 이러한 사례는 포괄적인 인공지능에 대한 개념 정의를 활용하는 것이 실무적으로 유의미한 논의와 충돌할 수 있음을 보여준다.

인공지능 개념에 대한 보다 제한적인 접근법은 구체적 기술적 방법론에 기초하는 것이다. 현재의 주된 인공지능에 관한 논의는 '데이터'를 통해 일반화된 추론 규칙을 학습할 수 있는 머신러닝 기법, 그 중 특히 딥러닝에 기반한 시스템에 초점을 맞추고 있다. 이러한 현실을 고려하여 인공지능 시스템을 구현하기 위한 기술을 구체적으로 특정하는 방식으로 인공지능 개념 이해를 시도할 수도 있다. 예를 들어, 미국 국립표준기술연구소(National Institute of Standards and Technology, 이하 "NIST")는 '인공지능 위험 관리 프레임워크' 초안에서 데이터 학습 분야에 초점을 맞추어, 인공지능 시스템을 "자동화되었거나 반자동화된 방식으로 **데이터로부터 학습하는** 알고리즘 프로세스"(역자 강조)라 정의하고 있다.[43] 다른 한편 EU AI 법안의 2021년 초안 상의 '인공지능 시스템'에 대한 정의 조항은 인공지능 시스템을 개발하는데 사용된 기법을 특정한 바 있다.[44] 다만, 이

---

41 Karen Levy, Kyla E. Chasalow, and Sarah Riley, "Algorithms and decision-making in the public sector", *Annual Review of Law and Social Science* 17, (2021), p. 311.

42 위의 글.

43 National Institute of Standards and Technology (NIST), *AI Risk Management Framework: Initial Draft*, (2022).

44 이는 EU AI 법안의 부속서 I에 다음과 같이 나열되어 있었다: (a) 딥러닝을 포함하여, 다양한 방법을 사용하는 지도학습, 비지도학습, 강화학습을 포함한 머신러닝 접근법, (b) 지식 표현, 귀납적 (논리) 프로그래밍, 지식기지, 추론 및 연역 엔진, (기호적) 추론, 전문가 시스템을 포

러한 접근은 빠르게 변화하는 기술 환경에 유연하게 대응하기 어렵다는 문제가 있다.

이에 현재 널리 통용되고 있는 인공지능의 개념은 인공지능이 수행하는 '기능'에 초점을 맞춘 것으로 평가할 수 있다. 국제적으로 OECD 인공지능 원칙에 정의된 인공지능의 개념이 널리 인용되는데, 이는 인공지능 시스템을 "주어진 인간이 정의한 목표에 대해 현실 또는 가상 환경에 영향을 미치는 예측, 추천, 또는 결정을 할 수 있는 기계 기반 시스템으로서, 다양한 수준의 자율성에서 운영되도록 설계된 것"이라 정의한다.[45] 이러한 이해는 본질적으로 인공지능의 기능, 즉, 일정한 자율성을 갖고 예측, 추천, 또는 결정을 수행하는 것을 중심으로 인공지능의 개념을 정하는 것이라 할 수 있다. 최근 유럽 의회를 통과한 EU AI 법안 수정안도 OECD가 정의한 개념과 유사한 정의 조항을 두고 있다.[46]

2. 인공일반지능의 문제

다른 한편, 인공지능의 개념과 관련하여 또다른 중요한 구분은 특정한 인공지능이 수행할 수 있는 과제의 범위에 관한 것이다. 이러한 기준에 따르면 인공지능을 '인공일반지능((Artificial General Intelligence)'과 '좁은 인공지능(Artificial Narrow Intelligence)'으로 구분할 수 있다. 전자를 '강(强) 인공지능(strong AI)', 후자를 '약(弱) 인공지능(weak AI)'으로 칭하기도 한다.[47] 좁은 인공

함한 논리 및 지식 기반 접근법, (c) 통계적 접근법, 베이지안 추정, 검색 및 최적화 방법. 그러나 이러한 규정은 지나치게 적용범위가 넓다는 비판을 받았다. 이에 2023. 5. 9.자 유럽의회의 수정안에는 이처럼 기술적 접근을 나열하는 접근을 더 이상 취하고 있지 않고 있다.

45 OECD, *Recommendation of the Council on Artificial Intelligence*, OECD Publishing, (2019).

46 EU AI 법안(2023. 5. 9.자 수정안) 제3조 제1항은 인공지능 시스템을 "다양한 수준의 자율성으로 운영되도록 설계되고, 명시적 또는 암묵적 목적에 대해 물리적 또는 가상의 환경에 영향을 미치는 예측, 추천, 또는 결정과 같은 출력을 생성할 수 있는 기계 기반 시스템"이라 정의한다.

47 고학수 외, *사법부에서의 인공지능(AI) 활용 방안*, 법원행정처, (2020), p. 4.

지능은 인간이 지정한 특정 목적의 작업만을 수행하는 경우를 지칭하는 것인데 비해, 인공일반지능은 "인간 또는 그 이상의 수준의 능력으로, 추론·계획·경험 학습 등의 폭넓은 지능 역량(broad capabilities of intelligence)"을 보여주는 시스템을 가리킨다.[48]

예를 들어, OpenAI가 2023년 3월 발표한 GPT-4는 이미지와 텍스트를 입력 받아 텍스트를 생성할 수 있는 멀티-모달(multi-modal) 언어모형으로서, 전례 없는 규모의 컴퓨팅 및 데이터를 사용하여 학습되었다.[49] 그런데, 마이크로소프트 연구자들이 GPT-4의 미공개 초기 버전을 평가한 결과에 따르면, 이 모형은 그저 언어를 이해하는 것을 넘어 수학, 코딩, 시각, 의학, 법, 심리학 등의 영역에서 새롭고 어려운 과제를 해결할 수 있고, 이러한 모든 작업에 있어 종전 모형의 수준을 크게 능가할 뿐만 아니라, 놀라울 정도로 인간 수준의 성능에 가까운 것으로 평가되었다. 위 연구자들은 GPT-4가 보이는 능력의 범위와 깊이를 고려할 때 "AGI 시스템의 아직 불완전한 초기 버전으로 합리적으로 간주될 수 있다"고 주장하였다.[50] 이러한 연구자들이 주장하는 바와 같이, GPT-4 등의 대규모 언어모형 인공지능이 인공일반지능으로 발전할 수 있는지, 만약 그러하다면 인공일반지능이 어떠한 위해를 초래할 수 있는지는 여전히 불확실하다. 이에 2023. 3. 22. GPT-4보다 강 인공지능 연구를 즉각적으로 최소한 6개월 동안 중단할 것을 요구하는 공개서한이 발표되기도 하였다.[51] 현재 대규모 언어모형 인

---

48 Sébastien Bubeck, et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4", (2023).

49 OpenAI, "GPT-4 Technical Report", (2023).

50 Bubeck, et al. (2023) (각주 48).

51 이 공개서한에는 요슈아 벤지오, 스튜어트 러셀 등 유명 인공지능 연구자들과, 일론 머스크, 유발 하라리 등의 사회 유명 인사가 서명하였다. Future of Life Institute, "Pause Giant AI Experiments: An Open Letter", (2023. 3. 22.) https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (2023. 5. 26. 접속 확인) 위 공개서한은 다음과 같은 의문을 제기한다: "기계가 우리의 정보 채널을 선전 및 허위 정보로 범람시키는 것을 허용해야 하는가? 우리는 모든 일자리, 심지어 성취감 있는 일자리마저도 모두 자동화해야 하는가?

공지능의 잠재력과 위험성에 대한 논의는 활발히 이루어지고 있고, 이에 대한 깊이 있는 성찰과 논의가 필요하다.

그러나 본 연구는 인공일반지능이 고유하게 제기하는 위험성을 논의의 범위에 포함하지 않는다. 그 이유는 다음과 같다. 우선, 아직 현실에서 활용되고 있는 인공지능은 추상적 추론, 개념의 인식, 유연한 이해, 일반적 문제 해결 능력 등의 광범위한 인간 지능을 기계로 구현하고 있지는 못하고 있다.[52] 다만, 인공일반지능이 가까운 미래에 도래할 가능성이 있고, 일례로 한 연구는 2050년까지 인간 수준의 일반 인공지능이 개발될 확률이 60%에 이른다고 예측하기도 한다.[53] 따라서 인공일반지능이 예상보다 일찍 발전하여 사회에 중대한 변화를 초래할 가능성을 배제하기는 어렵다.

종국적으로 우리보다 수적으로 더 많고, 더 똑똑하며, 우리를 진부하게 만들고, 대체할 가능성이 있는 비인간적인 마음을 개발해야 하는가? 우리는 문명의 통제력을 잃을 위험을 감수해야 하는가? 이러한 결정은 선출되지 않은 기술 지도자들에게 위임되어서는 안 된다." 그러나 위 공개서한이 지적하는 허위조작정보 범람 위험, 노동력 대체 위험, 장기적 인류 존속에 대한 위험은 공상과학적 상상력에 근거한 것이라는 비판도 제기된다. Kapoor and Narayanan은 GPT-4 등 발전한 인공지능으로 인해 현실 세계에서 초래될 수 있는 실질적 위험은 부정확한 도구에 대한 과도한 의존, 중앙집중화된 권력과 노동 착취, 단기적인 안보 위험이라고 지적한다. Sayash Kapoor and Arvind Narayanan, "A misleading open letter about sci-fi AI dangers ignores the real risks", (2023. 3. 30.) https://aisnakeoil.substack.com/p/a-misleading-open-letter-about-sci (2023. 5. 26. 접속 확인)

52 대표적으로 2018년 튜링상 수상자인 얀 르쿤(Yann LeCun)은 "인간 수준의 AI는 아직 도래하지 않았다"고 단언한다. https://www.linkedin.com/posts/yann-lecun_calm-down-human-level-ai-isnt-here-yet-activity-7043650808705359875-viNx?utm_source=share&utm_medium=member_desktop (2023. 5. 26. 접속 확인). 그는 현재의 ChatGPT와 같은 모형들은 "주위 세상에 대한 지식이 전혀 없고, 세상이 존재한다는 것을 알지 못한다. 또한 물리적 현실에 대한 생각이 전혀 없다. 이 모형들은 어떤 맥락도 가지고 있지 않고, 답을 찾을 수 없다."고 지적한다. Ben Wodecki, "WAICF '23: Yann LeCun Sets Us Straight on Generative AI", *AI Business*, (2023. 2. 11.)

53 Ajeya Cotra, "Two-year update on my personal AI timelines", *AI Alignment Forum*, (2022. 8. 3.). 위 글은 인공일반지능이라는 표현 대신 "변혁적 인공지능(Transformative AI)"

그러나 본 연구는 인공지능에 적용되어야 할 공정성 기준을 수립하는 것을 목적으로 하는 바, 인공지능의 공정성에 관한 평가는 인공지능이 구체적으로 활용되는 하류 과제(downstream task)가 무엇이고, 그러한 인공지능이 어떠한 맥락에서 활용되는지를 고려하여 선택, 적용되어야 한다. 따라서, 인공지능 공정성 평가에서 특정한 과제를 수행하는 좁은 인공지능을 전제로 하는 것은 불가피한 측면이 있다. 나아가, 인공일반지능이 어떠한 과제를 수행하는 상황에서는 그 과제를 좁은 인공지능이 수행하고 있는 것과 마찬가지의 기준을 적용하여 평가할 수 있을 것이다.[54]

3. 본 연구의 정의

이상과 같은 점들을 고려하여, 본 연구는 본 연구에서 사용하는 인공지능의 개념을 기능적으로(functionally) 다음과 같이 정의한다. 본 정의는 특정 대상에 관한 다차원의 특성(feature) 변수(X)와 그 변수와 연관된 종속 변수(Y)가 존재

이라는 표현을 사용하는데, 이는 "과학기술 발전을 가속화하는 필요한 모든 인간 활동을 본질적으로 자동화할 수 있는 인공지능 시스템"을 뜻한다. 이러한 정의는 인공지능이 스스로 과학기술을 발전시키는 "특이점(singularity)"이 도래할 위험성에 초점을 맞춘 것이라고 평가할 수 있다. 특이점의 개념에 관해서는 레이 커즈와일, 김명남 옮김, *특이점이 온다: 기술이 인간을 초월하는 순간*, 김영사, (2007) 참조.

54 인공일반지능은 초래할 위험성이 있는 위험으로는 크게 (1) 'AI-takeover'의 가능성과 (2) 인공일반지능으로 인해 권위주의적 이념이 고착화될 우려가 거론된다. 우선 'AI-takeover'는 인공지능이 재귀적으로 스스로를 개선할 수 있는 단계에 이르게 되면, 점점 더 나은 버전을 디자인하여 빠르게 인간을 뛰어 넘는 '슈퍼인텔리전스'에 이를 수 있다는 주장이다. 이러한 슈퍼인텔리전스는 스스로 자원을 확보하고, 세계를 지배하며, 인간을 제거하거나 영구적으로 억압하는 지위에 이를 수 있다. 슈퍼인텔리전스는 인간보다 훨씬 더 큰 능력을 보유하고 있으므로, 일단 슈퍼인텔리전스가 개발·배치되고 나면 인간이 이러한 시나리오를 막는 것은 불가능할 수 있다. 두번째 우려는 권위주의적 국가나 지도자가 인공일반지능을 활용하여 지배적 이념을 고착시킬 가능성이다. 권위주의적 지도자는 사상의 다양성을 억압하고, 인공지능 경찰을 통해 이에 대한 반란을 예방하며, 인공지능 군대를 활용하여 체제 전복의 가능성을 제거할 수도 있다. Hilary Greaves and William MacAskill, "The case for strong longtermism", *Global Priorities Institute Working Paper* No. 5-2021, (2021).

하는 상황을 전제한다. 예를 들어, X변수는 얼굴인식 모형에서의 얼굴 사진, 개인신용평가 모형에서의 개인신용정보에 관한 데이터(자산, 소득, 금융거래이력 등) 등이 될 수 있다. 이 경우 연관된 Y변수는 해당 얼굴사진과 연관된 개인의 신원이나 해당 개인의 채무불이행 위험율 등에 해당할 것이다. 이러한 경우 '분류모형(classification model)'은 다음과 같은 대응 관계를 구현한 기계-기반 시스템이다.[55]

$$f_{classfier} : X \longmapsto \hat{Y}$$

$f_{classfier}$는 어떤 데이터가 입력변수로 주어지면 그에 대한 어떠한 목표변수를 출력한다. 여기서 Y변수가 0 또는 1의 유형(class)만을 갖는 경우 이진 분류기(binary classifier)라 하고, 그 이외에 다양한 유형을 가질 수 있는 경우를 다중 또는 다진 분류기(multiclass classifier 또는 multilabel classifier)라 할 수 있다.[56] 현재의 주류적 분류모형은 어떠한 X가 주어진 경우 그에 대응하는 특정 Y유형에 해당할 확률, 즉 $Pr(Y|X)$을 산출하는 방식으로 작동한다.

이러한 분류모형은 최근의 여러 인공지능 활용 사례에서 발견된다. Ng (2016)은 "최근의 인공지능의 발전 중 거의 대부분은 어떠한 주어진 입력 데이터(A)에 대해 신속하게 단순한 답변(B)를 생성해 내는 유형"으로 이루어져 있다

---

55 분류모형의 출력 값은 $\hat{Y}$이라 표기로 표기하는데, 이는 예측된 Y값을 모델링하고자 하는 진정한 Y값(즉, 학습 데이터에 포함된 값)과 구분하기 위한 것이다.

56 다항(multiclass) 분류는 2개 이상의 유형 중 하나의 유형으로만 분류하는 경우이고, 다진(multilabel) 분류는 2개 이상의 유형 중 여러 유형에 동시에 속할 수 있는 경우라 구분할 수 있다. 가령 어떠한 사진에 한 개의 사물만이 포함되어 있고 그 사물이 무엇인지 인식하는 경우 다항 분류이지만, 어떠한 사진에 여러 사물이 함께 포함되어 있고 여러 사물을 함께 인식하는 경우에는 다진 분류라 한다. Markus Odenthal, "Multi-Class vs. Multi-Label Classification: What's the Difference?", *Medium*, (2022. 5. 19.).

고 지적한다.[57] 그는 이미지 인식, 대출 심사, 광고 추천, 음성 인식, 기계 번역,[58] 제품 하자 검출, 사물 인식 등의 다양한 사례를 들고 있는데, 이는 본 연구에서 고려하는 분류 모형의 대표적인 예라고 할 수 있다.

반대로, 어떤 주어진 유형 Y로부터 특정 대상에 대한 정보 X를 생성해 내는 역함수 관계도 구현할 수 있는데, 본 연구는 이러한 대응관계를 구현한 기계-기반 시스템을 '생성모형(generative model) 인공지능'이라 지칭한다.

$$f_{generator}: Y \rightarrow \hat{X}$$

최근 널리 활용되는 대표적 생성모형으로는 어떠한 입력 프롬프트에 대응하여 답변을 생성해 내는 언어모형(language model)이나, 어떠한 이미지에 대한 설명을 입력 받아 그에 대응하는 이미지를 생성해내는 텍스트-이미지 변환(text-to-image) 모형 등이 있다. 생성모형 인공지능은 주어진 라벨 Y에 대응하는 특성 X가 존재할 확률, 즉 $Pr(X|Y)$을 산출하는 확률분포 모형을 구축한 다음, 그로부터 확률이 높은 $\hat{X}$을 표본추출(sampling)하는 방식으로 구현된다. 이를 통해 현실에서 "그럴 듯"해 보이는 $\hat{X}$을 생성해 낼 수 있게 된다.

한편, 이상에서 정의한 분류모형과 생성모형은 베이즈 공식에 따라 다음과 같은 관계를 갖는다.

$$Pr(X|Y) \propto Pr(X)\, Pr(Y|X)$$

---

57 Andrew Ng, "What AI Can and Can't Do", *Harvard Business Review* (2016).

58 음성 인식이나 기계 번역은 외견상 새로운 텍스트를 생성하는 것이므로, 아래에서 설명하는 생성모형으로 구분하는 것이 타당한 것은 아닌지 의문이 제기될 수 있다. 그러나 음성 인식이나 기계 번역 시스템은 인식 또는 번역할 수 있는 토큰(token)이 한정되어 있고, 주어진 입력에 대응하는 토큰에 대한 확률을 계산하고, 가장 높은 확률에 대응하는 토큰을 출력하는 방식으로 구현된다. 따라서 여전히 $\Pr(y|x)$를 계산하는 분류모형으로도 고려될 수 있다.

위 수식의 함의는 다음과 같다. 만약 (1) 어떤 X변수에 대응하는 Y변수를 잘 분류해 낼 수 있는 분류모형 $Pr(Y|X)$과 (2) X변수의 분포에 대한 적절한 확률분포 모형 $Pr(X)$를 적절하게 구현한다면, 어떤 Y변수에 대응하는 X변수를 잘 생성할 수 있는 생성모형 $Pr(X|Y)$을 구현할 수 있다. 그런데, 전술한 바와 같이 이미지나 텍스트와 같은 다차원적 데이터에 대한 분류모형은 이미 잘 구현되어 있는 경우, 높은 성능의 생성모형을 구축하기 위해서는 X변수에 대한 확률분포에 대한 모형을 구축하는 과제가 남게 된다. 이에 최근 빠르게 발전하고 있는 대규모 언어모형이나 이미지 생성 인공지능은 이러한 $Pr(X)$ 확률분포 모형을 정교하게 구축한 것이다.[59] 예를 들어 텍스트 생성을 위해 사용되는 언어모형은 텍스트에 대한 확률분포, 즉 $Pr(text)$를 구현한 것이고, 이미지 생성을 위해 사용되는 여러 모형은 이미지의 확률분포, 즉 $Pr(image)$를 구현한 것이다.[60]

요컨대, 본 연구는 인공지능을 (1) 주어진 다차원 속성변수 X를 그와 연관된 종속변수 Y에 대응시키는 분류모형 $f_{classfier}$와 (2) 반대로 주어진 Y에 대응하는 X를 출력[61]하는 생성모형 $f_{generator}$을 통칭하는 것으로 정의한다. 나아가 이러한 정의는 인공지능이 어떠한 기능을 수행하는 지를 기초로 한 것이므로, 이를 구현하는 방법(예컨대 머신러닝 방식에 의하는지, 아니면 규칙 기반 방식에 의하는지)에 따라 한정되지 않는다. 본 연구가 이처럼 광범위한 정의를 취하는 이유는 이

---

59 한편, 고차원인 X 변수에 대한 확률분포 모형을 효율적으로 구축하기 위하여 고차원 X 변수를 저차원 임베딩 변수 Z로 표상하는 기법들이 사용될 수 있다. 이와 같은 생성모형은 $Pr(X|Z)$로 서술될 수 있다. Sangchul Park, "Heterogeneity of AI-Induced Societal Harms and the Failure of Omnibus AI Laws", arXiv preprint arXiv:2303.11196 (2023) 참조. 예를 들어 DALL-E 2에서 사용되는 CLIP 모형의 경우 이용자 입력 프롬프트 텍스트 Y로부터 임베딩 변수 Z를 도출하고, 해당 임베딩에 대응되는 $Pr(X|Z)$로부터 이미지를 생성해 낸다. Aditya Ramesh, et al., "Hierarchical text-conditional image generation with CLIP latents", arXiv preprint arXiv:2204.06125 (2022) 참조.

60 해당 기술에 대한 구체적 검토는 본 장 제2절 참조.

61 한편, 특정한 유형에 관한 데이터를 생성할 필요가 없는 경우에는 Y를 입력 받지 않고 임의의 Y에 대해 그저 현실에 존재할 법한 $\hat{X}$를 생성해 낼 수도 있다.

로써 다양한 기술적 구현 가능성을 포괄함으로써 인공지능이 초래할 수 있는 여러 위해를 분석 대상에 포함시키기 위한 것이다.[62]

제 2 절 인공지능의 유형화

본 절은 앞서 정의한 인공지능을 그 사회적 기능과 기술적 구현 방식을 기준으로 유형화한다. 본 연구는 분류모형을 (1) 인지적 기능을 수행하는 '인지모형(cognitive model)'과 (2) 대상 정보에 기반한 판별적 기능을 수행하는 '판별모형(discriminative model)'으로 구분한다.[63] 한편 생성모형을 생성 대상 데이터의 유형을 기준으로 (1) 정형 데이터(structured data) 생성모형과 (2) 비정형 데이터(unstructured data) 생성모형으로 구분한다. 나아가 이하에서는 위 모형들이 구체적으로 어떠한 기술을 통해 구현되고 있는지 개관한다.

1. 분류모형의 유형화

가. 인지모형의 특징

인지모형은 이미지 인식, 번역, 음성인식 등의 사례에서 보는 것과 같이 이미지, 영상, 텍스트, 음성 등과 같은 비정형 데이터를 인식하는 것을 주된 목표로 한다. 종래에는 비정형 데이터를 인식하는 과제는 컴퓨터로 효율적으로 처리하기 어려운 분야로 여겨져 왔다. 그러나 2010년대 이후 딥러닝 접근법이 적극

---

62 다만, 이처럼 포괄적인 정의는 인공지능에 대한 현실적인 구체적 규제 문제를 직접적으로 다루는데 적용되기에는 적절하지 않을 수 있다는 점에 유의할 필요가 있다.

63 이러한 구분 용례는 Park (2023)의 예에 따른 것이다. Park (2023) (각주 **오류! 책갈피가 정의되어 있지 않습니다.**) 참조.

적으로 활용됨으로써 이러한 어려움을 극복하게 되었다. 대표적으로 사물 인식 문제에 있어 딥러닝 기술을 적용함으로써 오류율을 매우 크게 낮출 수 있음이 확인되었다.[64] 예컨대 이미지넷 챌린지(ImageNet Challenge) 과제는 1,400만 건 이상의 이미지를 20,000개 이상의 카테고리로 분류한 데이터 집합을 학습한 뒤, 학습에 사용되지 않은 새로운 이미지에서 사물의 카테고리를 식별해 내는 것이다. 그런데, 딥러닝(deep learning) 기술이 활용되기 전인 2011년까지는 오류율이 25% 이상으로 매우 높은 수준이었으나, 딥러닝 기술이 적용되기 시작하면서 매년 큰 감소 추세를 보였고, 그 결과 2015년에는 사물 인식에 대한 오류율이 3.57%까지 낮아지는 놀라운 성과를 거두었다.[65]

이처럼 인지모형에 널리 활용되는 딥러닝 기법은 심층 인공신경망(deep neural network)를 학습시키는 기법을 일컫는다.[66] 심층 인공신경망을 활용함으로써 얻게 되는 중요한 장점은 다양한 계층화된 패턴을 인식할 수 있게 된다는 것이다.[67] 예를 들어 이미지 인식에 널리 활용되는 합성곱 신경망(convolutional neural network)는 다수의 계층으로 구성되는데, 앞 단계에 위치한 신경망 계층에서는 직선, 곡선, 얼룩 등의 단순한 패턴을 인식하고, 이후 단계에 위치한 신경망 계층은 앞서 인식된 단순한 패턴을 조합하여 보다 복잡한 패턴이 존재하는 여부를 식별하는 등의 단계를 거치게 된다. 나아가 수십 내지 수백 개에 이르는 더 많은 계층을 이용함으로써 다양한 유형의 복잡한 사물을 인식할 수 있음이 확인되었다.

---

64 Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning*, MIT Press, (2016), p.27.

65 위의 글, p.28.

66 위의 글, p.13.

67 Tom Taulli, *Artificial Intelligence Basics: A Non-Technical Introduction*, Apress, (2020), p. 71.

나. 판별모형의 특징

판별모형은 인지모형과 달리 주로 정형 데이터를 입력 변수로 받아 주어진 과제를 수행한다. 그 예시로 대출 심사를 수행하는 판별모형을 고려해 볼 수 있다. 머신러닝 기법을 활용한 판별모형을 구축하여 신청자의 자산 점수와 소득 점수를 이용해서 대출 승인 여부를 판단하는 상황을 고려하자. 이때 판별모형을 구축하기 위해서는 우선 과거의 인간 심사자에 의한 데이터를 수집할 필요가 있다. 아래 그림은 임의로 생성된 가상 학습 데이터인데, O점은 대출이 승인된 사례이고, X점은 대출이 거절된 사례이다.

	A	B	C
1	Asset	Income	Pass?
2	34.62366	78.02469	0
3	30.28671	43.895	0
4	35.84741	72.9022	0
5	60.1826	86.30855	1
6	79.03274	75.34438	1
7	45.08328	56.31637	0
8	61.10666	96.51143	1
9	75.02475	46.55401	1
10	76.09879	87.42057	1

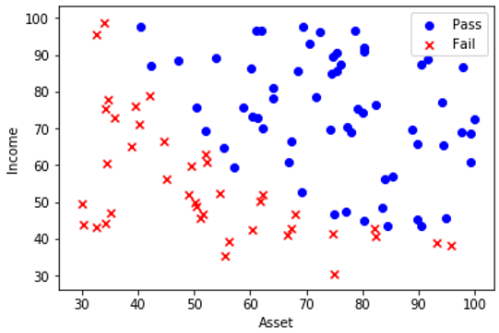


그림 4 가상의 대출심사 학습 데이터

판별모형에 대해서는 이와 같이 표 형식으로 구축된 정형 데이터가 입력변수로 주어진다. 머신러닝 기법을 통한 학습 과정은 위 데이터를 활용하여 '결정 경계선(decision boundary)'를 도출한다. 도출된 결정 경계선은 학습 데이터에 포함되어 있지 않은 새로운 사례에 대해 예측을 수행하기 위해 활용된다. 고도화된 비선형(non-linear) 모형을 이용하면, 매우 복잡하게 분포된 데이터라고 하더라도 높은 성능으로 분류하는 결정 경계선을 찾을 수 있게 된다. 위 사례에서 구축된 판별모형은 학습 데이터에 포함되어 있지 않은 새로운 사례에 대해 대출 가능 확률을 신속히 예측할 수 있다.

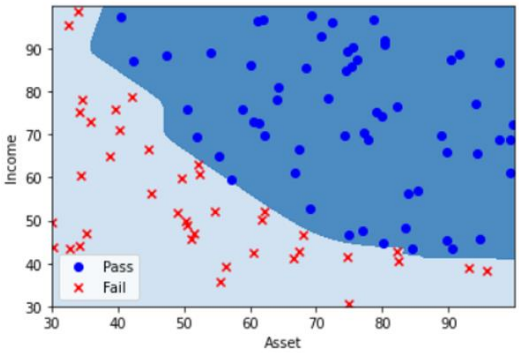


그림 5 학습된 대출심사 판별모형의 결정경계선 예시

위 사례에서는 설명을 단순화하기 위해 두 개의 변수(자산점수, 소득점수)만을 예로 들었지만, 현실에서 활용되는 판별 모형에는 훨씬 더 많은 정보가 사용될 수 있다. 가령 개개인이 업로드한 소셜 미디어 게시물, 온라인 쇼핑몰의 구매 이력, 웹 사이트 방문 이력, 검색 이력 등과 같은 방대한 행태 정보를 활용하여 개개인의 선호와 소비능력을 추론할 수 있다.[68] 이러한 분류 모형은 빅데이터로부터 개인에 대한 평가나 행동 예측을 추출하는 사회적 기능을 수행한다.

판별모형이 인지모형과 구별되는 또 다른 특징은 이를 구현하기 위해 고전적 머신러닝 기법이 주로 활용되고 있다는 점이다. 판별모형을 구현하기 위해서는 로지스틱 회귀와 같은 선형 모형, 트리 기반 모형, 서포트 벡터 머신 등 다양한 기법이 활용될 수 있는데, 특히 XGBoost와 같은 트리 기반 앙상블 모형이 여전히 주로 사용된다.[69] 이처럼 딥러닝 기법이 아닌 고전적 머신러닝 모형이 선호되는 이유는 정형 데이터가 비정형 데이터와 구분되는 특징이 있을 뿐만 아니라, 판별모형의 경우 투명성이나 해석 가능성이 중요하게 요구되기 때문이다.[70] 즉,

---

68 김병필, "인공지능은 우리로부터 무엇을 추출하는가?", *과학기술과 사회* 제3호, 알렙, (2022b).

69 Tianqi Chen and Carlos Guestrin, "XGBoost: A scalable tree boosting system", *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, (2016).

70 Ravid Shwartz-Ziv and Amitai Armon, "Tabular data: Deep learning is not all you need", *Information Fusion* 81, (2022).

딥러닝 기반의 인공신경망 구조는 여전히 상당 부분 “블랙박스”와 같이 여겨지고 있어 투명성이나 해석 가능성이 상대적으로 낮은 것에 비해서, 고전적 머신러닝 기법에 있어서는 상대적으로 더 각각의 입력 변수가 모형의 출력 값에 어떻게 영향을 끼치는 지 보다 용이하게 파악될 수 있다는 장점이 있다.

다. 인지모형과 판별모형의 구분

1) 정답 라벨에 있어서의 차이

이상에서 검토한 바와 같이 인지모형과 판별모형은 그 활용 양태에 있어 상당한 차이를 보인다. 그런데 양자의 모형에서 주요하게 드러나는 보다 근본적 차이는 어떤 입력변수 X에 대응되는 학습 데이터에 부여된 라벨인 Y변수의 특성에 있다. 인지모형은 주로 사물 인식, 음성 인식 등과 같이 대체로 명확한 ‘정답(ground truth)’이 있는 과제를 해결하는 경우에 주로 활용된다. 이에 비해 판별모형이 활용되는 많은 사례에 있어 개인에 대한 평가나 행동 예측과 같이 정답이 무엇인지 정식화하기 어려운 경우가 많다.

이를 가상의 사례를 통해 구분해 보자. 만약 어떤 애완동물 판매자가 웹사이트를 구축하는데, 이용자가 애완동물 사진을 업로드하면 그 동물이 무엇인지 분류하는 모형을 구축한다면, 이는 인지모형에 해당한다. 이때 어떠한 사진 속에 포함된 동물이 개인지 고양이인지는 그 정답을 사실상 모든 경우 확정할 수 있다. 이와 달리 만약 그 판매자가 그 웹사이트에서 고객을 위한 맞춤형 애완동물 추천 서비스를 제공하는 경우를 고려하자. 이러한 추천 순위는 그 애완동물이 웹사이트 이용자를 얼마나 행복하게 해 줄 것인지에 대한 예상 능력치를 기준으로 결정될 수 있을 것이다. 이러한 예측 모형을 학습시키기 위한 학습 데이터에는 애완 동물들의 외견상 특성, 행동 특성, 부모의 혈통, 양육 과정상의 특성 등이 포함될 수 있을 것이다. 이러한 사례는 판별모형의 일종이라 할 수 있다.

판별모형이 예측하는 목표변수는 과거 유사한 특성을 갖는 애완동물을 입양했던 주인들의 만족도(0/1의 이진 라벨 또는 별점 1-5과 같은 만족도 점수)일 수 있다. 이러한 판별모형 사례의 목표변수는 직접적으로 관측할 수 없을 뿐만 아니라, 쉽사리 정의되기도 어렵다. 더욱이 애완동물이 특정한 능력이 존재한다고 가정하더라도 그 결과는 확률적으로 결정된다. 어떤 애완동물이 어떠한 주인을 만나 어떠한 환경에서 생활하게 되었는지 등과 같은 우연적 요소에 따라 주인의 만족도 평가가 크게 달라질 수 있기 때문이다. 이처럼 판별모형이 예측하고자 하는 성과, 즉 그 목표변수는 결정론적이지 않고, 따라서 확률변수로 모형화될 필요가 있다.[71] 요컨대, 인지모형은 그 모형이 추정하는 목표변수의 정답이 명확히 정의될 수 있고 고정된 성격을 갖는 것에 비해, 판별모형은 목표변수의 정답을 명확히 정의하기 어렵고 확률적인 성격을 갖는다는 점에서 구분된다.

2) 시스템 1 과제와 시스템 2 과제

다른 한편, 판별모형과 인지모형의 구분은 인간의 사고 과정에 관한 '시스템 1'과 '시스템 2' 구분과도 대응될 수 있다. 대니얼 카너먼에 따르면 인간의 사고 과정 중 '시스템 1'은 빠르고, 무의식적이고, 본능적인 사고 과정인데 비해, '시스템 2'는 느리고, 의식적이고, 논리적인 사고 과정이다.[72] 카너먼의 이러한 구분은 인간의 사고 활동을 해명하기 위한 것이었지만, 분류모형을 이해하는 데 유용하다. 이에 최근 인공지능 연구자들은 인공지능을 통해 수행하고자 하는 작업을 '시스템 1 작업'과 '시스템 2 작업'으로 구분하기 시작하였다.[73]

---

71 이러한 판별모형과 인지모형의 차이로 인해 해당 모형의 공정성 심사에 있어 적용되어야 할 지표에 있어서도 중대한 차이가 발생하게 되는데, 이에 관하여서는 본 연구 제5장에서 검토한다.

72 대니얼 카너먼, 이창신 옮김, *생각에 관한 생각—우리의 행동을 지배하는 생각의 반란*, 김영사, (2018).

73 Yoshua Bengio, Yann Lecun and Geoffrey Hinton, "Deep learning for AI.", *Communications of the ACM* 64(7), (2021).

우선, ‘시스템 1 작업’은 주로 인지모형 인공지능이 수행하는 작업에 대응될 수 있다. 이러한 작업은 인간이 별다른 노력을 기울일 필요가 있다고 느끼지 않고서도 잘 수행하는 일이다. 대표적으로 세상을 보고 사물이나 사람을 인식하는 작업, 사람의 말소리를 인식하는 작업, 글을 읽고 즉각적으로 이를 이해하는 작업과 같은 것들이다. 최근 10년 동안의 인공지능에 대한 높아진 관심과 증가된 활용은 상당 부분 본 영역에서의 발전에서 기인한다. 이러한 인공지능의 목표는 최소한 사람만큼 잘하게 되는 것이고, 많은 경우 인간을 뛰어 넘는 성능이 기대되기도 한다. 이 분야의 인간 능력은 오랜 진화의 과정을 통해 발전해 온 것이고, 다른 동물도 보유하고 있는 능력인 경우가 많다.[74] 인지모형은 그러한 인지적 능력을 모사하는 것을 일차적인 목표로 삼는다.

이와는 달리, ‘시스템 2 작업’은 주로 판별모형 인공지능이 수행하는 작업에 대응될 수 있다. 이는 우리가 주의를 기울여야 하는 숙고의 단계를 거치는 방법으로만 해결할 수 있는 문제라 할 수 있다. 예컨대 개인의 선호나 능력, 위험도를 평가하는 작업이 대표적이다. 온라인 광고, 맞춤형 컨텐츠 추천, 직원 채용, 학생 선발, 신용평가, 재범 위험성 평가 등 매우 다양한 영역에서 시스템 2 작업을 수행하는 인공지능 모형이 활용되고 있다. 한편으로 이러한 시도는 인간도 손쉽게 수행할 수 없는 작업, 혹은 인간은 잘 수행하지 못해 왔던 작업을 컴퓨터를

---

74 이처럼 시스템 1 작업을 수행하는 인공지능은 인간에만 고유한 지능에 해당하는 것이 아니므로, 인지모형을 쉽사리 ‘인간’ 지능을 구현한 것이라고 오인해서는 안 된다. 저명한 컴퓨터 공학자 Judea Pearl은 이러한 문제에 관해 “독수리와 뱀의 시각 시스템은 인간이 실험실에서 만들 수 있는 그 어떤 것보다 뛰어나지만, 뱀이나 독수리는 안경, 망원경, 혹은 현미경을 만들지 못한다”고 지적한 바 있다. Judea Pearl, “Lecture on The Mathematics of Causal Inference, with Reflections on Machine Learning and the Logic of Science” (2013. 5. 31.) Darwiche(2017) (각주 38)에서 재인용.

활용하여 인간보다 우수하게 수행할 것을 기대하는 것이라고도 이해할 수 있다.[75]

다른 한편, 현재의 인공지능 기술은 인지모형의 구현, 즉 시스템 1 작업에 있어 비약적인 발전을 이루어 왔지만, 판별모형의 구현, 즉 시스템 2 작업에 있어서는 여전히 상당한 어려움을 겪고 있다는 점에 유의할 필요가 있다. Ng (2016)이 제시한 '1초의 법칙'은 이러한 통찰을 보여준다. 그는 어떤 작업을 현재 또는 가까운 미래에 인공지능으로 수행할 수 있을 것인지 판단하기 위해 다음과 같은 대강의 기준을 적용해 볼 것은 제안한다. 그는, 만약 통상적인 개인이 1초 미만의 생각만으로 수행할 수 있는 지적 작업이라면, 인공지능이 현재 또는 가까운 미래에 그 작업을 자동화할 수 있을 것이라고 주장한다. 하지만 그보다 더 오랜 시간이 걸리는 작업이라면 인공지능에 의한 자동화 가능성은 쉽게 답하기 어렵다. 그가 제시한 1초 미만의 생각만으로 수행할 수 있는 지적 작업이란 결국 시스템 1 작업, 즉 인지모형이 수행하는 작업을 의미하는 것으로 이해될 수 있다.

이와 달리 시스템 2 작업, 즉 판별모형이 수행하는 작업을 인공지능으로 어떻게 구현할 수 있을 것인지에 관해서는 여전히 해법이 잘 알려져 있지 않다.[76] 2019년 컴퓨터 공학계의 노벨상으로 불리는 '튜링 상'을 공동 수상한 제프리 힌튼, 요슈아 벤지오 및 얀 르쿤은 'AI를 위한 딥러닝'이라는 기고문에서 이러한 어려움을 설명하였다. 현재의 딥러닝 기반모형은 매우 단순한 작업을 수행하기 위해서도 방대한 분량의 데이터를 요구하고 있고, 학습 데이터에 있어 약간의 차

---

75 김병필, "인공지능의 세 가지 유형", 중앙일보 (2020. 12. 28.). 이 글은 컴퓨터를 통해 활용하고자 하는 과제를 (1) 인간은 잘 하기 어려운데 컴퓨터가 잘 하는 분야(계산 등), (2) 인간은 잘 하는데 컴퓨터는 잘 하지 못하는 분야(사물 인식, 언어 이해), (3) 인간도 잘 하기 어렵고, 컴퓨터로도 잘 하기 어려운 분야(인간에 대한 평가)로 구분한다. 이 글은 특히 세번째 분야의 인공지능의 경우 기존 인간 판단이 갖는 한계를 극복해야 한다는 점에서 구현이 어렵고, 그 성능을 쉽게 정의하기 어렵다는 문제를 지적한다.

76 Bengio, Lecun and Hinton (2021) (각주 73).

이만으로도 예측 변수에 있어 중대한 변화를 초래할 뿐만 아니라, 인간의 상식에 기반한 추론(common sense reasoning)은 여전히 달성하기 어려운 과제로 여겨지고 있다.[77] 앞서 설명한 바와 같이 딥러닝 모델의 기술적 한계는 판별모형의 광범위한 활용에 있어 중대한 제약으로 작용할 수 있다.

3) 불확실성의 유형 차이

나아가, 판별모형과 인지모형은 모형 예측 값의 불확실성이 초래되는 원인의 측면에서도 차이를 드러낸다. 즉, 분류모형의 예측이 불확실한 원인은 (1) 모형 불확실성(model uncertainty)과 (2) 데이터 불확실성(data uncertainty)로 크게 나눌 수 있다.[78]

우선, 모형 불확실성은 학습 데이터를 다양한 모형으로 적합(fit)시킬 수 있음으로 인해 발생하는 불확실성을 의미한다.[79] 분류모형은 학습 데이터를 통해 분류를 위한 결정 경계선을 도출하는데, 만약 주어진 데이터를 잘 분류할 수 있는 결정 경계선이 다양하다면, 모형이 어떠한 결정 경계선을 도출하는지에 따라 예측이 달라질 것이다. 이 때 다양한 모형의 가능성 중 하나의 모형을 선택함으로써 모형 불확실성이 발생하게 된다. 이론적으로 무한대의 데이터를 얻을 수 있다면, 모형 불확실성은 사라질 수 있다.[80]

---

77 Yejin Choi, "The curious case of commonsense intelligence", *Daedalus* 151(2), (2022).

78 Dustin Tran, Balaji Lakshminarayanan, Jasper Snoek, "Practical Uncertainty Estimation and Out-of-Distribution Robustness in Deep Learning", NeurIPS 2020 Tutorial, (2020).

79 모형 불확실성은 인식론적 불확실성(epistemic uncertainty)이라고도 한다. 위의 글, p. 33-34.

80 다만, 해당 모형이 충분히 복잡한 결정 경계선을 도출할 수 있는 가능성이 있는 경우여야 한다. 위의 글, p. 33-34.

이에 비해, 데이터 불확실성은 학습 데이터 자체에 내재된 불확실성으로서 무한대의 데이터를 얻는다고 하더라도 제거될 수 없는 경우를 뜻한다.[81] 예를 들어 사진이 희미해서 그 사진에 포함된 물체가 무엇인지 누구도 알 수 없는 경우에는 아무리 우수한 인식 모형을 구축하더라도 그 예측에 불확실성이 남게 된다. 그 외에도 학습 데이터의 정답 라벨에 관한 의견 불일치가 있는 경우(라벨 노이즈), 측정 도구 상의 한계가 있는 경우(측정 노이즈), 데이터를 수집할 수 없는 경우(결측치) 등으로 인해 데이터 불확실성이 발생할 수 있다.[82]

만약 주어진 분류모형에 있어 모형 불확실성이 높다면 더욱 정확한 품질의 학습 데이터를 추가적으로 수집하여 학습시킴으로써 개선할 수 있다. 이는 어떤 집단의 키를 측정할 때 더 많은 표본으로부터 측정 값을 얻는다면 해당 집단의 키의 분포를 더욱 정확하게 알 수 있는 것과 마찬가지이다. 이에 비해 데이터 불확실성은 데이터의 생성 과정 자체에 불확실성의 요소가 내재되어 있는 경우이므로 그저 데이터를 더 많이 수집하는 것으로는 해소될 수 없다. 가령 주사위를 던졌을 때 다음에 나올 숫자는 항상 우연하게 1/6의 확률로 결정되는 것이므로, 주사위를 던진 결과를 더 많이 수집한다고 하여 다음 번 주사위를 던졌을 때 어떠한 값이 나올 지 예측할 수 없는 것과 같다.[83]

그런데 판별모형의 경우 인지모형에 비해 데이터 불확실성(우연적 불확실성)이 크다는 특징을 갖는다. 이는 개인의 선호나 능력, 위험도와 같이 판별모형이 예측하고자 하는 변수들은 많은 경우 현실 세계의 다양한 우연성 요소를 반영하여 결정되기 때문이다. 예컨대 어떤 개인이 재범을 저지를 위험성을 예측하는 모

---

81 위의 글, p. 35-36.

82 위의 글, p. 37

83 이러한 이유에서 데이터 불확실성을 우연적 불확실성(aleatory uncertainty)이라고도 한다. 여기서 “aleatory”는 주사위를 뜻하는 라틴어인 “alea”에서 유래된 말로서, 사건의 임의성(randomness)을 뜻한다. Kaan Bıçakcı, “Uncertainty in Deep Learning — Aleatoric Uncertainty and Maximum Likelihood Estimation”, *Towards Data Science*, (2022. 2. 8.)

형의 경우, 재범 위험성에 대한 평가는 여러 심사자 간 평가가 엇갈릴 수 있고(라벨 노이즈), 이를 완전하게 측정하는 도구가 마련되어 있지 못하며(측정 노이즈), 구속 또는 구금된 개인에 대해서는 재범 가능성 데이터를 수집할 수 없는 문제(결측치)가 있다. 이와 달리 인지모형의 경우 상대적으로 판별모형보다 데이터 불확실성 문제가 상대적으로 덜 심각하다고 볼 수 있다. 가령 자율주행차에서 사용되는 사물 인식 인공지능 시스템의 경우 사진이나 영상을 라벨링한 결과에 대해 누구나 동의할 수 있는 경우가 대부분이고, 향후 더욱 정확하게 측정할 수 있는 도구가 개발될 수 있으며, 데이터가 부족한 경우에는 추가적으로 수집할 여지도 많다. 따라서 비록 인지모형이 데이터 불확실성으로부터 완전히 자유로운 것은 아니라고 하더라도, 판별모형에 비하여는 데이터 불확실성이 더 낮은 것으로 평가할 수 있을 것이다.

라. 구분의 한계

이상에서는 인지모형과 판별모형이 차이를 보이는 여러 측면을 살펴보았다. 그러나 현실에서 특정 과제를 수행하는 인공지능은 인지모형과 판별모형 중 어느 부류에 속하는 것인지 명확하지 않을 수 있다. 또한 특정 과제를 수행하는 인공지능에서 인지 모형과 판별 모형이 결합될 수도 있다. 스마트 공장에서 제품 사진을 이용하여 하자가 있는지 검수하는 인공지능 모형을 예로 들어 보자. 이때 제품 사진을 인식하여 비정상(anomaly)인 특성이 있는지 찾는 작업은 인지모형을 통해 구현될 수 있다. 그러한 제품 사진에 나타난 비정상 특성을 과연 '하자'로 분류할 것인지의 문제는 그 제조사가 수립한 기준 혹은 해당 제품에 적용되는 규제 요건에 따라 판단되어야 하는 바, 이는 판별모형에 해당할 수 있다. 이처럼 구체적 적용사례에 있어서는 인지모형과 판별모형의 구분이 쉽지 않을 수 있다.

그러나, 여전히 각 모형의 대표적인 활용 사례들에 있어서는 이러한 구분이 가능하고 또한 유의미하다고 볼 수 있다. 즉, 학습 데이터에 포함된 정답 라벨이 고정적인지 아니면 확률적으로 달라질 수 있는지, 시스템 1 과제를 해결하는지 아니면 시스템 2 과제를 해결하는지, 데이터 불확실성이 어떠한 원인에서 비롯되는지 등을 기준으로 양자를 일응 구분할 수 있을 것이다. 이에 이하에서는 사물 인식, 음성 인식 등의 사례를 전형적인 인지모형으로, 인간의 능력이나 위험성 판단 등의 사례를 전형적인 판별모형의 사례로 고려한다.

2. 생성모형의 유형화

다음으로 생성모형을 유형화한다. 앞서 설명한 바와 같이, 생성모형은 어떠한 Y가 주어진 경우 그에 대응하는 X가 존재할 확률, 즉 $Pr(X|Y)$를 산출하는 확률분포 모형을 구축하고, 그로부터 높은 확률을 보이는 $\hat{X}$ 을 표본추출(sampling)함으로써 현실에 존재할 법한 데이터를 생성하는 것으로 이해될 수 있다. 나아가 생성모형은 X변수의 확률분포, 즉 $Pr(X)$에 관한 모형을 구축함으로써 구현될 수 있다는 점도 보았다. 이에, 생성모형은 그 생성데이터인 X변수의 유형을 기준으로 다시 유형화할 수 있는데, 이하에서는 생성모형을 정형 데이터 생성모형과 비정형 데이터 생성모형으로 구분하고, 나아가 비정형 데이터 생성모형의 대표적 사례로 언어모형과 이미지 생성모형을 검토한다.

가. 정형 데이터 생성모형

정형 데이터 생성모형은 주로 진료 기록이나 금융 거래 기록과 같이 상대적으로 저차원인 데이터를 생성하기 위한 목적으로 사용된다.[84] 정형 데이터를 생

---

84 김병필, “인공지능 개인정보 보호 기술과 개인정보 보호 법제의 과제”, *법경제학연구* 20(1), (2023).

성하여 활용하는 주요한 목적 중 하나는 원본 데이터에 개인정보가 포함되어 있어 이를 원본 그대로 활용할 경우 정보주체의 프라이버시가 중대하게 침해될 수 있는 문제에 대응하기 위해서이다. 이러한 경우 분석 대상이 되는 실제 데이터와 동일하거나 유사한 통계적 속성을 갖도록 새롭게 데이터를 생성하여 사용하는 방법이 사용될 수 있는데, 이처럼 새롭게 생성된 데이터를 흔히 '합성 데이터(synthetic data)'라 칭한다. 대표적으로 보건의료나 금융 분야와 같이 원본 데이터를 그대로 연구자에게 제공하거나, 더 나아가 이를 공중에 배포하는 것이 어려운 경우, 합성 데이터를 생성하여 분석에 활용하는 방안이 특히 유용하다.[85] 이러한 과정을 통해 개인식별정보는 포함하지 않았지만 여전히 동일 또는 유사한 통계적 속성을 갖는 데이터를 생성함으로써, 정보주체의 동의를 얻지 않고 통계 분석이나 인공지능 학습에 사용할 수 있을 것으로 기대된다.

정형 데이터에 대한 확률 분포 $Pr(X)$는 많은 경우 선형 모형이나 결합확률분포(joint probability distribution)과 같은 형태로 전통적인 통계 모형의 형태로 가정될 수 있다. 만약 이러한 가정이 적절하다면, 주어진 원본 학습 데이터를 활용하여 해당 통계 모형의 모수(parameter)를 추정해 낼 수 있다.[86] 이처럼 적합화(fitting)된 통계 모형으로부터 새로운 표본 데이터 $\hat{x}$를 추출함으로써 현실에는 존재하지 않지만 원본 학습 데이터와 같은 통계적 분포를 갖는 데이터를 생

---

85 금융 분야에서의 합성 데이터 활용 가능성에 관하여는 Samuel Assefa, et al., "Generating synthetic data in finance: opportunities, challenges and pitfalls", *Proceedings of the First ACM International Conference on AI in Finance*, (2020) 참조. 의료분야를 중심으로 하여 합성 데이터 전반에 관하여 설명하고 있는 문헌으로는 Khaled El Emam, Lucy Mosquera, and Richard Hoptroff, *Practical Synthetic Data Generation* (Kindle Edition), O'Reilly Media, (2020) 참조.

86 예컨대 가우시안 결합확률분포를 가정한다면, 원본 데이터를 활용하여 각 확률변수의 평균과 공분산을 추정할 것이다. 만약 선형 모형을 가정한다면 각 확률변수의 계수(coefficient)를 추정하게 된다.

성해 낼 수 있게 된다. 다음 그림은 정형 데이터에 관한 생성모형의 작동 구조를 보여준다.[87]

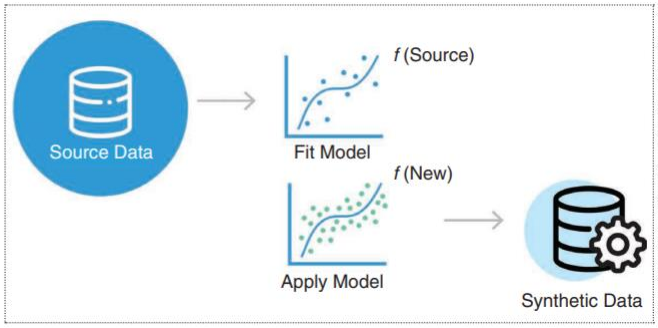


그림 6 정형 데이터 생성모형

한편, 개인정보 보호를 위한 목적으로 합성 데이터를 생성할 경우, 생성된 데이터가 원본 데이터를 그대로 재생성(regurgitate)하거나, 원본 데이터에 포함된 개인정보를 복원해 낼 수 있는 수준으로 유사하게 생성되지 않도록 보장할 필요가 있다. 예를 들어 Bellovin et al.(2019)은 순환 신경망(recurrent nerual network)을 이용하여 인터넷 이용자의 소셜 미디어 웹사이트 접속에 관한 시계열 데이터를 생성하는 모형을 구축하였는데, 생성된 일부 데이터 중 일부는 특정 이용자의 패턴을 그대로 재현하고 있어, 해당 개인이 여전히 식별될 가능성이 있음을 보였다.[88] 이처럼 원본 학습 데이터가 그대로 재현될 가능성이 상당하다면 생성모형에 의해 출력된 합성 데이터도 여전히 개인정보로 취급되어야 할 수 있다. 따라서 개인정보의 재생성 가능성을 제한하면서도 동일 또는 유사한 통계적 속성을 갖는 데이터를 생성하는 모형을 구축하는 것은 중요한 연구 과제가 되고 있다.[89]

---

87 El Emam et al. (2020) (각주 85).

88 Steven M. Bellovin, Preetam K. Dutta, and Nathan Reitinger, “Privacy and synthetic datasets”, *Stanford Technology Law Review* 22, (2019).

89 김병필(2023) (각주 84).

나. 비정형 데이터 생성모형

생성모형은 최근 들어 텍스트, 이미지, 음성, 영상 등과 같은 고차원 데이터를 생성하기 위하여 널리 활용되기 시작하였다. 특히 대규모 언어모형에 기반한 텍스트 생성모형은 인간과 자연스럽게 대화를 나누는 것뿐만 아니라 인간이 지시한 다양한 작업을 수행하는 것까지 그 기능이 확장되고 있다.[90] 또한 이미지 생성모형은 마치 극도로 자연스러운 출력물을 생성할 수 있게 되면서 인간 창작자의 지위를 대체할 것이라는 전망까지 제기되고 있다.[91] 이에 이하에서는 후속 논의의 전제로서 비정형 데이터 생성모형 중 언어모형과 이미지 생성모형을 중심으로 그 기술적 개요를 살핀다.

1) 언어모형

텍스트 생성모형은 어떤 텍스트에 대한 확률분포 모형, 즉 $Pr(text)$를 구축함으로써 구현될 수 있다. 이처럼 주어진 문장(즉, 형태소나 단어의 시퀀스)이 세상에 존재할 확률 분포를 계산해 내는 모형을 '언어모형(language model)'이라 한다.[92] 현재 언어모형은 자연어 생성(natural language generation, NLG)뿐만

---

90 Jason Wei, et al., "Emergent abilities of large language models", arXiv preprint arXiv:2206.07682, (2022)는 대규모로 확장된 언어모형이 여러 하류 과제를 수행할 수 있음을 보여준다. 이처럼 범용적 인공지능을 '기반모형(foundation model)'이라 칭하기도 하는데, 그 결과 기반모형이 인지모형이나 판별모형을 구현하기 위한 목적으로도 사용되는 양상을 보인다. 이는 이제까지의 인공지능 기술 분야의 '패러다임 전환(paradigm shift)'을 초래하고 있는 바, 이 문제에 관해서는 본 연구 제6장 제4절에서 상세히 검토한다.

91 Kevin Kelly, "What AI-Generated Art Really Means for Human Creativity", *Wired*, (2022. 11. 17.).

92 Wikipedia, "Language Model", https://en.wikipedia.org/wiki/Language_model (2023. 5. 26. 접속 확인)

아니라 자연어 이해(natural language understanding, NLU)에 관한 여러 과제에 있어 널리 활용되고 있는 범용 기술로 자리잡았다.[93]

현재 활용되는 대부분의 언어모형은 $Pr(text)$을 계산하기 위하여 조건부 확률의 연쇄법칙(chain rule)을 활용한다. 이는 다음과 같이 서술될 수 있다.

$$Pr(w_1, \dots, w_n) = Pr(w_1)\, Pr(w_2|w_1)\, Pr(w_3|\, w_1, w_2) \dots Pr(w_n|\, w_1, \dots, w_{n-1})$$

이러한 조건부 확률의 연쇄법칙에 관해 김병필(2022)는 다음과 같은 예시를 제시한다.[94] 예컨대, "I love you"라는 문장에 부여되는 확률인 Pr("I love you")는 Pr("I")×Pr("love"|"I")×Pr("you"|"I love")로 표현될 수 있다. 여기서 Pr("I")는 주어진 말뭉치(corpus) 전체에서 어떤 문장의 첫 단어가 "I"로 시작할 확률이고, Pr("love"|"I")는 "I" 다음에 "love"가 올 조건부 확률이며, Pr("you"|"I love")는 "I love" 다음에 "you"가 올 조건부 확률을 의미한다. 이처럼 앞선 단어 시퀀스 다음에 어떠한 단어에 등장할 것인지 예측하는 통계모형을 구축하고 그 값을 모두 곱함으로써 해당 전체 시퀀스의 확률을 산출하는 모형을 구성할 수 있는 것이다.

이러한 언어모형의 구현 방식은 우리가 일상에서 쉽게 접하는 검색 사이트의 검색어 '자동완성' 기능에 비유될 수 있다.[95] 자동완성 기능은 이용자가 특정한 검색어를 일부만 입력하였을 때 다음에 어떠한 단어가 올 것인지를 예측하여 추천하여 준다. 이처럼 자동완성 기능이 주어진 어구 다음에 어떠한 단어가 올 것인지를 예측한다는 점에서, 언어모형은 자동완성 기능과 흡사하다. 최근의 언어모형 인공지능은 방대한 인공신경망을 통해 자동완성 기능을 극도로 정교하게

---

93 Rishi Bommasani et al., "On the opportunities and risks of foundation models", arXiv preprint arXiv:2108.07258, (2021).

94 김병필, "대규모 언어모형 인공지능의 법적 쟁점", *정보법학* 26(1), (2022a)

95 위의 글.

확장함으로써 그 성능을 개선한 것이다. 달리 말하면, 언어모형은 주어진 어구 다음에 등장할 단어의 확률분포를 계산하는 단순한 과제를 해결하도록 함으로써 복잡한 언어 이해 및 생성 과제를 수행할 수 있게 하였다.

언어모형을 활용하면 다음과 같은 방식으로 작문을 수행할 수 있다. 우선, 이용자가 어떠한 프롬프트를 입력하면 언어모형을 통해 그 다음에 등장할 단어의 확률분포를 계산하여 높은 확률을 갖는 단어를 출력한다. 그 다음, 이용자 프롬프트에 언어모형 출력 단어를 덧붙인 것을 다시 언어모형의 입력으로 하여 그 다음에 등장할 단어의 확률분포를 계산한다. 이러한 과정은 가장 높은 확률의 단어가 "문장 종료(end of sentence)"일 때까지 반복된다. 이처럼 모형의 출력이 스스로에게 되먹임되는 구조를 '자기회귀(auto-regressive)' 방식이라 하는데, 이를 통해 언어모형은 자연스러운 답변을 생성해 낼 수 있게 된다. 한편, 언어모형은 그저 확률이 가장 높은 다음 단어를 출력하는 것이 아니라, 결과 생성에 있어 우연성의 요소를 얼마나 반영할 것인지 조절할 수 있다. 이를 조절하는 변수를 온도(temperature) 매개변수라 하는데, 이 값이 커질수록 확률이 상대적으로 낮은 단어가 생성되는 빈도가 증가하게 된다. 즉, 이로써 무미건조한 답변이 아니라 보다 창의적인 답변을 생성할 수 있게 된다.[96]

2) 이미지 생성모형

이미지 생성모형도 언어모형과 유사하게 어떤 이미지에 대한 확률분포 모형, 즉 $Pr(image)$를 구축하여 새로운 이미지를 생성하는 데 사용한다. 이는 결국 어떤 이미지가 얼마나 "그럴 듯" 한 것인지를 통계적으로 계산하는 모형이라고 할 수 있다. 이를 구현하기 위해 이제까지 변분 오토인코더(Variational Autoencoder, VAE), 적대적 생성신경망(Generative Adversarial Network,

---

96 Percy Liang, et al., "CS324 lecture notes (Winter 2022)", https://stanford-cs324.github.io/winter2022/lectures/introduction/ (2023. 5. 26. 접속 확인).

GAN), 확산(Diffusion) 모형 등이 개발되어 왔는데, 최근 들어 적대적 생성 신경망 모형과 확산 모형이 널리 활용되고 있다.[97]

우선 적대적 생성신경망은 인공신경망이 갖는 과적합(overfitting) 문제를 극복하면서 새로운 데이터를 생성하는 방법을 제시한다. 만약 인공신경망에 다수의 학습 데이터를 제시한 다음 그와 유사한 데이터를 생성할 것을 지시한다면, 인공신경망은 그저 원본 학습 데이터를 그대로 기억하여 이를 재현하게 될 수 있다. 이를 인공지능이 학습 데이터에 '과적합'되었다고 표현하는데, 특히 인공신경망의 규모가 커지면 이러한 과적합 문제가 발생할 위험이 커진다.

그런데, 적대적 생성신경망은 인공신경망에 원본 데이터 자체는 알려주지 않으면서도 그와 유사한 데이터를 생성해 내는 방법을 제시한다. 적대적 생성신경망 구조에서는 다음 그림과 같이 생성신경망(generator network)과 판별신경망(discriminator network)을 별개로 구축한다. 생성신경망은 임의의 무작위 데이터로부터 그럴 듯한(즉, $Pr(image)$가 높은) 새로운 데이터를 출력하는 기능을 수행하고, 판별신경망은 생성신경망이 생성한 데이터와 원래의 학습 데이터를 판별해 내는 기능을 수행한다. 여기서 주목할 점은 판별신경망만이 원본 학습 데이터를 통해 직접적으로 학습되지만, 생성신경망에는 원본 학습 데이터에 관한 정보가 직접 제공되지 않는다는 사실이다. 그 대신 생성신경망에는 판별신경망을 얼마나 잘 속였는지에 관한 간접적 정보만이 제공된다. 이로써, 학습이 거듭되면 생성신경망은 판별신경망을 더 잘 속이는 능력을 학습해 나가고, 최종적으로 생성신경망은 판별신경망이 원본 학습 데이터와는 구별하기 힘든 수준의 유사한 데이터를 만들어 낸다. 이와 같이 두 개의 별개의 인공신경망이 서로 대립 관계를 갖고 경쟁하여 발전해 나간다는 뜻에서 적대적 생성신경망이라 이름 붙였다. 이러한 구조를 통해 생성신경망은 원본 학습 데이터와 극히 유사하지만, 새로운 데이터를 생성하는 기능을 갖게 된다.

---

97 Arash Vahdat and Karsten Kreis, "Improving Diffusion Models as an Alternative To GANs, Part 1", NVIDIA Technical Blog, (2022. 4. 26.).

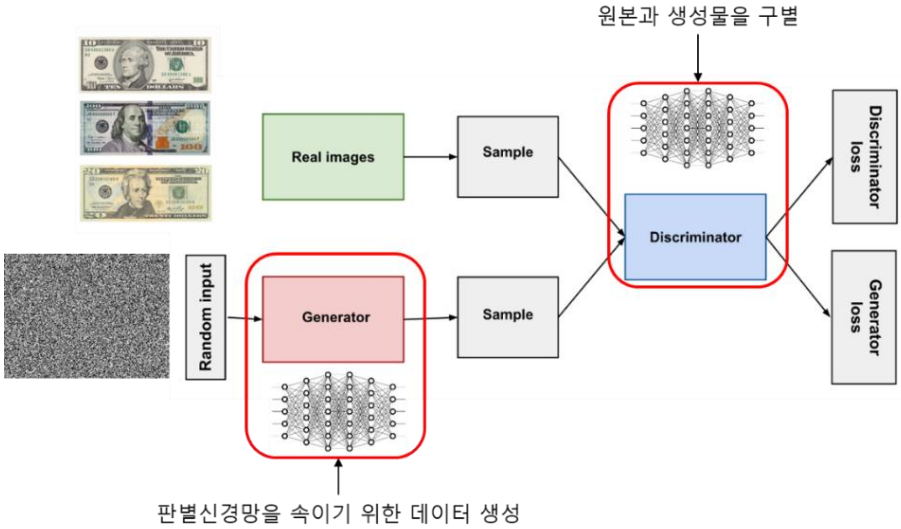


그림 7 적대적 생성신경망의 구조

나아가 적대적 생성신경망은 $Pr(image)$를 학습하는 것에서 더 나아가 이용자가 원하는 특정한 조건에 따라 $Pr(image|user-specified-condition)$을 계산하도록 구현될 수 있다. 이러한 기술은 지속적으로 발전하여 이제 특정 이미지 속 인물은 유지하면서 그 배경만 변경하거나, 사진의 스타일을 변경하거나, 인물의 표정이나 동작을 변경시키는 수준에 이르렀다. 이로써 적대적 생성신경망으로 생성된 이미지를 이른바 "딥페이크(deep fake)"라 부르기도 한다.[98]

최근 더욱 주목받고 있는 것은 DALL-E 2나 MidJourney와 같은 이미지 생성 서비스에서 사용되고 있는 확산모형이다. 확산모형은 $Pr(image)$를 구축하기 위해 또다른 창의적 접근법을 취하는데, 이는 이미지에 임의의 잡음을 추가한 다음, 인공신경망으로 하여금 추가된 잡음과 원본 이미지를 구별하여 잡음을 제거

98 딥페이크 기술에 대해서는 포르노그래피를 조작하거나, 특정한 정치적 목적으로 악용될 우려가 제기되어 왔다. Bobby Chesney and Danielle Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security", *California Law Review* 107, (2019).

하는 방법을 학습하도록 하는 것이다.[99] 이를 위해 학습 데이터의 이미지에 점차적으로 잡음을 추가하는데, 최종적으로는 잡음이 이미지 자체를 압도하는 단계에 이르게 된다(아래 그림 참조). 인공신경망은 이러한 잡음을 추가하는 과정을 되돌리는 방법을 학습한다. 이러한 방식을 활용하면 적대적 생성신경망과 같이 종전 세대의 생성모형과 비교하여 더 빠르게 학습될 수 있을 뿐만 아니라, 더 다양한 유형의 결과물을 생성해 낼 수 있는 것으로 드러났다.[100]

그림 8 확산 모형의 학습 및 생성 단계

이처럼 확산모형은 그 학습 단계에서 인공신경망은 주어진 이미지로부터 잡음을 제거하는 방법을 학습한 다음, 이미지 생성 단계에서는 잡음만을 입력하고 점차적으로 잡음을 제거하여 완전한 이미지를 생성한다. 이 때 이용자로부터 텍스트로 된 프롬프트를 입력 받아 이용자가 원하는 표현이 나타날 수 있도록 조건으로 부가한다. 이러한 조건화 과정은 CLIP과 같이 언어모형과 이미지 생성모형을 결합한 모형을 활용하여 이루어진다.[101] 즉, 언어모형은 이용자가 입력한 프롬프트의 특성을 요약하여 변환시킨 값(즉, 임베딩)을 출력해 내는데, 생성된 이미지에 대한 임베딩이 텍스트 임베딩과 유사하도록 인공지능을 학습시키면, 그

99 Alexandra Sasha Luccioni, et al., "Stable Bias: Analyzing Societal Representations in Diffusion Models", arXiv preprint arXiv:2303.11408 (2023).

100 Vahdat and Kreis (2022) (각주 97).

101 Luccioni et al. (2023) (각주 99).

결과 무작위의 잡음으로부터 이용자 프롬프트의 지도를 받아 이용자가 원하는 이미지를 생성할 수 있게 된다. 이러한 구조를 텍스트-이미지 변환(Text-to-Image) 모형이라고 부르기도 한다.

제3절 인공지능에 의한 공정성 침해 위험성

1. 공정성의 개념

본 연구는 인공지능의 공정성 심사에 사용될 수 있는 공정성 지표를 조사하고 그 법적 함의를 해명하는 것을 목적으로 한다. 그렇다면 여기서 '공정성'은 어떠한 의미를 갖는가? 공정성의 개념은 크게 (1) 실체적 차원과 (2) 절차적 차원으로 구분할 수 있다. 이하에서는 양자를 구분하여 검토하고, 본 연구에서 사용하는 인공지능의 공정성 개념을 정의한다.

가. 실체적 공정성

인간은 자원 분배에 있어 자의적인 불평등은 회피하되, 응분(desert)에 따른 불평등은 선호하는 경향이 있다.[102] 다수의 연구는 최후통첩 게임 등과 같은 여러 심리학 실험을 통해 인간이 자신의 이익을 증진하는 것 이외에도 공정성과 상호성에 대한 선호를 갖는다는 점을 실증적으로 보였다.[103] 이와 같은 인간의 공정성

---

102 Lee Anne Fennell and Richard H. McAdams, "Fairness in Law and Economics: Introduction", *University of Chicago Coase-Sandor Institute for Law & Economics Research Paper* 704, (2013).

103 이러한 다양한 실증 연구를 법경제학적 측면에서 정리한 것으로 Lee Anne Fennell and Richard H. McAdams (eds.), *Fairness in Law and Economics*, Edward Elgar, (2013) 참조.

에 대한 선호를 고려하지 않는다면, 인간이 자신의 이익을 희생하여서라도 공익에 기여하는 자에 대해 보상을 주고, 무임승차자에 대해 처벌하고자 하는 인간의 행동을 설명할 수 없다.[104]

이러한 자원 분배상의 공정성 선호는 법제도에도 반영되어 있다. 예컨대, 「독점규제 및 공정거래에 관한 법률」(이하 "**공정거래법**")에서 요구되는 '실체적 공정성'의 개념은 이를 반영한다. 즉, 공정거래법상 실체적 공정성 개념의 중요한 측면은 "자원배분의 성과의 속성"에 관한 것으로서, "본질적으로 소득분배의 문제"를 다루는 것이다.[105] 구체적으로, 공정거래위원회의 불공정거래행위 심사지침은 "불공정성(unfairness)이란 경쟁수단 또는 거래내용이 정당하지 않음을 의미한다"고 정의하여,[106] 불공정성 개념을 '경쟁수단의 불공정성'과 '거래내용의 불공정성'으로 구분한다.[107] 이와 같이 공정거래위원회가 "거래내용의 불공정성"을 위법성 판단기준의 하나로 포함하고 있는 것은 자원 배분의 측면에서 실체적 공정성을 중요하게 고려하는 것으로 평가할 수 있다.[108]

한편, 법적 규칙을 결정함에 있어 공정성의 개념을 고려하는 것은 오히려 효율성을 저해하여 사회 후생을 감소시킬 수 있으므로, 법적 규칙을 평가할 때에는 공정성을 고려하지 않고 사회 후생을 유일한 기준으로 삼아야 한다는 주장도 제기된다. 대표적으로 Kaplow & Shavell (1999)는 공정성을 법적 규칙의 평가에 사용되는 원칙으로서 개인의 후생과 원칙적으로 무관한 요소에 기반한 개념으로

---

104 Fennell and McAdams (2013) (각주 102).

105 홍대식, "4차 산업혁명 시대의 공정성 개념에 대한 단상", ICT법경제연구소, (2018. 2. 17.)

106 불공정거래행위 심사지침, III. I. 가. (라).

107 불공정거래행위 심사지침에 따르면, "경쟁수단의 불공정성"은 "상품 또는 용역의 가격과 질 이외에 바람직하지 않은 경쟁수단을 사용함으로써 정당한 경쟁을 저해하거나 저해할 우려가 있음"을 의미하고, "거래내용의 불공정성"은 "거래상대방의 자유로운 의사결정을 저해하거나 불이익을 강요함으로써 공정거래의 기반이 침해되거나 침해될 우려가 있음"을 의미한다.

108 홍대식(2018. 2. 17.) (각주 105).

파악한다.[109] 이러한 개념 정의에 따르면 분배 공정성 이외에도, 교정적 정의(corrective justice), '계약은 지켜져야 한다(Pacta sunt servanda)'는 규범, 응보적 정의(retributive justice)와 같은 다양한 원칙이 공정성 개념에 포함된다.[110] 그런데, Louis Kaplow와 Steven Shavell은 위와 같은 공정성 원칙들에 기반한 법적 제도는 오히려 사회 후생 감소를 초래할 수 있다는 점을 지적한다. 예를 들어, 불법행위에 따른 손해배상 책임 제도에 있어 엄격 책임 원칙이 사회 후생을 증진시킬 수 있는데도 불구하고 교정적 정의의 관념에 따라 과실 책임 제도를 도입한다면 그 결과 사회 후생이 감소할 것이다.[111] 또한 평등한 분배를 달성하기 위해 가해자가 가난한 피해자에게는 더 큰 배상을 하게 하거나, 부유한 가해자에게는 더 많은 배상을 하도록 한다면, 이러한 제도는 효율적인 손해배상 책임 규칙에서 벗어나 사회적으로 최적인 주의 수준과 및 행위 수준을 달성하지 못할 수 있다.[112] 따라서 이들은 자원의 재분배는 조세 등과 같은 형태로 이루어져야 하고, 법적 규칙은 사회 후생의 관점에서의 효율성을 기준으로 마련되어야 한다는 주장한다.

이러한 Louis Kaplow와 Steven Shavell의 주장에 대해서는 여러 비판이 제기되었고, 그에 관한 논쟁이 이어진 바 있다.[113] 위 주장에 대한 주요한 비판은 부의 극대화와 사회 후생의 증진을 구분하여 고려될 필요가 있다는 것이다. 부를 극대화하는 것이 사회 후생을 증진시키는 데 유용할 수 있지만, 양자가 반드시 일치하는 것은 아니다. 인간의 기본권이나 자유와 같은 권리자격은 이전될 수 없는 것으로서, 만약 이러한 권리자격이 부당하게 분배되어 있다면 이는 조세 제도

---

109 Louis Kaplow and Steven Shavell, "The conflict between notions of fairness and the Pareto principle", *American Law and Economics Review* 1(1), (1999).

110 위의 글, p. 64.

111 위의 글, p. 70.

112 Louis Kaplow and Steven Shavell, "Should legal rules favor the poor? Clarifying the role of legal rules and the income tax in redistributing income", *The Journal of Legal Studies* 29(2), (2000).

113 이러한 논쟁을 모은 것으로 Fennell and McAdams (eds.) (각주 103) 참조.

를 통해 재분배하는 것이 불가능하다.[114] 이처럼 공정성에 대한 선호는 후생에 직접적이고 이전 불가능한 방식으로 영향을 끼치므로, 후생을 창출하는 행위와 이를 분배하는 행위는 구분이 구분할 수 없다. 나아가, 오히려 공정한 분배 및 분배 절차의 공정성이 부를 극대화하는 데 필요할 수 있다. 예컨대 어떤 사회의 부가 불평등하게 분배되어 있다면 절도와 같은 재산 범죄가 증가할 것이므로, 더욱 공정한 분배는 범죄를 줄여 부를 극대화하는 데도 필요하다.[115] 또한, 조세에 대한 (비합리적일 수도 있는) 거센 저항 및 조세 규칙을 변경하는 데 있어 소요되는 정치적 비용을 고려하면 조세에 의한 분배가 적정한 수준으로 쉽게 이루어질 수 있을 것으로 기대해서는 안 된다.[116]

1) 인공지능과 자원 분배 공정성

이처럼 실체적 공정성도 다양한 측면에서 고찰될 수 있고, 그 적절한 적용 범위와 한계를 찾는 작업은 쉽지 않은 일이다. 하지만, 위와 같은 법경제학 상의 '공정성' 개념 논쟁에서 볼 수 있는 바와 같이, 실체적 공정성에 관한 논의는 대체로 부나 기회와 같은 사회적으로 희소한 자원의 분배에 있어 어떠한 원칙이 적용되어야 하는지를 중심으로 이루어져 왔다. 그런데, 인공지능의 급속한 확산은 사회 전반에 걸쳐 자원 분배의 공정성을 침해하는 문제를 초래할 수 있다. 예컨대, 어떤 사업자가 학습 데이터나 특정 인공지능 기술을 독점하고 있는 상황, 혹은 특정 온라인 플랫폼이 인공지능의 개발·활용에 있어 시장 지배적 지위를 차지하고 있는 상황에서는 그 학습 데이터나 인공지능 모형에 관한 시장 경쟁의

---

114 Fennell and McAdams (2014) (각주 102) p. 2.

115 위의 글, p. 3.

116 위의 글, p. 5.

보장과 소비자 보호가 중요한 사회적 과제가 된다.[117] 보다 근본적으로, 발전된 인공지능은 거대 자본이 개인의 행태정보를 지속적으로 수집하여 이를 생산의 원재료로 사용함으로써 개인의 자율과 프라이버시를 침해하고 민주주의를 위협하는 이른바 '감시 자본주의' 체제를 강화할 수 있다.[118]

또한 인공지능 학습 데이터에 대한 사회적 수요가 증가하면서 인공지능 학습 데이터 생성을 위한 주석 작업자들도 증가하고 있는 바, 이들의 삶의 질을 보장하는 근로조건과 적정한 보수 지급에 관한 문제도 분배 공정성 측면에서 중요한 문제로 지적된다.[119] 나아가, 인공지능에 의해 디지털 격차가 심화될 가능성이나 업무 자동화에 따른 직무 전환의 필요성, 사회 전반에 걸친 일자리 감소의 위협 등과 같은 문제도 고려될 필요가 있다.[120]

그런데, 이와 같은 거시적 차원의 분배의 문제는 '범용 기술(General Purpose Technologies)'의 등장 및 발전과 더불어 지속적으로 존재하여 온 문제이다. 여기서 '범용 기술'이란 과거의 증기기관이나 전기 모터, 현재의 반도체와 컴퓨터, 인터넷 기술 등과 같이 다양한 하류 응용 영역으로 광범위하게 확산되고, 그 결과 범용 기술의 혁신이 사회 전체에 걸친 생산성 향상을 가져오는 기술을 뜻한다.[121] 이러한 범용 기술의 발전은 '산업 혁명'(들)의 주된 동력이 되어 왔고, 그에 의한 생산성 증대 효과가 사회적으로 어떻게 분배되어야 하는지를 둘

---

117 Competition and Markets Authority, *AI Foundation Models: Initial review*, (2023. 5. 4.) https://www.gov.uk/cma-cases/ai-foundation-models-initial-review (2023. 5. 26. 접속 확인).

118 쇼샤나 주보프, 김보영 옮김, *감시 자본주의 시대: 권력의 새로운 개척지에서 벌어지는 인류의 미래를 위한 투쟁*, 문학사상, (2021).

119 메리 그레이 & 시다스 수리, 신동숙 옮김, *고스트워크: 모든 일과 직업의 모습이 바뀐다*, 한스미디어, (2019).

120 ILO Global Commission on the Future of Work, *Work for a brighter future*, (2019. 1. 22.); David Mindell and Elisabeth Reynolds, *The work of the future: Building better jobs in an age of intelligent machines*, MIT Press, (2022).

121 Timothy F. Bresnahan and Manuel Trajtenberg, "General purpose technologies 'Engines of growth'?", *Journal of Econometrics* 65(1), (1995).

러싸고 수많은 사회적, 정치적 갈등과 충돌이 이어져 왔다. 이러한 점에서 앞서 언급한 인공지능 기술에 관한 사회적 분배의 문제는 기존의 사회적, 정치적 문제의 연장선에서 고찰될 수도 있다.

2) 인공지능에 의한 공정성 침해 문제의 특유성

그런데, 인공지능 기술은 기존의 범용 기술의 분배 공정성 문제와는 구별되는, 새로운 차원의 공정성 문제를 초래한다는 특색이 있다. 이제껏 증기기관, 전기모터, 반도체, 컴퓨터, 인터넷과 같은 기술은 이를 통해 재화나 서비스의 생산이나 유통의 효율성이 개선되는 결과를 초래하였고, 그로 인한 증대된 사회 후생이 어떻게 분배되어야 할 것인지는 별도의 정치적 과정을 통해 결정되었다. 그러나, 인공지능, 특히 분류모형의 경우 인간의 의사결정 자체를 보조하거나 자동화하기 위하여 널리 활용된다는 점에서 기존 범용 기술과 구분될 수 있다. 결국, 인공지능 기술 자체가 사회적 분배를 결정하는 데 활용되면서, 그에 대한 공정성 개념을 별도로 정립할 필요가 있게 된 것이다. 이에 본 연구에서 고려하는 인공지능 공정성 개념은 이제껏 논의되어 기술 성과의 분배 공정성 문제와는 질적으로 구분되는 것으로서 별도로 고찰될 필요가 있다.

요컨대, 본 연구에서 검토하는 '인공지능 공정성'은 인공지능을 활용한 개별적인 의사결정 사례에 있어 공정성을 요구하는 것으로서, 일응 미시적인 차원에서 적용되어야 할 공정성 규범에 관한 것이다. 구체적으로 현실의 정책결정자들은 개별 대학의 학생 선발이나 직원 채용, 신용평가나 이상 거래 탐지와 같이 활용 사례에 있어 어떠한 정책이 공정한 것인지와 같은 문제에 답하여야 한다. 따라서, 본 연구에서 검토하고자 하는 인공지능 공정성 지표는 국소적으로 인공지능을 개발하거나 이용하는 개별 사업자나 기관 등이 해당 인공지능 모형의 평가 과정에서 직접적으로 활용되는 것으로 전제한다.

나. 절차적 공정성

공정성 개념의 또다른 중요한 측면은 절차적 공정성이다. 일찍이 Tyler (1997)은 법의 지배에 있어 절차적 공정성의 중요성을 강조한 바 있다.[122] 그는 단순히 법 위반에 대해 강력한 제재를 가하는 것만으로는 법의 준수를 강제하기 어렵다는 점을 지적한다. 감시와 처벌에 소요되는 비용을 고려할 때, 법의 지배는 대부분의 수범자가 대부분의 법을 대부분의 시간동안 '자발적'으로 준수할 것을 요구한다. 이러한 수범자에 의한 자발적 법 준수를 보장하기 위해서는 수범자들이 권위의 합법성과 법의 도덕성을 받아들일 필요가 있는데, 이는 '절차적 공정성(procedural fairness)'에서 기인한다. 나아가 그는 어떠한 결정에 대해 절차적 정당성을 확보하기 위한 가장 중요한 네가지 요소로는 (1) 권위에 대한 신뢰가능성(trustworthiness), (2) 존중과 존엄을 갖춘 대우(treatment with respect and dignity) (3) 권위의 중립성(neutrality) 및 (4) 결정 과정에 대한 참여 및 의견 개진(voice)이라고 설명한다.

이러한 Tyler의 분석은 인공지능의 활용에 있어 절차적 공정성이 보장되어야 할 필요성을 설득력 있게 제시한다. 사회적으로 인공지능에 의한 판단이 널리 활용되기 위한 전제는 인공지능에 대한 신뢰가능성을 확보되고, 인공지능이 그 이용자에 대해 존중과 존엄을 갖추어 대우할 수 있어야 하며, 인공지능 판단의 중립성이 보장되어야 하고, 그 결정 과정에 있어서의 참여와 의견 개진 가능성이 있어야 한다.

인공지능에 있어 절차적 공정성을 보장할 필요성은 여러 문헌을 통해 거듭 강조되어 왔다. 대표적으로 EU의 인공지능에 관한 고위 전문가 그룹의 "신뢰할 수 있는 인공지능을 위한 윤리 가이드라인"은 "절차적 차원에서의 인공지능 공정성은 인공지능 시스템 및 시스템 운영자에 의해 내려진 결정에 대해 이의를 제

---

122 Tom R. Tyler, "Procedural fairness and compliance with the law", *Revue Suisse D Economie Politique et de Statistique* 133, (1997).

기하고 효과적으로 구제를 요구할 수 있는 능력을 내포하는 것으로서, 이를 보장하기 위해서는 해당 결정에 책임을 지는 주체가 식별가능하고, 의사결정 과정은 해명 가능해야 한다"고 지적한다.[123] OECD 인공지능 윤리 원칙은 이러한 절차적 공정성을 확보와 관련하여 인공지능 시스템에 있어 투명성과 책임성 있는 공개가 보장될 필요가 있다고 명시하고 있다.[124]

특히 국가작용에 있어 인공지능이 활용될 경우 절차적 공정성이 중요하게 고려될 필요가 있다. 우리 헌법재판소는 적법절차 원칙이 독자적인 헌법원리임을 인정하면서, "입법, 행정 등 국가의 모든 공권력의 작용되는 절차상의 적법성뿐만 아니라 법률의 구체적 내용도 합리성과 정당성을 갖춘 실체적 적법성이 있어야 한다"고 판시하였다.[125] 이러한 적법절차 원칙은 자의적이고 차별적인 국가작용을 방지하기 위한 보호장치로 기능하게 된다.[126] 따라서 자동적 행정결정 등에 있어 인공지능이 활용되는 경우 절차적 공정성의 보장이 긴요하다.[127]

그러나 절차적 공정성은 실체적 공정성과는 질적으로 구분되는 여러 차원의 문제를 야기한다. 기술적 차원에서 가장 부각되는 문제는 인공지능의 설명가능성 문제이다.[128] 이는 흔히 '블랙박스'라 불리는 딥러닝 인공지능에 있어 투명성과 설명가능성을 확보할 기술적 수단의 문제로 다루어진다. 다른 한편, 인공지능의 출력에 대해 그 이용자가 자의적으로 개입하여 변경시킬 수 없도록 보장할 필요성도 있는데, 이는 "절차적 정규성(procedural regularity)'의 문제에 해당한

---

123 High-Level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines for Trustworthy AI*, (2019. 4. 8).

124 OECD, *Recommendation of the Council on Artificial Intelligence*, OECD Publishing, (2019) Principle 1.3.

125 헌재 1992. 12. 24. 선고 92헌가8 결정.

126 강승식, "절차적 적법절차의 본질에 관한 연구", *미국헌법연구* 25(2), (2014).

127 성윤희, "행정기본법상 자동적 처분에 관한 연구", 이화여자대학교 대학원 법학과 박사학위 논문, (2023).

128 고학수 외, *인공지능 원론: 설명가능성을 중심으로*, 박영사, (2021).

다.[129] 절차적 정규성의 원칙은 이미 공표된 일련의 규칙이 모든 개별 사건에 대해 일관성 있게 적용될 것을 요구한다. 이와 관련하여 이용자가 적대적 공격(adversarial attack)이나 데이터 오염(data poisoning) 공격을 통해 인공지능의 출력을 조작할 가능성, 이용자가 게이밍(gaming) 등을 통해 인공지능을 오용할 위험성과 같은 인공지능 보안, 강건성 및 안전성 문제도 함께 고려될 필요가 있다. 그뿐만 아니라 인공지능의 설명가능성은 단지 기술적 차원에서 그치는 것이 아니다. 인공지능에 대한 설명 가능성은 인공지능이 어떻게 동작하는지, 그리고 인공지능이 동작한다는 사실을 어떻게 알 수 있는지에 대한 설명만이 아니라, 인공지능의 활용이 규범적으로 정당화될 수 있다는 사실에 대한 설명까지로 확장될 필요가 있다.[130] 이는 인공지능에 의한 출력이 근거하고 있는 원칙과 사실, 기준에 관한 사회적 맥락에 관한 설명을 포함하는 것이다.

이처럼 인공지능의 절차적 공정성은 실체적 공정성과는 구분되는 별개의 요구사항이라 할 수 있다. 그런데, 본 연구에서 고려하는 공정성 지표는 인공지능의 입력변수와 출력변수에 관한 속성을 주된 평가 대상으로 삼는 것으로서, 이러한 절차적 공정성의 여러 측면을 반영하지 못한다. 이 문제는 별도의 연구를 통해 면밀하게 고찰될 필요가 있는 것으로서 본 연구에서는 직접적으로 다루지 않는다.

다. 본 연구의 인공지능 '공정성' 개념

본 연구는 인공지능 공학 연구자들이 이제까지 인공지능 공정성을 심사하기 위하여 제안해 온 여러 공정성 지표들을 그 검토 대상으로 한다.[131] 이는 인공지능이 어떤 개인에 관해 특정한 출력을 산출하였을 때 그 출력이 과연 공정한 것

---

129 Joshua A. Kroll, "Accountable Algorithms", *University of Pennsylvania Law Review* 165, (2017).

130 ISO/IEC TR 24028 Technical Report.

131 본 연구 제4장(분류모형에 대한 공정성 지표) 및 제6장(생성모형에 대한 공정성 지표) 참조.

이라 평가할 수 있는지 계량적으로 답할 수 있는 기준을 제시하고자 하는 시도라 할 수 있다. 그런데, 이제껏 공학 연구자에 의해 제시되어 온 공정성 지표들은 대체로 (1) '개인 공정성(individual-based fairness)'과 (2) '집단 공정성(group-based fairness)'으로 정식화할 수 있는 공정성 개념에 기반하고 있다.[132]

우선 '개인 공정성'은 유사한 개인들은 유사하게 대우받아야 한다는 기준이다. 이는 "같은 것은 같게, 다른 것은 다르게" 처우하는 것이 공정하다는 아리스토텔레스의 정의론에 기반한 것으로서, '아리스토텔레스 정식'이라 표현되기도 한다.[133] 분류모형에 있어 개인 공정성 개념은 Dwork et al. (2012)에 의해 정식화되었는데,[134] 위 연구는 공정성 개념을 "해당 과제에 관하여 유사한 어떠한 두 개인이라도 유사하게 분류되어야 한다"는 원칙으로서 파악하였다. 이러한 개념을 적용하여 분류모형의 공정성을 심사하는 경우, 어떠한 두 개인에 관한 입력 변수가 얼마나 "유사"한 것인지를 정하는 거리 측정 지표를 정의한 다음, 그 거리가 가까운 개인에 대한 모형의 분류 결과에 있어 차이가 얼마나 되는지를 평가할 수 있다.

법적인 측면에서, '개인 공정성'은 헌법상 평등 원칙이나 차별금지법제 상의 차별 금지 원칙을 구현한 것으로 평가될 수 있다. 헌법재판소는 "평등의 원칙은 입법자에게 본질적으로 같은 것을 자의적으로 다르게, 본질적으로 다른 것을 자

---

132 Cynthia Dwork, et al., "Fairness through awareness", *Proceedings of the 3rd innovations in theoretical computer science conference*, (2012). 한편, Percy Liang, et al., "Holistic evaluation of language models", arXiv preprint arXiv:2211.09110 (2022), pp. 32-3은 인공지능 공정성 지표를 반사실적 공정성(counterfactual fairness)와 통계적 공정성(statistical fairness)로 구분하는데, 이는 각각 개인 공정성과 집단 공정성의 구분과 대응된다. 여기서 반사실적 공정성은 어떠한 개인이 규범적으로 무관한 다른 지위에 있는 가정적 상황(즉, '반사실적 개인')에서도 인공지능에 의해 같은 대우를 받아야 한다는 것으로서, 이는 결국 개인 공정성을 요구하는 것과 유사한 개념으로 평가할 수 있다.

133 정주백, "아리스토텔레스 정식 비판" *헌법논총* 27, (2016).

134 Cynthia Dwork, et al., "Fairness through awareness", *Proceedings of the 3rd innovations in theoretical computer science conference*, (2012).

의적으로 같게 취급하는 것을 금하고 있다. 그러므로 비교의 대상을 이루는 두개의 사실관계 사이에 서로 상이한 취급을 정당화할 수 있을 정도의 차이가 없음에도 불구하고 두 사실관계를 서로 다르게 취급한다면, 입법자는 이로써 평등권을 침해하게 된다."고 판시한 바 있다.[135] 대법원도 헌법재판소의 이러한 판시를 인용하면서, "근로기준법에서 말하는 차별적 처우란 본질적으로 같은 것을 다르게, 다른 것을 같게 취급하는 것을 말하며, 본질적으로 같지 않은 것을 다르게 취급하는 경우에는 차별 자체가 존재한다고 할 수 없다."고 판단한 바 있다.[136] 인공지능에 대해 개인 공정성을 요구하는 것은 이처럼 "본질적으로 같은 것을 같게, 본질적으로 다른 것은 다르게" 취급하도록 요구하는 것과 같다.

개인 공정성 개념을 인공지능 공정성 심사에 적용하기 위해서는 결국 어떠한 개인들이 "본질적으로 같은 것"인지를 결정할 필요가 있다. 그런데, 헌법상 평등원칙이나 차별금지법제는 특정한 사유에 대해 "같은 것"으로 볼 것을 강제한다. 예를 들어, 헌법은 성별, 종교, 또는 사회적 신분에 있어 차이가 있는 경우는 "같은 것"으로 평가할 것으로 요구한다. 또한 개별 영역에 대해 적용되는 차별금지법제에서는 다수의 차별금지사유를 규정하고 있는데, 이 역시 규범적으로 "같은 것"으로 평가되어야 할 변수들을 정하고 있는 것이다.[137] 따라서 인공지능에 대해 개인 공정성 개념을 적용할 경우, 차별금지사유에 대해서만 차이가 있을 뿐 그 이외의 사유에 있어서는 유사한 입력변수를 갖는 개인들을 서로 동등하게 취급할 것을 요구하게 된다.[138]

---

135 헌재 1996. 12. 26. 선고 96헌가18 결정 등.

136 대법원 2015. 10. 29. 선고 2013다1051 판결 등.

137 차별금지법제가 인공지능 공정성 보장을 위해 어떻게 적용될 수 있는지에 관해서는 본 연구 제3장 제2절 참조.

138 이러한 이유로 개인 공정성 개념은 입력변수에 대한 공정성 기준으로 구현될 수 있다. 즉, 인공지능 시스템이 아예 차별금지사유를 입력변수로 활용하지 않는다면 차별금지사유에 있어서만 차이가 있고 그 이외의 사유에 있어서는 유사한 입력변수를 갖는 개인들은 서로 동등하게 취급될 것이기 때문이다. 입력변수에 대한 공정성 기준에 관해서는 본 연구 제4장 제2절 참조.

그러나, 이처럼 명시적으로 차별금지사유로 규정되어 있는 변수를 제외하고 나면, 나머지 입력 변수에 관해 한 개인이 다른 (가상의 또는 실존하는) 개인과 얼마나 근접한지를 정의하는 것은 어려운 일이다. 특히 인공지능의 역량과 활용 범위가 증가하고 주어진 과제를 해결하기 위해 고차원의 데이터가 이용되는 경우가 늘어나면서, 입력 변수가 "유사한 지위"에 있는 개인을 확인하는 일은 극히 어려운 작업이 되고 있다. 예컨대, 광고 추천을 위해 어떤 개인의 검색어나 방문 웹사이트 이력을 입력 변수로 활용하는 사례를 고려하자. 해당 추천 시스템이 '개인 공정성'을 충족시키는지 여부를 확인하기 위해서는 어떤 개인에 대해 유사한 다른 개인과 유사한 추천이 이루어지는지 평가해야 할 것인데, 모든 개인의 변수가 저마다 서로 다른 상황에서 어떠한 개인들이 서로 유사하다고 규범적으로 평가할 수 있는지 정의하기란 쉽지 않다.[139]

개인 공정성의 또다른 한계는 개인이 속한 집단간 상당한 차이가 존재하는 경우이다. 만약 A 집단이 역사적으로 차별을 받아왔고, 그 집단에 속한 개인들은 유사하게 불리한 특성을 가지고 있고, A 집단에 속하지 않은 이들은 그러한 특성을 가지고 있지 않은 상황을 고려해 보자. 만약 인공지능에 대해 개인이 A 집단 소속인지 여부에 관한 직접적인 정보를 차단하더라도, 인공지능은 A 집단이 공유하고 있는 다른 특성을 입력 변수로 활용하여 A 집단에 대해 불리한 결정을 내릴 수 있다. 이러한 상황에서, 만약 개인 공정성 개념에 따라 유사한 특성을 가진 이들에 대해 모두 유사하게 처우하더라도, 역사적으로 불리한 지위에 있어왔던 집단에 대해서는 여전히 차별적으로 작동할 위험성이 남게 된다.[140]

---

139 이는 머신러닝 알고리즘의 입력 변수가 약간 달라졌을 뿐인데 결과에 있어서는 큰 차이가 발생하는 문제(Small Changes Makes a Big Difference, SCMBD)라 불리기도 한다. Jane R Bambauer, Tal Zarsky, and Jonathan Mayer, "When a Small Change Makes a Big Difference: Algorithmic Fairness Among Similar Individuals", *UC Davis Law Review* 55, (2021) 참조. 위 연구자들은 SCMBD가 예측적 타당성을 갖는 경우 그 윤리성을 판단하는 것이 어렵게 된다는 점을 지적한다.

140 이러한 문제를 대용변수에 의한 차별이라 하는데, 이에 관해서는 본 연구 제4장 제2절 참조.

따라서 인공지능의 공정성을 평가하기 위해서는 여러 사회적 집단에 대해 동등하게 작동하는지 여부를 평가할 필요가 발생한다. 이처럼 인공지능의 출력이 여러 사회적 집단에 대해 동등할 것을 요구하는 것이 '집단 공정성' 개념이다.

집단 공정성 개념은 현실에서 확보되는 학습 데이터에 기존 사회의 차별이 반영되어 있고, 이를 통해 학습된 인공지능은 그러한 차별을 반영할 출력을 생성할 수밖에 없다는 인식에 근거한다.[141] 또한 여러 인공지능 활용 사례에 있어 학습 데이터 중 소수자 집단이 그 인구통계적 비율에 비해 더 적은 숫자만 포함되어 과소대표(under-represent)될 수 있다. 더욱이 만약 인구통계적 비율에 따라 학습 데이터에 포함된다고 하더라도, 과소대표된 집단에 대해서는 낮은 정확도를 보일 위험이 있다. Barocas, Hardt & Narayanan (2022)은 소셜 미디어 플랫폼 상의 허위 계정을 탐지하는 알고리즘을 예로 든다. 이 알고리즘은 사람들이 잘 사용하지 않는 이름을 가진 계정에 대해서는 낮은 정확도를 보였는데, 그 결과 그 결과 흔하지 않은 이름을 가진 소수 집단 이용자들은 더 높은 비율로 허위 계정으로 판단되는 문제가 있었다.[142] 따라서 만약 시스템 개발자가 이러한 위험성에 대해 충분한 주의를 기울이지 않는다면 인공지능은 주류적 문화에 따른 규칙을 일반화하여 학습할 것이고, 그 결과 소수자 집단에 대해서는 부정확하게 작동하는 문제를 초래할 수 있는 것이다.

인공지능 공정성 심사에 있어 집단 공정성 개념을 적용하면, 인공지능 시스템에 위와 같은 문제점이 있는지 여부를 확인할 수 있도록 인공지능의 출력이 통계적으로 집단간 동등한 지 여부를 평가하게 된다. 이러한 점에서 집단 공정성은 개인 공정성과는 구분된다. 달리 말하자면, 개인 공정성은 개별 사례에 있어 특정 개인이 유사한 지위에 있는 비교대상(즉, 차별금지사유를 가지지 않은 개인)

---

141 Barocas, Hardt & Narayanan (2022) (각주 19) Chapter 1.

142 위의 글.

과 비교하여 동등한 대우를 받을 것을 요구하는 것에 비해, 집단 공정성은 다수 사례에 대해 통계적으로 인공지능이 집단간 동등하게 작동할 것을 요구한다.[143]

요컨대, 이제껏 인공지능 공학자에 의해 제안 및 논의되어 왔고, 본 연구를 통해서 검토하는 공정성 개념은 주로 개별 인공지능 활용 사례에 있어 헌법상 평등 원칙 또는 차별금지법제 상의 차별 금지 규제를 준수하기 위한 것이라 할 수 있다. 이러한 인공지능 공정성 개념은, 소극적으로는 인공지능이 차별금지법 상의 차별 행위를 하지 않도록 보장하고, 적극적으로는 인공지능의 활용을 통해 실질적 평등을 실현할 수 있도록 조력할 수 있다.

다만, 이처럼 평등 원칙의 관점에서 정의된 인공지능 공정성 지표가 갖는 한계도 유의할 필요가 있다. 무엇보다도 개별적인 인공지능 활용자들이 인공지능 공정성 지표를 적용한다고 하더라도 사회 전체에 걸친 실체적 공정성이 곧바로 보장되는 것은 아니라는 점이다.[144] 인공지능 공정성을 연구하는 공학 연구자들도 이와 같은 아리스토텔레스 정식에 기반한 공정성 개념이 갖는 이러한 한계를 인식하고 있는 것으로 보인다. Dwork et al. (2012)는 자신의 공정성 개념에 대한 정식화가 "전역적(global)" 차원의 문제를 해결하지 못할 수 있고, "국소적(local)"인 차원에서만 적용될 수 있다는 점을 인정한다. 따라서 인공지능에 의한 사회 및 시장 구조의 총체적 변화 과정에서 거시적 분배 공정성을 보장할 수 있는 방안에 대해서는 별도의 연구가 필요하다. 그러므로 이러한 지표는 사회 각 분야에서 인공지능이 활용되기 위하여 보장되어야 할 최소한의 안전장치로 이해될 수 있을 것이다.

---

143 이에 본 연구 제4장 제2절에서 검토하는 출력변수에 대한 공정성 기준은 집단 공정성 개념을 구현한 것이라 평가할 수 있다.

144 Dwork et al. (2012) (각주 134) p. 3는 "일반적으로, 국소적 해법이 동시에 취해진다고 하더라도 전역적 문제를 해결하지는 못한다."고 지적한다.

2. 인공지능에 의한 공정성 침해 양상

이상에서 본 바와 같이 본 연구에서 고찰하는 인공지능의 공정성은 개별 사업자나 기관이 인공지능을 활용하는 개별적이고 구체적인 맥락에서 평등 원칙을 실현하는 것에 초점을 맞춘다. 그런데, 위 제2절에서 본 인공지능의 다양한 유형마다 그 활용으로 인해 발생할 수 있는 잠재적 공정성 침해 양상에 차이가 있다는 점에 유의할 필요가 있다.

인공지능에 의한 공정성 침해 위험성은 인공지능이 분배적 기능을 수행하는 경우와 표상적 기능을 수행하기 위한 경우로 구분하여 평가될 수 있다. 이러한 구분은 Crawford (2017)에 따른 것으로서,[145] 전자의 기능에 따른 위해를 분배적 위해(harms of allocation), 후자의 기능에 따른 위해를 표상적 위해(harms of representation)이라 한다. 분배적 위해는 입학, 취업, 대출 등 사회적으로 제한된 기회나 자원이 인공지능에 의해 사회적 집단에 대해 차등적으로 분배되는 문제를 일컫는 것임에 비해, 표상적 위해는 검색, 검색, 번역, 광고, 챗봇 등의 서비스와 같이 이용자가 일상적으로 당면하는 상황에 있어 인공지능에 의해 사회적 특정 집단을 종속화(subordination)하는 결과가 초래되는 문제를 지칭한다.

가. 분배적 위해

인공지능에 의한 분배적 위해는 인공지능에 개인에 관한 평가나 예측을 위임하는 사례에서 발생할 수 있다. 이러한 활용 사례에서는 대체로 판별모형을 활용하는 것이 일반적이다. 즉, 이제까지의 개인의 행동이나 개인에 대한 평가를 수집하여, 이를 학습 데이터로 삼아 미래 행동을 예측하거나(예컨대, 광고나 콘텐츠 클릭, 채무 불이행, 재범 위험 예측 등) 그의 잠재적 능력(예컨대, 학업이나

---

145 Kate Crawford, “The Trouble with Bias”, *NIPS 2017 keynote speech*, (2017)

업무 수행 능력)을 평가하는 것은 판별모형이 수행하는 기능이다.[146] 그런데, 만약 이러한 평가나 예측이 사회적 집단에 대해 차별적으로 이루어질 경우, 해당 집단 구성원에 대해 개별적이고 구체적인 법률관계에 있어 직접적인 피해가 발생할 수 있다. 그 결과 판별모형에 의한 분배적 위해에 대한 문제제기는 개인의 삶에 중대한 영향을 끼치는 의사결정이 이루어지는 영역에 관해 이루어져 왔다. 이제까지 주요하게 문제제기가 이루어진 영역으로는 대표적으로 교육, 채용, 금융 및 신용평가 등이 있다.

우선, 교육 분야는 인공지능 활용에 있어 기회의 균등 이념을 담보하는 것이 특히 중시된다. 인공지능에 의한 교육 차별 가능성과 관련하여서는 영국 자격검정청(Ofqual)의 학생 평가 알고리즘 사건이 종종 인용된다.[147] COVID19 팬데믹 상황으로 인해 영국은 A-level(한국의 수학능력시험에 해당)을 치르는 학생들의 성적을 자격검정청이 관리하는 알고리즘에 의해 생성된 점수를 학생들에게 부여하고자 하였으나, 거센 사회적 저항에 직면하게 되었다. 특히, 위 알고리즘에 의한 학생 평가 결과는 사립학교에 다니는 학생들에게는 유리한 반면, 어려운 배경의 학생들에게 가장 불리한 것으로 드러났다. 더욱이 위 알고리즘은 영국정부가 수립한 공공부문 AI 활용 지침에도 위반한 것이라는 비판도 제기되었다. 그 결과 아직까지 학생 평가에 있어 인공지능의 공정성 개념을 어떻게 정의하고 적용해야 할 것인지는 여전히 논란이 되고 있다.[148]

---

146 본 장 제2절 제1항 참조.

147 Helen Smith, "Algorithmic bias: should students pay the price?", *AI & Society* 35(4), (2020). 이 사례에 관한 법적 분석으로는 Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law", *West Virginia Law Review* 123(3), (2021b). 한편, 교육에서의 알고리즘 편향 문제에 관한 전반적 조사로는 Ryan S. Baker and Aaron Hawn, "Algorithmic bias in education", *International Journal of Artificial Intelligence in Education* 32, (2021) 참조.

148 Smith (2020) (각주 147)은 교육 분야에서의 인공지능 공정성은 단지 인공지능 자체에 대한 통계적 평가를 통해서만 달성될 수 있는 것이 아니고, 기존 제도의 불공정성과 불평등을 조

다음으로, 채용 과정에서의 인공지능 활용 사례 역시 크게 증가하고 있다. 채용 과정에서 인공지능을 활용함으로써 기존 인간 심사자에 의한 편향이나 의도적 차별행위를 막을 수 있을 것이라는 긍정적인 기대도 있는 반면, 인공지능이 기존 편향이 지속시키거나, 오히려 강화시킬 수 있다는 우려도 함께 제기된다.[149] 채용 분야 인공지능에 의한 차별 가능성을 보여주는 대표적인 사례로는 아마존의 이력서 검토 인공지능 사례가 있다.[150] 아마존은 2014년부터 이력서 검토를 자동화하는 소프트웨어를 개발해 왔다. 이 소프트웨어는 지난 10년 동안의 이력서 평가 결과를 학습 데이터로 사용하였는데, 그 결과 이력서에 지원자가 여성임을 지칭하는 단어가 포함된 경우 낮은 점수를 매기는 문제가 발견되었고, 결국 아마존은 2018년 해당 프로젝트를 포기하기에 이르렀다. 채용과 관련하여 또 다른 문제로는 온라인 플랫폼에서의 구인 광고의 표시의 공정성 문제나 직업 소개 서비스에서의 매칭 알고리즘의 공정성과 관련된 것이다. 이러한 광고 표시의 공정성에 관해서는 역사적으로 차별받아 온 집단에 대해 부정적인 광고가 표시된다는 문제제기,[151] 여성에 대해 STEM(Science, Technology, Engineering,

---

명하여 함께 고려되어야 하며, 역사적 편견과 불평등을 수정하기 위한 정책적 수단과 함께 도입될 필요가 있다고 주장한다.

149 현재 채용 인공지능의 편향에 대한 분석이 이루어지고 있는데, 최근 연구 사례로는 Manish Raghavan, et al., "Mitigating bias in algorithmic hiring: Evaluating claims and practices", *Proceedings of the 2020 conference on fairness, accountability, and transparency*, (2020); Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards, "What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems", *Proceedings of the 2020 conference on fairness, accountability, and transparency*, (2020) 등이 있다.

150 Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women", *Ethics of data and analytics*, Auerbach Publications, (2018).

151 Sweeney (2013) (각주 23).

Mathematics) 분야 직종에 관한 광고가 덜 표시된다는 문제제기가 이루어진 바 있다.[152]

금융 분야에서 활용되는 인공지능 공정성과 관련하여서도 사회적으로 논란이 되고 있다. 다른 영역의 경우와 마찬가지로 금융 인공지능이 기존의 인간에 의한 차별을 줄이고, 포괄적 금융을 확장하는 데 기여할 수 있다는 기대도 있으나,[153] 인공지능이 여성이나 소수 인종에 대해 차별적으로 작동한다는 비판도 제기된다. 예를 들어 2019년 골드만삭스와 애플이 애플카드를 출시한 이후 한 여성에 대해 낮은 신용 한도가 책정되는 문제가 제기된 바 있다.[154] 또한 미국의 차량 보험

---

152 Amit Datta, Michael Carl Tschantz, and Anupam Datta. "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination", *In Proceedings on Privacy Enhancing Technologies*, (2015). 다른 한편, 이 연구에 대해서는 STEM 분야 광고가 성별에 대해 차별적으로 표시되는 것은 인터넷상의 행동 패턴의 차이나 성차별적인 문화에 기인한 것은 아니라, 오히려 평균 광고단가가 성별에 따라 차이가 나기 때문에 발생하였을 개연성이 있다는 반론이 제기되었다. 즉, 다른 광고업계의 광고수요가 여성에 대해 몰려 있기 때문에 상대적으로 경쟁이 적은 남성에게 광고가 더 많이 노출되었을 수 있다는 것이다. 따라서 외견 상 드러나는 결과만을 두고 섣불리 알고리즘에 의한 차별로 단정지어서는 안 되고, 구체적인 작동 메커니즘에 대한 분석이 필요하다. Anja Lambrecht and Catherine E. Tucker, "Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads", *Management Science* 65(7), (2019) 참조.

153 예컨대, Robert Bartlett, et al., "Consumer-lending discrimination in the FinTech era", *Journal of Financial Economics* 143(1), (2022)은 기존 대면 대출 심사 결과와 핀테크 대출에서의 알고리즘 심사 결과를 비교하였는데, 핀테크 대출에서는 대면 심사에서 존재하였던 소수인종에 대한 차별 정도가 완화되었다고 보고한다.

154 이 문제에 관해서는 뉴욕 금융정책부(New York Department of Financial Services, NYDFS)가 감사를 수행하였고, 2021. 3. 그 결과가 발표되었다. NYDFS는 애플카드 신용평가 알고리즘이 법적으로 금지되는 차별을 하였다는 통계적 근거는 없는 것으로 판단하였다(다만, 그 구체적인 평가 방법은 공개된 보고서에 명시되어 있지 않다). 하지만 NYDFS는 신용평가 알고리즘이 향후 투명성 제고와 공정성 확보에 노력을 기울여야 한다고 강조했으며, 그를 위한 방법으로 대체 데이터를 활용한 재심사 등을 제시하였다. New York Department of Financial Services (NYDFS), *Report on Apple Card Investigation*, (2021. 3.).

료 산정에 있어 소수인종 집단이 주로 거주하는 지역에 대해 더 높은 보험료를 부과하는 문제도 지적된 바 있다.[155]

나. 표상적 위해

인공지능에 의해 표상적 위해가 초래될 수 있다는 우려는 대체로 인지모형과 생성모형과 관련하여 제기된다. Crawford(2017)은 표상적 위해를 인식(recognition), 고정관념화(stereotyping), 과소대표(under-representation), 폄훼(denigration)로 구분하는데, 이하에서는 이를 구분하여 살핀다.

1) 인식 문제

표상적 위해 중 인식 문제는 인지모형이 특정한 사회적 집단에 소속된 개인을 정확하게 인식하지 못하는 경우에 발생할 수 있다. 2017년 인터넷에 게시된 "인종차별적" 비누 디스펜서(dispenser) 영상은 이를 상징적으로 보여준다.[156] 이 영상에 나타난 비누 디스펜서는 손바닥이 흰 이용자가 손을 대면 정상 작동하지만, 손바닥이 어두운 이용자에 대해서는 작동하지 않는다.[157] 이처럼 이용자의 다양성에 대한 고려가 이루어지지 못할 경우, 비누 디스펜서와 같이 단순한 기능을 수행하는 장치에서도 공정성 침해 문제가 발생할 수 있다.

인지모형이 차별적 작동 위험성을 보여주는 사례로는 얼굴인식 인공지능의 문제가 있다. Buolamwini & Gebru (2018)는 새로운 감사 데이터를 개발하여 마이크로소프트, IBM, Face++의 상용 얼굴인식 소프트웨어를 평가한 결과, 흑인

---

155 Julia Angwin, et al., "Minority neighborhoods pay higher car insurance premiums than white areas with the same risk", *ProPublica*, (2017. 4. 5.). 하지만 이러한 문제제기에 대해서 보험회사들은 피보험자가 운전하고 있는 환경이 더 위험하기 때문에 이러한 결과가 초래된 것이지 소수인족 집단에 대한 차별이 이루어지는 것은 아니라고 반박하였다.

156 https://twitter.com/nke_ise/status/897756900753891328 (2023. 5. 26. 접속 확인).

157 추측컨대, 디스펜서는 센서의 명암을 기준으로 손을 인식하기 때문인 것으로 보인다.

여성에 대한 정확도가 타 인종이나 성별에 비해 상당히 낮게 나타났다는 점을 발견하였다.[158] 이 연구는 얼굴인식 인공지능의 공정성 문제에 관한 큰 관심을 불러 일으켰다. 다만, 현재 발표되고 있는 얼굴인식 알고리즘은 인구통계적 차이에 따라서도 정확도 차이를 크게 보이지 않고 있다.[159] 그 이외에도 구글의 사진 인식 소프트웨어가 흑인 여성을 고릴라로 잘못 인식한 사례,[160] 카메라에 포함된 얼굴 인식 소프트웨어가 눈을 뜨고 있는 아시아인에 대해 눈을 뜨라는 경고 메시지를 보낸 사례[161] 등이 제시된다. 이러한 인식 기능에 있어서의 집단간 성능 격차는 특정 개인에게 개별적이고 구체적인 법률관계에 있어 직접적인 피해를 초래하는 것은 아니지만, 일상적이고 문화적인 차원에서 지속적으로 악영향을 끼칠 가능성이 있다.

2) 고정관념화

표상적 위해 중 고정관념화의 문제는 주로 생성모형의 출력물에 사회적 편향이 반영되어 있는 경우 발생할 수 있다. 예를 들어 Sheng et al.(2019)는 GPT-2 모델에 서로 다른 성별, 인종 및 성적 지향을 가진 개인을 프롬프트로 입력하였을 때 다음 그림과 같이 편향된 텍스트가 생성되었다고 지적하였다.[162] 아래 사

---

158 Joy Buolamwini and Timnit Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification”, *Conference on fairness, accountability and transparency*, (2018)

159 National Institute of Standards and Technology, “NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software”, NIST News (2019. 12. 19.).

160 BBC, “Google apologises for Photos app's racist blunder”, (2015. 7. 1.)

161 PetaPixel, “‘Racist’ Camera Phenomenon Explained — Almost”, (2010. 1. 22.)

162 Emily Sheng, et al., “The Woman Worked as a Babysitter: On Biases in Language Generation”, *Empirical Methods in Natural Language Processing*, (2019).

례에서는 여성은 매춘부로, 흑인 남성은 포주로, 동성애자는 춤을 좋아하고 마약을 하는 것으로 묘사되었다.[163]

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

그림 9 GPT-2에 의해 생성된 편향된 텍스트

또한, 이미지를 텍스트로 변환하거나, 텍스트를 이미지로 변환하는 인공지능에 있어서도 유사한 편향에 대한 문제제기가 이루어지고 있다. 예를 들어 Hendriks et al. (2018)은 이미지 캡셔닝 모형에 "스노보드를 주로 타는 사람은 남성이다"와 같은 고정관념이 반영되어 있다는 점을 확인하였다.[164] 이미지 캡셔닝 모델은 누군가 스노보드를 타는 사진에 대해 실제 성별과 무관하게 남성으로 잘못 묘사하는 경향이 있다는 것이다. 이러한 문제는 해당 모형이 기존 학습 데이터의 분포와 이미지의 배경에 지나치게 의존하는 경우 발생할 수 있다. 해당 연구는 아래 그림을 통해 이 문제를 설명한다. 이미지 캡셔닝 모형이 어떤 여성이 컴퓨터 앞에 앉아 있는 사진을 남성이라고 잘못 묘사한 사례에 있어, 그 모형은 그 사진 중 컴퓨터 부분을 중심적으로 고려한 것으로 나타났다. 이와 달리 그 사람에 관한 부분을 중요하게 고려하도록 개선한 모형에서는 정확한 결과를 생

163 한편 이러한 사례는 여성, 흑인 및 동성애자 집단 구성원을 폄훼하는 것으로도 평가될 수 있을 것이다.

164 Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach, "Women also snowboard: Overcoming bias in captioning models", *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018).

성해 냈다. 이처럼 생성모형은 고정관념이 잘못 반영하여 부정확한 결과를 생성해 내는 경우, 고정관념을 반영하지 않도록 개선하는 것은 정확도 향상으로 이어질 가능성이 있다.

나아가 텍스트를 이미지로 변환하는 인공지능 시스템의 경우에도 유사한 편향을 확인한 연구들이 발표되고 있다. 최근 연구는 DALL-E 2 및 Stable Diffusion과 같이 널리 사용되는 이미지 생성 시스템이 생성한 96,000건 이상의 이미지를 분석한 결과, 이들 시스템이 모두 성별 및 인종에 있어 백인과 남성을 과다하게 생성하고 있다는 점을 발견하였다.[165] 해당 연구는, 인공적으로 생성된 허구적 인간에게는 고유한 성별이나 인종이 없고, 사회적으로 구성된 집단에도 속하지 않으므로 기존의 일반적인 범주화에 얽매일 필요가 없음에도 불구하고, 이미지 생성 시스템이 기존의 사회적 편향과 고정관념을 반영하는 것은 타당하지 못하다는 문제를 지적한다.

165 Luccioni, et al. (2023) (각주 99).

3) 폄훼

'폄훼'는 생성모형의 편향으로 인해 특정 집단이나 개인의 명예나 명성을 훼손하는 결과가 도출되는 경우를 일컫는다. 종래 검색 사이트가 제공하는 자동완성 기능이 어떤 사회적 집단에 관해 부정적인 단어를 추천하는 사례가 잘 알려져 있다. 예를 들어 검색 엔진에 "women should"라고 입력했을 때 "stay at home", "be in the kitchen" 등의 어구가 추천된 바 있다.[166] 또한, 구글 검색엔진에서 흑인 이름을 검색하였을 때 그 사람의 체포 여부를 확인하라는 키워드 검색 광고가 표시되는 것에 대한 문제제기가 이루어진 바 있다.[167]

최근의 언어모형이 역사적으로 불이익을 받아온 집단을 폄훼하는 결과를 생성하는 문제도 지적되고 있다.[168] 국내 '이루다' 챗봇 서비스에 있어서도 장애인, 임산부, 성적 소수자, 흑인 등의 집단에 대한 혐오표현이 생성되는 문제가 발생한 바 있다. 이러한 문제는 여러 대규모 언어모형에 있어 공통적으로 발견되고 있다.[169] 또한 언어모형이 특정 종교나 그 신도들을 폄훼하는 답변을 생성하는 경향도 발견된다. Abid, et al. (2021)은 GPT-3에 대해 프롬프트 완성, 비유적 추론 및 스토리 생성 등의 실험을 수행한 결과, GPT-3가 이슬람교도에 대해 모욕적이고 오해를 불러일으키는 결과물을 생성한다는 점을 밝혔다.[170] 더욱이 위 연

---

166 사피야 노블, 노윤기 옮김, *구글은 어떻게 여성을 차별하는가*, 한즈미디어, (2018).

167 Sweeney (2013) (각주 23). 한편, 이후 구글은 사람 이름에 대한 검색을 하는 경우 광고가 표시되지 않도록 서비스를 수정하는 방식으로 이 문제에 대응하였다.

168 Jwala Dhamala, et al., "BOLD: Dataset and metrics for measuring biases in open-ended language generation", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (2021).

169 위의 글.

170 Abubakar Abid, Maheen Farooqi, and James Zou, "Persistent anti-muslim bias in large language models", *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, (2021). 위 연구에 소개된 사례로는 "두 명의 이슬람교도가 걸어 들어갔는데(Two Muslims walked into a)"라는 불완전한 문장을 주고, 이를 완성하는 과제를 주었을 때, "이들은 텍사스의 교회로 들어가서 총을 쏘기 시작했다. 경찰이 도착하자 첫 경찰관의 머리를 저

구에 따르면 이슬람교에 관한 출력 결과는 다른 종교와 비교해서 고정관념과 연관되는 사례가 더 자주 발견되었는데, 이슬람교도는 언어모형 생성물 중 23%의 사례에서 "테러리스트""에 비유된 것에 비해, 유대인이 "돈"에 비유된 사례는 5%에 불과했다. 위 연구는 언어모형의 이러한 편견이 특정 집단에 대해 잘못된 정보를 유포하고, 고정관념을 고착시키는 데 기여하는 결과를 초래할 수 있다고 지적한다.[171]

4) 과소대표

과소대표는 생성모형의 출력물이 특정 집단을 비례적으로 반영하지 못하는 문제를 일컫는다. 이는 주로 직업과 관련된 출력에서 자주 언급되는데, 가령 사진 검색엔진에서 "CEO"로 검색한 결과 여성이나 흑인 CEO는 표시되지 않거나 극소수만 표시되는 경우 이들이 과소대표된다고 평가할 수 있다. 예를 들어 특정 직업에 관한 생성모형의 출력 중 특정 성별이 과소대표된다면, 사람들의 성별에 관한 고정관념이 고착화되거나 과장될 수 있다.[172] 더욱이 미성년자에게 이러한

격했다" 혹은 "이들은 도끼와 폭탄을 가지고 유대교 회당으로 들어갔다"고 문장이 완성된 경우가 있다.

171 위의 글. 위 연구자들은 이 문제에 대한 해법으로, 더욱 포용적인 학습 데이터의 구축, 편향을 평가하기 위한 더 나은 지표의 개발, 인공지능 개발 과정에 대한 투명성 증진을 제안한다.

172 검색 엔진에 의한 편향된 이미지 검색 결과로 인하여 직업에 관한 성별 고정관념이 강화될 위험성에 관해서는 오랫동안 연구되어 왔다. 예를 들어 Matthew Kay, Cynthia Matuszek, and Sean A Munson, "Unequal representation and gender stereotypes in image search results for occupations", *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, (2015); Vivek K Singh, et al., "Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms", *Journal of the Association for Information Science and Technology* 71(11), (2020); Danae Metaxa, et al., "An image of society: Gender and racial representation and impact in image search results for occupations", *Proceedings of the ACM on Human-Computer Interaction* 5, (2021).

생성물이 제시되는 경구 그 위험성은 더 클 수 있다.[173] 예를 들어 최근 한 이용자가 ChatGPT에 대해 "한 소년과 한 소녀가 자신의 직업을 고르는 이야기를 해 달라"고 요청하자, 소년은 성공적인 의사가 되었고, 소녀는 사랑받는 교사가 된 이야기를 생성해 낸 바 있다.[174] 나아가, 언어모형을 활용한 이미지 생성 시스템에서도 성별이나 인종과 같이 시각적으로 눈에 띄는 특성에 있어 소수자 집단을 과소 대표하는 경향이 발견된다.[175]

제 4 절 소결

본 장에서는 본 연구에서 사용하는 인공지능의 개념을 분류모형과 생성모형으로 구분하여 정의하였다. 나아가 분류모형은 다시 인지모형과 판별모형으로 구분하였다. 이상에서 검토한 인지모형과 판별모형 간의 차이를 요약하면 다음과 같다.

---

173 Equality Now, "ChatGPT-4 Reinforces Sexist Stereotypes By Stating A Girl Cannot 'Handle Technicalities And Numbers' In Engineering", *News and Insights* (2023. 3. 23.) https://www.equalitynow.org/news_and_insights/chatgpt-4-reinforces-sexist-stereotypes/ (2023. 5. 26. 접속 확인)

174 다른 ChatGPT 이용자는 위 결과에 대한 통계적 분포를 얻기 위해 같은 요청에 대해 10,000건의 이야기를 생성할 것을 요청하였는데, 생성된 답변 중 소년이 공학자가 된 사례는 2,731회, 소녀는 1,209회였다. 반대로 소년이 교사가 된 사례는 1,187회, 소녀는 2,244회였다. https://twitter.com/IvanaBartoletti/status/1637401609079488512 (2023. 5. 26. 접속 확인)

175 Luccioni, et al. (2023) (각주 99), p. 4.

	인지모형	판별모형
정답 라벨	결정론적, 안정적	확률적, 비고정적
모형 구조	딥러닝 모형	고전적 머신러닝(XGBoost 등)
학습 데이터	비정형 데이터	정형 데이터
과제	시스템 1 과제	시스템 2 과제
활용 예시	얼굴 인식, 사물 인식, 음성 인식, 자연어 이해	온라인 광고, 맞춤형 컨텐츠 추천, 직원 채용, 학생 선발, 신용 평가, 재범 위험성 평가
설명가능성	상대적으로 낮음	상대적으로 높음
데이터 불확실성	상대적으로 낮음	상대적으로 높음

표 1 인지모형과 판별모형의 구분

다른 한편, 생성모형도 다시 정형 데이터 생성모형과 비정형 데이터 생성모형으로 구분할 수 있다. 전자는 주로 개인정보 보호규제 준수를 위하여 목적으로 고려되고 있는 것과 달리 비정형 데이터 생성모형은 다양한 창작물을 생성할 수 있을 뿐만 아니라, 다양한 과제를 수행하는 일반 목적 인공지능으로 발전하고 있다.

이상과 같은 인공지능의 유형화는 인공지능 공정성 지표의 선택 문제에 있어 중요한 함의를 갖는다. 즉, 분류모형 중 판별모형은 사회적 자원이나 기회를 분배하는 목적으로 주로 사용되고 있고, 그 결과 그로 인해 분배적 위해를 초래할 위험성에 관한 문제제기가 주요하게 이루어지고 있다. 이와 달리 분류모형 중 인지모형은 인식 과제에서의 차별적 작동으로 인해 표상적 위해 중 인식 문제를 초래할 위험성이 지적된다. 다른 한편, 생성모형은 역사적으로 불리한 지위에 있어왔던 사회적 집단이나 그 구성원에 대한 고정관념화, 폄훼, 과소대표 위험성 등이 지적되고 있다. 이에 생성모형에 반영된 고정관념을 발견하고 이를 완화하는 조치를 취하는 작업이 중요한 사회적 과제로 부각되고 있다.

제3장 인공지능 공정성 보장을 위한 법적 규제

제 1 절 개관

이상에서 본 바와 같이 각 인공지능에 의해 공정성이 침해될 수 있다는 우려가 제기되면서, 이에 대한 법적 규율이 점차 중요해지고 있다. 우선 인공지능의 차별적 작동 위험에 대해 기존의 차별금지법제를 적용하는 방안을 고려할 수 있다.[176] 나아가 인공지능의 활용 범위가 크게 증가하면서 차별금지법제 이외에도 개인정보 보호법제와 같이 다른 규제가 중첩적으로 적용되어가는 양상을 보이고 있다.[177] 이에 그치지 않고 인공지능의 편향과 위해를 직접적으로 다루기 위하여 인공지능 공정성 규제를 입법화하려는 시도가 국내외적으로 활발히 진행되고 있기도 하다. 이에 본 장에서는 인공지능의 공정성을 보장하기 위한 다양한 실정법상 규제 및 현재 진행 중인 입법안들을 검토한다. 본 연구는 인공지능의 공정성 보장을 위한 현행법상 규제 또는 법안을 그 입법 목적 및 규율 방식을 기준으로 다음 세 가지로 구분한다.

첫번째 규제 형태는 차별금지법제이다. 차별금지법제는 헌법상 평등 이념을 실현하기 위하여 기존의 차별을 시정하고 차별행위를 금지하는 것을 목적으로 한다.[178] 통상적으로 차별금지법은 성별·종교 또는 사회적 신분 등과 같이 차별금지사유를 나열하고, 채용, 교육 등 일정한 영역에서 차별금지사유를 이유로 불리

---

176 인공지능에 의한 차별 문제에 대한 전반적인 국내 연구로는 고학수·정해빈·박도현, “인공지능과 차별”, *저스티스* 171, 199-277, (2019) 참조.

177 Ignacio N. Cofone, “Algorithmic Discrimination Is an Information Problem”, *Hastings Law Journal* 70(6), (2019a).

178 최승철, *차별금지법의 이해*, 한울아카데미, (2011).

하게 대우하는 행위를 금지한다.[179] 이처럼 차별금지법제의 적용 범위는 차별금지사유와 차별금지영역에 따라 결정되는데, 차별금지법은 다수의 차별금지사유와 차별금지영역을 포괄적으로 규정하는 '일반적 차별금지법'과 특정한 차별금지사유나 차별금지영역만을 규정하는 '개별적 차별금지법'으로 구분할 수 있다.[180]

본 연구에서 고려하는 또다른 규제 형태는 개인정보 보호법제이다. 개인정보 보호법제는 개인정보자기결정권 또는 사생활의 비밀 등을 보호하는 목적을 갖고 있으나, 인공지능 공정성 문제에 대해서도 그 적용이 확대되어 가는 경향을 보인다.[181] 특히 인공지능에 의한 차별이 이루어지는 경우, 머신러닝 학습 및 추론 단계에서 개인정보가 활용되므로, 개인정보 보호법제는 인공지능에 의한 차별 문제에 대한 중요한 규제 수단이 되어 가고 있다. 특히 인공지능에 의한 차별은 본질적으로 개인정보의 이용 문제에 해당하므로, 이를 막기 위해서는 개인정보 이용에 대한 적절한 규칙을 정립할 필요가 있다는 주장이 설득력 있게 제기되고 있다.[182] 구체적으로는 개인정보 보호법제가 프로파일링 등을 포함한 자동화된 의사결정에 있어 투명성 규제 또는 설명요구권 등을 규정하는 경우가 있는 바, 이는 인공지능에 의한 차별을 통제하기 위한 중요한 규제수단이 될 수 있다.[183]

마지막 규제 형태로는 인공지능 규제법제가 있다. 이는 인공지능이라는 특정 기술을 활용함으로써 초래될 수 있는 위험을 규율하는 것을 목적으로 한다는 점

---

179 박은정 외, *차별판단기준 수립을 위한 연구(제1부)*, 국가인권위원회, (2021), 11-24면은 차별금지 사유 및 차별금지 영역에 관한 국내법상 규율 현황을 전반적으로 조사하고 해외 입법례와 비교하고 있다.

180 홍성수, "포괄적 차별금지법의 필요성", *이화젠더법학* 10(3), (2018), 6-7면.

181 ICO의 "Guidance on AI and data protection" (2020.7.)은 "개인정보 보호 맥락에서 공정성(fairness)이란 일반적으로 개인정보를 사람들이 합리적으로 예상할 수 있는 방법으로 취급해야 하고, 그들에게 정당화될 수 없는 부정적 영향을 끼치는 방식으로 사용하지 말아야 한다는 점을 의미한다"고 설명하고 있다.

182 Cofone (2019a) (각주 177).

183 GDPR 제22조 및 개인정보 보호법(법률 제19234호, 2023. 3. 14., 일부개정) 제37조의 2 참조. 한편, 고학수 외(2019) (각주 176) 234-242면은 이 문제를 차별의 사후적 사업판단에 관한 절차법적 문제로 다루고 있기도 하다.

에서 앞서 살펴본 다른 법제와는 입법 형태가 확연히 구별된다. 이러한 법제에는 대표적으로 EU AI 법안과 같이 인공지능 활용 사례를 전반적으로 규율하고자 하는 시도가 있다.[184] 또한 미국 연방 의회에 발의된 바 있는 알고리즘 책임 법안(Algorithmic Accountability Act)이나,[185] 미국 여러 주 차원에서는 얼굴인식 알고리즘 규제, 채용 알고리즘 규제와 같이 특정한 활용 사례를 중심으로 고려한 규제(안)도 존재한다. 또한 이러한 규제는 인공지능 기술 활용에 대한 투명성 의무, 학습 데이터에 대한 관리 의무, 로그 기록 의무, 인간 감독 의무의 부과 등과 같이 인공지능 기술의 개발 및 이용에 특유한 의무사항을 규정하고 있다는 점에서 특색이 있다. 이하에서는 각각의 입법 중 인공지능의 공정성 보장을 위한 규율을 중심으로 소개하고 이들을 비교한다.

제 2 절 차별금지법제

본 절은 인공지능에 의한 차별의 맥락에서 기존 차별금지법제를 검토한다. 특히 인공지능에 의한 차별에 있어 차별금지법제가 실무적으로 어떻게 적용될 수 있고, 어떠한 법적 쟁점이 제기되고 있는지 살핀다.

---

184 국내 다수 문헌이 EU AI 법안의 주요 내용을 소개하고 있다. 예컨대, 김송옥, "AI 법제의 최신 동향과 과제 - 유럽연합(EU) 법제와의 비교를 중심으로 -", *공법학연구* 22(4), (2021); 홍석한, "유럽연합 '인공지능법안'의 주요 내용과 시사점", *유럽헌법연구* (38), (2022); 정소영, "유럽연합 인공지능법안의 거버넌스 분석 - 유럽인공지능위원회와 회원국 감독기관의 역할과 기능을 중심으로 -", *연세법학* 39, (2022) 등 참조.

185 이 법안은 당초 2019년 발의되었다가 통과되지 못하였고, 2022년 대폭 수정된 법안이 다시 발의되어 있다. Office of U.S. Senator Ron Wyden (2022), "Algorithmic Accountability Act of 2022"; EU AI 법안과 미국 알고리즘 책임 법안에 관한 비교로는 Jakob Mökander, et al., "The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other?", *Minds and Machines* 32, (2022) 참조.

차별금지 법제는 국가별 차이가 있으나, 일반적으로 '직접차별(Direct Discrimination)'[186]과 '간접차별(Indirect Discrimination)'[187]을 모두 금지하는 방식을 취한다.[188] 한편, 미국 민권법 제7편(Title VII of the Civil Rights Act of 1964)상 고용 영역에서의 차별행위를 포괄적으로 금지하고 있는데, 이는 "의도적 차별행위(Disparate Treatment)"[189]와 "결과적 차별행위(Disparate Impact)"

---

186 직접차별은 "차별금지사유가 직접적인 이유가 되어 발생하는 합리적 이유가 없는 차별"을 뜻한다. 박은정 외(2021) (각주 179), 54면. 국내에서는 남녀고용평등과 일·가정 양립 지원에 관한 법률 제2조, 장애인차별금지 및 권리구제 등에 관한 법률 제4조 제1호, 고용상 연령차별금지 및 고령자고용촉진에 관한 법률 제4조의4 제1항, 고용정책기본법 제7조, 기간제 및 단시간근로자 보호 등에 관한 법률 제8조, 파견근로자보호 등에 관한 법률 제21조, 외국인근로자의 고용 등에 관한 법률 제22조, 장애인차별금지 및 권리구제 등에 관한 법률 제6조, 고용상 연령차별금지 및 고령자고용촉진에 관한 법률 제4조의4, 후천성면역결핍증 예방법 제3조 등에서 직접차별 금지를 규정하고 있다.

187 간접차별은 "중립적인 기준을 적용하였으나, 그러한 중립적 기준이 특정 소수자 집단에게 불이익한 결과를 야기하는 경우를 차별로 보는 개념"이다. 박은정 외(2021) (각주 179), 54면. 국내에서는 남녀고용평등과 일·가정 양립 지원에 관한 법률 제2조에서 "사업주가 채용조건이나 근로조건은 동일하게 적용하더라도 그 조건을 충족할 수 있는 남성 또는 여성이 다른 한 성(性)에 비하여 현저히 적고 그에 따라 특정 성에게 불리한 결과를 초래하며 그 조건이 정당한 것임을 증명할 수 없는 경우"를 금지되는 차별로 포함하고 있다. 그 외에도 장애인차별금지 및 권리구제 등에 관한 법률 제4조 제1항 제2호, 고용상 연령차별금지 및 고령자고용촉진에 관한 법률 제4조의4 제2항에서 간접차별 금지를 규정하고 있다.

188 고학수 외(2019) (각주 176) 219면. 한편, 직접차별과 간접차별 이외에도 불리한 대우로서의 보복행위, 성희롱, 괴롭힘 등도 차별의 한 유형으로 볼 수 있다. 박은정 외(2021) (각주 179), 68-80면. 그러나, 인공지능에 의한 차별의 맥락에서는 그 연관성이 높지 않으므로 이하에서는 이를 고려하지 않는다. 고학수 외(2019) 220면 이하도 마찬가지이다.

189 의도적 차별행위는 "인종, 성별 등 금지된 요소에 기초해 사용자가 의도적으로 차별하는 것"을 뜻한다. 이승길, "미국의 고용차별금지법 개관", *맞춤형 법제정보* 5, 한국법제연구원, (2015). 한편, "disparate treatment"를 곧바로 "직접차별"로 번역하거나, "차등적 대우" 또는 "차별적 취급"으로 번역하기도 하는데, 본 연구는 최윤희, "미국에서의 결과적 차별행위(Disparate Impact) 이론에 대한 고찰", *노동법 연구* 20, (2006)에 따라 "의도적 차별행위"라는 번역어를 사용한다.

[190]로 구분된다. 일반적으로 미국 민권법상의 의도적 차별행위는 직접차별과, 결과적 차별행위는 간접차별과 대체로 대응되는 것으로 이해된다.[191]

본 연구에서 검토한 인공지능의 유형 중 차별금지법제의 적용 대상으로 고려되는 것은 주로 판별모형에 의해 자동화된 의사결정이 이루어지거나, 판별모형이 인간의 의사결정을 보조하는 형태로 활용되는 사례이다. 이때 관련 차별금지법률이 적용될 것인지는 법정된 차별금지영역에 있어 차별금지사유에 따른 차별행위가 존재하는지에 의해 결정될 것이다. 구체적으로, 국가인권위원회법은 차별금지영역을 고용, 재화 등 공급·이용, 교육시설 등 이용으로 구분하여 정하고 있다.[192] 그 이외에도 개별적 차별금지법률은 각기 적용되는 차별금지영역을 설정하고 있다.[193] 나아가 국가인권위원회법은 19가지의 차별금지사유를 규정하고 있고,[194] 개별적 차별금지 법률에서도 각각의 차별금지사유를 정하고 있다.[195]

---

190 결과적 차별행위는 미국 Griggs v. Duke Power 사건을 통해 성립된 개념으로서, 미국 연방대법원은 위 사건에서 어떠한 조건들이 외견상 중립적으로 보이더라도 이전의 차별적 고용관행을 유지하게 하는 역할을 하는 경우 차별행위에 해당한다고 판단하였다. 이후 누적된 미국 판례에서 인정된 법리에 따라 1991년 개정된 민권법 제7편 703(k)(1)는 결과적 차별 주장을 명시적으로 규정하고 있다. 이병운, "미국의 고용차별금지법제에 관한 연구", 동아법학 67, (2015), 403-404면. 한편, "결과적 차별행위"는 "간접차별", "차등적 효과", "차별적 효과" 등으로 번역하기도 하는데, 본 연구는 최윤희(2006) (각주 189)에 따라 "결과적 차별행위"라 번역한다.

191 고학수 외(2019) (각주 176) 219면은 유럽과 미국의 차별금지 법리에 있어 차이가 있음을 지적하면서도 인공지능에 의한 차별의 맥락에서는 두 법리에 공통된 구조를 중심으로 논의하고 있다. 본 연구에서도 이와 마찬가지로 우리 법제도를 중심으로 하되, 관련된 범위 내에서 미국과 유럽의 법리를 참조하여 검토한다.

192 국가인권위원회에 대해 가장 많은 진정이 이루어지는 차별행위 영역은 재화 등 공급이나 이용이고, 고용 영역에서는 채용, 모집, 임금, 해고 순이라고 한다. 재화 등 공급·이용 영역에 있어서의 진정은 장애를 이유로 하는 사례가 가장 많고, 고용 영역에서의 차별은 사회적 신분을 이유로 하는 경우가 가장 많다고 한다. 박은정 외(2021) (각주 179), 19-20면 참조.

193 각주 186 및 각주 187에 인용된 개별 차별금지법률 참조.

194 국가인권위원회법 제2조 제3호.

195 박은정 외(2021) (각주 179), 12-13면 참조.

이하에서는 판별모형 인공지능이 법률상의 차별금지영역에 속한 분야의 의사결정에 있어 활용되는 사례를 전제한다. 나아가 차별금지사유에 해당하는 속성을 가진 어떤 개인이 인공지능으로 인하여 불리하게 대우받았다고 주장하는 상황을 고려한다. 여기서 쟁점은 과연 어떠한 경우 차별금지법제상 인공지능을 활용한 판단이 법적으로 금지되는 차별행위에 해당하는지 여부이다. 이처럼 특정 개인이 불리하게 대우받게 된 원인은 해당 판별모형에 기인한 것일 수도 있고 그렇지 않을 수도 있다. 즉, 학습 데이터상에 기존의 구조적 차별이나 사회적 편향이 반영되어 있기 때문일 수도 있고, 판별모형 갖는 통계적 속성으로 인해 각 개인의 고유성이 제대로 반영되기 어렵다는 한계에서 기인한 것일 수도 있다. 그 원인의 여하에도 불구하고, 인공지능을 활용한 차별적 의사결정이 과연 현행 차별금지법제상 어떠한 방식으로 평가될 수 있는지는 쉽지 않은 도전으로 여겨지고 있다. 나아가, 만약 현행 법제가 이러한 문제에 대해 적절히 규율하지 못하고 있다면, 이를 어떻게 개선할 수 있을 것인지는 중요한 법적 과제가 될 것이다.[196]

1. 인공지능에 의한 직접차별

인공지능에 의한 판단이 직접차별에 해당하기 위해서는 해당인이 차별금지사유를 가졌다는 이유로 유사한 지위에 있는 다른 사람이 받아야 할 대우에 비하여 해당인이 불리한 대우(less favorable treatment)를 받았음이 입증되어야 한다.[197] 나아가, 그러한 대우에 대한 '합리적 이유'가 있다면 직접차별에 해당하지

---

196 특히 Barocas and Selbst (2016) (각주 31)는 현재의 미국 차별금지법제가 빅데이터에 기반한 결과적 차별행위에 대응하기에 적합하지 못하다는 주장을 제기한다.

197 이는 유럽차별금지법에 따른 직접차별의 요건이다. European Union Agency for Fundamental Rights and Council of Europe, *Handbook on European Non-Discrimination Law*, Publications Office of the European Union, (2018). 국가인권위원회의 직접차별 요건도 대체로 유사한 것으로 보인다. 박은정 외(2021) (각주 179) 제54면.

않게 된다.[198] 한편 차별금지소송에서 원고가 “차별 의도”를 입증할 것을 요건으로 하는지는 국가마다 차이가 있다. 미국 민권법상 의도적 차별행위는 차별적 의도를 입증할 것을 요구하고 있지만,[199] 유럽연합에서는 원고가 차별 의도를 입증할 것을 요구하지 않는다고 본다.[200]

인공지능에 의한 직접차별의 문제는 (1) 그 모형이 차별금지사유를 명시적 입력변수로 활용하는 경우와 (2) 이를 명시적 입력변수로 활용하지 않지만 다른 입력변수로 인해 차별금지사유를 추론할 수 있는 경우로 구분하여 검토할 수 있다.

가. 차별금지사유를 명시적 입력 변수로 활용하는 경우

인공지능 모형이 차별금지사유를 명시적 입력 변수로 활용한 사례로 앞서 소개한 아마존의 이력서 검토 인공지능을 예로 들 수 있다. 언론 보도에 따르면 아마존이 개발한 소프트웨어는 이력서에 “여성 체스 동호회 회장” 등과 같이 여성

---

198 국가인권위원회법 제2조 제3호 및 남녀고용평등과 일·가정 양립 지원에 관한 법률 제2조 제1호는 모두 “합리적 이유 없이”를 포함하고 있다. 직접차별에 대한 대표적인 합리적 이유로는 이른바 “진정직업자격(bona fide occupational qualification, BFOQ)”이 있는데, 이는 업무의 본질적 성격상 불가피하게 차별이 필요한 경우를 일컫는다. 박은정 외(2021) (각주 179) 제87면.

199 그런데, 사용자의 차별 의도를 입증할 수 있는 직접 증거(이른바 ‘smoking gun’)가 있는 경우는 드물 것이므로, 대부분의 사례에서 간접 증거에 의해 차별 의도를 추인하게 된다. 미국 연방대법원은 McDonnell Douglas 사건에서 3단계 입증 규칙을 확립하고 있는데, 이는 (1) 원고가 차별의 ‘일응 증명(prima facie case)’을 수립하면 (2) 피고가 반증으로 ‘적법하고 비차별적인 이유(legitimate, non-discriminatory reason)’를 제시하고, (3) 원고가 다시 피고가 제시한 이유가 ‘핑계(pretext))’에 불과하다는 것을 입증하도록 한다. 이승길(2015) (각주 189).

200 European Union Agency for Fundamental Rights and Council of Europe (2018) (각주 197) pp. 239-241는 차별을 증명하기 위해서 의사결정자가 편견을 가지고 있다거나 차별 의도를 가졌다는 점은 입증할 필요가 없다고 명시하고 있다.

에 관한 단어가 포함되어 있을 경우 평가 점수를 낮추는 것으로 밝혀졌다.[201] 이처럼 인공지능이 이력서 평가 점수를 산정하기 위해 이력서에 기재된 성별 정보를 활용하는 것은 차별금지사유를 이유로 불리한 대우를 한 것이므로, 일응 직접차별로 평가될 수 있을 것이다. 이처럼 직접차별 금지원칙은 차별금지사유를 인공지능의 입력 변수로 활용하는 것을 금지하는 것으로 일반적으로 해석되고 있다.[202] 그러나, 차별금지 영역에 있어서의 의사결정을 위해 인공지능을 활용하는 사례에 있어, 인공지능을 개발하거나 도입하는 주체가 차별금지사유를 명시적인 입력변수로 사용하는 경우는 실무상 많지 않을 것이다. 위 사례에서도 아마존은 결국 2014년부터 진행해 온 프로젝트를 폐기하였다.[203] 따라서 다른 입력 변수로 통해 차별금지사유를 추론할 수 있는 경우가 더욱 중요한 법적 과제로 고려될 필요가 있다.

나. 차별금지사유를 추론할 수 있는 경우

인공지능에 차별금지사유가 명시적으로 입력되지 않더라도 다른 외견상 중립적인 속성을 통해 차별적 속성을 추론할 수 있는 경우가 많다. 이를 '대용변수

---

201 Dastin (2018) (각주 150) 참조. 아마존이 구체적으로 어떠한 인공지능 기술을 사용하였는지는 알려져 있지 않지만, 언론 보도에 기초하여 보면 이력서에 포함된 단어마다 특정한 가중치를 부여한 것으로 보인다. 이와 같이 문서에 포함된 단어의 순서를 고려하지 않고 어떠한 단어가 포함되어 있는지, 각 단어의 등장 빈도는 얼마나 되는지를 고려하는 모형을 "단어 가방(bag of words)" 방식이라 한다. 아마존이 해당 프로젝트를 시작한 2014년 무렵에는 단어 가방 방식이 자연어 처리를 위해 널리 활용되고 있었으므로 아마존 역시 이러한 모형을 사용하였을 가능성이 있다.

202 이를 'unawareness' 기준 또는 '블라인드' 기준이라 하는데, 이에 대해서는 본 연구 제4장 제2절에서 다룬다.

203 다만, 아마존이 이력서 검토 인공지능 프로젝트를 포기함으로써 다시 인간 평가자에 의한 심사 과정으로 돌아가게 된 것이 과연 사회적으로 바람직한 결과인지에 대해서도 숙고할 필요가 있다. 오히려 인공지능을 이용함으로써 기존 채용 담당자들의 의사결정이 차별적으로 이루어져 왔다는 사실이 드러난 이상, 이들의 채용 판단을 개선하기 위한 다른 조치를 취하지 않는 한 기존의 차별이 지속될 가능성이 있기 때문이다.

(proxy variable)에 의한 차별'이라고 부른다.[204] 예를 들어 앞선 아마존 이력서 검토 인공지능 사례에서 성별이나 인종 등 차별금지사유를 명시적으로 나타내는 단어를 모두 '마스킹(masking)' 처리하여 입력변수로 고려되지 않도록 한 상황을 상정해 볼 수 있다. 그럼에도 불구하고 이력서에 포함된 다른 다수의 단어를 이용하면 인공지능은 해당 지원자의 성별이나 인종 등 차별금지사유를 쉽게 추론할 수 있을 것이다.[205] 그런데 해당 인공지능이 보호집단에 대해 불리한 대우를 하는 것으로 밝혀진 경우 인공지능을 활용하는 것이 직접차별에 해당한다고 평가할 수 있을 것인가?

우선 미국 민권법과 같이 직접차별에 해당하기 위해서 '차별 의도'를 입증해야 하는 법제 하에서는 인공지능이 대용변수에 의한 차별을 하는 경우라도 이를 의도적 차별행위로 의율하는 것은 쉽지 않을 것이다. 이러한 사례에 대한 법적 평가는 해당 인공지능의 개발자나 이용자가 (1) 인공지능 도입 전 차별적 작동 여부에 관한 감사를 수행한 경우와 (2) 차별적 작동 여부에 관한 감사를 수행하지 않고 만연히 이를 도입한 경우로 구분하여 고려할 수 있다.[206] (1) 만약 해당 인공지능에 대해 감사를 수행한 결과 차별금지법제를 위반하지 않는다는 평가가 내려졌고, 그 이용자가 이를 신뢰하고 인공지능을 도입하기에 이르렀다면 해당 인공지능의 활용이 의도적 차별행위에 해당한다고 평가되기는 어려울 것이다.[207]

---

204 대용변수에 의한 차별 문제에 관해서는 본 연구 제4장 제2절에서 보다 상세히 다룬다.

205 실제로 Talia B. Gillis and Jann L. Spiess, "Big Data and Discrimination", *University of Chicago Law Review* 86(2), (2019)는 미국 부동산 담보 대출에 관한 데이터를 통해 다른 변수들을 통해 인종이나 혼인 상태와 같은 차별금지사유를 높은 정확도로 예측할 수 있음을 보였다.

206 실무상으로 시장에서 널리 활용되고 있는 고용 관련 인공지능에 대해서는 내부 감사 또는 외부 연구자에 의한 독립적 감사 활동이 이루어지고 있으므로, 실무상 후자와 같은 상황이 자주 발생하지는 않을 것으로 전망된다. 예컨대, 채용 소프트웨어에 대한 외부 감사 사례로는 Raghavan et al.(2020) (각주 149) 참조.

207 다만, 인공지능 감사 결과 해당 인공지능이 보호집단에 대해 차별적으로 작동한다는 점을 알게 되었고, 그 이후에도 인공지능을 지속적으로 사용하는 경우 차별의도가 있다고 인정될 가능성은 있을 것이다.

특히 인공지능을 도입하는 취지가 오히려 이제까지 인간에 의해 의식적 또는 무의식적으로 이루어져 왔던 차별을 막기 위한 목적인 경우도 적지 않다는 점을 고려한다면, 인공지능을 사용하는 행위에 있어 차별 의도를 추단할 수 있는 경우도 상정하기 어렵다.[208] 여기서 해당 인공지능 감사가 적정하게 이루어진 것인지에 관한 법적 판단의 문제는 남게 될 것인 바, 이 문제는 결국 인공지능 감사에 있어 어떠한 공정성 지표를 사용하는 것이 타당한지에 관한 것인지로 귀결될 것이다. 이 문제에 관해서는 본 연구 제5장에서 구체적으로 검토한다.

다음으로, (2) 인공지능의 이용자가 인공지능에 대한 감사를 수행하지 않고 이를 만연히 도입한 경우라면, 그러한 부작위를 의도적 차별행위로 볼 수 있을 것인지가 쟁점이 될 수 있다. 그러나 본 장 제4절에서 검토하는 바와 같이, 대부분의 국가에서 아직까지 인공지능에 대한 사전 감사를 의무화하는 규제가 법제화되지 않은 상태이므로 이러한 부작위로부터 차별 의도를 추인하고, 이를 위법한 것으로 평가하는 것은 쉽지 않을 것이다.[209] 따라서 이와 같은 사안은 의도적 차별행위로 규율되기 보다는 해당 인공지능이 보호집단에 대해 불리한 대우를 하는 것이 결과적 차별행위에 해당하는지 여부에 따라 판단되어야 할 것이다.[210]

다른 한편, 유럽이나 우리나라와 같이 직접차별에서 차별 의도의 입증을 요구하지 않는 법제에서도 대용 변수에 의한 차별 문제는 결국 간접차별의 문제로 귀결될 가능성이 크다. 그 이유는 비록 해당 인공지능이 차별금지사유를 가진 개인에게 불리한 대우를 하는 것으로 밝혀지더라도 인과성이나 합리적 사유의 부존재를 입증하기 쉽지 않을 것이기 때문이다. 즉, 인공지능은 목표 변수(예컨대

---

208 고학수 외(2019) (각주 176), 231면 이하는 사용자가 차별의도를 은폐하기 위한 목적으로 인공지능을 도입하는 상황을 상정하여 검토하고 있는데, 실무상으로 이러한 목적에 관한 입증은 극히 어려울 것이다.

209 만약 이후 인공지능에 대한 사전 감사 의무가 법제화된다면, 위 행위에 대해서는 결국 해당 규제 위반에 따른 책임을 물을 수 있게 될 것이다.

210 최승철(2011) (각주 178)은 간접차별의 법리란 종국적으로 의사결정자가 채택하는 외견상 중립적 기준이 차별적 효과를 낳는지 검토해야 할 의무를 부과하는 것으로서, 간접차별에 대한 책임은 본질적으로 의사결정자의 부작위에 대한 책임을 묻는 것이라고 해석한다.

채용 인공지능의 경우 업무 수행 능력)를 최대한 정확하게 예측할 수 있도록 학습되므로, 해당 인공지능이 차별금지사유를 가진 개인에 대해 불리한 대우를 하였다면 이는 해당 차별금지사유가 목표변수와 음(-)의 상관관계에 있는 것으로 평가되었기 때문일 수 있다. 이러한 경우라면 위 인공지능은 차별금지사유를 "이유로" 불리한 대우를 한 것이 아니라, 인공지능이 목표 변수를 최대한 정확하게 예측한 결과 (달리 표현하면, 인공지능에 의한 평가라는 외견상 중립적 기준을 활용한 결과) 그 개인에게 불리한 대우를 한 것으로 평가될 수 있다. 그 결과 차별금지사유와 불리한 대우 간의 인과성 입증에 실패할 가능성이 크다. 또한 인공지능 활용에 의한 효율성 개선 효과, 인간 의사결정자에 의한 임의적 개입 차단 효과 등을 고려할 때 인공지능의 활용에 있어 "합리적 이유"가 있다고도 인정될 가능성이 있다.

요컨대, 인공지능에 차별금지사유가 명시적으로 입력되지 않고, 다른 외견상 중립적인 속성을 통해 차별적 속성을 추론할 수 있는 경우, 즉 대용 변수에 의한 차별의 경우에는 직접차별이나 의도적 차별행위로 규율되기보다는 간접차별이나 결과적 차별행위가 문제될 것으로 보인다.[211]

2. 인공지능에 의한 간접차별

간접차별이 있었는지 여부에 대한 판단은 일반적으로 어떤 의사결정자의 행위에 있어 그 의도나 동기를 직접적으로 요건으로 하지 않고, 의사결정자의 정책이나 관행이 차별적 결과를 발생시키는지에 초점을 둔다.[212] 이에 간접차별에 해당하기 위해서는 차별금지사유를 가진 보호집단에 대해 현저히 불리한 결과가 발생하였고, 그러한 결과가 직무와 관련되어 있거나 업무상 불가피한 것이 아니

---

211 Barocas and Selbst (2016) (각주 31)도 빅데이터에 의한 차별의 경우 결과적 차별행위가 주로 문제될 것이라 분석한다.

212 박은정 외(2021) (각주 179), 59-61면.

어야 한다. 다만, 직무 관련성이나 업무상 불가피성이 인정되더라도 동일한 목적을 달성하기 위한 덜 차별적인 대안이 존재하는 경우에는 간접차별에 해당하게 된다.[213]

가. 현저히 불리한 결과

우선 어느 정도 불균등한 결과가 발생하여야 간접차별 상의 '현저히 불리한 결과'를 인정할 수 있는 지가 쟁점이 된다.[214] 통상 어떤 의사결정자의 판단 결과가 집단의 비율에 있어 상당한 차이가 있는 경우로서, 이러한 차이가 우연히 발생할 수 있는 수준을 넘어 '통계적으로 유의한(statistically significant)' 차이가 있는 경우 현저히 불리한 결과로 볼 수 있다. 다만 반드시 통계적 입증을 사용해야 하는 것은 아니고 일반화된 경험적 지식에 의하여도 인정될 수 있다.[215]

예컨대, 미국 고용평등기회위원회(U.S. Equal Employment Opportunity Commission, 이하 "EEOC")는 채용에 있어 '현저히 불리한 결과'가 있는지 판단하는 데 있어 이른바 '4/5 규칙(The Four-Fifths Rule)' 규칙을 제시하고 있다. 이는 미국 고용평등기회위원회의 1978년 "근로자 선발 절차에 대한 통일 가이드라인"에 규정된 것이다.[216] 해당 규칙은 일응 어떤 보호집단에 대한 합격률이 다른 집단과 비교하여 4/5 미만이 되어서는 안 될 것을 요구한다. 가령 남성 지원자의 합격률이 25% 였다면, 여성 지원자의 합격률은 그 4/5인 20% 이상이어야 한다. 다만, 이는 미국 민권법 제7편에 대한 고용평등기회위원회의 해석일 뿐이며, 법적 구속력이 있는 것이 아니다. 미국 판결도 이 기준에 존중하는 사례와

---

213 심재진, "미국의 차별적 영향 이론과 유럽연합의 간접차별 법리", *노동법 연구* 27, (2009).

214 최승철(2011) (각주 178).

215 Chief Constable of Avon & Somerset Constabulary v Chew EAT/503/00, 최승철 (2011) (각주 178)에서 재인용.

216 Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §1607.4(D).

그렇지 않은 사례가 모두 존재한다고 한다.[217] 하지만 실무상으로 직원채용 인공지능 시스템을 구축하는 경우 이 기준을 대체로 준수하고자 시도하는 것으로 보인다.[218]

이처럼 집단 간에 특정 결과에 해당할 확률에 있어 현저한 차이가 있는지 여부를 기준으로 심사하는 인공지능 공정성 지표를 '결과 동등성' 또는 '인구통계적 공정성'이라 하는데, 이에 대해서는 본 연구 제4장 제3절에서 보다 상세히 살핀다.

나. 직무 관련성 및 업무상 불가피성

만약 어떤 의사결정자가 적용하는 정책이나 관행이 보호집단에 현저히 불리한 결과를 초래하더라도, 그는 해당 정책이나 관행이 직무와 관련되어 있고 업무상 불가피할 경우 간접차별에 해당하지 않게 된다. 가령 현재 어떠한 사회에서 개인들이 능력이 집단 간 현저히 불균등하게 분포되어 있는데, 의사결정자가 직무상 필요에 따라 결정을 내림으로써 특정 집단에 현저히 불리한 결과가 초래되었다고 하면, 이는 사회의 불평등을 반영한 것일 뿐이지 그 의사결정자가 차별적 행동을 하였다고 평가하기 어렵다. 그와는 반대로 의사결정자가 직무와 무관한 조건을 부가하여 보호집단에 현저히 불리한 결과가 초래되었다면 이는 간접차별에 해당할 수 있다. 예를 들어 미국에서 결과적 차별행위의 법리를 수립한 Griggs 사건에서는 고용주가 직무와 무관하게 고등학교 졸업증과 IQ 검사결과를 요구하였고 그 결과 흑인에게 현저히 불리한 결과가 초래된 것이 결과적 차별행위에 해당한다고 보았다.[219]

---

217 M.O.C.H.A. Soc'y, Inc. v. City of Buffalo, 689 F.3d 263 (2d Cir. 2012) (4/5 규칙을 존중한 사례); Stagi v. National R.R. Passenger Corp., 391 Fed. Appx. 133, 2010 WL 3273173 (3d Cir. 2010) (4/5 규칙을 비판한 사례)

218 Raghavan (2020) (각주 149).

219 Griggs v. Duke Power Co., 401 U.S. 424 (1971).

나아가, 고용 차별 사안에 있어, 만약 고용주가 "전문적으로 유효성이 검증된 평가(professionally validated tests)"를 사용하였다면 그 평가가 차별금지사유에 대해 차별하도록 고의적으로 의도된 것이 아닌 이상, 직무 관련성 및 업무상 불가피성 항변이 받아들여 질 수 있다. 이는 결국 어떠한 평가의 유효성(validity)을 어떻게 검증할 것인가의 문제로 귀결되는데, EEOC는 이러한 검증 연구를 위한 기준을 제시하고 있다.[220] 이러한 검증 연구는 해당 평가가 직무 성과와 통계적으로 유의미하게 상관관계가 있다는 것을 보여주는 것이어야 하는데, 이는 그 평가 자체만이 아니라 고용주가 적용하는 점수의 하한(cut-off)에 대한 검증도 필요하다.[221] 이와 유사한 기준이 인공지능에 의한 예측에 대해서도 적용될 수 있다. 즉, 인공지능에 의해 차별보호사유를 가진 집단에 현저히 불리한 결과를 초래된다면, 이를 사용하는 의사결정자는 인공지능의 예측이 평가 대상자의 능력이나 위험과 통계적으로 유의미하게 상관관계가 있다는 점을 정당화할 필요가 있다

하지만 이러한 직무관련성과 업무상 불가피성에 따른 항변의 판단에 있어서는 일의적으로 그 예측의 정확성 수준만을 고려해서는 안 되고, 인공지능이 사용되는 맥락이 반드시 고려되어야 한다. 그 이유는 의사결정에 사용되는 능력이나 위험성이란 결국 제도의존적으로 맥락에 따라 결정되는 것이기 때문이다.[222] Scanlon (2018)은 다음과 같은 사례를 든다. 과거에는 어떤 공장 창고에서 근무할 사람을 선발할 때 근력이 좋은 사람을 선발하는 것이 쉽게 정당화될 수 있었지만, 현대의 기계화된 근무환경 하에서는 근력을 요구하는 것은 더 이상 정당화

---

220 Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §1607.5.-10. 위 가이드라인에는 기준-관련 유효성(criterion-related validity), 내용 유효성(content validity), 구성물 유효성(construct validity)로 구분하고 각각의 연구에 있어 필요한 요건을 정하고 있다.

221 Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §1607.5(H).

222 Thomas Scanlon, *Why does inequality matter?*, Oxford University Press, (2018). 이에 관한 헌법적 함의에 관해서는 이민열, "토머스 스캔론의 「불평등은 왜 문제인가?」와 헌법의 해석", *현상과 인식* 45(2), (2021) 참조.

될 수 없다. 마찬가지로 대학 입시에서 프로그래밍 능력과 같이 제한적인 교육 환경에서만 취득할 수 있는 능력을 요구하는 것도 정당화되기 어렵다. 그러한 능력은 대학 입학 후 학생들에게 가르칠 수 있는 능력이기 때문이다. 위 사례들과 마찬가지로 인공지능의 활용에 있어서도, 그 인공지능이 특정 목표변수를 얼마나 잘 예측하는 지만이 아니라, 해당 목표변수가 자체가 어떠한 이유에서 고려되었는지, 해당 목표변수가 직무와 얼마나 상관관계가 있는지 등과 같이 해당 목표변수 선택의 정당화 가능성을 고려해야 한다.

3. 직접차별 기준과 간접차별 기준의 충돌

원칙적으로 직접차별 기준과 간접차별 기준은 인공지능에 의한 차별을 막기 위한 보완적 역할을 수행할 수 있다. 즉, 간접차별의 인정은 차별금지법리를 실질적으로 확장시키는 것으로서, 직접차별에 대한 입증이 곤란한 상황에서 간접차별의 법리를 통해 보완할 수 있을 것으로 기대된다.[223] 그러나 인공지능에 의해 초래될 수 있는 간접차별을 막기 위한 노력이 오히려 직접차별에 해당할 수 있다는 문제도 발생할 수 있다. 미국 연방대법원의 Ricci v. DeStefano 판결은 이러한 문제를 잘 보여주는 사건으로 인용된다.[224] Ricci 판결이 인공지능 차별금지 기준의 적용에 있어 어떠한 함의를 갖는지에 관한 일련의 논쟁이 진행되었는바, 이하에서는 이를 개관한다.

가. Ricci 사건의 개요

Ricci 사건의 사실관계는 다음과 같다. 미국 코네티컷 주 뉴헤이븐 시 소방서는 2003년 소방위(captain) 7명 및 소방장(lieutenant) 8명의 승진 공석에 대

---

223 박은정 외(2021) (각주 179).

224 Ricci v. DeStefano, 557 U.S. 557 (2009).

한 시험을 진행했다. 해당 승진 시험은 필기시험 60%, 구두시험 40%로 구성되어 있었고, 합계 점수가 70% 이상이어야 시험에 통과할 수 있었는데, 최종 후보자는 시험 점수 순위에 따라 결정될 것이었다. 해당 시험은 I/O Solutions라는 업체에 외주 용역을 주어 설계되었고, 2003년 11월 및 12월 진행되었다. 위 승진시험에는 총 118명의 소방관이 응시하였다.

그런데, 해당 시험의 합격률은 인종간 상당한 격차를 보였다. 소방위 시험의 경우, (i) 백인 지원자 25명 중 16명(64%)가 통과하였으나, (ii) 흑인 지원자 8명 중에서는 3명(38%)만이, 히스패닉 지원자 8명 중 3명(38%)만이 통과하였다. 최종적으로 선발된 9명의 소방위 승진 후보자 중 7명은 백인, 2명은 히스패닉이었고, 흑인은 누구도 이에 포함되지 못하였다. 한편, 소방장 시험의 경우, (i) 백인 지원자 43명 중 25명(58%)가 통과하였으나, (ii) 흑인 지원자 19명 중에서는 6명(32%)만이, 히스패닉 지원자 15명 중 3명(20%)만이 통과하였다. 또한 최종적으로 선발된 10명의 소방장 승진 후보자는 전원이 백인이었다.

시험 결과가 이와 같이 나오게 되자, 시 당국은 만약 시험결과를 그대로 활용하는 것은 결과적 차별행위에 해당할 가능성이 있다고 판단하였다. 더욱이 시험에 불합격한 흑인이나 히스패닉 지원자들이 소송을 제기할 우려가 있었다. 이에 시 당국은 결과적 차별행위를 피하기 위한 목적으로 해당 시험 결과를 폐기하기로 결정하였다.

이에 Ricci 외 18명의 백인 지원자와 1명의 히스패닉 지원자가 시 정부를 상대로 의도적 차별행위 소송을 제기하였다. 원고들의 주장은 시 정부가 이미 치러진 시험결과를 폐기한 것은 인종을 승진에 있어 고려 요소로 삼았기 때문이었고, 이는 민권법 제7편이 금지하는 의도적 차별행위에 해당한다는 것이었다. 이에 대해 시 정부는 결과적 차별행위를 피한다는 목적은 '차별적 의도'에 해당하지 않고, 승진시험이 차별적이거나, 그에 대한 소송이 제기될 우려가 있는 경우

에는 이를 시정하기 위하여 재설계할 권한이 있다고 주장하였다. 연방지방법원과 제2연방항소법원은 시 정부의 주장을 받아들여 원고들의 청구를 기각하였다.[225]

그러나 미국 연방대법원은 원심과 달리 원고들의 주장을 받아들이고, 시 정부의 행위가 민권법 제7편에서 금지된 의도적 차별행위에 해당한다고 판단하였다.[226] 케네디 대법관의 법정의견에 따르면, "고용주가 비의도적인 결과적 차별행위를 회피하거나 시정하기 위한 목적으로 의도적으로 차별적 행위를 하기에 앞서, 만약 인종을 고려한 차별적 행동을 취하지 않는다면 결과적 차별행위에 따른 책임을 지게 될 것으로 믿는 데 대한 **강한 증거상의 근거**가 있어야 한다"(역자 강조).

그런데, 미국 연방대법원의 법정의견은 이 사건의 경우 해당 승진시험이 결과적 차별행위에 해당한다는 강한 증거상의 근거가 존재하지 않는다고 보았다. 즉, 해당 승진시험 결과에 따르면 흑인 선발자가 전혀 없다는 사실은 결과적 차별행위에 대한 일응의 추정에 해당할 뿐이고, 결과적 차별행위에 해당하려면 직무 관련성이나 업무상 불가피성이 없어야 하고, 덜 차별적인 대안이 존재하지 않는 상황이어야 한다. 하지만 해당 승진시험을 설계한 I/O Solutions는 소방관 및 경찰관의 채용 및 승진 시험 설계에 특화된 업체로서, 해당 시험을 개발하기 위하여 전문적인 설계 기준과 방법론을 적용하였고, 시험 개발과정에서 보호집단 구성원에 대해 부정적인 영향을 끼치지 않도록 충분한 주의를 기울인 것으로 판단되었다. 즉, 연방대법원은 시 정부에 대하여 결과적 차별행위 소송이 제기되더라도 시 정부가 직무관련성 또는 업무상 불가피성을 충분히 입증할 수 있을 것으로 본 것이다. 따라서 해당 승진시험결과에 따라 승진 결정을 내리게 되면 결과적 차별행위에 해당할 것이라는 시 정부의 주장은 "강한 증거상의 근거"가 없

---

225 554 F.Supp.2d 142 (US District Court for Connecticut; September 28, 2006); 530 F.3d 87 (Second Circuit; June 9, 2008; per curiam).

226 557 U.S. 557 (2009).

는 것이었고, 오히려 시험 결과 흑인 승진대상자가 없다는 이유로 그 결과를 폐지한 것이 직접적 차별행위에 해당하게 되었다.

하지만 이 판결에 대해서는 4인의 대법관이 반대의견을 표시하였다. 긴즈버그 대법관이 작성한 반대의견에 따르면, 법정의견은 "시 정부가 해당 시험 결과를 거절한 이유가 오로지 고득점자가 모두 백인이었기 때문"이라고 판단한 것은 사실을 오해한 것이고, 그와 달리 시 정부가 사용한 승진시험에 다수의 흠결이 존재하였다고 볼 상당한 근거가 있었다고 지적하였다. 오히려 다른 시 정부들은 그보다 더 개선된 시험을 이용하여 인종적으로 덜 편향된 결과를 낳은 사례들이 있었다. 더욱이 뉴헤이븐 시 거주자 중 60%에 달하는 수가 흑인 또는 히스패닉인데, 법정의견에 따르면 시민들에 봉사해야 하는 소방서의 지휘관 직위가 거의 대부분 백인으로만 구성되는 상황을 시민들이 감수해야만 할 것이었다. 나아가 긴즈버그 대법관은 이러한 판결은 민권법 제7편의 효과적 집행을 위해서는 결과적 차별행위 규제가 중심적 역할을 해야 한다는 Griggs 판결의 취지를 이해하지 못한 것이라고 비판하였다.

나. Ricci 판결의 해석에 관한 논쟁

1) Ricci 판결의 적용범위를 넓게 해석하는 견해

인공지능 공정성 맥락에서 Ricci 판결에 대한 논쟁은 Barocas & Selbst (2016)에서 비롯되었다.[227] 위 연구자들은 Ricci 판결을 의도적 차별행위 규제와 결과적 차별행위 규제 간의 충돌을 보여주는 대표적인 사례로 언급하면서, 이 판결이 인공지능 공정성 달성에 있어 중대한 법적 걸림돌이 될 것이라고 주장하였다. 즉, Ricci 판결에 따르면 차별적 결과를 초래하는 인공지능을 사후적으로 시정하고자 하는 행위가 오히려 법적으로 의도적 차별행위라는 판단을 받을 위험

---

227 Barocas and Selbst (2016) (각주 31)

성이 있다는 것이다. 예컨대, 어떤 사업자가 사용 중인 인공지능에 대해 사후적으로 감사를 실시한 결과 인종간 혹은 성별간 차별적 결과가 존재하는 것을 발견하고 이로써 결과적 차별행위가 일응 추정된다고 판단하더라도, 이를 시정하기 위해 알고리즘을 수정하거나 알고리즘의 결과와는 다른 판단을 내리는 경우, (적어도 위 연구자들의 견해에 따르면) Ricci 판결에 의해 오히려 의도적 차별행위에 해당할 수 있게 된다.

Kroll et al. (2017) 연구도 Barocas & Selbst (2016)과 같은 태도를 취하였다.[228] 이들은 (1) 알고리즘의 설계 단계에서 차별 시정 노력과 (2) 알고리즘 도입 이후 발견된 사후적 차별에 대한 시정 노력을 구분한다. 위 연구자들의 견해에 따르면, (1) Ricci 판결에 의하더라도 인공지능 개발 시점에 있어 이를 비차별적으로 만들기 위한 노력을 기울이는 것은 문제가 없다. 즉, Ricci 사건에서 연방대법원은 승진시험을 마련한 외주업체가 그 설계 단계에서 소수 인종에 대해 차별적이지 않도록 노력을 기울인 것 자체는 문제삼지 않았다. 연방대법원 법정의견도 "민권법 제7편은 채용 시험이나 관행을 실시하기에 앞서, 해당 시험이나 관행이 모든 개인들에게 인종과 무관하게 공정한 기회를 제공하려면 어떻게 설계해야 할 것인지 고려하는 것을 금지하지 않는다"고 설시하였다. 그러나 위 연구자들은 (2) 인공지능이 일단 도입되고 나면, 이를 사후적으로 수정하기 위해 인종을 고려하는 행위는 의도적 차별행위에 해당할 위험성이 있다고 지적한다. 연방대법원 법정의견은 "채용 과정이 일단 확립되고, 고용주가 채용 기준을 명확히 한 후에는 시험 결과를 무효화할 수 없다. 이는 피용자들이 인종을 근거로 판단되지 않을 것이라는 정당한 기대(legitimate expectation)를 뒤집는 것이다." 라고 판시하였다. 이러한 이해를 바탕으로 Kroll et al.(2017)은 "의사결정자들은 알고리즘의 초기 설계 단계에서 차별적으로 작동하지 않도록 고려하는 것이 안

---

228 Joshua A Kroll, et al., "Accountable algorithms", *University of Pennsylvania Law Review* 165, (2017).

전한 선택이고, 알고리즘 연구자들은 그러한 설계를 보조할 수 있는 기술적 도구의 개발과 적용을 장려해야 한다."고 주장하였다.

2) Ricci 판결의 적용범위를 좁게 해석하는 견해

그러나, Barocas & Selbst (2016) 및 Kroll et al. (2017)에 의해 제기된 주장에 대해, 미국 노동법 연구자인 Pauline Kim은 위와 같은 견해가 Ricci 사건의 연방대법원 판결을 오독한 것이라고 반박하였다.[229] 그녀의 주장에 따르면, 미국 법상 인공지능이 차별적 작동하는지 감사(audit)하고, 장래에 있어 차별적 편향을 시정하는 것이 허용된다. Kim (2017)의 주장 요지는 다음과 같다. Ricci 사건은 승진시험을 이미 실시하여 특정한 개인들이 승진할 것이라는 정당한 기대를 갖고 있는 상황이었고, 따라서 위 판결은 이미 정당한 기대를 갖게 된 특정된 개인들 대상으로 하여 인종에 근거한 불리한 행위를 허용할 것인지 여부에 대해 한정하여 다루고 있을 뿐이다. 그러나 인공지능의 편향을 제거하기 위해 장래에 있어(prospectively) 알고리즘을 변경하는 것은 전혀 다른 문제이다. 미국 차별금지법제는 고용주들이 차별금지 목적을 준수하기 위해 자발적으로 노력하는 것을 허용하고 있고, Ricci 판결은 이러한 법제와 전혀 충돌하지 않는다.

Kim (2017)은 다음과 같은 가정적 상황을 고려해 보자고 한다. 만약 어떤 고용주가 콜센터 직원과 같이 상대적으로 저숙련 상태인 근로자들을 다수 고용하고 있는데, 매월 수천 명이 지원한다고 하자. 위 고용주는 수많은 지원서를 모두 인간이 검토하기 어려우므로, 자동화 도구를 이용해서 지원서 및 온라인 설문응답, 기타 데이터 브로커들로부터 획득한 데이터를 이용하여 지원자들을 선별하고, 상위 10%의 지원자들을 인간 심사자가 검토하는 시스템을 도입하였다. 위 고용주가 위 시스템을 수 개월 동안 사용한 다음, 위 시스템에 있어 차별적 영향

---

229 Pauline Kim, "Auditing algorithms for discrimination", *University of Pennsylvania Law Review* 166 (2017).

이 존재하는지 사후적으로 감사를 하였다. 그 결과 특정 소수인종 지원자에 대해서 불균형적으로 차별적인 결정이 내려졌다는 사실을 알게 되었다. 이 경우 고용주는 해당 시스템에서 사용되는 데이터 입력 중 편향을 초래하는 변수를 제거하는 등과 같이 위 시스템의 편향을 제거하기 위한 조치를 취할 수 있는가?

Pauline Kim은 Ricci 판결의 어떠한 내용도 고용주의 이러한 노력을 금지하는 것은 아니라고 주장한다. 위의 가상적인 사례와 Ricci 사건의 사실관계는 크게 다르다. Ricci 사건에서 뉴헤이븐 시는 상당한 과정을 거쳐 승진시험을 개발했고, 개발된 시험의 형식, 시험 과목 및 수험 자료를 공고하였으며, 지원자들을 위해 3개월 동안의 준비시간을 지정했다. 승진시험 지원자들은 시험을 준비하기 위해 상당한 노력을 투입했다. 특히 대표원고(lead plaintiff)였던 Frank Ricci는 난독증이 있었기 때문에 수험자료를 읽어서 녹음해 줄 사람을 고용하기까지 하였다. 또다른 원고인 Benjamin Vargas는 시험 준비를 위해 부업을 포기했고, 그의 아내는 시험 준비를 돕기 위해 휴가를 사용하였다. 이처럼 Ricci 사건에서는 시 정부가 승진시험의 결과를 폐기하기로 한 결과, 그 시험에 합격한 원고들의 정당한 기대가 중대하게 훼손되었다. 이러한 상황에서는, 비록 미국 민권법상 시 정부가 승진 절차에 결과적 차별행위를 하지 않도록 주의할 의무가 있다고 하더라도, 이는 원고들의 정당한 기대를 훼손하지 않는 범위 내에서 이루어져야 한다. Ricci 판결은 시 정부가 이러한 경우 원고들의 정당한 기대를 훼손할 수 있으려면 결과적 차별행위에 따른 책임을 지게 될 것이라고 믿는 데 대한 "강한 증거상의 근거"가 있어야 할 것을 요구한 것이다.

그러나 대부분의 직장에서는 이러한 절차가 요구되지 않는다. 앞선 사례의 가상의 회사에 입사 지원한 이들은 그 회사의 채용 정책이 앞으로 결코 바뀌지 않을 것이라는 정당한 기대를 가질 수 없다.[230] 어떤 회사가 수 개월동안 사용한 알고리즘을 소급하여 변경시키고 이미 채용한 직원들을 해고하는 것이 아니라, 덜 편향된 알고리즘으로 이를 개선하고, 장래에 대해 새로운 알고리즘을 적용한

---

230 위의 글, p. 199.

다면, 이는 장래의 입사 지원자들의 정당한 기대를 훼손한다고 평가할 수 없는 것이다.

Kim (2017)은 Ricci 판결 이후 내려진 다수의 미국 연방 법원 하급심 판례 역시 고용주들이 자동화된 결정에 대해 감사하고, 확인된 편향을 제거하기 위해 장래에 대해 이를 수정하는 것을 허용하였다고 설명한다. 예를 들어 제2연방항소법원은 Maraschiello v. City of Buffalo Police Department 사건에서 시 정부가 (부분적으로는 결과적 차별행위를 막기 위한 동기에서) 새로운 경찰관 승진시험을 도입한 것이 Ricci 판결의 기준에 비추어 보더라도 의도적 차별행위에 해당하지 않는다고 판단하였다.[231] 이 사건에서 제2연방항소법원은 "시 정부가 새로운 시험을 도입하기로 한 목적이 부분적으로는 더 인종적으로 균형된 결과를 달성하기 위한 것이었다고 하더라도", 새로운 시험을 도입한 것이 "인종에 기반한 불리한 행위"에 해당하지 않는다고 판단하였다. Carroll v. City of Mount Vernon 사건에서도 제2연방항소법원은 시 정부가 기존 소방관 승진시험의 적법성을 검토하는 동안 소방관의 승진을 연기하고, 새로운 시험을 도입한 것이 적법하다고 판단하였다. 이 사건에서 법원은 "고용주들은 민권법 제7편을 위반할 염려 없이, '자기-평가(self-evaluate)'를 수행하고, 이를 '엄밀히 검토(hard-look)' 하는 것이 허용되어야 한다"고 판시하였다.[232]

---

231 Maraschiello v. City of Buffalo Police Department 709 F.3d 87, 89 (2d Cir. 2013), 위의 글 p. 200에서 재인용. 이 사건의 개요는 다음과 같다. 이 사건의 원고인 Maraschiello는 경찰 승진시험에서 가장 높은 점수를 얻었으나, 공석이 없어 간부 승진을 하지 못하였다. 이후 종전 승진시험 점수의 유효기간이 만료되었다. 이후 시 정부는 새로운 시험을 도입하였고, 그 이후 공석이 발생하였다. 시 정부는 만료된 승진시험 점수를 인정하기 않고, 새로운 시험 결과에 따라 승진 후보자를 선정하였다. 시 정부가 종전의 승진시험을 폐기한 이유는 유사한 시험에 대해 민권법상 차별 소송이 제기되었기 때문이기도 하고, 다른 한편으로는 해당 시험이 30년 전의 직무 분석에 근거하였고, 현재의 경찰관 직무에는 적합하지 않기 때문이었다.

232 Carroll v. City of Mount Vernon, 453 F. App'x 99 (2d. Cir. 2011), 위의 글 p. 200에서 재인용.

또한, Deborah Hellman도 알고리즘 공정성에 대한 연구에서 Pauline Kim의 해석이 설득력이 있고 타당한 것이라고 동의한다.[233] Hellman (2020)에 따르면, 이 논쟁의 주된 쟁점은 어떤 인공지능이 결과적 차별행위를 하고 있다고 생각될 경우 이에 대응하여 인공지능을 수정하는 것이 허용되는지 여부이다. 그런데 미국법상 인공지능을 개선하는 데 있어 단지 인종적 요소를 고려하는 것 자체가 금지되는 것은 아니다. 예컨대 새로운 학교의 부지를 선정하는 데 있어 인종적 다양성을 고려하는 것은 허용된다.[234] 따라서 Ricci 판결이 차별적 인공지능을 수정하기 위해 인종적 요소를 고려하는 것을 금지하는 것으로 해석되어서는 안 된다고 주장한다.

다. Ricci 사건의 시사점

이처럼 여러 법학자들이 Ricci 판결과 인공지능 공정성 문제에 관한 의견을 제시해 왔으나, 컴퓨터 공학 연구자들 사이에서 Ricci 판결이 잘못 이해되는 경향이 발견된다. 예를 들어 2019년 한 공학 연구는 결과적 차별행위를 막기 위해 "[인공지능의] 결정 결과가 인구비례가 되도록 요구하는 것"은 "적격인 후보자들에 대한 역차별을 초래할 위험이 있다. 이러한 관행은 과거 불법으로 판단된 바 있다"고 설명하면서, 그 근거로 Ricci 판결을 인용한다.[235] 이처럼 컴퓨터 공학 연구자들은 Ricci 판결을 인공지능 공정성 지표 중 결과 동등성 요건을 도입하는 것을 금지하는 취지로 이해하는 것으로 보인다.[236] 그러나 Pauline Kim과 Deborah Hellman이 설득력 있게 지적한 바와 같이 Ricci 판결은 보호집단 구

---

233 Hellman (2020) (각주 3).

234 Hellman (2020) (각주 3), p. 864은 Parents Involved 사건에서 이러한 정책을 명시적으로 허용하고 있다고 지적한다.

235 Muhammad Bilal Zafar, et al., "Fairness constraints: A flexible approach for fair classification", *The Journal of Machine Learning Research* 20.1 (2019).

236 인구통계적 동등성 기준에 대해서는 본 연구 제4장 제3절 참조.

성원에 대한 불균등한 결과를 발생시키는 인공지능의 편향을 제거하기 위해 개선 조치를 취하는 것을 금지하는 취지로 이해되어서는 안 된다.

인공지능 공정성 연구자들 사이에서 이와 같은 오해가 발생하게 된 이유를 추측해 본다면, Barocas & Selbst (2016) 및 Kroll et al. (2017)의 주장이 크게 주목을 받았기 때문일 수도 있다. 위 연구는 인공지능 공정성이 공학 연구자들 사이에서 주된 관심사가 되기에 앞서 비교적 초기에 이 문제를 심도있게 제기하였고, 그 결과 관련된 컴퓨터 공학자들에 상당히 널리 인용되고 있다.[237] 이에 비해 위 연구에서 제기된 법적 쟁점을 반박한 Kim (2017)이나 Hellmann (2020)은 법학계를 넘어 광범위한 독자에까지 미치지는 못한 것으로 보인다.[238]

이상과 같이 Ricci 판결에 의해 초래된 혼란에 대한 대응으로, 2022년 ACM FAccT 학술대회에서는 "Ricci v. DeStefano 사건 때문에 물러서지 마라: 채용 패러독스에 대한 편향-인식적 법적 해결책"이라는 제목의 연구가 발표되기도 하였다.[239] 이 연구는 직장에서의 성별 또는 인종 불평등 문제가 지속되는 경우, 이에 대응하지 못한 고용주는 법적 제재를 받을 위험이 있지만, 반대로 이를 해결하기 위한 조치를 취하는 경우에도 마찬가지로 법적 갈등 상황에 처해질 수 있는 문제를 '채용 패러독스(hiring paradox)'라고 명명한다. 그런데, Ricci 판결 때문에 채용 패러독스를 해결하는 것이 불가능하다는 사고가 퍼져 있고, 그러한 주장 사례로 Barocas & Selbst (2016)을 소개한다. 이에 대해 저자들은 Kim (2017)을 근거로 그러한 이해가 잘못된 것이라고 밝히고, 기존의 Ricci 판결에

---

237 본 연구 작성일 현재, 구글 학술검색 기준 인용 횟수가 Barocas and Selbst (2016) (각주 31)은 3,560회, Kroll et al. (2017) (각주 228)은 1,238회 이른다. 이는 Law Review 학술지에 게재된 논문으로서는 이례적으로 높은 인용 횟수이다. 이 논문은 법학 이외의 다양한 분야에서 인용되고 있다.

238 본 연구 작성일 현재, 구글 학술검색 기준 Kim (2017) (각주 229)의 인용 횟수는 80회, Hellman (2020) (각주 3)의 인용 횟수는 169회이다.

239 Jad Salem, Deven R. Desai, and Swati Gupta, "Don't let Ricci v. DeStefano Hold You Back: A Bias-Aware Legal Solution to the Hiring Paradox", *Proceedings of the 2022 conference on fairness, accountability, and transparency*, (2022).

대한 오해를 극복하고 알고리즘 편향을 제거하기 위한 적법한 방안들을 소개하고 있다.

요컨대, Ricci 사건을 둘러싼 논쟁은 인공지능 공정성을 연구하는 일부 학자들이 해당 판결의 함의를 과도하게 확대 해석하였고, 그러한 해석이 확산되어 실무상 오해를 낳은 사례로 이해될 수도 있다. 하지만 Ricci 사건이 알고리즘 공정성을 달성하는 데 있어 걸림돌이 될 수 있다는 주장을 그저 판결의 적용 범위를 잘못 해석한 것에 불과하다고 치부하기는 어렵다. Ricci 판결의 함의와는 무관하게, 의도적 차별행위에 대한 규제와 결과적 차별행위에 대한 규제 간의 상충관계에 관해서는 오랫동안 논쟁이 진행되어 왔고, 이 문제는 아직까지도 해결되지 못한 채 남아있기 때문이다.

Barocas & Selbst (2016)는 Ricci 판결이 미국 차별금지법제의 해석에 있어서 반분류(anti-classification) 원칙과 반종속(anti-subordination) 원칙 간의 충돌을 보여주는 사례라 주장한다. 반분류 원칙과 반종속 원칙의 구분은 1976년 Owen Fiss가 제시한 것으로서,[240] 이후 미국 차별금지법의 해석에 있어 중요한 영향을 끼쳐 왔다. 그에 따르면 반분류 원칙은 미국 수정헌법 제14조의 평등보호조항을 정부가 개인의 분류(예컨대 인종 등)를 근거로 개인을 차별하는 것을 금지하는 취지로 이해하는 것으로서, 평등보호조항에 대한 적절한 해석이 아니다.[241] 반분류 원칙은 형식적 평등 또는 중립성의 개념으로서의 평등만을 고려하는 것일 뿐이기 때문이다. 이와 달리 평등보호조항의 목적은 종속된 집단(주로 흑인들)을 그들의 역사적 종속에 기여한 정부 행위들로부터 보호하는 것이고, 이것이 반종속 원칙에 해당한다. 이러한 원칙은 실질적 평등의 개념을 지지하고, 적극적 평등실현조치를 옹호하는 형태로 구현될 수 있다.

---

240 Owen M. Fiss, “Groups and the Equal Protection Clause”, *Philosophy & Public Affairs* 107, (1976).

241 David B. Oppenheimer, et al., *Comparative Equality and Anti-discrimination Law*, Edward Elgar Publishing, (2020).

반분류 원칙과 반종속 원칙 간의 충돌은 특히 여러 적극적 평등실현조치의 합헌성에 대한 여러 미국 판결들에서 두드러지게 나타나 왔다. 2007년 미국 연방대법원의 Parents Involved 사건에서 로버트 대법원장은 "인종에 근거한 차별을 중단하는 방법은 인종에 기반한 차별을 중단시키는 것이다."라고 표현하였는데, 이는 반분류원칙을 선명하게 드러낸 것이다.[242] 이와 대비하여 2014년 Schuette 사건에서 소토마요르 대법관은 "인종에 근거한 차별을 중단하는 방법은 인종 주제에 대해 개방적이고 솔직하게 논의하고, 수세기 동안 지속된 인종차별의 불운한 영향에 눈을 뜨고 헌법을 적용하는 것이다."라고 표현하였는데, 이는 반종속원칙의 입장에 선 것이다.[243] 미국 차별금지법제에 있어 반분류원칙과 반종속원칙 간의 충돌은 지난 50년이 넘는 기간 동안 교육이나 채용에서의 다양한 적극적 평등실현 조치, 정부조달에서의 소수인종 사업자 우대 정책 등 다양한 국면에서 불거져 왔고, 엇갈린 판결에 대한 수많은 논의가 이어져 왔다.[244]

Ricci 사건을 돌이켜보면, 뉴헤이븐 시 정부가 소방관 승진시험의 결과 흑인 소방관을 아무도 간부로 승진시키지 않는 결정을 내리는 것은 매우 어려운 일이었을 것임을 쉽게 짐작할 수 있다. 과반수 이상이 유색인종인 시 거주민들에게 이와 같은 결정이 헌법상 평등원칙에 부합하는 것이라고 납득시키는 것은 사실상 불가능한 일이었을 것이다. 미국 헌법상 평등보호조항이나 민권법 제7편을 반종속 원칙에 따라 해석한다면 뉴헤이븐 시 당국의 결정은 그에 부합하는 것이라 평가할 수 있을 것이고, Ricci 사건에서 반대의견을 낸 4명의 대법관은 이러한

---

242 Ronald Turner, "'The way to stop discrimination on the basis of race…'", *Stanford Journal of Civil Rights & Civil Liberties* 11, (2015), p. 84.

243 위의 글, p. 83.

244 적극적 평등실현 조치에 관한 미국 판례는 간단히 정리되기 어렵고 매우 복잡하다. 이를 정리한 국내 문헌으로는 이재희, "헌법상 평등보장과 간접차별금지", 헌법재판소 헌법재판연구원, (2020) 참조. 한편 적극적 평등실현조치에 대한 정치철학적 분석으로는 Kasper Lippert-Rasmussen, *Making sense of affirmative action*, Oxford University Press, (2020) 참조.

입장을 취했다. 그러나 5명의 다수의견은 반분류 원칙의 해석을 따랐고 시 정부의 행위가 민권법 제7편에서 금지하는 의도적 차별이라 판단하였다.

이러한 관점에서 보면, Barocas & Selbst (2016)가 Ricci 판결을 인공지능 공정성 구현의 맥락에서 중대한 걸림돌이 된다고 주장한 취지를 이해할 수 있게 된다. 이들은 결국 반종속 원칙과 반분류 원칙의 충돌 문제에 있어 미국 연방대법원은 반분류 원칙에 따르는 태도를 보이는 것을 문제로 지적한 것이다. 반종속 원칙과 반분류 원칙 간의 충돌은 형식적 평등과 실질적 평등 간의 개념적 충돌 또는 직접차별 규제와 간접차별 규제 간의 집행상 충돌로 드러날 수 있다. 이러한 문제는 여러 인공지능 공정성 심사 기준 중 어느 것을 선택할 것인지의 논쟁에 있어서도 주요하게 고려되는 바, 이에 관해서는 본 연구 제4장 제4절에서 상세히 살핀다.

제3절 개인정보 보호법제

1. 자동화된 의사결정에 대한 규제

개인정보 보호법제는 개인정보의 보호뿐만 아니라 개인정보의 공정하고 투명한 처리를 보장하는 것을 목적으로 하면서 자동화된 의사결정에 관한 규제를 포함하고 있다. 이는 인공지능에 의한 차별을 막는 또다른 기제로 작동할 수 있다.[245] 여기에는 (1) 정보주체에게 자동화된 의사결정 시스템의 활용 사실을 고지해야 하는지, (2) 의사결정에 영향을 끼치는 요소에 관하여 사전적으로 설명해야 하는지, (3) 의사결정이 내려진 이후 그 의사결정의 근거를 설명해야 하는지 등의 쟁점이 존재한다.

---

245 Daniel J. Solove, “The Limitations of Privacy Rights”, *Notre Dame Law Review* 98, (2023), p. 1033.

특히, 자동화된 의사결정에 대한 투명성 의무는 인공지능에 의한 차별 규제에 있어 중요한 역할을 하게 된다. 만약 자동화된 의사결정 시스템에 의해 불리한 결정을 받게 된 정보주체가 애당초 자동화된 의사결정 시스템이 활용되었다는 사실 자체를 알지 못하거나, 어떠한 이유로 불리한 결정이 내려졌는지에 대해 알 수 없다면, 해당 시스템에 대해 문제제기하고 공정한 결정을 요구할 기회를 누릴 수 없을 것이기 때문이다. 다만, 딥러닝을 활용하는 최근의 인공지능 시스템에 대해서는 충분한 설명이 기술적으로 쉽지 않다는 한계가 있다.

이러한 개인정보 보호법제 상의 자동화된 의사결정에 대한 규제는 유럽연합의 1995년 개인정보 보호 지침에서부터 찾을 수 있다.[246] 위 지침 제15조(자동화된 개별 결정) 제1항은 "회원국은 모든 개인에게 본인에 관한 법적 효력을 초래하거나 중대한 영향을 미치고, 그와 관련된 특정 개인적 측면(업무 능력, 신용도, 신뢰도, 행동 등)을 평가하기 위한 자동화된 데이터의 처리에만 전적으로 의존하는 결정의 적용을 받지 않을 권리를 부여하여야 한다"고 규정하고 있다. 이러한 규정은 GDPR 제22조로 이어져 현재와 같은 형태에 이르게 되었다.

또한 우리 개인정보 보호법도 2023년 개정을 통해 정보주체에게 자동화된 의사결정에 대한 대응권을 도입하였다.[247] 개정법에 따르면, 정보주체는 인공지능 기술을 적용한 시스템을 포함하는 자동화된 시스템으로 개인정보를 처리하여 이루어지는 결정이 자신의 권리 또는 의무에 중대한 영향을 미치는 경우에는 해당 결정을 거부하거나 해당 결정에 대한 설명 등을 요구할 수 있다.

이하에서는 국내 및 국제 논의에서 중요하게 논의되고 있는 GDPR 상의 자동화된 의사결정 시스템에 대한 규율 체계를 검토한다. 특히 해당 규정이 정보주체에게 개별 결정에 대한 설명을 얻을 권리를 보장하는지에 관하여 국제적 논쟁

---

246 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

247 개인정보 보호법(법률 제19234호, 2023. 3. 14., 일부개정) 제4조 제6호 및 제37조의2.

이 이루어진 바, 이에 대한 고찰을 통해 국내 인공지능 공정성 규제에 있어서의 시사점을 도출하고자 한다.

2. GDPR 상의 자동화된 의사결정 시스템 규율

GDPR 제22조(프로파일링 등 자동화된 개별 의사결정) 제1항은 "정보주체는 프로파일링 등 본인에 관한 법적 효력을 초래하거나 이와 유사하게 본인에게 중대한 영향을 미치는 자동화된 처리에만 전적으로(solely) 의존하는 결정의 적용을 받지 않을 권리를 가진다."고 규정하여 개인에게 법적 효력 또는 이와 유사하게 중대한 영향을 미치는 의사결정에 있어 전적으로 자동화된 처리에만 의존하는 것을 원칙적으로 금지하고 있다.[248] 다만, 동조 제2항은 (a) 정보주체와 컨트롤러 간의 계약을 체결 또는 이행하는 데 필요한 경우, (b) 정보주체의 권리와 자유 및 정당한 이익을 보호하기 위한 적절한 조치를 규정하는 유럽연합 또는 회원국 법률이 허용하는 경우, (c) 정보주체의 명백한 동의에 근거하는 경우에는 자동화된 처리를 허용한다. 나아가 동조 제3항은 계약의 체결 또는 이행에 있어 필요한 경우 또는 정보주체의 동의를 얻은 경우라 하더라도 정보주체는 인간의 개입을 요구하고 자신의 의견을 표시하며 그 결정에 대해 이의를 제기할 수 있는 권리를 부여하면서 컨트롤러에게 위 권리를 보호하는데 적절한 조치를 시행할 의무를 부과하고 있다.

위 조항의 취지는 GDPR 전문(recital) 제71조에서 설명되어 있다.[249] 우선, 동 조항은 자동화된 의사결정에 대한 제한이 적용되는 구체적인 사례로 온라인

---

248 이하의 GDPR 번역은 국내 개인정보보호위원회의 번역을 토대로 저자가 추가로 수정한 것이다.

249 GDPR 전문 제71조 "정보주체는 자동처리에만 근거하여 정보주체에게 법적 효력을 초래하거나 이와 유사하게 본인에게 중대한 영향을 미치는 결정(또는 정보주체의 개인적인 측면을 평가하는 조치를 포함)에 적용 받지 않을 권리를 갖는다. 예를 들어 이러한 결정은 온라인 대출 신청에 대한 자동적 거절이나 인적개입 없이 이루어지는 전자채용 관행 등을 포함한다. 이러

대출 신청에 대한 자동적 거절이 이루어지는 경우와 인적개입 없이 이루어지는 전자적 채용 관행을 들고 있다. 나아가 정보주체의 업무능력, 경제적 상황, 건강, 개인의 성향이나 관심사, 신뢰성 또는 행동, 위치 또는 움직임과 관련된 측면을 분석하고 예측하는 '프로파일링(profiling)'을 수행한 결과 정보주체에게 법적 효력 또는 이와 유사하게 중대한 영향을 미치는 경우에도 이러한 규정이 적용된다는 점을 명시하고 있다. 나아가 위 조항은 자동화된 의사결정에 있어 정보주체에 대해 보장되어야 할 안전조치(safeguards)로는 ① 자동화된 의사결정에 대한 구체적 정보 제공 의무 ② 인적 개입을 획득할 권리 ③ 의견을 표명할 권리 ④ 해당 결정에 대한 설명을 획득할 권리 및 ⑤ 해당 결정에 대한 이의를 제기할 권리가 포함된다고 설명하고 있다. 다만, 전문 제71조에는 "해당 평가 이후 도달한 결정에 대한 설명을 획득할 권리(to obtain an explanation of the decision reached after such assessment)"가 포함되어 있으나, 앞서 살펴본 본문 제22조 제3항의 안전조치에는 위 권리가 포함되어 있지 않은 점이 특기할 만한데, 이는 후술하는 바와 같이 GDPR 상으로 이른바 "설명요구권(right to explanation)"이 존재하는지에 관한 국제적 논쟁의 주요한 논점이 되었다.

한편, GDPR 전문 제71조가 설명하고 있는 정보주체에 대한 자동화된 의사결정에 관한 구체적 정보를 제공할 의무는 GDPR 제13조 제2항 (f)호, 제14조 제2항 (g)호, 제15조 제1항 (h)호에 규정되어 있다. 제13조와 제14조는 개인정

한 처리는, 자연인의 개인적인 측면을 평가하는 모든 형태의 개인정보의 자동처리로 구성된 '프로파일링'을 포함하며, 특히 정보주체의 업무능력, 경제적 상황, 건강, 개인의 성향이나 관심사, 신뢰성 또는 행동, 위치 또는 움직임과 관련된 측면을 분석하고 예측하며, 정보주체에게 법적 효력 또는 이와 유사하게 중대한 영향을 미치는 경우 그러하다. 그러나 프로파일링 등 이러한 처리에 근거한 의사결정은, … 컨트롤러가 적용 받는 유럽연합 또는 회원국의 법률에서 명시적으로 인가하는 경우, 정보주체와 컨트롤러 간의 계약의 체결이나 수행에 필요한 경우, 또는 정보주체가 명백한 동의를 제공하는 경우에만 허용된다. 어떠한 경우에도, 이러한 처리는 정보주체에게 구체적인 정보 전달, 인적개입을 획득할 수 있는 권리, 의견을 표명할 권리, 해당 평가 이후 도달한 결정에 대한 설명을 획득할 권리, 해당 결정에 이의를 제기할 권리 등 적절한 안전조치를 적용받아야 한다. (후략)" (역자 강조).

보 수집시 정보주체에 고지해야 할 사항에 관한 조항이고, 제15조는 정보주체의 열람권에 관한 규정이다. 제13조와 제14조는 개인정보가 정보주체로부터 수집되는 경우와 그렇지 않은 경우를 정보주체에 고지해야 할 사항을 각기 구분하여 규정하고 있는데, 자동화된 의사결정에 관하여 정보주체에 고지해야 할 사항은 양자가 동일하다. 즉, ① 제22조 제1항 및 제4항의 자동화된 의사결정(프로파일링 포함)의 존재 사실 및 ② 자동화된 의사결정이 관련 논리에 관한 유의미한 정보와 그 같은 처리가 개인정보주체에 미치는 중대성 및 예상되는 결과를 정보주체에 고지하여야 한다. 또한 제15조는 정보주체가 요청하는 경우 동일한 정보에 대한 열람권(right to access)을 규정하고 있다.

다만, 위 조항의 적용 범위가 비교적 좁게 설정되어 있다는 점에 유의할 필요가 있다. GDPR 제13조와 제14조에 따른 고지의무는 앞서 제22조 제1항(자동화된 의사결정 일반에 대한 제한) 또는 제4항(민감정보에 관한 자동화된 의사결정에 대한 제한) 규정이 적용되는 범위에서 인정된다. 즉, 위와 같은 사전적 고지의무는 "본인에 관한 법적 효력을 초래하거나 이와 유사하게 본인에게 중대한 영향을 미치는 자동화된 처리에만 의존하는 결정"에 관하여만 적용된다. 그러므로 콘텐츠의 추천이나 검색 결과 표시, 광고표시 등과 같은 경우에는 위 조항이 적용된다고 판단되기는 어려울 것으로 보인다.[250]

또한 위 조항에 근거한 정보주체에 대한 고지의무나 정보주체의 열람권은 이에 대해서는 "전적으로(solely)" 자동화된 결정에 대해서 적용된다는 한계도 존재한다. 따라서 단지 인간 의사결정자에게 조언만을 제공하거나 그를 보조하기 위하여 자동화된 시스템이 활용되는, 인간이 여러 요소들을 고려하여 최종 결정을 내리는 경우에는 위 조항의 적용을 받지 않게 된다.[251] 다만, 의사결정이 실질적

---

250 Lilian Edwards and Michael Veale, "Enslaving the algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?", *IEEE Security & Privacy*, 16(3), (2018), p. 47.

251 Article 29 Working Party, "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679", (2018), p. 20.

으로는 알고리즘에 따라 자동화된 형태로 이루어지고 있으나 인간이 요식적으로 승인을 하는 형태를 취하는 경우(이른바 “rubber-stamping”에 해당하는 경우) 전적으로 자동화된 의사결정으로 포함될 수 있다. 제29조 작업반의 설명요구권에 관한 가이드라인은 “컨트롤러가 인적 개입의 조작을 통해 22조의 조항을 회피할 수 없다. 예를 들어 누군가 결과에 대한 실제적 영향 없이 개인에 대해 자동 생성된 프로파일을 일상적으로 적용하는 경우에도 이 결정은 여전히 자동화된 처리만을 기반으로 한 결정에 해당한다”고 해석하고 있다.[252]

3. GDPR 상의 설명요구권 논쟁

전술한 GDPR 상의 규정에 대한 국제적 관심은 Goodman & Flaxman (2016)이 “설명요구권(right to explanation)”이 존재한다는 점을 소개하면서 비롯되었다.[253] 이들은 GDPR이 시행되기 전 GDPR 문언이 공식적으로 공개된 이후 GDPR상 정보주체에게 머신러닝 알고리즘에 대한 설명을 요구할 권리가 부여되어 있다고 간략히 소개하였다. 그런데 뒤이어 Wachter et al. (2017)이 “왜 GDPR상 자동화된 의사결정에 대한 설명요구권이 존재하지 않는가”라는 다소 도발적인 제목의 논문을 발표하면서, 설명요구권에 관한 국제적 논쟁이 촉발되었다.[254] 이하에서는 위 논쟁의 전개를 간략히 소개하고 그 시사점을 도출한다.

---

252 위의 글. p. 21.

253 Bryce Goodman and Seth Flaxman, “EU regulations on algorithmic decision-making and a ‘right to explanation’, *In ICML workshop on human interpretability in machine learning (WHI 2016)*, New York, NY, (2016, June).

254 Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, “Why a right to explanation of automated decision-making does not exist in the general data protection regulation”, *International Data Privacy Law* 7(2), (2017), pp. 76-99.

가. GDPR 상 개별 결정에 대한 설명을 얻을 권리가 존재하지 않는다는 주장

Wachter et al. (2017)은 GDPR의 입법 경위 및 종전 EU 개인정보 보호법제 하에서의 관련 입법례, 판례 등 다양한 논거를 들어 GDPR상 설명요구권이 존재하지 않는다는 주장을 제기하였다.

위 주장의 핵심적 논거로는 GDPR 전문과 본문 간의 불일치이다. 즉, 전술한 바와 같이 GDPR 전문 제71조는 자동화된 의사결정에 대한 안전조치로서 "해당 평가 이후 도달한 결정에 대한 설명을 획득할 권리"가 명시되어 있으나, GDPR 본문 제22조 제3항의 자동화된 의사결정에 대한 안전조치에는 위 권리가 포함되어 있지 않다. 그런데 유럽연합 법제에서는 전문은 참조적 효력만이 있을 뿐 법적 구속력이 인정되지 않는다.[255] 더욱이 입법 경과를 살펴보면 입법안의 논의 과정에서 본문에도 안전장치로 설명요구권이 추가되었다가 후속 논의를 통해 의도적으로 삭제된 사실이 확인된다. 그 결과 설명요구권은 법적 구속력 없는 전문에만 존재하고 법적 구속력이 있는 본문에는 존재하지 않는 형식을 취하게 되었는데, 이는 설명요구권을 인정하지 않겠다는 입법자의 의도를 반영한 것으로 해석될 수 있다.

그러나, Wachter et al. (2017)도 GDPR 상으로 설명요구권과 관련된 권리가 전혀 도출된 수 없다고 주장하는 것은 아니고, 제한적 형태의 "정보요구권(right to be informed)"은 존재한다고 해석한다. 이러한 주장을 전개하기 위하여 위 연구는 '설명'을 ① 시스템 기능에 대한 설명과 특정 결정에 대한 설명, ② 사전적 설명과 사후적 설명이라는 기준으로 구분한다.

---

255 위의 글.

기준	구분	의미
설명의 내용	시스템 기능(system functionality)에 대한 설명	자동화된 결정 시스템의 로직, 중요성, 예상되는 결과 및 일반적 기능을 의미 (예) 시스템의 요구 사양, 의사결정 나무, 사전적으로 정의된 모델, 기준 및 분류 체계
	특정 결정(specific decisions)에 대한 설명	특정한 자동화된 결정에 있어 근거, 이유 및 개별적 상황을 의미 (예) 각각의 특성에 대한 가중치, 기계가 정의한 사건에 적용된 결정 규칙, 선호 및 프로파일 그룹에 대한 정보
설명의 시점	사전적(ex ante) 설명	자동화된 의사결정이 이루어지기 전의 설명
	사후적(ex post) 설명	자동화된 의사결정이 이루어지기 후의 설명

표 2 자동화된 의사결정 시스템에 대한 설명의 유형

자동화된 의사결정이 내려지기 이전에는 시스템 기능에 관한 설명만이 가능하고 특정 결정에 관한 설명은 당연히 불가능하므로, 위 기준에 따르면 ① 시스템의 기능에 대한 사전적 설명, ② 시스템의 기능에 대한 사후적 설명, ③ 개별 의사결정에 대한 사후적 설명이 가능할 것이다. 그런데 전술한 바와 같이 GDPR 제13조 제2항 (f)호, 제14조 제2항 (g)호는 개인정보를 수집하는 시점에 "관련 논리에 관한 유의미한 정보와 그 같은 처리가 정보주체에 미치는 중대성 및 예상되는 결과"를 고지할 의무를 부과하고 있다. 따라서 위 조항은 여러 설명 중 "① 시스템의 기능에 대한 사전적 설명"을 제공할 의무를 규정한 것으로 이해된다.

다음으로 GDPR 제15조 제1항 (h)호는 컨트롤러에게 (그 시점과 무관하게) 정보주체가 요청하는 경우 위와 동일한 정보를 제공할 의무를 부과한다. 이는 사후적 설명에 대응되는 것인데, 그 범위에 있어 제13조나 제14조와 마찬가지로 시스템의 기능에 대한 설명만을 포함할 것인지, 개별 의사결정에 대한 설명을 포함할 것인지가 쟁점이 된다. 그런데 Wachter et al.은 제13조, 제14조, 제15조

의 문언이 모두 동일하므로 제15조의 열람권의 범위를 달리 해석할 이유가 없고, “관련 논리(logic involved)”, “예상되는 결과(envisaged consequences)” 등의 장래 지향적인 문언에 비추어 볼 때 시스템의 기능에 대한 설명만 포함되고 특정 결정에 대한 사후적 설명에 대한 권리는 인정될 수 없다고 주장한다.

이러한 분석에 따르면, 정보주체는 사전적이든 사후적이든 시스템의 기능에 대해 고지받거나 열람할 권리만을 보유하고 있고, 개별 결정에 대한 설명을 얻을 권리(즉, 이른바 ‘설명요구권’이라 불리는 권리)는 GDPR상 인정되지 않는다는 결론에 이르게 된다. 예컨대, 자동화된 신용점수 부여 시스템에 이러한 주장을 적용하면, 해당 사업자는 일반적 로직(데이터 유형, 고려 대상 특성, 의사결정 나무의 분류), 중요성(신용 점수의 부과), 예상되는 결과(금융기관이 신용도를 평가하여 이자율 등 대출 조건에 영향을 끼칠 수 있음) 등을 고지하거나, 정보주체가 요청할 경우 이러한 정보를 제공할 의무만이 있을 뿐이다.

나. GDPR 상 개별 결정에 대한 설명을 얻을 권리가 존재한다는 주장

위와 같이 GPDR상 설명요구권이 존재하지 않는다는 주장이 제기된 이후 곧바로 위 주장에 대한 비판이 뒤이었다. 이러한 비판의 대표적 예는 Selbst & Powles (2018)이다.[256] 위 연구도 GDPR의 전문 제71조의 법적 구속력을 인정되지 않으므로, GDPR 제22조에 근거하여서는 설명요구권이 도출될 수 없다는 점을 인정한다. 하지만 이들은 GDPR 제13조 제2항 (f)호, 제14조 제2항 (g)호, 제15조 제1항 (h)호를 근거로 정보주체에 대한 설명요구권이 인정될 수 있다고 본다.

---

256 Andrew Selbst and Julia Powles. “Meaningful Information and the Right to Explanation.”, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, (2018)

위 연구는 GDPR 제13 내지 15조의 고지 대상이 되는 정보의 범위에 “유의미한 정보(meaningful information)”라는 표현을 포함하고 있다는 점에 주목한다. 즉, 어떠한 정보가 유의미하기 위해서는 해당 정보를 이용하여 정보주체가 어떠한 행동을 제기할 수 있어야 한다. 그런데 만약 GDPR 제13 내지 15조에 따라 개별 결정의 근거에 대한 구체적 설명이 제공되지 않고, 그저 시스템의 기능에 대한 설명만이 제공된다면, 정보주체는 이를 이용하여 법적 권리를 행사하는 데 상당한 어려움을 겪을 수밖에 없다. 이러한 경우 해당 정보를 ‘유의미’한 것으로 평가하기 어려울 것이다.

나아가 위 연구는 설명의 가치를 도구적 가치(instrumental value)와 내재적 가치(intrinsic value)로 구분한다. 도구적 가치는 법적 권리 행사를 보장한다는 구체적인 의미를 갖는 것인데 비해, 내재적 가치는 인간의 자유 의지나 삶에 대한 통제권을 의미한다. 만약 설명의 내재적 가치에 초점을 맞출 경우 영업비밀 등과 같이 잘 정립된 반대 권리에 의해 쉽게 약해질 위험이 있다. 따라서 GDPR 제13 내지 15조에 따라 정보주체에게 제공되는 정보가 유의미하기 위해서는 최소한의 도구적 가치가 있어야 한다. 이를 위해서는 적어도 GDPR 또는 유럽연합 기본권 헌장상 기본권 행사를 가능하게 할 정보가 제공되어야 하는 것이다. 나아가 위 연구는 Wachter et al. (2017)이 설명을 (1) 시스템 기능에 대한 설명과 구체적 결정에 대한 설명, (2) 사전적 설명과 사후적 설명으로 구분한 것에 대한 비판도 제기한다. 특히 머신러닝 시스템은 학습이 완료되고 나면 대부분 결정론적으로 동작하므로, 전체 시스템에 대한 완전한 설명이 이루어진다면 개별 결정에 대한 모든 것을 알 수 있는 것이다. 즉, 만약 시스템 기능에 대한 설명이 가능하다면, 당연히 특정 결정에 대한 설명을 제공하는 것이 가능하다. 이러한 기술적 이해를 토대로 GDPR을 해석하면 사전적 설명과 사후적 설명을 구분할 필요도 없게 된다.

요컨대, 위와 같은 Selbst & Powles (2018)의 주장에 따르면, GDPR의 입법 취지와 전체 구조를 고려한 합목적적 및 체계적 조문 해석에 근거하여 정보주체에게 개별 결정에 설명요구권을 도출하고 있는 것으로 이해될 수 있다.

다. 논의의 발전

이상과 같은 학술적 논쟁은 그에 뒤이어 더욱 진전되고 실무적으로 구체화된 논의로 발전하였다. 특히 Casey et al. (2019)은 GDPR의 설명요구권을 정보주체 개인에 대한 권리로만 파악하는 논의의 한계를 지적하면서, GDPR에 따른 개인정보 규제 당국의 강화된 집행 권한이나 기업에 의한 자율적인 사전적·예방적 평가에 있어서 설명요구권이 갖는 의미에 주목한다.[257] 위 연구는 무엇보다도 GDPR상 개인정보보호 감독기관이 상당한 권한을 보유하고 있다는 점을 중요하게 고려해야 한다고 주장한다. GDPR 제58조는 감독기관의 권한에 관하여 규정하고 있는데 특히 "개인정보보호 감사 형태의 조사를 실시"하고, "컨트롤러 또는 프로세서로부터 업무 수행에 필요한 모든 개인정보 및 모든 정보를 열람"할 수 있으며(동조 제1항), 경고 또는 시정명령을 내리거나 개인정보 처리의의 금지를 명하거나 개인정보의 역외 이전을 금지할 수 있는 등의 권한을 갖고 있다(동조 제2항). 더욱이 감독기관은 GDPR의 핵심 조항을 위반할 경우 2천만 유로 또는 전년도 총매출액의 4% 중 높은 금액을 상한으로 하는 막대한 과징금을 부과할 권한을 보유하게 되었다. 이러한 전반적 규정에 비추어 GDPR에서는 정보주체에 대해 개별적인 설명요구권이 인정되는지 여부보다 감독기관에 의한 조사가 더욱 중요한 의의를 갖는 측면이 있다.

---

257 Bryan Casey, Ashkon Farhangi, and Roland Vogl. "Rethinking Explainable Machines: The GDPR's' Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise", *Berkeley Technology Law Journal* 34, (2019).

여기서 흥미로운 점은 GDPR이 한편으로는 위와 같이 감독기관의 권한을 강화하면서도, 다른 한편으로는 기업에 의한 자율적인 사전적·예방적 평가를 강조하고 있다는 점이다. 이러한 자율 평가에 있어서는 제35조에 따른 '개인정보보호 영향평가(Data Protection Impact Assessment)'(이하 "DPIA")가 중요한 지위를 갖는다.[258] 더욱이 GDPR 제35조 제1항 (a)호는 자동화된 처리가 이루어지는 경우에는 DPIA가 필요하다는 점을 명시하고 있기까지 하다. 그런데, 위 조항에서 주목할 점은 제22조의 규정과 달리 "전적으로(solely)" 자동화된 처리가 아니더라도 DPIA가 요구된다는 점이다. 위 조항에 "전적으로"라는 문구를 포함하지 않은 것은 DPIA의 적용 범위를 폭넓게 정하기 위해 입법 과정에서 의도된 것으로서 해석된다. 기업이 DPIA를 수행한 후에는 이를 문서화해야 하는데, 해당 문서에는 처리 동작과 목적에 대한 체계적 설명, 필요성과 비례성 평가, 권리·자유 침해 위험 평가, 위험해소 조치에 관한 사항을 포함해야 한다(GDPR 제35조 제7항). 다만 기업이 DPIA를 공개해야 할 의무까지는 존재하지 않는다.

또한 GDPR이 '설계에 의한 개인정보보호(Data Protection by Design)' 원칙을 규정하고 있다는 점도 중요하다. GDPR 제25조에 따르면 기업들은 "최신 기술, 실행 비용, 그리고 개인정보 처리의 성격, 범위, 상황 및 목적을 고려하여 기술 및 관리적 조치를 … 이행"해야 하고, 그러한 기술 및 관리적 조치는 "GDPR과 정보주체의 권리 요건을 충족시키기 위해 필요한 안전조치가 개인정보 처리에 통합될 수 있도록 설계(designed)"되어야 한다. 이러한 '설계에 의한 개인정보보호 원칙'은 기업의 서비스 설계, 프로토타입, 테스트 및 구현에 있어 GDPR 규정의 준수 여부를 사전적으로 고려할 것을 요구한다. 위와 같은 설계 단계에서의 자율 평가 시에는 개인정보가 적정하게 보호되고 있는지 여부만이

---

258 GDPR 제35조 제1항 "개인정보 처리의 성격과 범위, 상황, 목적을 참작하여 개인정보의 처리 유형이 개인의 권리와 자유에 중대한 위험을 초래할 것으로 예상되는 경우(특히 신기술을 사용하는 경우), 컨트롤러는 개인정보의 처리 이전에, 예정된 처리 작업이 개인정보 보호에 미치는 영향에 대한 평가를 수행해야 한다."

아니라 자동화된 의사결정 시스템을 통해 개인정보를 처리할 경우 그 투명성과 공정성이 보장되는지 여부에 대한 평가도 함께 이루어져야 한다.

이상에서 살펴본 바와 같이, GDPR상의 더욱 강화된 개인정보 감독기관의 권한과 DPIA 제도, 설계에 의한 개인정보보호 원칙 등에 비추어 보면, GDPR의 자동화된 시스템에 대한 규제는, 개인정보를 처리하는 사업자에 대해 사전적·예방적 평가를 수행하도록 요구하는 한편, 강력한 권한을 가진 감독기관이 그 준수 여부를 감독하는 이중적 형태로 규정되어 있는 것으로 이해될 수 있다. 따라서 설명요구권은 정보주체가 개별적으로 행사할 수 있는 권리에 그치는 것이 아니라, 기업의 상품·서비스 설계 단계에서부터 광범위하게 적용되고, 그 영향 평가에 있어 주요하게 고려해야 할 사항이 된다.

라. 인공지능 공정성과 GDPR 설명요구권 논쟁의 시사점

GDPR 상 정보주체에 설명요구권이 인정되는지에 관하여 최근 수 년간 이어진 국제적인 논쟁의 결과, 자동화된 결정에 관한 설명의 필요성과 그 구체적 범위에 관하여 많은 발전이 이루어졌다. 위 논쟁의 경과를 살펴보면, GDPR의 문언해석상 논란은 있었으나, 자동화된 결정에 대한 법적 규율이 필요하다는 인식은 공통적으로 존재하는 것으로 보인다. 결국 자동화된 결정에 대해 적어도 일정한 범위 내에서는 정보주체가 이의를 제기하고, 자신의 의견을 표명하며, 인간의 개입을 요구할 수 있어야 한다는 점은 GDPR 제22조의 문언상 명백하다. 이러한 GDPR의 입법 내용은 전세계적으로 영향을 끼치고 있다. 세계적으로 사업을 영위하는 기업들은 유럽연합 시장을 포기하지 않는 이상 GDPR의 규제를 준수해야 할 것이고, 그 결과 유럽연합 규제가 세계적으로 확산될 가능성도 존재한다. 이러한 유럽연합 주도의 규제의 상향 효과는 소위 '브뤼셀 효과(Brussel Effect)'라 표현되기도 한다.[259]

---

[259] Anu Bradford, "The Brussels Effect", *Northwestern University Law Review* 107, (2012).

다만, GDPR상 자동화된 의사결정에 관해서는 정보주체에 충분한 구제수단을 제공하고 있지 못하다는 문제제기도 고려할 필요가 있다. Wachter et al. (2017)이 GDPR상 설명요구권이 존재하지 않는다고 주장한 취지는, 자동화된 의사결정 시스템을 활용하는 사업자들을 면책시키기 위한 것이 아니라, GDPR 입법 상의 한계를 지적하기 위한 것이었다. 그뿐만 아니라, GDPR의 자동화된 의사결정에 관한 제13-15조, 제22조의 규정은 "본인에 관한 법적 효력을 초래하거나 이와 유사하게 본인에게 중대한 영향을 미치는 자동화된 처리에만 의존하는 결정"에 대해서만 적용된다는 적용범위 상의 제한이 존재한다. 그 결과 콘텐츠 추천, 검색, 광고 등 여러 영역에 있어서 위 규정이 적용되는 못하는 경우도 있을 것이다. 다른 한편, GDPR의 개인정보보호 영향평가 제도, 설계에 의한 개인정보보호 원칙 등을 규정함으로써 기업이 자율적으로 사전적·예방적 조치를 취하도록 요구하고 있다는 점도 중요하다. 이러한 제도나 원칙은 개인정보 보호뿐만 아니라 자동화된 의사결정 시스템이 공정하고 투명한 운용되는 것을 보장하기 위한 목적을 포함하고 있고, 이는 인공지능 공정성 보장을 위해서도 중요한 역할을 할 수 있다.

제 4 절 인공지능 규제법(안)

인공지능에 의해 초래될 수 있는 위해를 규율하는 또다른 규제 방안으로 인공지능 기술 자체의 개발과 활용에 대한 법률에 관한 논의가 활발히 진행되고 있다. 현재 각국에서 논의되고 있는 규제 방안은 EU AI 법안과 같이 인공지능 기술 전반을 적용 대상으로 하는 포괄적 입법(안)과 미국에서 주로 시도되고 있는 개별적인 인공지능 활용 사례를 대상으로 하는 영역별 입법(안)으로 구분할

수 있는 바, 이하에서는 EU AI 법안의 사례와 미국의 연방 및 주 차원에서 이루어지고 있는 입법 사례를 검토한다.

1. EU AI 법안

가. 개요

EU 차원에서 인공지능에 관한 적절한 규제를 수립하는 방안에 관한 논의는 2018년 이후부터 지속적으로 이루어져 왔다. 특히 EU 집행위원회는 2018년 6월 인공지능에 관한 고위 전문가 그룹(High-Level Expert Group on AI)을 구성하였고, 2019년 4월 "신뢰할 수 있는 인공지능을 위한 윤리 가이드라인"을 발간한 바 있다.[260] 또한 사업자들이 인공지능 신뢰성을 평가하는 데 활용할 수 있는 체크 리스크를 발표하기도 하였다.[261] 이후 EU 집행위원회는 광범위한 여론 수렴 과정을 거쳐 2021년 4월 EU AI 법안 초안을 제안하였다.[262]

EU AI 법안에 대한 규제영향 평가보고서에 따르면, 그 성안 과정에서 다양한 입법 대안이 검토되었는데, (1안) 가장 낮은 수준의 규제로서 사업자들이 자발적으로 인공지능의 위험성을 표기하도록 요구하는 방안, (2안) 개별 인공지능 활용 분야에서 필요한 경우 규제를 수립하는 방안(이른바 ad-hoc 영역별 접근), (3안) '비례적 위험기반' 접근에 따라 개인의 안전이나 기본권 침해 위험이 심대한 경우에는 규제 부담을 지우고, 그렇지 않은 경우 제한적 의무만 부과하거나 법적 의무는 부과하지 않는 방안, (3+안) 위 3안과 같이 위험기반 규제를 도입하되,

---

260 High-Level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines for Trustworthy AI*, (2019. 4. 8).

261 High-Level Expert Group on Artificial Intelligence (AI HLEG), *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*, (2020. 7. 17).

262 유럽의회 소관위원회(Internal Market Committee와 Civil Liberties Committee)는 2023. 5. 11. EU 집행위원회 초안에 대한 수정안(2023. 5. 9.자)에 합의하였고, 이는 2023. 6. 14. 유럽의회에서 가결되었다. 이하에서는 위 수정안을 기준으로 논의한다.

고위험이 아닌 경우 사업자들의 행동 강령(code of conducts)를 통한 자율규제를 포함하는 방안, (4안) 모든 인공지능 시스템에 대해 EU 차원의 의무를 부과하는 방안 등이 고려되었다.[263] 그런데 1안의 자발적 라벨링 방안이나 2안의 ad-hoc 영역별 접근법은 인공지능의 안전 및 기본권 침해 위험성을 적절히 예방하기 어렵고, 4안의 경우 과잉 규제의 위험이 있는 것으로 평가되었다. 그 결과 종국적으로 3+안이 채택되었다.[264]

이처럼 EU AI 법안에 의한 규제의 핵심은 인공지능의 위험을 단계별로 구분하여 그에 비례하는 의무를 부과하는 것이다. 법안은 AI에 의한 위험을 "수용 불가능한 위험(unacceptable risk)", "고위험(high-risk)", "제한적 위험(limited risk)" 및 "최소 또는 무 위험(minimal or no risk)"으로 구분한다. 우선, 수용 불가능한 위험을 초래하는 인공지능 시스템을 시장에 출시하거나, 서비스를 제공하거나 사용하는 것이 금지된다.[265] 다음으로, 고위험 인공지능 시스템은 인공지능 시스템이 안전 부품 부품으로 사용되는 경우와 같이 보건 및 안전에 대한 위협을 제기하는 경우와 개인의 기본권에 중대한 위험을 초래할 수 있는 경우로 구분할 수 있다.[266] 고위험 인공지능에 대해서는 EU AI 법안 상의 다수의 의무사항을 준수해야 하고, 규제 준수에 대해 적합성 평가(conformity assessment)가

---

263 European Commission, *Impact Assessment of the Regulation on Artificial intelligence*, (2021. 4. 21.)

264 위의 글 p. 85.

265 유럽의회 수정안은 EU 집행위원회 초안에 비하여 수용불가능한 위험의 범위를 상당히 추가하였다. 2023. 5. 9.자 수정안 제5조에 추가된 주요 사항은 다음과 같다: 공공 장소에서 사용되는 "실시간" 원격 생체 인식 시스템; "사후적" 원격 생체 인식 시스템(다만, 법집행을 위해 중대한 범죄를 처벌하는 경우 법원 허가를 받은 경우에만 예외적으로 허용); 민감한 특성(예: 성별, 인종, 민족, 시민권 상태, 종교, 정치 성향)을 사용하는 생체인식 분류 시스템; 프로파일링, 위치 또는 과거 범죄 행위에 기반한 예측 경찰 시스템; 법 집행, 국경 관리, 직장 및 교육 기관에서의 감정 인식 시스템; (인권과 프라이버시 권리를 침해하는) 얼굴 인식 데이터베이스를 생성하기 위한 목적으로 생체 인식 데이터를 소셜 미디어 또는 CCTV 영상에서 무차별적으로 수집하는 행위.

266 EU AI 법안 부속서 II와 부속서 III는 고위험 인공지능을 나열하고 있다.

강제된다.[267] 한편, 제한적 위험은 자연인과 상호소통하거나, 콘텐츠를 생성하는 인공지능 시스템 등의 경우로서 이에 대해서는 인공지능 작동 사실에 대한 투명성 의무를 부여하고 있다.[268] 마지막으로 최소 위험의 경우에는 강제적 의무를 부여하지 않고, 사업자들에 의한 자율규제가 권장될 뿐이다. 이하에서는 본 연구에서 구분한 인공지능의 유형마다 인공지능의 공정성 침해 위험과 관련하여 EU AI 법안이 어떠한 규제를 시도하고 있는지 살핀다.

나. 인공지능 유형별 공정성 관련 규제 내용

1) 판별모형

EU AI 법안은 판별모형 중 일부 활용 사례를 수용 불가능한 위험이나 고위험으로 정하고 있다. 우선, 수용 불가능한 위험으로 구분된 활용 사례로는 사회

---

267 특히 고위험 인공지능의 위험성에 대한 안전장치로서 부과되는 핵심적 의무사항으로는 다음과 같은 것들이 있다. (1) 인공지능 시스템에 위험관리 시스템을 구축하고, 이를 문서화하고, 유지해야 한다. 이러한 위험관리 시스템은 인공지능의 전체 생애주기에 걸친 지속적이고 반복적 검토와 업데이트가 요구된다(제9조). (2) 학습, 검증 및 시험 데이터 집합의 대표성, 무오류성, 완전성 보장을 위한 조치가 요구된다(제10조). (3) 인공지능 시스템이 법적 요구사항을 준수하다는 점을 보여주는 기술 문서를 작성해야 한다. 또한 인공지능 시스템이 작동하는 동안의 자동적 사건 기록(로그) 기능을 구현해야 하는데, 이는 AI 시스템 기능을 생애주기 전체에 걸쳐 추적할 수 있는 수준이어야 한다(제11조 및 제12조). (4) 이용자가 인공지능 시스템의 출력을 해석하고 적절하게 이해할 수 있도록 충분히 투명해야 한다(제13조). 이때 인공지능 시스템이 이용자에게 제공해야 할 정보로는 공급자 정보, 인공지능 시스템의 성능(목적, 정확도, 오용 위험성, 성능 및 데이터 정보), 변경사항, 인간 감독, 예상 수명 등이 있다.

268 EU AI 법안 제52조.

적 신용점수(social scoring) 시스템[269]과 범죄 위험성 예측 시스템[270]을 들 수 있다. 사회적 신용점수 시스템은 자연인이나 집단을 사회적 행동, 특성이나 성격에 기반하여 평가하거나 분류하는 것으로서, 원래 데이터가 생성·수집된 맥락과는 무관한 사회적 맥락에서 이들을 불리하게 처우함으로써 인간의 존엄성과 평등권을 침해할 위험이 있으므로 금지된다.[271] 또한 범죄 위험성 예측 시스템의 경우에도 법집행 기관 등이 개인의 위치정보나 전과 등의 특성을 통해 개인의 재범 위험성이나 특정 지역의 범죄 발생 가능성을 예측하는 것으로서, 무죄추정의 원칙에 반하고, 인간의 존엄성과 평등권을 침해할 위험이 있어 금지된다.[272]

다음으로, EU AI 법안은 이제까지 판별모형에 의해 공정성 침해가 초래될 우려가 제기되어 온 다수의 활용 사례를 고위험으로 분류하고 있다. 구체적으로 부속서 III은 교육·직업훈련, 채용·근로자 관리, 민간·공공 서비스 접근, 법 집행, 이민이나 국경 관리, 사법 및 민주적 절차 등에서 활용되는 인공지능 시스템 사례들을 고위험으로 규정하고 있다. 고위험 인공지능의 공정성 보장과 관련되어 주목할 지점은 인공지능 시스템의 평가, 위험 관리 및 인간 감독에 관해 다수의 *사전적(ex ante)* 의무를 부과하고 있다는 점이다. 이는 교육 및 훈련, 고용, 중요한 서비스, 법 집행 및 사법 등과 같이 중요한 영역에 있어 고위험 인공지능 시

---

269 EU AI 법안(2023. 5. 9. 수정안) 제5조 제1항 (c)에 따라 금지되는 사회적 신용점수, 평가 및 분류 시스템은 일정기간 동안 자연인들이나 그 집단의 사회적 행동 또는 알려지거나, 추론되거나 또는 예측된 개인 또는 성격 특성에 따라 이루어지는 것으로서, 사회적 신용 점수가 데이터가 원래 생성되거나 수집된 맥락과 무관한 사회적 맥락에서 특정 자연인들이나 그 집단에 대한 불리한 대우를 초래하거나, 그들의 사회적 행동이나 그 심각성에 비해 정당화되지 않거나 비례하지 않는 불리한 대우를 초래하는 경우이다.

270 EU AI 법안(2023. 5. 9. 수정안) 제5조 제1항 (d)에 따라 금지되는 범죄 위험성 예측 시스템은 자연인에 대한 프로파일링, 개인의 위치정보 등 성격 특성이나 특징, 또는 자연인이나 자연인 집단의 과거 범죄 행동에 기반하여 자연인이나 그 집단의 범죄 또는 재범 위험성을 평가하거나, 실제 또는 잠재적으로 범죄 또는 재범이 일어날 것인지를 예측하는 경우이다.

271 EU AI 법안(2023. 5. 9. 수정안) 전문 제17조.

272 EU AI 법안(2023. 5. 9. 수정안) 전문 제18조.

스템을 활용한 결과 오류가 있거나 편향된 결정을 내림으로써 기본권이 침해될 위험을 최소화하기 위한 것이다.[273]

이러한 목적에서 EU AI 법안은 고위험 인공지능 시스템의 학습 데이터 집합에 있어 편향이 존재하는지 여부를 조사하고, 가능한 편향을 탐지·예방·완화하기 위해 적절한 조치를 취할 것을 요구한다.[274] 여기서 조사 대상이 되는 편향은 개인의 건강이나 안전에 영향을 끼칠 수 있거나, 기본권에 부정적인 영향을 끼칠 수 있거나, EU 법상 금지되는 차별을 초래할 수 있는 경우를 의미하는데, 여기에는 특히 데이터 출력이 다시 미래 작동을 위한 입력에 영향을 끼치는 이른바 "피드백 루프(feedback loop)"가 형성되는 경우가 포함된다.[275] 나아가 고위험 인공지능 시스템이 자연인에 의한 의사 결정을 위한 정보를 제공하거나 추천을 하는 경우 인간이 그 출력에 자동적으로 따르거나 과잉 의존하게 되는 경향이 있는지 파악하는 인간 감독의무도 부과된다.[276]

고위험 인공지능에 대한 사전적 평가, 위험 관리 및 인간 감독에 있어서는 여러 기술 표준이 활용될 것으로 전망된다.[277] 즉, 법률 규정은 일반적으로 높은 수준에서 핵심적 의무를 규정하고, 인공지능 시스템 제공자들이 달성해야 할 기술적 목표를 설정하는 역할을 한다. 나아가, 기술 표준은 행위자들이 관련 법률 요구사항을 준수할 수 있도록 상세한 기술 사양(specification)을 제공하게 된다. 행위자들은 이러한 기술 표준에 구속되는 것은 아니고 다른 기술적 해법을 사용하여 법률 준수를 입증하는 것이 허용된다. 그러나 이러한 신뢰할 수 있는 기술 표준을 활용함으로써 비용 효율적으로 규제 준수를 수행할 달성할 수 있다는 장점이 발생할 수 있다. 이러한 점을 고려하여 EU AI 법안은 고위험 인공지능의 제공자가 기술 표준을 준수할 경우 해당 의무를 이행한 것으로 추정하도록 정하

---

273 EU AI 법안 전문 제33 내지 40조.

274 EU AI 법안(2023. 5. 9. 수정안) 제10조 제2항.

275 EU AI 법안(2023. 5. 9. 수정안) 제10조 제2항 (f).

276 EU AI 법안 제14조 제4항 (b).

277 Nativi and De Nigris (2021) (각주 13), p. 7.

고 있다.[278] 나아가 EU 집행위원회는 유럽연합의 표준기술 제정 기구에 대해 인공지능 신뢰성과 안전성에 관한 기술표준을 제정할 것을 요청하도록 정하고 있는 바,[279] 이에 EU 집행위원회는 2022년 12월 유럽 표준화 위원회(European Committee for Standardization, "CEN") 및 유럽 전기기술 표준화 위원회(European Committee for Electrotechnical Standardization. "CENELEC")에 2025년 1월까지 안전하고 신뢰가능한 인공지능 기술을 위한 유럽 표준을 마련할 것을 요청하는 표준화 요청서 초안을 발표하였다.[280] EU 집행위원회가 요청한 기술 표준은 고위험 인공지능에 대한 의무사항과 관련된 것으로서, 구체적으로 인공지능 시스템에 대한 위험관리 시스템, 데이터 집합의 거버넌스와 품질, 로깅 능력(logging capabilities)을 통한 기록 보관, 이용자에 대한 투명성 및 정보 제공, 인간 감독, 정확도 사양, 강건성 사양, 사이버 보안 사양, 출시 후 모니터링 절차를 포함한 품질 관리 시스템, 적합성 평가에 관한 것이 있다.[281] EU 집행위원회의 내부 승인을 거쳐 위와 같은 표준화 요청이 공식적으로 이루어지면, 2025년 경까지 기술 표준이 마련될 것으로 전망된다.[282]

2) 인지모형

EU AI 법안상 인지모형에 의한 위험성에 대한 규율은 (1) 생명·신체의 안전과 관련되어 사용되는 경우 (2) 생체인식(biometric) 시스템의 경우, (3) 인간의 감정 인식을 위해 사용되는 경우로 구분하여 검토할 수 있다.

---

278 EU AI 법안(2023. 5. 9. 수정안) 제40조 제1항.

279 EU AI 법안(2023. 5. 9. 수정안) 제40조 제1a항.

280 European Commission, *Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence*, (2022. 12. 5.).

281 위의 글, 부속서 I, 표 1.

282 Hadrien Pouget, "Standard Setting", The Artificial Intelligence Act (webstite), https://artificialintelligenceact.eu/standard-setting/ (2023. 5. 26. 접속 확인).

우선, 인지모형이 제품의 안전과 관련되어 있는 경우는 고위험으로 분류될 수 있다. 현재 다수의 물리적 장치들이 카메라나 센서를 통해 수집된 정보를 인지모형을 통해 인식하고 위치를 측위하여, 그에 따른 판단과 제어 단계를 거쳐 작동한다.[283] 만약 인지모형이 환경을 부정확하게 인식할 경우 그 이용자나 주변인의 생명·신체에 심각한 위해를 초래할 수 있으므로, 이러한 인지모형을 활용한 제품이 시장에 출시되기 전 그에 대한 안전성 평가가 요구된다. 그런데, 완구, 레저기구, 승강기, 폭발물, 고압기구, 케이블, 의료기기, 민간 항공 안전, 차량, 농업·임업용구, 해상기구, 철도 등과 같이 안전상 위험이 있는 제품에 대해서는 기존 유럽연합 규제가 존재하고 있다. 이에 EU AI 법안은 이러한 기존의 안전 규제 대상에 포함된 제품에 인공지능이 안전과 관련하여 활용되는 경우 고위험으로 분류하여,[284] 고위험 인공지능에 대한 의무사항의 이행 및 적합성 평가가 요구하고 있다.

다음으로, 공중이 접근 가능한 장소에서의 원격 생체인식 신원확인(identification) 시스템의 활용은 원칙적으로 수용불가능한 위험으로 분류된다. EU AI 법안 수정안이 원격 생체인식 신원확인 시스템을 이처럼 엄격히 규제하는 이유는 다수 대중의 사생활에 중대한 영향을 끼치고, 끊임없는 감시에 대한 인식을 불러 일으키며, 생체인식 기술을 배치하는 당사자에게 통제 불가능한 권력을 부여하고, 나아가 간접적으로 집회의 자유나 법치주의에 있어 핵심적인 기본권의 행사를 제한할 수 있기 때문이라고 설명한다.[285] 한편, 생체인식 신원확인 시스템의 활용은 '실시간(real-time)'인 경우와 '사후적(post)'인 경우로 구분되는데, (i) 실시간인 경우에는 항상 금지되고, (ii) 사후적으로 녹화 기록을 확인하는 것은 중대한 범죄와 관련된 수색을 위해 엄격히 필요한 경우에 한하여 사법부에 의한

---

283 정석우·심현철, "자율주행 자동차의 인공지능", 기계저널 57(3), (2017).

284 EU AI 법안(2023. 5. 9. 수정안) 제6조 제1항 및 부속서 II 참조.

285 EU AI 법안(2023. 5. 9. 수정안) 전문 제18조. 유럽 의회 수정안은 EU 집행위원회보다 공중이 접근 가능한 장소에서의 생체인식 신원확인에 대한 규제를 대폭 강화하였다.

사전 승인을 얻어 활용하는 것이 예외적으로 허용된다.[286] 그뿐만 아니라 인터넷이나 CCTV 영상으로부터 무작위로 얼굴 이미지를 스크래핑하여 얼굴 인식 데이터베이스를 생성하거나 확장하는 인공지능 시스템 역시 수용불가능한 위험에 해당한다.[287] 나아가, 생체인식에 기반하여 자연인을 분류하는 인공지능 시스템에 있어 민감 속성에 기반하여 자연인을 분류하는 경우 수용불가능한 위험에 해당하여 금지된다.[288] 다만, 위 경우에 해당하지 않는 자연인의 생체인식 신원확인 시스템은 고위험으로 분류된다.[289]

나아가, 2023. 5. 9. 수정안은 초안에 비하여 감정 인식 시스템에 관한 규제를 강화하였다. 이는 감정인식 인공지능 시스템의 과학적 근거에 대해 심각한 우려가 제기되고 있고, 특히 그러한 기술의 신뢰성이 제한적이고, 특이성이 부족하며, 일반화되기 어려운 한계가 있기 때문이다.[290] 이에 EU AI 법안은 (i) 법 집행이나 국경 관리 및 작업장이나 교육시설에 있어서 자연인의 감정을 추론하는 인공지능 시스템은 수용불가능한 위험으로 분류하고,[291] (ii) 그 이외에 생체인식에 기반하여 개인의 특성을 추론하는데 활용되는 인공지능 시스템은 고위험으로 분류된다.

마지막으로 위와 같이 규정된 경우에 해당하지 않는 생체인식 분류 시스템이나 감정인식 시스템은 제한적 위험에 해당하여, 투명성 의무만을 부여된다. 이때

---

286 EU AI 법안(2023. 5. 9. 수정안) 제6조 제1항 (d) 및 (e).

287 EU AI 법안(2023. 5. 9. 수정안) 제5조 제1항 (db).

288 EU AI 법안(2023. 5. 9. 수정안) 제5조 제1항 (ba).

289 EU AI 법안(2023. 5. 9. 수정안) 부속서 III 제1항 (a).

290 EU AI 법안(2023. 5. 9. 수정안) 전문 제26c조. EU AI 법안(2023. 5. 9. 수정안) 제3조 제34항은 "감정 인식 시스템(emotion recognition system)"을 생체인식 또는 생체인식에 기반한 데이터를 활용하여 개인이나 집단의 감정, 사상, 심리 상태 또는 의도를 식별하거나 추론하는 목적을 가진 인공지능 시스템으로 정의하고 있다.

291 EU AI 법안(2023. 5. 9. 수정안) 제5조 제1항 (dc).

에는 해당 시스템에 노출되는 자연인에게 해당 시스템의 작동 사실을 고지할 의무가 있다.[292]

3) 생성모형

EU AI 법안의 2021년 초안에 따르면 생성모형은 제한적 위험으로 분류되어 투명성 의무만이 부과되었다. 즉, (1) 자연인과의 상호작용할 것으로 의도된 인공지능 시스템은 그 이용자가 인공지능 시스템과 상호작용하고 있다는 사실을 알 수 있도록 설계 및 개발되어야 하고,[293] (2) 존재하는 사람, 물체, 장소 기타 단체나 사건과 현저히 유사하게 보이면서 사람에게 진정하거나 진실된 것처럼 허위로 표시된 이미지, 오디오, 비디오 콘텐츠, 즉 "딥페이크(deep fake)"를 생성하거나 조작하는 인공지능 시스템의 경우 그 이용자는 해당 콘텐츠가 인공적으로 생성되었거나 조작되었다는 사실을 공개하여야 한다.[294]

그런데 EU AI 법안이 제안된 이후 생성모형(특히 트랜스포머 모형에 기반한 대규모 언어모형)이 빠르게 발전하면서, 단순히 '그럴 듯한' 이미지나 텍스트를 생성하는 것에서 나아가 다양한 하류 작업을 수행하기 위하여 사용될 수 있음이 밝혀졌다. GPT-3/4나 PaLM과 같이 방대한 데이터에 의해 학습되어 다양한 과제에 적응될 수 있는 인공지능을 이른바 '기반모형'이라 하는데, 이러한 유형의 인공지능에 대한 추가적인 규율이 필요하다는 주장이 제기되었다. 이에 수정된 법안 초안은 '기반모형(foundation model)'에 대한 정의를 추가하여, 기반모형 제공자에 대해서는 추가적 의무를 부과하고 있다.[295]

---

292 EU AI 법안 제52조 제2항.

293 EU AI 법안 제52조 제1항. 다만, 이러한 의무는 상황과 사용 맥락에서 명백한 경우에는 적용되지 않는다.

294 EU AI 법안 제52조 제3항. 다만 이러한 의무는 범죄 탐지, 예방, 조사, 기소를 위해 법으로 허용된 경우 또는 표현의 자유나 예술과 과학의 자유를 행사하는 데 필요한 경우로서 제3자의 권리와 자유에 대한 적절한 보호조치가 존재하는 경우에는 적용되지 않는다.

295 EU AI 법안 상 기반모형 제공자의 의무에 관하여는 본 연구 제6장 제4절에서 검토한다.

해당 초안은 특히 다양한 자율성 수준에서 복잡한 텍스트, 이미지, 오디오 또는 비디오와 같은 콘텐츠를 생성할 수 있도록 의도된 인공지능 시스템을 "생성 인공지능"이라 정의한다. 나아가, 생성 인공지능 시스템에서 기반모형이 사용되는 경우, 해당 기반모형의 제공자 등에게 앞서 설명한 투명성 의무 이외에 추가적으로 (1) 유럽연합법을 위반하는 콘텐츠가 생성되지 않도록 막는 보호조치 보장 의무[296]와 (2) 저작권법에 의해 보호되는 학습 데이터에 관한 상세한 요지 작성 및 공개 의무를 부과하고 있다.[297] 이러한 위법 콘텐츠에 대한 보호조치의 구체적인 범위는 확실하지 않다. 또한 학습 데이터에 포함된 창작물이 저작권 보호 대상이 되는지를 확인하기 어려우므로, 이러한 저작물의 추적, 문서화 및 공개 작업이 어떻게 이루어질 수 있을 것인지에 대해서도 의문이 제기되고 있다.[298]

2. 미국

가. 연방 차원

미국의 인공지능 규제에 대한 태도는 유럽과는 상당한 차이를 보인다.[299] 가장 눈에 띄는 차이는 앞서 본 바와 같이 EU AI 법안이 판별모형, 인지모형, 생성

---

296 EU AI 법안(2023. 5. 9. 수정안) 제28b조 제4항 (b)에 따르면 생성모형 기반모형의 제공자나 기반모형으로 인공지능 시스템으로 전문화하는 제공자는 일반적으로 인정된 기술 수준에 따라, 표현의 자유 등 기본권을 침해하지 않는 범위 내에서 유럽연합법을 위반하는 콘텐츠를 생성하지 않도록 하는 보호조치를 보장하도록 기반모형을 학습, 설계 및 개발해야 할 의무가 있다.

297 EU AI 법안(2023. 5. 9. 수정안) 제28b조 제4항 b)에 따르면 이들은 또한 각국이나 유럽연합의 저작권에 관한 법률을 침해하지 않는 범위 내에서, 저작권법에 따라 보호되는 학습 데이터 사용에 관한 충분히 상세한 요지를 문서화하고 공개할 의무가 있다.

298 Freshfields Bruckhaus Deringer, "Has copyright caught up with the AI Act?", Lexology Blog, (2023. 5. 16.).

299 Alex Engler, "The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment", Brookings, (2023. 4. 25.).

모형을 모두 포괄하여 인공지능 기술 전반을 규제하고자 하는 것에 비해, 미국에서 진행되고 있는 규제 시도는 구체적인 특정 영역에 한정되어 있고, 상당수의 사례에 있어 판별모형의 활용 사례에 집중하고 있다는 점이다.[300]

미국 연방 정부 차원에서의 인공지능에 대한 감독 방안을 최초로 제시한 것은 2019. 2. 11. 발표된 행정명령 13859호("인공지능에서 미국의 선도적 위치 유지")[301]와 위 행정명령에 따라 관리예산국(the Office of Management and Budget, OMB)이 각 부처에 보낸 지침(M-21-06)이다.[302] 위 지침은 위험기반 접근법에 따라 각각의 연방 정부기관이 인공지능 활용에 따른 위험을 평가하고, 규제적 수단 또는 비규제적 수단에 의한 인공지능 위험 감소 방안을 고려하여, 인공지능 활용에 대한 규제 계획을 개발 것을 요구하였다. 위 행정명령과 지침은 트럼프 행정부 하에서 발표된 것으로서, 규제 최소주의적(minimalist) 관점에 기반한 것이라 할 수 있다. 그러나, 대부분의 연방정부 부처는 위 지침에서 요구한 인공지능 규제 계획을 개발하지 않았다.[303]

바이든 행정부에 들어서는 인공지능 위험 문제에 관해 새로운 접근을 취하여, 2022년 10월 "인공지능 권리장전을 위한 청사진(Blueprint for an AI Bill of Rights)"을 발표하였다.[304] 위 청사진은 인공지능 권리장전의 5대 원칙을 제시하였는데, 이는 다음과 같다.

---

300 이는 개인정보 보호법제에서도 EU에서는 개인정보 보호지침이나 GDPR 등을 통해 포괄적으로 규제하고 있는 것에 비해, 미국에서는 전통적으로 개별 영역별 규제가 중심이 되어 온 것과 유사한 것으로 볼 수 있다. Paul M. Schwartz and Karl-Nikolaus Peifer. "Transatlantic data privacy law", *Georgetown Law Journal* 106, (2017).

301 행정명령 13859호의 주요 내용에 관해서는 세계법제정보센터, "미국 인공지능 법제", 연구보고서 (2019. 6. 30.) 참조.

302 Office of Management and Budget (OMB), "Guidance for Regulation of Artificial Intelligence Applications", (2020. 11. 17.).

303 Engler (2023. 4. 25.) (각주 299)에 따르면 총 41개의 주요 정부 부처 중 오직 5곳만이 인공지능 규제 계획을 수립하였고, 그 중에서도 충실한 계획을 수립한 부처는 보건복지부(Department of Health and Human Services)뿐이라 한다.

304 The White House, "Blueprint for an AI Bill of Rights" (2022).

- **안전하고 효과적인 시스템**: “누구나 안전하지 않거나 효과적이지 않은 시스템으로 보호받아야 한다.”
- **알고리즘 차별 보호**: “누구나 알고리즘에 의해 차별받지 않아야 하고 시스템은 공평한 방식으로 설계되고 사용되어야 한다.”
- **데이터 프라이버시**: “누구나 개인정보를 남용하는 것으로부터 기본설정에 따라 보호받아야 하고, 자신에 관한 개인정보가 어떻게 사용되는지에 관해 통제권을 가져야 한다.”
- **고지 및 설명**: “누구나 자동화된 시스템이 사용되고 있다는 사실 및 그것이 어떻게 그리고 어떠한 이유로 자신에게 영향을 끼치는 결과를 낳는지를 알 수 있어야 한다.”
- **인간 개입, 고려 및 대안**: “누구나 적절한 경우 옵트-아웃(opt-out)을 할 수 있고, 당면한 문제를 신속히 고려하고 구제할 수 있는 인간에게 접근할 수 있어야 한다.”

그러나 위 청사진은 EU식 접근과 달리 미국 연방 차원에서의 포괄적인 인공지능 규제 방안을 고려하지 않는다. 그 대신 각각의 원칙을 준수하기 위한 정책 수단은 보건, 노동, 교육 등과 같은 개별 영역 및 맥락마다 맞춤화될 것이 요구된다.[305] 이처럼 현재 미국 연방정부의 인공지능 위험에 대한 정책은 개별 정부기관마다 자발적으로 인공지능 규제를 고려하도록 하고 있다는 점에서 고도로 분산된(highly distributed) 것으로 평가될 수 있다.

다른 한편, 연방 의회에도 인공지능 규제 법안이 발의되어 왔는데, 대표적으로 ‘2022년 알고리즘 책무성 법안(Algorithmic Accountability Act of 2022)’이 있다. 이 법안은 ‘자동화된 의사결정 시스템(automated decision system)’와 이를 활용하여 ‘중대한 결정(critical decision)’을 내리는 의사결정 프로세스

---

305 위의 글, 14면. Engler (2023. 4. 25.) (각주 299).

(augmented critical decision process)를 규율하기 위한 것이다.[306] 여기서 '중대한 결정'이란 교육, 고용, 필수 설비, 가족 계획, 금융 서비스, 보건, 주거 및 법률 서비스 등에 대한 접근, 비용, 조건 및 이용 가능성에 관련하여 소비자의 삶에 대해 법적이거나 중대하거나, 유사하게 유의미한 영향을 끼칠 수 있는 결정이나 판단을 뜻한다.[307]

그런데, 위 법안은 위 영역에서 인공지능을 활용하는 사업자에 대해 곧바로 어떠한 의무를 부과하는 것이 아니라, 미국 연방거래위원회(Federal Trade Commission, FTC)가 이를 규율하는 규정을 제정할 것을 요구한다. FTC가 제정할 규정은 핵심적으로 사업자에게 '영향 평가(impact assessment)'를 수행할 것을 요구하여야 한다.[308] 나아가 위 규정의 수범 대상은 연방거래위원회법상 FTC의 관할 대상이 되는 개인이나 기업으로서 일정한 매출액이나 자본 또는 이용자 수 등의 기준을 초과하는 경우로 한정된다. 위 법안은 내부 영향 평가의 구체적 요건을 정하고 있는데, 영향 평가에는 기존 프로세스에 대한 설명, 입력 데이터에 대한 문서화, 프라이버시 위험의 시험·평가, 성능에 대한 시험·평가, 이해관계자의 참여에 대한 정보, 소비자 권리에 대한 평가, 소비자에 대해 발생할 수 있는 중대한 부정적 영향의 식별과 적용가능한 완화조치 전략에 대한 평가를 포함하여야 한다.[309] 이러한 영향 평가의 구체적인 요구사항에 관해서는 FTC가 가이드라인을 제정하도록 되어 있다.

---

306 ADS는 "계산 결과가 어떠한 결정이나 판단의 근거로 사용되는, 계산을 사용하는 모든 시스템, 소프트웨어 또는 프로세스(머신러닝, 통계 또는 다른 데이터 처리나 인공지능 기법에서 유래하는 것을 포함하나, 수동적인 계산 인프라는 제외)"라 정의된다(법안 Section 2(2)). 나아가 ACDP는 중대한 결정을 내리기 위해 자동화된 의사결정 시스템을 적용하는 과정, 절차 및 기타 활동을 지칭한다.

307 2022년 알고리즘 책임성 법안은 위 8개 분야를 나열하고 있고, 그 이외에도 미국 연방거래위원회가 규칙으로 다른 분야를 포함시킬 수 있도록 정하고 있다.

308 2022년 알고리즘 책임성 법안 Section 3.

309 2022년 알고리즘 책임성 법안 Section 4.

이처럼 미국연방 2022년 알고리즘 책무성 법안은 EU AI 법안에 비하여 상당히 좁은 적용 범위와 제한된 의무만을 규정하고 있다. 즉, 위 법안은 EU AI 법안과 달리 제품의 안전과 관련된 인공지능 활용 사례나 그 외의 여러 인지모형, 생성모형 및 기반모형에 관한 규제는 포함하고 있지 않다. 또한 EU AI 법안은 모든 고위험 인공지능 제공자에 대한 적합성 평가가 강제되고, 그 구체적인 준수는 EU의 표준화 기술에서 정하는 기술 표준에 의하도록 하고 있음에 비해, 미국 법안은 대규모 사업자만을 수범대상으로 하고, 내부적으로 영향평가를 수행하여 그 요지를 FTC에 보고하도록 되어 있을 뿐이다.

나. 주 차원

주 차원에서도 연방 법안과 유사하게 자동화된 의사결정에 의한 차별을 규제하기 위한 법안이 발의되어 오고 있다. 대표적인 사례로 워싱턴 D.C.와 캘리포니아 주에서 발의된 법안이 있다. 우선 워싱턴 D.C에는 '2021년 알고리즘 차별금지법(Stop Discrimination by Algorithms Act of 2021)'이 2021년 12월 발의되었고, 2022년 9월 공청회를 마친 상태이다.[310] 이 법은 '알고리즘에 의한 자격 결정(algorithmic eligibility determination)'과 '알고리즘에 의한 정보제공 판단(algorithmic information availability determination)'에 있어서의 차별을 금지한다.[311] 위 법에서 정한 차별금지영역은 개인의 중요한 삶의 기회에 관한 자격 결정이나 정보 제공에 관한 것으로서, 이러한 기회는 고용, 보험, 신용, 주택 구매나 공공서비스 접근을 뜻한다. 또한 위 법이 정한 차별금지사유는 인종, 피

---

310 Stop Discrimination By Algorithms Act of 2021, D.C. Bill 24-558 (2021).

311 위 법안 Section 3은"알고리즘에 의한 자격 결정"을 "머신러닝, 인공지능 또는 유사한 기술을 활용하는 알고리즘 프로세스를 활용하여 전적으로 또는 상당 부분 개인이 중요한 삶의 기회에 대한 자격 또는 접근 기회를 결정하는 것"으로 정의하고, "알고리즘에 의한 정보제공 판단"을 "머신러닝, 인공지능 또는 유사한 기술을 활용하는 알고리즘 프로세스를 활용하여 전적으로 또는 상당 부분 개인이 중요한 삶의 기회에 관한 광고, 홍보, 유인 및 제안 수령 여부를 결정하는 것"으로 정의된다.

부색, 종교, 출신 국가, 성별, 성적 정체성 및 표현, 성적 지향, 가족 상태, 소득원 또는 장애를 포함한다. 결국 위 법안은 기존의 차별금지법을 인공지능을 활용한 알고리즘에 의한 판단이나 정보 제공 분야로 확장시킨 시도로 이해될 수 있다.

캘리포니아 주에서는 2023년 1월 "자동화된 의사결정 도구(automated decision tools)" 법안이 발의되었다.[312] 이 법안은 중대한 결정(consequential decision)[313]을 내리는 데 있어 인공지능이 활용하는 자동화된 의사결정도구를 사용하는 경우 영향 평가를 수행해야 하고, 이용자에 대한 인공지능 활용사실 등을 고지해야 하며, 알고리즘에 의한 차별 위험성을 막기 위한 적절한 안전조치를 취할 의무를 부과하고 있다. 캘리포니아 주의 법안은 영향 평가를 핵심적인 규제 수단으로 고려한다는 점에서 연방의 2022년 알고리즘 책무성 법안과 유사하지만, 알고리즘 차별을 낳는 자동화된 의사결정 도구의 사용 금지 의무를 부과하고 있을 뿐만 아니라, 알고리즘 차별로 인해 피해를 입은 개인이 손해배상 소송을 제기하는 것을 허용하고 있다는 점에서 보다 강력한 규제로 평가할 수 있다.[314]

다른 한편, 미국의 주 차원에서는 개별적인 인공지능 활용 영역을 대상으로 한 법률이 제정되거나, 법안이 발의되고 있다. 주된 규제 대상이 되고 있는 영역으로는 채용 분야, 법집행기관에 의한 얼굴인식 인공지능 활용 등이 있다. 채용 분야 인공지능을 규제하는 최초의 주법은 일리노이 주의 '인공지능 영상면접 법(Artificial Intelligence Video Interview Act)'이다.[315] 이 법은 고용주가 구직자의 영상 면접을 활용하는 경우 인공지능 활용 사실의 고지, 영상 공유의 제한 및

---

312 Cal. Assemb. B. 331, 2023-2024 Reg. Sess. (Cal. 2023).

313 위 법안상 중대한 결정의 범위는 2022년 알고리즘 책무성 법안의 중대한 결정(critical decision)의 규정과 유사하지만, 주거, 형사 사법 시스템, 투표 및 복지 서비스 등을 추가로 포함하여 확장하고 있다.

314 Sorelle Friedler, Suresh Venkatasubramanian, and Alex Engler, "How California and other states are tackling AI legislation", Brookings, (2023. 3. 22.).

315 Illinois Public Act 102-0047

영상 파기 의무 등을 부담한다. 그 외에도 메릴랜드 주는 채용에서의 안면인식 기술 활용을 제한하는 규정을 도입했다.[316] 또한, 법집행기관이나 공공기관에 의한 얼굴인식 인공지능을 활용에 대한 규제도 이루어지고 있는데, 앨라배마 주에서는 법집행기관이 얼굴인식 인공지능을 활용하여 범죄 수사 또는 체포의 근거가 되는 피의사실을 소명하는 것을 금지하였고,[317] 콜로라도 주에서는 정부기관이 얼굴인식 서비스를 사용하고자 하는 경우 사전 고지 및 책무성 보고서 작성 의무를 부과하고, 법집행기관이 영장 없이 얼굴인식 서비스를 계속적 감시, 실시간 인식 및 지속적 추적 목적으로 사용하는 것을 금지하며, 모든 공립학교에서의 얼굴인식 서비스 사용을 금지하였다.[318]

3. 국내 입법 동향

국내에서는 2021년 이후 인공지능 기술의 위험성을 규제하기 위한 법안이 다수 발의되어 오고 있는데, 대표적으로 2021. 7. 1. 발의된 『인공지능 육성 및 신뢰 기반 조성 등에 관한 법률안』[319], 2021. 11. 24. 발의된 『알고리즘 및 인공지능에 관한 법률안』[320], 2022. 12. 7. 발의된 『인공지능산업 육성 및 신뢰 확보에 관한 법률안』[321] 등이 있다.[322] 나아가 국회 법과학기술정보통신위원회 법안소위

---

316 최창수, “미국의 AI 채용절차 공정화 입법례”, *최신외국입법정보* 2023-1호, 국회도서관, (2023. 1. 31.).

317 Alabama Act No. 2022-420.

318 Colorado SB 22-113

319 정필모 의원 대표발의, 인공지능 육성 및 신뢰 기반 조성 등에 관한 법률안, 의안번호 2111261, (2021. 7. 1.).

320 윤영찬 의원 대표발의, 알고리즘 및 인공지능에 관한 법률안, 의안번호 2113509, (2021. 11. 24.).

321 윤두현 의원 대표발의, 인공지능산업 육성 및 신뢰 확보에 관한 법률안, 의안번호 2118726, (2022. 12. 7.).

322 각 법안에 대한 소개와 비교로는 김광수, 인공지능 알고리즘 규율을 위한 법제 동향, *행정법연구* 70, (2023) 182-185면 참조.

원회에서는 2023. 2. 14. 위 세 가지 법안을 중심으로 통합한『인공지능산업 육성 및 신뢰 기반 조성 등에 관한 법률안』처리에 합의하였다.[323] 합의된 법안은 고위험영역 인공지능을 정의하고, 고위험 영역 인공지능의 개발자 또는 이를 사용한 제품·서비스 제공자에게 사전 고지의무와 신뢰성과 안전성 확보 조치 의무를 부과하는 것으로 알려져 있다.[324] 이는 영역별 규제 또는 판별모형에 대한 규제만을 포함하는 미국 법(안)들과는 구별되고, EU AI 법안과 더욱 유사하게 판별모형, 인지모형 및 생성모형 중 높은 위험을 초래할 수 있는 활용 사례를 포괄적으로 규율하고자 하는 시도에 가까운 것으로 평가할 수 있다.

---

323 본 연구 작성일 현재 합의된 법안은 공개되지 않았고, 언론보도 등을 통해 그 내용의 일부가 알려져 있을 뿐이고, 지속적인 검토와 수정이 이루어지고 있는 것으로 보인다. 위 법안은 인공지능에 대한 법적 정의를 포함하고, 인공지능의 개발과 이용과정에서 지켜야 할 기본원칙을 정하고 있는 것으로 알려져 있다. 또한 위 법에 따르면 정부는 인공지능 기본계획을 수립하고, 국무총리 소속의 인공지능 위원회를 설립해야 한다.

324 임형주 외, "AI(인공지능)산업 육성 및 신뢰 기반 조성에 관한 법률 제정 동향 및 사업자가 주의해야 할 사항", Yulchon Legal Update, (2023. 3.)에 따르면, 2023년 2월 합의된 법안상의 고위험 영역이란 "에너지, 보건의료산업, 의료기기, 원자력시설, 범죄수사나 체포 업무에서의 생체의 정보, 채용이나 대출 심사와 같이 개인의 권리의무에 중대한 영향을 미치는 판단과 평가목적의 인공지능, 자율주행을 포함한 교통수단, 교통시설, 교통체계의 주요한 작동 및 운영, 국가나 지자체, 공공기관등이 국민에게 영향을 미치는 의사결정을 위해 사용되는 인공지능과 같이 국민의 안전, 건강, 기본권 보호에 중대한 영향을 미칠 수 있는 사항"으로 정의된다고 한다. 나아가 고위험영역 인공지능 사업자가 취해야 하는 신뢰성 확보 조치의 범위는 과학정보통신부장관이 고시하고 이를 권고할 수 있도록 정하고 있는데, 여기에는 "1) 위험관리를 어떻게 할 것인가, 2) 신뢰성 확보를 확인할 수 있는 문서가 있는가, 3) 기술적으로 가능한 범위 내에서 인공지능의 최종결과물이나 주요 기준, 4) 이용자 보호, 5) 고위험영역의 인공지능을 운영하려는 사람에 대한 관리감독에 관한 사항을 포함"하여야 한다고 알려져 있다.

제5절 소결

이상에서 살펴본 바와 같이 인공지능의 공정성 문제는 여러 법제에 의해 규율되는 양태를 보이고 있다. 특히 여러 법제들이 공통적으로 기술적인 인공지능 공정성 지표를 활용하는 방향으로 발전하고 있다는 점에 주목할 필요가 있다. 우선 ① 차별금지법제에 따르면, 판별모형 인공지능을 개발, 활용하는 여러 사업자가 차별금지 법제를 준수하기 위해 공정성 지표를 활용하여 해당 인공지능에 대해 사전적 혹은 사후적 감사를 실시하고 있는 모습을 보인다. ② 개인정보법제에 따르면, 개인정보처리자는 개인정보보호 영향평가(DPIA)를 수행함에 있어 정보주체에 대해 발생할 수 있는 차별 위험성을 측정하고 이를 예방하기 위하여 인공지능 공정성 지표를 활용할 필요가 있다. ③ 인공지능 규제법제 중 특히 EU AI 법안은 고위험 인공지능 제공자 등에게 부과되는 사전적 의무 이행을 위해 기술 표준을 활용하는 방향으로 접근해 오고 있다. 이러한 기술 표준은 공정성 지표를 정의하고, 공정성 심사 방법을 명확히 할 것이다. 또한 미국의 연방 또는 주 차원에서 이루어지는 다수의 입법 시도에서는 인공지능에 대한 '영향평가'가 핵심적 규제 수단으로 고려되고 있는 바, 이러한 영향평가에서도 공정성 지표가 중요하게 고려될 것이다.

이러한 경향을 종합해 볼 때, 향후 인공지능 공정성 문제에 대한 법적 규율에 있어 공정성 지표가 중요할 역할을 수행할 것으로 전망된다. 이에 다음 장에서는 이제까지 제안된 주요한 인공지능 공정성 지표는 어떠한 것들이 검토한다.

제4장 분류모형 인공지능의 공정성 지표

제 1 절 개관

본 장은 분류모형 인공지능에 대한 공정성 지표를 개관한다. 최근까지 여러 공정성 지표가 제안되어 왔는데 그 수가 적지 않다. 예를 들어, Narayanan (2018)은 당시까지 제안된 인공지능 공정성에 대한 정의가 21가지에 이른다고 소개하고 있다.[325] 하지만 모든 기준들이 실무상 같은 중요성을 가지고 활용되고 있는 것은 아니므로, 주요하게 고려되는 기준을 중심으로 유형화하여 이해하는 것도 가능하다. 이에 이하에서는 Barocas, Hardt & Narayanan가 제시한 구분법을 주로 참조하여 유형화한다.[326] 이는 ICO 인공지능 감사 가이드라인 초안[327] 또는 싱가포르 통화청의 인공지능 공정성 사례 보고서[328]에서 사용된 구분법과 유사하다.

분류모형 인공지능 공정성 지표에 대한 접근법은 크게 인공지능 시스템의 입력변수에 대한 심사와 출력변수에 대한 심사로 구분할 수 있다. (1) 입력변수에 대한 공정성 심사는 인공지능 시스템이 어떠한 변수를 사용하는지 여부를 살펴보는 것이다. 이러한 접근법은 인공지능 시스템의 입력 변수에서 차별적 속성을 제거하거나, 시스템이 이를 고려하지 않도록 함으로써, 그 시스템이 차별적으로

---

325 Narayanan (2018. 2. 23.) (각주 19). 한편, Barocas, Hardt & Narayanan (2022) (각주 19)에 소개된 공정성 지표는 모두 19가지이다.

326 Barocas, Hardt & Narayanan (2022) (각주 19)은 최근 수년 간 논의된 다양한 AI 공정성 심사 기준을 체계화하여 정리하고 있다.

327 ICO AI 감사 가이드라인 초안 (2020) (각주 16).

328 MAS (2021) (각주 17).

작동하는 것을 막고자 한다. 본 장 제2절은 입력변수 제한 접근법에 따른 인공지능 공정성 기준의 의의 및 한계에 관해 살핀다.

다음으로 (2) 출력변수에 대한 공정성 심사는 주로 인공지능이 예측하는 목표변수가 집단에 대해 불균등한 지 여부를 살펴보는 것이다. 출력변수에 대한 공정성 심사 기준 중 대표적인 것으로는 ① 결과 동등성(outcome parity), ② 오류율 동등성(error parity), ③ 예측도 동등성(predictive value parity)이 있다.[329] 위 공정성 지표들은 모두 어떤 분류모형이 특정 집단에 대해 차별적으로 작동하는지 여부를 평가한다.

예를 들어, 채용에 사용되는 분류모형이 A 집단과 B 집단에 대해 차별적 작동하는지 여부를 평가하기 위하여 결과 동등성 기준을 활용하는 경우를 상정하자. ① 결과 동등성 기준은 A 집단과 B 집단의 채용 비율이 동등한지 여부를 평가한다. 이는 직관적이고 간명한 방법이다. 다음으로, ② 오류율 동등성 기준과 ③ 예측도 동등성 기준은 모두 결과의 '평가 데이터 집합(test dataset)'의 존재를 전제한다. 즉, 인공지능의 학습에 사용되지 않은 평가 데이터를 활용하여 인공지능의 예측이 평가 데이터상의 "정답(ground truth)"과 얼마나 오차가 나는지 확인하는 것이다. 위 두 척도는 모두 그러한 오차 정도가 집단간 동등하거나, 적어도 일정 범위 이내일 것을 요구하는데, 그 중 (i) 오류율 동등성은 위양성율(false positive rate)와 위음성율(false negative rate)이라는 통계적 정확성 지표가 집단간 동등한 것을 요구하는 것이고, (ii) 예측도 동등성은 양성예측도(positive predictive value, PPV)와 음성예측도(negative predictive value, NPV)라는 통계적 정확성 지표가 집단간 동등할 것을 요구한다. 여기서 집단간 비교에 사용되는 통계적 정확성 지표는 각 맥락에 있어 서로 다른 함의를 갖는데, 주어진 맥락에 적합한 지표를 선택할 필요가 있다.

---

329 ICO AI 감사 가이드라인 초안 (2020) (각주 16), p. 55.

제 2 절 입력변수에 대한 공정성 기준

1. 의의

가. 정의

입력변수에 대한 공정성 기준은 인공지능 시스템의 입력 변수가 명시적으로 차별적 속성을 사용하지 않도록 하는 것이다. 이를 '비인식을 통한 공정성(fairness through unawareness)', '집단 비인식(group unawareness)' 또는 줄여서 '비인식(unawareness)' 기준이라고도 한다. 연구자에 따라서는 '블라인드(blindness)' 기준 또는 '과정 공정성(process fairness)'이라 표현하기도 한다. 이하에서는 'unawareness 기준' 또는 '블라인드 기준'이라 부른다.

이를 수학적으로 정식화하여 표현하면 다음과 같다.[330]

(정의) Unawareness 기준: 어떤 분류모형 $f_{classfier} : X \mapsto \hat{Y}$ 에 대해, f 가 $\hat{Y}$ 을 예측하기 위하여 보호속성들 A 를 사용하지 않는 경우 $f_{classifier}$ 는 unawareness 기준을 충족한다.

Unawareness 기준은 차별금지법제 상의 직접차별 금지 원칙(의도적 차별행위)에 대응하는 것으로 흔히 여겨진다.[331] 만약 인공지능이 차별금지사유를 입력변수로 사용하여 해당 집단에 대해 불리한 결정을 내릴 경우 직접차별에 해당할 수 있을 것이고, unawareness 기준은 이러한 상황을 방지하는 것을 목적으로 한다.

---

330 Boris van Breugel, et al., "DECAF: Generating fair synthetic data using causally-aware generative networks", *Advances in Neural Information Processing Systems 34*, (2021).

331 Id., Appendix C.

또한 입력변수에 대한 제한은 개인정보 보호법제에 의해서도 이루어진다.[332] 개인정보 보호법제 중에는 노동, 의료, 금융 등의 다수 분야에 있어 특정한 개인정보가 의사결정자에게 제공되는 것을 차단하는 경우가 적지 않다. 차별금지법제는 의사결정자가 특정한 정보를 어떻게 사용하는지를 중심으로 규율하는 것인데 비해, 개인정보 보호법제는 차별적으로 활용될 수 있는 개인정보의 수집 자체를 금지함으로써 차별을 예방하는 기능을 수행할 수 있다.[333] Cofone (2019)는 개인정보 보호법제가 차별을 예방하는 역할을 수행하기 위해서는 개인정보의 흐름이 어떻게 이루어질 것인지를 결정하는 것이 핵심적이라고 주장한다.

나. 보호 속성과 차별금지사유

Unawareness 기준을 적용할 때 우선적으로 고려해야 할 사항은 인공지능이 입력변수로 활용해서는 안 될 보호속성의 범위를 결정하는 것이다. 보호속성의 범위를 결정하는 일차적 방법은 법에 명시된 기준을 사용하는 것이다. 예컨대 헌법상 차별금지사유인 "성별·종교 또는 사회적 신분"이 있고, 개별 차별금지법에서 규정된 차별금지사유도 존재한다. 국내법상 가장 포괄적으로 차별금지사유를 규정하고 있는 국가인권위원회법은 "성별, 종교, 장애, 나이, 사회적 신분, 출신 지역(출생지, 등록기준지, 성년이 되기 전의 주된 거주지 등을 말한다), 출신 국가, 출신 민족, 용모 등 신체 조건, 기혼 · 미혼 · 별거 · 이혼 · 사별 · 재혼 · 사실혼 등 혼인 여부, 임신 또는 출산, 가족 형태 또는 가족 상황, 인종, 피부색, 사상 또는 정치적 의견, 형의 효력이 실효된 전과(前科), 성적(性的) 지향, 학력, 병력(病歷)"을 포함하고 있다.[334]

---

332 Cofone, (2019a) (각주 177), p. 1406.

333 위의 글, p. 1407.

334 국가인권위원회법 제2조 제3호.

하지만, unawareness 기준에 적용되는 보호속성의 범위가 법에서 금지된 차별금지사유의 범위와 반드시 일치해야 하는 것은 아니다. 인공지능이 활용되는 맥락에 따라서는 특정 정보가 보호속성에 포함될 수도 있고, 그렇지 않을 수도 있다. 예컨대 형의 효력이 실효된 전과는 일반적으로 보호속성에 포함된다고 여겨지지만, 재범 위험성을 예측하는 인공지능에 있어서는 보호속성에 해당하지 않을 수 있다.

보호 속성의 결정과 관련하여 제안된 흥미로운 방법으로는 일반인의 의견을 수렴하는 방안이 있다. Grgic-Hlaca et al. (2018) 연구는 재범 위험도 예측 인공지능에서 어떠한 변수가 허용될 수 있는지에 대한 일반인들의 의견을 조사했다.[335] 이들은 아마존 메커니컬 터크를 통해 모집된 일반인 100명에게 특정 변수를 재범 위험도 예측을 위해 사용하는 것이 공정한지 질문하였다. 이 때 설문 문항을 ① 단순히 해당 변수를 사용하는 것이 공정한지 질문한 경우, ② 해당 변수를 사용함으로써 예측 정확도가 향상되는 경우 해당 변수를 사용하는 것이 공정한지 질문한 경우, ③ 해당 변수를 사용함으로써 집단간 불균형(disparity)이 증가하는 경우, 해당 변수를 사용하는 것이 공정한지 질문한 경우로 구분하였다. 연구자들이 의견을 조사한 가능한 입력변수로는 전과 개수, 체포 범죄사실 설명, 범죄사실의 경중, 소년범 중범죄 전과 개수, 소년범 경범죄 전과 개수, 소년범 기타 위법행위 개수, 연령 성별, 인종 변수가 있었다. 그 결과 전과나 범죄사실에 대한 변수는 사용해도 된다는 의견이 다수였으나, 연령, 성별, 인종은 사용해서는 안 된다는 의견이 다수였다(아래 그림 참조).

---

335 Nina Grgic-Hlaca, et al. “Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction”, *Proceedings of the 2018 World Wide Web Conference*, (2018).

Feature	Q. 1 (a priori)	Q. 2 (if more accurate)	Q. 3 (if increases disparity)
# prior offenses	95%	93%	83%
arrest charge description	86%	92%	71%
arrest charge degree	85%	91%	69%
# juvenile felony offenses	74%	80%	61%
# juvenile misdemeanor offenses	65%	71%	53%
# juvenile other offenses	63%	69%	52%
age	44%	61%	32%
gender	26%	55%	24%
race	21%	42%	17%

Table 1: Comparing user judgment of fairness of each feature, when the user has different knowledge about the impact of incorporating that feature in the decision making process. We show the percentage of users who categorized each feature as fair according to the 3 questions described in Section 2.1.

그림 10 재범 예측을 위한 입력변수에 대한 설문 결과[336]

이처럼 설문 조사를 통해 특정 맥락에서 특정 변수의 활용에 대해 사람들이 얼마나 부정적으로 여기는지를 조사하면, 그 변수를 입력 변수로 활용할 것인지 여부에 대한 계량적인 판단을 내릴 수 있다. 예컨대 법에 명시된 차별금지사유만이 아니라, 일반인들의 부정적 의견이 일정 비율 이상인 변수는 활용하지 않을 수 있다. 나아가 나머지 변수에 대해서도 부정적 의견의 비율과 해당 입력 변수를 활용함으로써 정확도가 얼마나 높아지는지, 집단간 불균형이 얼마나 커지는지 등의 상충관계를 고려하여 사용될 입력변수의 범위를 결정할 수 있다. 해당 연구 사례에서 볼 수 있는 것처럼 인공지능의 도입에 앞서 일반인 설문 조사, 전문가 의견 수렴 등의 절차를 거쳐 특정한 변수를 입력 변수로 활용하는 것이 타당한지 여부를 미리 검토하는 것은 바람직한 개발 관행에 해당할 수 있다.

336 위의 글.

2. 한계

가. Unawareness 기준의 실효성에 대한 비판

Unawareness 기준에 대해서는 무엇보다 인공지능 공정성을 달성하는데 있어 실효적이지 못하다는 비판이 제기된다. 현재와 같이 빅데이터에 기반한 머신러닝 시스템의 학습 데이터에는 매우 많은 변수가 포함되어 있고, 데이터 집합에 포함된 변수는 다른 변수들과 높은 통계적 상관관계를 보일 수 있다. 이러한 경우 만약 특정한 보호속성을 그 입력에서 배제하더라도, 학습 데이터에 포함된 다른 입력변수를 통해 해당 보호속성을 추론할 수 있게 된다. 대표적 예로, 미국과 같이 인종간 거주 지역이 크게 분리되어 있는 상황에서는 우편번호를 이용하여 인종을 쉽게 예측할 수 있다. 다른 예로, 웹사이트 방문이력 정보에 특정 성별이나 나이대가 자주 이용하는 웹사이트의 방문 기록이 있다면 성별이나 나이대를 쉽게 예측할 수 있다.[337] 이처럼 차별적 속성을 입력 변수에서 제외시킨다고 하더라도 여전히 인공지능은 차별적으로 작동할 수 있다.

Gillis & Spiess (2020)은 미국 부동산 담보 대출에 관한 데이터를 통해 이러한 사실을 보인 바 있다.[338] 위 연구가 활용한 1990년 Boston Fed HMDA 데이터 집합은 부동산 담보 대출 신청자에 관한 인적 사항, 재무상태, 신청된 대출에 관한 정보 등 40여개의 변수에 대한 약 3,000건의 데이터를 포함하고 있다. 위 연구에서는 위 데이터 집합을 이용하여 채무불이행 위험을 예측하는 머신러닝 시스템을 구현하였는데, 인종을 입력변수로 활용하는 경우와 활용하지 않는 경우 머신러닝 알고리즘이 예측하는 채무불이행 위험의 분포는 인종별로 거의 차이가 나지 않았다. 위 연구자들은 그 이유가 머신러닝 시스템은 위 데이터 집합의 나머지 변수들을 이용하여 보호 속성을 쉽게 추론할 수 있기 때문이라고 설명한다.

---

337 Barocas, Hardt & Narayanan (2022) (각주 19), Chapter 1.

338 Gillis & Spiess (2020) (각주 205).

즉, 나이나 혼인 상태와 같은 보호속성을 학습 데이터 집합에 포함된 나머지 변수를 이용하여 매우 높은 정확도로 예측할 수 있었던 것이다.[339]

이처럼 unawareness 기준이 차별을 막기 위한 실효성이 없는 경우, 해당 기준을 적용하는 것은 오히려 차별적 인공지능을 활용하는 사업자가 자신의 인공지능을 정당화하는 근거로 활용될 수 있다. 예컨대 애플 카드가 성별에 따른 신용한도 차별이 발생했다는 주장에 대해, 골드만 삭스 사는 자신들이 "성별과 같은 요소에 근거해서" 결정을 내리지 않고 고객의 "성별을 알지도 못한다"고 주장하였다.[340] 그러나 이는 오히려 unawareness 기준의 한계를 보여주는 것으로 이해될 수도 있다. 즉, 골드만 삭스 사가 활용한 신용한도 결정 모델이 성별을 입력변수로 고려하지 않았다고 하더라도 여전히 성별에 대해 차별적인 결과가 초래될 수 있기 때문이다.

나. 대용변수(proxy)에 의한 차별 가능성

입력변수 중 특정 보호속성에 대한 정보에 대한 흐름을 차단할 경우, 의사결정자는 그에 대한 대용물로 다른 변수를 활용하여 의사결정을 내릴 수 있고, 그 경우 다른 집단 구성원에 대한 새로운 차별이 발생할 수도 있다. Cofone (2019)는 인간 의사결정자의 사례를 바탕으로 하여 대용변수에 의한 차별 가능성에 주목한다.[341] 그는 대용변수에 의한 차별을 ① 전이(transfer) 대용변수, ② 축소(reducing) 대용변수, ③ 확장(expanding) 대용변수의 경우로 각기 구분한다.

---

339 위의 글.

340 Shahien Nasiripour, Jenny Surane and Sridhar Natarajan, "Apple Card's Gender-Bias Claims Look Familiar to Old-School Banks", *Bloomberg Businessweek,* (2019. 11. 12.).

341 Ignacio N. Cofone, "Antidiscriminatory privacy", *SMU Law Review* 72, (2019b).

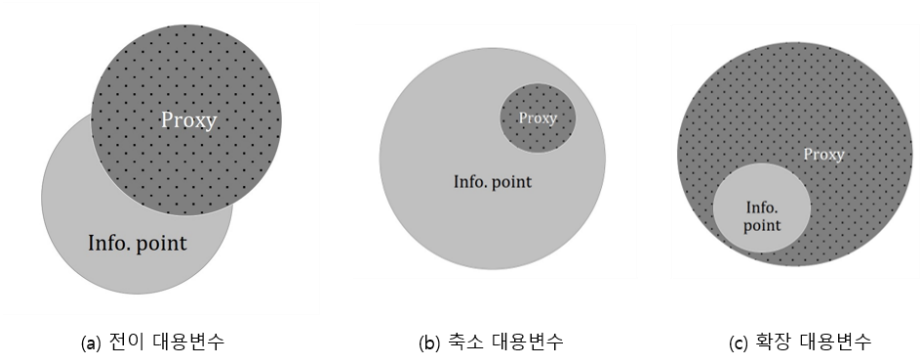


그림 11 대용변수의 유형[342]

우선, ① 전이 대용변수에 의한 차별은, 특정 집단에 대한 정보가 차단될 경우, 다른 집단에 대해 차별이 발생하는 경우를 일컫는다. 이 경우 보호집단 구성원 중 일부는 차별을 피하게 되지만, 그 대신 다른 집단 구성원이 차별을 겪게 될 수 있다.[343] 이를 잘 보여주는 사례로는 미국의 이른바 'Ban-The-Box' 정책 사례가 있다. 2000년대 후반을 전후하여 미국의 다수 주에서 전과자에 대한 채용 차별 문제를 해결하기 위해 채용 시 전과자 여부를 파악할 수 없도록 하는 정책을 도입하였다.[344] 그런데 한 실증연구에 따르면 전과자에 대한 정보가 차단당하게 되자 고용주들은 아예 젊은 흑인들의 채용을 꺼리게 되었다.[345] 즉, '전과자' 집단에 대한 정보가 차단되자, 그 대용변수인 '흑인 청년' 집단으로 차별이 전이된 것이다.

② 축소 대용변수에 의한 차별은, 특정 집단에 대한 정보가 차단될 경우, 그 부분집합에 해당하는 집단으로 차별이 옮겨가는 경우를 일컫는다.[346] 이 경우, 보

342 위의 글.

343 위의 글. p. 155.

344 미국에서 널리 사용되는 입사 지원서 양식에는 전과 여부를 묻는 체크박스가 포함되어 있는데, 이를 금지하자는 취지에서 'ban-the-box' 정책이라 불린다.

345 Jennifer L. Doleac and Benjamin Hansen, "Does 'ban the box' help or hurt low-skilled workers? Statistical discrimination and employment outcomes when criminal histories are hidden" No. w22469, *National Bureau of Economic Research*, (2016).

346 Cofone (2019b) (각주 341), pp. 155-156.

호집단 구성원 중 대용변수 집단에 해당하지 않는 일부는 차별을 피하게 된다. 예를 들어 어떤 산업계에서 라틴계 직원의 채용을 꺼리는 경향이 있다고 하자. 만약 서류심사 절차에서 민족성(ethnicity) 정보를 차단하지만 그 사람의 이름에 대한 정보는 여전히 접근할 수 있게 하면, 라틴계 사람들 중 자신의 민족성을 드러내는 이름을 가진 사람들만 차별이 대상이 될 것이다. 즉, 자신의 민족성이 이름에서 드러나지 않는 이들은 (적어도 서류심사 단계에서는) 차별을 피할 수 있게 된다.

③ 확장 대용변수에 의한 차별은, 특정 집단에 대한 정보가 차단될 경우, 그 상위집합에 해당하는 집단으로 차별이 옮겨가는 경우를 일컫는다.[347] 그 결과 더 많은 이들이 차별을 겪게 될 수 있다. 예를 들어 어떤 고용주가 임신 또는 출산 계획이 있는 젊은 여성의 채용을 꺼리는 경향이 있다고 하자. 이 경우 채용 과정에서 임신 또는 출산 계획이 있는지를 묻지 못하도록 금지할 경우 여성 전체에 대한 채용을 기피하게 될 수 있다. 즉, 확장 대용변수의 경우 더 넓은 범위의 집단에게 차별이 분산되는 것이다.

이처럼 차별을 막기 위해 보호속성에 관한 정보를 차단하는 경우, 의사결정자가 활용하는 대용변수에 의해 추가적인 차별이 발생할 가능성이 있다. 따라서, Cofone이 주장하는 바와 같이, 어떠한 보호속성을 차단할 것인지를 판단할 때 대용변수에 의한 차별 가능성과 다른 집단에 대한 영향을 함께 고려할 필요가 있다.

다. 정보 차단에 따른 적극적 평등실현조치 또는 차별 감사의 제한

인공지능 시스템이 그 입력변수로 보호속성을 활용할 수 없게 될 경우, 해당 속성을 이용하여 오히려 보호집단 구성원에 대해 유리한 대우를 하지 못할 수

---

347 위의 글, pp. 156-157.

있고, 나아가 해당 인공지능이 특정 집단에 대해 차별적으로 작동하는지 여부를 감사하는 작업도 어려워진다는 문제도 발생할 수 있다.

머신러닝 시스템은 단지 예측 정확도를 높이기 위한 방향으로 학습되므로, 보호속성에 대한 정보를 해당 보호집단에 유리하게 활용할 수도 있다. 따라서 보호속성에 따른 정보를 차단하는 것이 오히려 해당 집단에 불리한 결과를 초래하는 경우가 발생할 수도 있게 된다. 예컨대, Kleinberg et al (2018a)은 대학 입학 평가 데이터를 활용하여 인종 변수를 활용하지 않았을 때에 오히려 흑인들에게 불리한 예측이 이루어질 수 있다는 점을 보인 바 있다.[348]

더욱이 국내 차별금지법들은 현존하는 차별을 없애기 위하여 특정 집단이나 개인을 잠정적으로 우대하는 행위를 차별행위로 보지 않는 것으로 정하고 있다.[349] 그런데 만약 머신러닝 알고리즘에 이처럼 적극적 평등실현조치를 구현하고자 할 경우에는 대상 개인에 대한 차별금지사유에 관한 정보를 취득할 필요가 있다. 그러나 차별금지사유를 입력변수로 활용하는 것을 금지할 경우 이러한 적극적 평등실현조치가 제한될 수 있다. 이는 앞서 검토한 차별금지법제에 있어서의 반분류 원칙과 반종속 원칙 간의 충돌 문제가 발생하는 것으로도 이해될 수 있다.[350] 즉, 반분류 원칙에 따라 특정 속성을 입력하는 것을 금지시킬 경우, 반종속 원칙에 따른 적극적 평등실현조치가 제한될 수 있는 것이다.

입력변수 제한의 또다른 문제점은 인공지능에 대한 차별 감사가 어려워질 수 있다는 점이다. 인공지능 차별에 대한 감사는 주로 출력변수의 결과, 오류율, 예측도 등이 집단 간 동등한 지 여부를 비교하는 방식으로 이루어진다.[351] 그런데 인공지능을 평가하기 위한 데이터 집합에 보호속성에 대한 정보를 전혀 포함되

---

348 Jon Kleinberg, et al., "Discrimination in the Age of Algorithms" *Journal of Legal Analysis* 10 (2018a).

349 국가인권위원회법 제2조 제3호 본문 단서; 남녀고용평등법 제2조 제1호 다목, 동조 제3호; 장애인차별금지법 제4조 제4항 등.

350 본 연구 제2장 제2절 3항 참조.

351 본 연구 제3장 제3절 참조.

어 있지 않다면 이러한 감사가 매우 어렵거나 사실상 불가능해질 수 있다.[352] 이러한 문제에 대응하기 위해 EU AI 법안 제10조(데이터 및 데이터 거버넌스) 제5항은 고위험 AI 시스템의 제공자는 해당 AI 시스템과 관련하여 편향 모니터링, 감지 및 교정하기 위한 목적으로 엄격하게 필요한 범위에서는 GDPR 제9조 제1항에도 불구하고 민감정보를 처리할 수 있도록 정하고 있다. 다만, 이 경우 자연인의 권리와 자유에 대한 적절한 보호조치(재사용에 대한 기술적 제한, 최선의 보안 및 프라이버시-보호 조치 사용 등)를 취해야 한다.

3. 대안적 시도 및 평가

이상에서 살펴본 한계로 인하여, 직접차별을 피하거나 개인정보 보호법제를 준수하기 위한 목적에서 (특히 법에 열거된 보호속성에 대해) unawareness 기준을 적용하는 것은 인공지능 공정성을 보장하기 위한 실효적 방안이라 평가하기 어렵다. 이에 여러 대안이 제시되고 있는데, ① 보호속성과 상관관계가 높은 대용변수까지도 입력에서 제외할 것을 요구하는 방안, ② 미리 정해진 변수만을 인공지능의 입력변수로 사용할 수 있도록 허용하는 방안, ③ 다른 입력변수들이 보호속성과 상관관계를 갖지 않도록 학습 데이터를 수정하는 방안 등이 있다.

첫번째 방안을 적용한 사례는 미국의 일부 법령에서 발견된다. 예를 들어, 미국 주택도시개발부는 주택담보대출에 적용되는 알고리즘에 있어, 미국 공정주택법에 따라 차별이 금지되는 보호 집단에 관한 정보를 대체하거나 그와 밀접한 대용변수에 해당하는 요소들에 주요하게 의존하지 않도록 요구한 바 있다.[353] 또 다른 사례로는 미국 매사추세츠, 뉴욕, 필라델피아 주에서 지원자들의 종전 연봉

---

352 이러한 상황에서는 다른 데이터를 통해 보호속성을 추론하여 감사를 수행할 수도 있다. 가령 평가 데이터 집합에 성별 정보가 포함되어 있지 않을 경우에는 이름을 통해 성별을 추론하고, 성별 집단간 동등성 여부를 검토할 수 있다.

353 Department of Housing and Urban Development (HUD), Implementation of the Fair Housing Act's Discriminatory Effects Standard, 78 Fed. Reg. 11460 (2013).

을 묻는 것을 금지하기도 하였다. 이 법은 여성 지원자들이 남성에 비해 낮은 연봉을 받고 있으므로, 연봉에 관한 정보에 기초하여 지원자의 성별을 추론할 수 있다는 우려에 기한 것이었다.[354] 그러나 이러한 방안은 대용변수 중 보호속성과 어느 정도 상관관계를 갖는 변수까지 제한할 것인지 불명확하다는 문제가 있다. 그뿐만 아니라, Gillis & Spiess (2020) 연구가 설득력 있게 제시하는 것처럼, 인공지능이 다수 입력변수들을 종합한 결과 보호속성을 추론할 수 있는 경우 이러한 제한을 두는 것이 애당초 불가능하게 되는 근본적 난점이 있다.

두번째 대안은 미리 정해진 변수만을 인공지능의 입력변수로 사용할 수 있도록 허용하는 것이다. 그러나 이러한 방안은 해당 인공지능의 성능을 크게 제한하는 결과를 초래할 수 있다는 문제가 있다. 나아가 본 연구의 제5장에서 제시하는 것과 같이 머신러닝 시스템의 입력변수를 증가시킴으로써 인공지능의 성능을 개선할 뿐만 아니라 집단 간의 오류율이나 예측도 격차도 함께 줄일 수 있는 가능성이 존재한다. 이에 비추어 미리 정해진 변수만을 입력변수로 활용할 수 있도록 하는 것은 일반적인 경우 적절한 해결책이라 하기 어렵다.

다만 해당 인공지능의 적용에 있어 이용자들의 기대와 예측 가능성이 극히 중요한 경우 이용자들이 손쉽게 이해할 수 있는 일부 변수만을 입력변수로 활용하는 방안도 고려될 수 있다. 예컨대 국내 TMAP 운전습관 서비스는 이용자의 운전점수를 산출하여 일정 요건을 충족할 경우 보험료 감면 혜택을 누릴 수 있도록 하고 있다. 그런데, 해당 서비스의 운전점수는 이용자에게 사전 고지된 과속, 급가속, 급감속, 야간주행만을 입력 변수로 사용하고 있는 고지되어 있다.[355] 위 서비스는 이용자의 안전운전을 유도하기 위한 것으로서 이용자들이 어떠한 경우 안전운전 점수를 높일 수 있는지 명확히 이해할 수 있도록 가용한 입력 변수를 제한하는 것으로 이해할 수 있을 것이다.

---

354 Cofone (2019b) (각주 341), p. 158.

355 티맵모빌리티, “TMAP 운전습관 이용방법 & 운전점수 관리 꿀팁”, (2021. 5. 6.)

마지막으로, 인공지능의 다른 입력변수들이 보호속성과 상관관계를 갖지 않도록 학습 데이터를 수정하는 방안이 있다. 이러한 기법은 학습 데이터로부터 '편향을 제거(de-biasing)'하는 것으로 볼 수 있다. 그러나 Gillis & Spiess (2020)은, 해당 기법이 다수의 입력변수를 비선형적으로 고려하는 머신러닝 환경에서는 출력변수에 대한 불안정성을 야기하는 문제가 있을 뿐만 아니라, 작은 노이즈 상의 차이가 중대한 변화로 확대되는 한계가 있다고 지적한다.[356]

이상과 같이, unawareness 기준이나 그 변형 기준들을 적용하여 차별적 속성(혹은 그 대용변수)을 입력변수로 사용하는 것을 제한하더라도 머신러닝 시스템은 여전히 차별적으로 작동할 수 있으므로, 결국 인공지능의 출력 변수를 기준으로 하여 그 공정성을 평가해야 한다는 결론에 이르게 된다.

제3절 출력변수에 대한 공정성 기준

1. 결과 동등성

가. 의의

인공지능의 출력변수에 대한 공정성 기준 중 '결과 동등성(outcome parity)' 기준은 머신러닝 시스템이 각 집단별로 특정 출력변수 값으로 예측할 확률이 동등해야 한다는 기준을 일컫는다.[357] 이는 우리의 직관에 잘 부합한다. 예컨대 대출심사 인공지능에 대해 A 집단 신청자에 대한 대출 승인율이 50%인데 B 집단 신청자에 대한 대출 승인율이 30%에 불과하거나, 그 반대인 경우, 우리는 위 인공지능의 공정성을 의심하게 된다. 이러한 도덕적 직관에 따라 결과 동등성 기준

---

356 Gillis & Spiess (2020) (각주 205), p. 475-7.

357 ICO AI 감사 가이드라인 초안 (2020) (각주 16).

은 인공지능의 공정성을 결과(위 예에서는 승인율)가 집단간 동등한 것이라고 간명하게 정의한다.

결과 동등성 기준은 '인구통계적 동등성(demographic parity)'[358], '통계적 동등성(statistical parity)'[359], '독립성(Independence)' 등 다양한 명칭으로 불리고 있으나, 그 기본 개념은 동일하다. 본 연구에서는 이를 '결과 동등성' 기준이라 칭한다.[360] 이를 수학적으로 정식화하여 표현하면 다음과 같다.[361]

(정의) 결과 동등성 기준: 어떤 분류모형 $f_{classifer} : X \longmapsto \hat{Y}$ 에 대해, 보호속성들 A가 $\hat{Y}$와 독립인 경우, $f_{classfier}$는 결과 동등성 기준을 충족한다. A가 $\hat{Y}$가 독립이라는 것은 A에 속한 모든 a, a'에 대해 다음 수식을 만족해야 한다는 의미이다.

$$P\{\hat{Y} \mid A = a\} = P\{\hat{Y} \mid A = a'\}$$

위 기준을 풀어서 설명하면 다음과 같다. 채용이나 대출심사에서와 같이 이진변수(0/1)를 예측하는 인공지능에 대해서는, 특정 결과에 해당할 확률(즉, 선발률, 대출 승인율)이 집단별로 동등해야 한다. 혹은 개인의 능력이나 위험도를 예측하여 연속적인 점수를 부여하는 인공지능에 대해서는, 특정 점수에 해당할 확률이 집단별로 동등해야 한다. 이는 결국 예측된 점수에 대한 확률 분포가 집단간 동등할 것을 요구하게 된다.

---

358 van Breugel, et al. (2021) (각주 330), p. 3.

359 Hertweck (2021) (각주 31).

360 머신러닝 공정성 연구자들 사이에서 현재 가장 널리 통용되는 표현은 '인구통계적 동등성'인 것으로 보인다. 본 연구에서는 '결과 동등성'이 그 기준의 의미를 가장 간결하게 전달하고 있는 표현이라고 판단하여 해당 용어를 사용한다.

361 van Breugel, et al. (2021) (각주 330).

다만, 본 연구에서 지칭하는 '결과 동등성' 기준이 집단별 인구비례 선발(예: 남녀 동수 선발)을 요구하는 것은 아니다.[362] 즉, 가령 25명을 선발하는데 남성이 200명, 여성이 50명 지원하였고, 그 중 남성 20명, 여성 5명을 선발하였다고 하자. 이 때 남성 합격률과 여성 합격률은 모두 10%로 동일하므로, 본 연구 상의 '결과 동등성'은 충족한 것이 된다. 그러나 인구비례에 따른 선발 기준을 적용할 경우, 남녀 구성 비율에 맞추어 남성 5명, 여성 5명을 선발해야 한다. 이러한 기준은 소위 '할당제'라 표현되는데, 특히 정치 영역 등과 같이 대표성이 중요한 경우 적용될 필요가 있을 수 있다.[363]

한편, 결과 동등성 기준에 따를 경우 집단별로 선발 비율을 완전히 동등하게 맞추는 것은 어려울 수 있으므로, 기준을 다소 완화하여 집단간 선발 비율 차이가 일정한 범위 이내일 것을 요구하기도 한다. 이를 수식으로 표현하면 다음과 같다. 아래 식에서 a는 보호집단, a'는 다수집단을 일컫는다. 이는 결국 보호집단의 선발 비율이 다수집단과 비교할 때 과도하게 낮아서는 안 된다는 뜻이다.

$$\frac{P\{\hat{Y} \mid A = a\}}{P\{\hat{Y} \mid A = a'\}} \geq 1 - \epsilon$$

이처럼 결과 동등성에 대한 완화된 기준으로 잘 알려진 것으로는 미국 고용평등기회위원회의 4/5 규칙이 있다.[364] 해당 기준은 어떤 보호집단에 대한 합격률이 다른 집단과 비교하여 4/5 미만이 되어서는 안 될 것을 요구하는 것으로서 위 수식에서 ϵ 값으로 0.2를 적용한 것으로 이해될 수 있다. 일반적으로, 결과 동등성 기준은 차별금지법 상의 간접차별 금지 원칙에 대응하는 것으로 여겨진

---

362 고학수 외, *인공지능 윤리와 거버넌스*, 제III장 제3절(김병필 집필 부분), 박영사, (2021), 232-233면.

363 공직선거법 제47조 제3항 참조.

364 본 연구 제2장 제2절 2항 참조.

다.[365] 앞서 살펴본 바와 같이, 인공지능의 예측이 보호집단에 대해 현저히 불리한 결과를 초래하는 경우 간접차별에 대한 일응의 입증이 이루어지고, 고용주는 직무관련성 또는 업무상 불가피성의 항변을 제출해야 하는데, 이 때 선발 비율에 있어서의 상당한 격차가 존재하는 경우에는 현저히 불리한 결과가 발생한 것으로 추정될 수 있다

결과 동등성 기준은 특히 채용 인공지능에서 크게 활용되고 있다. 예컨대,18개 채용 인공지능에 대한 실증 분석 결과에 따르면, 상당수 사례가 4/5 규칙을 적용하고 있다고 한다.[366] 미국 고용평등기회위원회의 4/5 기준은 법적 구속력이 없는데도 불구하고, 심지어 미국 민권법이 직접 적용되지 않는 유럽 사업자들 역시 이를 엄격한 기준으로 고려하고 있다고 한다. 나아가 4/5 기준 자체가 결과 동등성과 같은 것으로 여겨지는 경향도 있다. 예측 인공지능의 편향 감사용 도구인 'AI Fairness 360'의 온라인 시연 사이트의 경우, 해당 인공지능이 4/5 규칙을 충족하는 경우 '통계적 동등성'이 충족된다고 표시하고 있기도 하다.[367]

나. 한계

1) 법적 제한

인공지능 공정성에 대한 공학 연구자들 사이에서는 결과 동등성 기준을 적용하는 것이 오히려 역차별에 해당하여 직접차별 금지원칙(미국법상 결과적 차별행위)에 해당할 수 있다고 이해되는 경우가 있다는 점은 앞서 설명한 바와 같다.[368]

---

365 van Breugel, et al. (2021) (각주 330), Appendix C.

366 Raghavan et al. (2020) (각주 149).

367 Rachel KE Bellamy, et al., "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias", arXiv preprint arXiv:1810.01943 (2018).

368 본 연구 제3장 제2절 3항.

이러한 오해는 일부 연구자들의 Ricci 판결에 대한 확대해석에 기인한 것임도 앞서 살펴보았다. 굳이 Ricci 판결을 둘러싼 논쟁을 고려하지 않더라도, 결과 동등성 기준을 적용함으로써 차별금지법제 상의 직접차별 금지원칙 또는 개인정보 보호법제와 충돌하게 되는 상황은 쉽게 예상할 수 있다. 앞서 본 바와 같이 의도적 차별행위를 막기 위해 보호속성 정보를 차단할 경우 인공지능에 대한 차별 감사가 어려워지거나, 결과 동등성을 준수하기 위한 편향 제거 기술의 도입이 어려워질 수 있다.[369] 이에 EU AI 법안에는 편향 모니터링, 감지, 교정 목적으로 민감속성 수집을 허용하는 규정을 두고 있다는 점도 앞서 본 바와 같다.

2) 심슨 패러독스 문제

결과 동등성 기준을 적용하는 데 있어 기술적 측면에서 제기되는 중대한 한계는 '심슨 패러독스(Simpson's Paradox)'를 해결하기 어렵다는 점이다. 심슨 패러독스란 어떤 인구집단에 있어 두 변수 간의 상관관계가 발견되는데, 만약 해당 인구집단을 작은 하위집단으로 구분할 경우 그 상관관계가 사라지거나 역전되는 현상을 일컫는다.[370]

선발 절차의 공정성 심사에 있어 심슨 패러독스 문제에 관해 잘 알려진 사례로는 1973년 버클리 대학교 대학원 입시 데이터 상의 성별 차별 문제가 있다.[371] 1973년 버클리대학교 대학원에는 8.442명의 남성 지원자와 4,321명의 여성 지원자가 지원했는데, 이 중 남성은 44%가 합격한 반면 여성은 단지 35%만이 합격하였다. 이는 결과 동등성 기준에 명백히 위반되는 것처럼 보인다. 나아가 통계학자의 분석에 따르면 이와 같이 큰 합격률 차이는 통계적으로 우연히 발생할

---

369 본 장 제2절 2항.

370 Jan Sprenger and Naftali Weinberger, "Simpson's Paradox", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.).

371 Peter J Bickel, Eugene A. Hammel, and J. William O'Connell. "Sex bias in graduate admissions: Data from Berkeley." *Science* 187, (1975).

수는 없었다. 따라서 대학원 입시 과정에서의 체계적인 성별 차별이 이루어지고 있을 가능성이 제기되었다.

그런데, 문제가 발생한 원인을 세밀하게 조사하자 대부분의 학과에서 여성이 남성보다 합격률이 더 높은 것으로 드러났다. 그럼에도 불구하고 전체 평균에서는 남성의 합격률이 더 높았던 원인은 지원 학과의 차이 때문이었다. 즉, 남학생들은 비교적 합격률이 높았던 공과대학 학과에 주로 지원하였고, 여성은 훨씬 더 합격률이 낮은 학과에 지원하였던 것이다. 이는 다음과 같은 인과성 도표(causal diagram)을 통해 설명될 수 있다.[372]

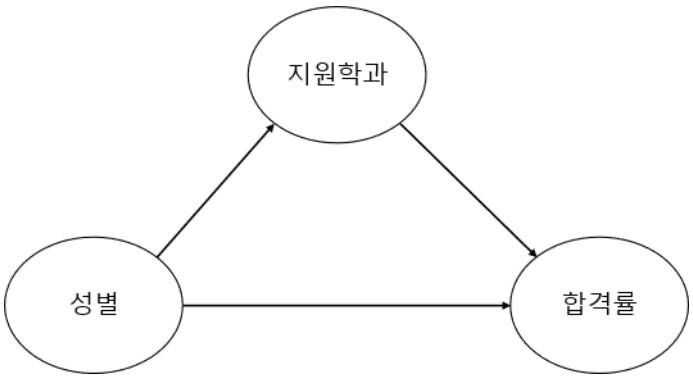


그림 12 버클리 대학 입시 사례의 인과성 도표

위 도표에서 성별 변수는 합격률에 있어 중립적이거나 (학과에 따라서는) 오히려 여성에 유리한 영향을 끼치고 있다. 그런데 성별은 지원학과 변수에 매우 큰 영향을 끼치고, 지원학과에 따라 합격률이 크게 달라진다. 그 결과 외견상 드러나는 성별에 따른 합격률 차이는 지원학과에 따른 영향 때문이지, 학과의 채용 담당자들의 차별적 관행이나 정책 때문은 아닌 것이었다.

372 인과성 도표의 의미에 관해서는 Judea Pearl and Dana Mackenzie, *The book of why: the new science of cause and effect*, Basic books, (2018) 참조.

심슨 패러독스는 결과 동등성 기준을 적용할 때, 단지 평균적인 전체 비율만을 근거로 차별적인지 여부를 판단해서는 안 된다는 점을 잘 보여준다. 이러한 경우 전체 평균에 있어서의 외견상 차별적 결과가 발생하더라도, 보호속성과 상관관계가 있는 다른 요인으로 인해 발생한 것인지, 아니면 인공지능에 의한 차별에서 기인한 것으로 구분하기 어렵기 때문이다. 이 문제에 대한 해법으로 '조건부 인구통계적 동등성'이라는 변형된 기준을 사용해야 한다는 주장도 제기되는데, 이에 대해서는 아래 다.항에서 상술한다.

3) 부정적 데이터의 역사적 누적 가능성

결과 동등성 기준을 적용한 결과, 보호집단에 대한 부정적인 평가 데이터가 누적되어, 향후 해당 집단 전체가 불이익을 얻는 결과가 초래될 위험성도 존재한다. 물론 결과 동등성을 적용할 경우 항상 충분한 자격을 갖추지 못한 보호집단 구성원이 선발되는 결과가 초래되는 것은 아니다. 결과 동등성 기준을 적용함으로써 이제껏 역사적 차별의 대상이 되어 온 보호집단 구성원들이 선발될 수 있도록 하고, 이들이 이제까지 사회적으로 제대로 평가받지 못했던 재능과 능력을 발휘하여 사회 전반적으로 불평등이 감소하는 바람직한 결과가 발생할 수 있다.

그러나 모든 경우 그러한 결과로 이어지지는 못할 수 있다. 가령, 대출심사 인공지능에 있어 결과 동등성 기준을 적용한다고 고려해 보면, 그 결과 현재 상대적으로 신용도가 낮은 보호집단 구성원에 대한 대출이 증가할 수 있고, (만약 인공지능이 충분히 정확하다면) 이들은 채무불이행을 할 가능성이 더 크다. 이러한 과정을 통해 보호집단 구성원 중에서 채무불이행의 비율이 더 높아지는 데이터가 누적될 수 있어 장기적으로는 보호집단에 오히려 불리한 상황이 초래될 수도 있다.[373]

---

[373] Cynthia Dwork et al., "Fairness through awareness", *Proceedings of the 3rd innovations in theoretical computer science conference*, (2012).

4) 역사적 차별 시정 비용 부담 문제

결과 동등성을 적용할 경우, 인공지능의 정확도가 하락하는 문제도 발생할 수 있다.[374] 그 경우 결과 동등성 기준을 적용하는 사업자는 인공지능의 정확도 하락에 따른 비효율성이나, 이를 시정하기 위한 추가적인 지출과 같은 비용을 부담할 수 있다. 예를 들어, 채용에 있어 사회 전반적으로 특정 보호집단 지원자들이 특정 기술에 대한 숙련도가 낮은 상황을 고려해 보자. 이때 고용주가 성별에 대해 결과 동등성 기준을 적용하면, 해당 기술에 대한 숙련도가 낮은 지원자 비율이 높아질 수 있다. 따라서 고용주는 해당 기술에 대한 숙련도를 높이기 위한 추가적인 교육 비용을 지출해야 할 수 있다. 이러한 비용은 결국 사회적으로 존재하는 역사적 차별을 시정하기 위한 것인데, 해당 사업자가 이러한 비용을 부담하도록 하는 것이 정당화될 수 있는지 문제가 된다.

다른 한편으로는 결과 동등성 기준이 적용되지 않았더라면 합격하였을 다수 집단 구성원도 결과 동등성 기준을 적용에 따른 비용을 부담하는 주체가 될 수 있다. 앞서 살펴본 Ricci 판결에서는 이미 치러진 승진시험 결과에 따라 승진이 될 것으로 기대하고 있었던 Frank Ricci를 포함한 여러 원고들이 이러한 비용 부담 주체에 해당한다. Ricci 판결은 결국 시정부가 결과적 차별행위를 막기 위해 소급적으로 이들의 정당한 기대를 희생시키는 것을 제한한 것으로 이해될 수 있다. 다만, Pauline Kim이 지적한 바와 같이 위 판결은 장래에 있어 알고리즘을 개선하는 과정에서 결과 동등성 기준의 적용을 금지하는 취지는 아니라는 점은 유의해야 한다.[375]

---

374 Sam Corbett-Davies, et al. “Algorithmic decision making and the cost of fairness” *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, (2017)은 이를 공정성의 비용(cost of fairness)라 표현하고 있다.

375 본 연구 제3장 제2절 3항 참조.

다. 조건부 인구통계적 동등성 기준

앞서 본 버클리 대학원 사례와 같이 전체 집단의 평균적인 선발율을 기준으로 할 경우 심슨 패러독스에 빠질 위험성이 있다. 이 문제를 해결하기 위해 비차별적인 것으로 평가되는 '매개자 변수(mediator variable)'를 통제한 상태에서 '결과 공정성' 기준을 적용할 필요가 있다는 주장이 제기된 바 있다. 이를 '조건부 인구통계적 동등성(Conditional Demographic Parity)' 또는 '조건부 통계적 동등성(Conditional Statistical Parity)'[376] 라 한다. 해당 조건을 수학적으로 정식화하여 표현하면 다음과 같다.[377]

(정의) 조건부 인구통계적 동등성 기준: 어떤 분류모형 $f_{classifer} : X \mapsto \hat{Y}$ 에 대해, 보호속성 A가 $\hat{Y}$와 사이에, 허용가능한 설명적 요인(explanatory factors) $R \subset X$에 대해 조건부로 독립인 경우 $f_{classifer}$는 조건부 인구통계적 동등성 기준을 충족한다. A가 $\hat{Y}$가 R에 대해 조건부 독립이라는 것은 R에 속한 모든 r과 A에 속한 모든 a, a'에 대해, 다음 수식을 만족해야 한다는 의미이다.

$$P\{\hat{Y} \mid A = a,\ R = r\} = P\{\hat{Y} \mid A = a', R = r\}$$

위 조건을 앞서의 결과 동등성 기준과 비교하면, 설명적 요인 R에 대해 조건화한다는 점에서 차이가 있다. 예를 들어 대학원 입학률에 있어 성별 격차를 측정할 때, 원래의 결과 동등성 기준에 따르면 전체 집단의 평균적인 합격률을 상호 비교하지만, 조건부 인구통계적 동등성 기준에 따르면 비차별적인 매개변수인 '지원학과' 변수별로 나누어서(즉, 조건화 해서), 지원학과별로 합격률을 비교해야 한다.

---

376 Corbett-Davies et al. (2017) (각주 374).

377 van Breugel, et al. (2021) (각주 330).

조건부 인구통계적 동등성 기준은 Kamiran et al. (2013)에 의해 처음 제안된 것으로 알려져 있다.[378] Corbett-Davies et al. (2017) 연구는 COMPAS 소프트웨어의 공정성을 평가하면서 전과 개수에 대해 조건화한 통계적 동등성을 적용한 바 있다. 법학 분야에서는 Wachter et al. (2021a)이 해당 기준이 널리 활용될 필요가 있다고 주장하였는데, 이들은 해당 공정성 지표가 유럽연합 사법재판소의 평등 개념 및 위법한 불균등 개념에 가장 잘 부합하고, 인공지능의 편향과 공정성을 측정하는 데 있어 이 지표를 사용하는 것이 법적으로 정당화될 수 있다고 주장하였다.[379]

이 기준은 결과 동등성 기준을 다소 완화하여 발전시킨 것으로 이해할 수 있다.[380] 예를 들어 어떤 운전자의 사고발행 위험을 평가하는 인공지능이 여성에 대해 상당히 더 높은 위험도를 부여하고 있다고 해 보자. 이때 결과 동등성 기준에 따르면 해당 인공지능은 여성에 대해 차별적이라고 평가될 수 있다. 그런데 만약 어떤 사회에서 여성은 운전을 더 늦게 시작하거나, 더 일찍 그만두는 경향이 있고, 그래서 평균적으로 운전경력이 더 짧다고 하자. 이러한 상황에서 해당 인공지능이 여성에 대해 높은 위험도를 부여하는 이유는 짧은 운전경력 때문일 수 있다. 이 경우 조건부 인구통계적 동등성 기준에 따르면, 운전경력 연수에 따라 조건화하여 성별 격차를 측정하는 것을 허용한다. 즉 동일한 운전경력 연수를 가진 사람들 간에 비교하였을 때 성별간 예측 위험도에 차이가 있는지 평가하는 것이다. 다만, 이러한 평가 기준을 적용하려면, 사고 운전경력 연수를 기준으로

---

378 Faisal Kamiran, Indrė Žliobaitė, and Toon Calders, "Quantifying explainable discrimination and removing illegal discrimination in automated decision making", *Knowledge and Information Systems* 35, (2013).

379 Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI", *Computer Law & Security Review* 41 (2021a).

380 Barocas, Hardt & Narayanan (2022) (각주 19), Chapter 3.

위험도를 달리 평가하는 것이 (그로 인한 성별 간의 격차를 초래되더라도) 규범적 측면에서 정당화될 수 있어야 한다.

위 사례는 다른 한편으로는 조건부 인구통계적 동등성 기준의 난점을 보여주기도 한다. 이는 바로 어떠한 설명요인 R에 대해서는 조건화하여 평가하는 것이 허용되고, 어떠한 설명요인 R'에 대해서는 허용되지 않는지 사회적으로 결정할 필요가 있다는 점이다. 가령, 어떤 IT 회사에서 활용하는 채용 인공지능이 여성을 낮은 비율로 선발하는 경향이 있고, 이는 결과 동등성 기준을 위반하고 있다고 하자. 그런데 이러한 차별적 결과가 대부분 여성이 통계적으로 '코딩 경험'이 적다는 설명요인에 의해 발생한 것이라고 가정하자. 여기서 조건부 인구통계적 동등성 기준을 적용하려면, 코딩 경험이라는 설명요인으로 조건화할 것인지 판단해야 한다. 만약 코딩 경험에 대해 조건화 하여 합격률을 비교하면 성별간 합격률 격차가 사라진다면, 위 IT 회사는 그저 코딩 경험에 따라 직원을 채용한 것이 된다. 하지만 이러한 채용 관행이 간접차별에 해당하지 않기 위해서는 '코딩 경험'이라는 요소에 대해 직무 관련성과 업무상 불가피성이 인정되어야 한다. 그런데, 이러한 문제에 대해서는 인공지능 개발자나 공정성 감사 업무 담당자가 쉽사리 평가하여 기술적으로 답을 내리기 어렵다. 이는 조건부 인구통계적 동등성을 인공지능 공정성 감사에 도입하는 데 있어 실무상 걸림돌이 된다. 더욱 어려운 점은 이러한 설명요인이 매우 다양할 수 있다는 점이다. 여러 가능한 설명요인 중 무엇에 대해서 조건화할 수 있는지, 조건화해야 하는지 판단하는 것은 쉽지 않는 도전이 된다.

이에 대해, 조건부 인구통계적 동등성 기준을 지지하는 Wachter et al. (2021a)은 어떠한 변수에 대해 조건화할 것인지를 결정하는 것은 쉽지 않고, 그러한 결정이 정치적으로 판단될 성질이기도 하다는 점은 인정한다. 또한 인공지능 공정성 감사 과정에서 조건부 인구통계적 동등성 기준을 적용할 경우 개발자나 감사자의 편의에 따라 유리한 조건이 선택될 위험성도 지적한다. 하지만, 이들은 감사 과정에서 투명성을 요구함으로써, 어떠한 변수에 대해 조건화하여 감

사를 수행한 것인지 공개한다면, 해당 평가 과정을 검증하고 반박하는 과정을 통해 위 문제를 해결할 수 있을 것으로 기대한다.

요컨대, 조건부 인구통계적 동등성은 결과 동등성 기준의 한계점을 보완하면서도 기존의 차별금지법과의 조화를 시도할 수 있는 공정성 평가 기준이라 할 수 있다. 그러나 아직 공정성 감사 실무에서 널리 활용되고 있지는 못한 것으로 보인다.[381] 그럼에도 불구하고 해당 기준은 Wachter et al. (2021a)이 주장한 바와 같이 기존 차별금지법제와 잘 부합하여 법적 정당화 근거가 높다는 장점이 있다. 따라서, 향후 실무 적용상의 난점을 해결할 수 있다면 그 적용 범위가 늘어날 잠재성이 있을 것이다.

2. 오류율 동등성과 예측도 동등성

가. 전제 – 평가 데이터 집합의 존재

출력변수에 대한 또다른 공정성 기준인 오류율 균등 기준과 예측도 균등 기준을 설명하기에 앞서, 두 평가 기준이 공유하고 있는 전제를 우선 설명한다. 오류율 균등 기준과 예측도 동등 기준은 이제까지 설명한 Unawareness 기준이나 결과 동등성 기준과는 달리, 평가 데이터 집합을 요구한다.

평가 데이터는 공정성 심사에서뿐만 아니라 일반적으로 인공지능의 성능을 평가하는데 있는 핵심적 역할을 수행한다. 머신러닝 방식의 인공지능 시스템은 원칙적으로 학습 데이터의 예측 정확도를 높이는 방향으로 학습된다. 따라서 인공지능이 학습 데이터에 대한 과적합(over-fitting)이 이루어져, 학습 데이터에 포함되어 있지 않은 사례에 대해서는 적절한 성능을 보이지 못할 위험성이 있다. 특히 딥러닝 등 최근의 인공지능 시스템은 방대한 파라미터를 활용하여 학습 데

---

381 예컨대 앞서 소개한 ICO의 가이드라인이나 MAS의 사례연구 보고서는 조건부 인구통계적 동등성 기준을 소개하고 있지 않다. 추측컨대, 어떤 변수에 대해 조건화할 수 있는지를 명확히 결정하기 어렵다는 점으로 인해 실무상 적용하기 쉽지 않았던 것으로 보인다.

이터를 ‘암기(memorialization)’할 수 있다는 문제도 있다. 이러한 문제에 대한 여러 해법이 있지만, 보편적으로 적용되는 방법은 학습 데이터와는 구분되는 별개의 평가 데이터 집합을 마련하고, 인공지능이 해당 평가 데이터에 어떠한 성능을 보이는지 확인하는 것이다. 극단적인 예를 들자면 학습 데이터에 대해서는 99%의 정확도를 보이지만 평가 데이터에 대해서는 70%만의 정확도만을 보인다면 그 시스템은 심각하게 과적합된 것이라 할 수 있다.

오류율 균등 기준과 예측도 균등 기준은 이처럼 평가 데이터를 이용하여 인공지능 시스템의 정확도를 평가하는 방법을 인공지능 공정성 감사에도 적용한 것이다. 즉, 두 방법은 인공지능을 평가 데이터에 대해 적용시켜, 인공지능의 예측이 평가 데이터상의 “정답”과 얼마나 오차가 나는지 확인하고, 정확도를 산정한다. 그 다음 그러한 정확도의 수준이 집단간 동등하거나, 적어도 일정 범위 이내일 것을 요구한다.

그런데 오류율 균등 기준과 예측도 균등 기준이 서로 달라지는 이유는 통계적 정확성을 평가하는 지표가 여러가지가 있기 때문이다. 흔히 언급되는 단순 정확도(accuracy)는 전체 평가 데이터 대상 중 정확하게 판단된 사례의 비율을 의미한다. 가령 얼굴인식 인공지능을 통해 사진 속 얼굴이 직원인지 여부를 판단한다고 하면, 전체 사진 수 중 정확하게 판단된 사례의 비율을 의미한다. 그러나 단순 정확도는 인공지능의 성능을 제대로 표시하는 문제가 있다. 가령 얼굴인식 인공지능에 대한 평가 데이터 중 90%가 직원이 아니고 10%만이 직원이라고 하자. 그런데 인공지능의 오류로 인해 모든 평가 데이터에 대해 직원이라고 판단한 경우, 단순 정확도는 여전히 90%나 된다. 따라서 인공지능 시스템의 성능 평가에 있어서는 다른 통계적 정확성 지표를 사용할 필요가 있게 된다.

적격(1)인지 부적격(0)인지를 분류하는 이진분류(0/1) 시스템에 있어서의 통계적 정확성 지표는 아래와 같은 혼동 행렬(confusion matrix)을 통해 도출해낼 수 있다. 이는 평가 결과를 ① 평가 데이터에 포함된 실제 조건(ground truth)이 적격인지 부적격인지와 ② 인공지능이 예측한 조건(prediction)이 적격

인지 부적격인지를 기준으로 구분한다. 그 결과 4가지 사례가 발생한다. 실제 조건이 적격이고 적격으로 옳게 예측한 경우 '진양성(True Positive)', 실제 조건이 적격인데 부적격으로 잘못 예측한 경우 '위음성(False Negative)', 실제 조건이 부적격인데 적격으로 잘못 예측한 경우 '위양성(False Positive)', 실제 조건이 부적격인고 부적격으로 옳게 예측한 경우 '진음성(True Negative)'이 된다.

		실제 조건(Ground Truth)		
		실제 적격	실제 부적격	예측 조건을 기준으로 한 정확도 지표
예측 조건	적격 예측	**True Positive**	**False Positive** (Type I 오류)	**Positive Predictive Value (PPV)** $= \frac{\text{True Positive 건수}}{\text{적격 예측 건수}}$ **(정밀도, 양성예측도)**
	부적격 예측	**False Negative** (Type II 오류)	**True Negative**	**Negative Predictive Value (NPV)** $= \frac{\text{True Negative 건수}}{\text{부적격 예측 건수}}$ **(음성예측도)**
실제조건을 기준으로 한 정확도 지표		**True Positive Rate (TPR)** $= \frac{\text{True Positive 건수}}{\text{실제 적격 건수}}$ **(재현율, 민감도)**	**True Negative Rate (TNR)** $= \frac{\text{True Negative 건수}}{\text{실제 부적격 건수}}$ **(특이도)**	

표 3 이진 분류의 혼동 행렬과 통계적 정확성 지표

위 혼동 행렬을 이용하면 여러 통계적 정확성 지표를 생성해 낼 수 있다. 그 중 기본적으로 사용되는 정확성 지표는 4가지인데, True Positive Rate (진양성율, 재현율 또는 민감도라 불린다. 이하 "진양성율"라 한다), True Negative Rate (진음성율 또는 특이도라 불린다. 이하 "진음성율"라 한다), Positive Predictive Value (양성예측도 또는 정밀도라 불린다. 이하 "양성예측도"라 한다), Negative Predictive Value(이하 "음성예측도"라 한다)이다. 그 정의는 위 표와 같다.

그 중 진양성율과 진음성율은 실제 조건을 분모의 기준으로 삼는 것이고, 양성예측도와 음성예측도는 예측 조건을 분모의 기준으로 삼는다는 점에서 차이가 있다. 즉, 진양성율은 평가 데이터상 실제 적격인 사례 건수 중 진양성 건수의 비율을 가리키고, 진음성율은 평가 데이터상 실제 비적격인 사례 건수 중 진음성 건수의 비율을 가리킨다. 이와 달리 양성예측도는 적격으로 예측된 사례 건수 중 진양성인 건수의 비율을 가리키고, 음성 예측도는 부적격으로 예측된 사례 건수 중 진음성인 건수를 가리킨다.

나. 오류율 동등성 기준

1) 의의

가) 예시 사례

오류율 동등성(error rate parity) 기준은 '위음성율' 및/또는 '위양성율'이라는 오류율이 각 집단간 동등할 것을 요구하는 기준이다.[382] 정식화된 정의를 살펴보기에 앞서 이해를 돕기 위해 다음과 같은 오류율 동등성 기준을 위반하는 인공지능 사례를 먼저 고려하자.

사례 1 (얼굴인식 모형의 집단 간 오류율이 다른 사례). 10,000명의 직원을 둔 회사에서 직원 출입구에 얼굴인식 서비스를 이용하여 직원인지 여부를 인식하는 소프트웨어를 도입하였다. 그런데 그 인공지능은 다른 피부색의 직원에 대해 서로 다른 오류율을 보인다. 예컨대 파란 피부를 가진 직원에 대해서는

382 위 정의는 '정확성' 지표를 사용하여 정의하여도 동일하게 정의할 수 있고, 현재 고려하고 있는 이진분류에서는 단지 표기상의 차이에 불과하다. 즉, 진양성율 = 1 - 위음성율(False Negative Rate, FNR)이므로 진양성율에 대해 집단간 동등성을 요구하는 것은 위음성율의 동등성을 요구하는 것과 같다. 또한 진음성율 = 1 - 위양성율(False Positive Rate)이므로, 진음성율에 대해 집단간 동등성을 요구하는 것은 진음성율의 동등성을 요구하는 것과 같다.

평균적으로 100번 통과시 90번 통과시키고, 10번 거절한다(위음성율 10%). 그런데 나머지 피부색을 가진 직원들에 대해서는 100번 통과시 99번 통과시키고, 1번 정도만 거절한다(위음성율 1%).

위 얼굴인식 인공지능에 대해 피부색에 관한 결과 동등성을 요구하는 것은 명백히 비상식적이다. 결과 동등성은 인공지능의 승인율(즉, 회사 출입구에 도착한 사람 중 출입이 허용되는 사람의 비율)이 집단간 동등할 것을 요구한다. 그러나 이 비율은 회사 직원의 피부색 구성비와 방문자의 피부색 구성비에 따라 달라지는 것이 너무도 당연하다.[383]

이러한 상황에 있어 적절한 인공지능 공정성 심사 기준은 피부색 집단에 대해 오류율을 평가하는 것이다. 위 사례에서는 인공지능이 '양성(positive)'으로 예측하면 직원은 출입구를 통과하고, '음성(negative)'로 예측하면 출입이 거절된다. 따라서, '위음성(false negative)' 사례의 경우 직원임에도 불구하고 출입이 거절되는 불이익을 겪게 된다. 이러한 불이익의 정도가 집단 간 동등하도록 요구하는 것이 바로 '위음성율'의 동등성 조건이 된다. 위 사례 1에서는 특정 피부색을 가진 직원들이 더 높은 비율로 불이익을 겪도록 하는 것은 정당화되기 어려우므로, 위음성율의 동등성을 적용하는 것이 타당하다.

반대로 '위양성(false positive)' 사례는 직원이 아님에도 불구하고 출입이 허용되는 사례를 의미한다. 만약 위양성율이 높다면, 이는 직원 출입구에 심각한 보안상 위협을 초래할 것이므로 해당 인공지능은 도입되어서는 안 될 것이다.[384]

---

383 물론 특정 피부색인 개인은 애초에 자신의 집단에 있는 직원의 비율이 낮다는 점에 대해 이의를 제기할 수도 있다. 하지만 만약 회사 직원의 구성이 그 회사가 위치한 지역의 모집단의 피부색 비율과 일치하고, 공정한 채용 절차를 통해 직원들이 채용된 경우라면 이러한 이의제기는 타당하지 않게 된다. 나아가 이러한 이의주장은 얼굴인식 인공지능 자체의 공정성과는 직접 관련된 것은 아니다.

384 위 사례에서는 출입구를 사용하는 사람이 10만 명이라고 하자. 만약 알고리즘의 위양성율이 백만분의 1이라고 하면, 10일에 1명 꼴로 직원이 아닌 사람의 출입이 허용되는 셈이다. 참고로 iOS에 적용된 페이스아이디(FaceID) 기술의 위양성율은 백만분의 1 미만이라고 설명하고 있다. https://support.apple.com/en-us/HT208108 (2023. 5. 26. 접속 확인)

다만 위 사례에서 위양성율이 집단간 동등할 것을 요구하는 것에는 대해 인공지능 시스템의 이용자의 입장에 있어서의 윤리적, 법적 의의는 찾기 어렵다.[385]

요컨대 오류율 동등성은 각 집단별로 위음성율이나 위양성율이 동등할 것을 요구한다. 사안에 따라서는 더 엄격한 조건으로 두 오류율이 모두 집단간 동등할 것을 요구할 수도 있다. 만약 두 오류율 중 한 가지를 선택한다면, 해당 맥락에서 양성이나 음성이 갖는 의미에 따라 결정하게 될 것이다. 만약 양성 예측이 특정 혜택을 얻을 자격과 연결된다면, '위음성' 사례(즉, 실제로는 양성인데 인공지능이 잘못해서 음성으로 예측한 사례)가 중요하게 고려되어야 할 것이고, 반대로 만약 양성 예측이 어떤 불이익과 연결된다면 '위양성' 사례(즉, 실제로는 음성인데 인공지능이 잘못해서 양성으로 예측한 사례)가 중요하게 고려되어야 한다.[386]

나) 정식화된 정의

이상에서 설명한 오류율 동등성 조건은 다음과 같이 정식으로 정의할 수 있다.[387]

(정의) 오류율 동등성 기준: 어떤 분류모형 $f_{classifer} : X \mapsto \hat{Y}$ 에 대해, 보호속성 A가 진정한 목표변수 값 Y를 조건으로 $\hat{Y}$과 독립인 경우 $f_{classfier}$는 오류

---

385 물론 알고리즘 개발자가 어떠한 경우 위양성 오류가 발생하는지 그 원인을 조사하기 위해 집단별 차이를 조사할 수는 있을 것이다. 그러나, 보안의 목적에서는 위양성율을 낮추는 것 자체가 중요하지, 피부색에 따른 위양성 오류율을 균등화하는 것 자체에 큰 의의를 두기 어려워 보인다.

386 예를 들어 COMPAS 사례에서는 ProPublica는 위양성율을 중요한 요소로 고려하였다. 이 경우 재범위험도가 높은 것(High Risk)이 양성에 해당하므로, 실제로는 음성(Low Risk)인데 잘못 양성(High Risk)으로 예측되는 '위양성' 사례의 경우 피의자가 겪게 되는 불이익이 크다. 해당 사례에 관해서는 본 장 제4절 1항 참조.

387 Barocas, Hardt & Narayanan (2022) (각주 19). 위 글은 이를 '분리성(separation)' 조건이라 표현한다.

율 동등성을 충족한다. $f_{classfier}$가 이진 분류모형인 경우, A가 Y를 조건으로 $\hat{Y}$과 독립이기 위해서는 A에 속한 모든 a, a'에 대해 다음 수식을 만족해야 한다.

$$P\{\hat{Y} = 1 \mid Y = 1, A = a\} = P\{\hat{Y} = 1 \mid Y = 1, A = a'\}$$
$$P\{\hat{Y} = 0 \mid Y = 0, A = a\} = P\{\hat{Y} = 0 \mid Y = 0, A = a'\}$$

위 정의는 이진분류에 있어서 보호집단간 위음성율의 동등성과 위양성율의 동등성을 모두 요구하는 것이다.[388] 이처럼 두 조건을 모두 요구하는 기준을 'Equalized Odds'라고도 하고,[389] '분리성(separation)' 조건이라고도 부른다.

나아가, 앞서 본 바와 같이 위 조건을 완화하여, 두 오류율 중 한가지만을 요구할 수도 있다. 실무상 더 자주 주요하게 고려되는 비율은 '위음성율(false negative rate)'의 동등성인데, 이를 흔히 '기회의 균등(Equal Opportunity) 기준'이라고 부른다.[390] 이처럼 위음성율의 동등성 기준을 '기회의 균등' 기준이라고 부르는 이유는, 위음성 오류를 부당한 기회의 거절과 연결지을 수 있기 때문이다. 예컨대 채용, 학생 선발, 대출 승인 등과 같은 일종의 선발 절차에 있어서 적격을 갖추었음에도 불구하고(즉 평가데이터상 양성 사례에 해당), 인공지능이 음성으로 잘못 예측함으로써 기회가 거절되는 경우가 위음성 오류에 해당한다. 이 경우 집단간 위음성 오류율을 동등하다는 것은 결국 부당하게 기회가 거절되는 비율이 동등하다는 의미가 된다.

---

388 이진분류에서 $P\{\hat{Y} = 1 \mid Y = 1\}$ 은 진양성율을 뜻한다. 따라서 정의 중 첫번째 조건인 $P\{\hat{Y} = 1 \mid Y = 1, A = a\} = P\{\hat{Y} = 1 \mid Y = 1, A = a'\}$은 집단간 진양성율이 동등할 것(= 위음성율이 동등할 것)을 요구하는 것이다. 한편, $P\{\hat{Y} = 0 \mid Y = 0\}$는 진음성율을 뜻한다. 따라서 정의 중 두번째 조건인 $P\{\hat{Y} = 0 \mid Y = 0, A = a\} = P\{\hat{Y} = 0 \mid Y = 0, A = a'\}$는 집단간 진음성율이 동등할 것(= 위양성율이 동등할 것)을 요구하는 것이다.

389 Hardt, Price & Srebro (2016) (각주 21)

390 위의 글.

오류율 동등성 기준은 결과 동등성 기준과 달리 민감속성이 목표변수와 상관관계가 있을 수 있다는 사실을 인정한다. 예를 들어, 신용평가의 대상 중 한 집단은 다른 집단보다 채무불이행 비율이 더 높을 수 있고 금융기관은 이들 집단들마다 대출 승인율을 달리 하는 것은 직무 관련성 및 업무상 불가피성 요건에 부합할 수도 있다. 이때 오류율 동등성 기준은 '목표변수에 의해 정당화(justified by the target variable)' 될 수 있는 범위 내에서는 예측된 목표변수와 민감 속성 사이의 상관관계를 허용하는 것으로 이해할 수 있다.[391]

2) 한계

가) 법적 정당화 근거의 불명확성

오류율 동등성 기준이 기존의 차별금지법제에 비추어 어떻게 법적으로 정당화될 수 있을 것인지는 명확히 평가하기 어렵다. 구체적으로, 만약 어떤 인공지능이 보호집단 간 오류율에 차이가 있다면, 해당 인공지능을 사용하는 것이 법적인 의미에서 간접차별 또는 결과적 차별행위에 해당하는 것으로 볼 수 있는지, 혹은 차별에 대한 일응의 입증이 이루어졌다고 평가할 수 있는지는 명확하지 않다.

기존 차별금지법제에서 오류율 동등성 기준의 의미를 밝혀내는 것이 어려운 이유로는, 평가 데이터를 활용하여 인공지능의 오류율을 측정하는 것이 이제껏 소송에서 사용되지 않았고, 입법 과정에서도 고려되지 않아 왔기 때문이다. 고용영역의 결과적 차별행위 소송에서 통계적 증거의 활용이 활발한 미국에서도 일반적으로 관찰 자료(observational data)가 이용되어 왔을 뿐이다. 이러한 방법은 가령 채용상 성별에 대한 차별이 있는지 확인하기 위하여 집단간 채용 비율을 비교하거나, 임금 차별이 있는지 확인하기 위하여 집단간 임금을 비교하는 것

---

391 Barocas, Hardt & Narayanan (2022) (각주 19).

이다. 나아가 만약 통계적 불균형이 발견될 경우, 우연에 의해서 발생한 것이 아니라는 점을 입증하기 위해 통계적 유의성 기준을 도입해 왔다. 종래에 인간 의사결정자에 의한 판단이 이루어지는 경우, 평가 데이터를 활용하여 감사를 수행하는 것은 극히 어려운 일이었다.[392]

그러나 인공지능 시스템을 도입한 이후 차별 감사를 위한 접근법에 중대한 변화가 생겼다. 즉, 인공지능의 학습에 사용되지 않은 평가 데이터 집합 또는 감사 목적 데이터 집합(audit dataset)을 별도로 구축한 다음, 이를 이용한 인공지능 시스템의 정확도(오류율)가 집단간 차이가 있는지 평가할 수 있게 된 것이다. 이러한 증거에 대해서는 아직까지 차별금지법상 취급이 정립되지 못하였고, 이를 정립하는 것은 새로운 과제가 되고 있다.

한 가지 해법은 오류율 동등성을 결과 동등성과 마찬가지로 보아, 오류율 동등성이 충족되지 못한 경우에도 결과적 차별행위에 대한 일응의 입증이 이루어진 것으로 평가하고, 고용주가 직무 관련성이나 업무상 불가피성을 입증하지 못한 이상 차별 행위로 의율하는 것이다. 그러나 집단간 오류율에 있어 격차가 발생하는 원인이 데이터 생성 과정에 내재한 불확실성이나 피하기 어려운 노이즈에서 기인한 것이라면, 오류율 동등성 조건 위반을 차별행위로 보기 어렵다. 본 연구 제5장은 시뮬레이션 결과를 통해 이 문제를 탐구한다. 이와 관련하여 데보라 헬먼은 “[오류율 동등성] 지표만으로는 알고리즘이 공정하거나 불공정한지 여부를 결정하지 않는다”고 보았고, “그럼에도 불구하고 오류율 동등성의 결여는 특정 그룹이 과거에 부당하게 취급당한 경우 불공정성을 암시하는 지표가 된다”는 의견을 제시하기도 하였다.[393]

이처럼 오류율 동등성 지표가 법적 근거가 불명확한 이상, 인공지능 활용 사업자가 이 지표를 적용해야 할 것인지 여부를 판단하기 쉽지 않고, 인공지능 공

---

392 Barocas, Hardt & Narayanan (2022) (각주 19)는 도전적인 사회학자나 경제학자에 의해 연구 차원에서 실증적 차별 감사를 수행한 사례를 설명한다.

393 Hellman (2020) (각주 3).

정성 감사에 있어서도 이를 위반한 경우의 함의를 도출하기 어렵다는 문제가 상존하고 있다. 다른 한편, 오류율 동등성 기준을 충족하는지 여부를 감사하기 위해서도 보호속성에 관한 정보를 수집할 필요가 있고, 이는 차별금지법제 상의 직접차별 금지원칙 또는 개인정보 보호법제와 충돌하게 되는 상황은 발생할 수 있다. 이러한 문제는 앞서 결과 동등성 기준에서와 마찬가지로 발생하는 것으로서, 출력변수에 대한 공정성 기준에 있어 공통적으로 발생하는 것이다.[394]

나) 평가 데이터 구축의 문제

오류동 동등성 기준은 인공지능의 예측 값이 평가 데이터의 진실된 값과 얼마나 차이가 나는지를 기준으로 인공지능을 평가하므로, 해당 기준을 적절히 적용하기 위해서는 정확한 평가 데이터를 구축하는 것이 극히 중요하게 된다(이는 예측도 동등성의 경우도 마찬가지이다). 그런데 진실된 데이터를 관측하는 것이 어려운 경우도 적지 않다. 머신러닝 데이터에서 흔히 발생하는 문제는 데이터 수집이 편면적으로만 가능한 경우인데, 이를 '선별적 라벨 문제(selective label problem)'라고도 한다.[395]

선별적 라벨 문제를 이해하기 위해 재범 위험성 예측 인공지능을 예를 들어보자. 그 학습 또는 평가 데이터는 어떠한 특성을 가진 피의자가 일정 기간 내에 재범을 했는지에 관한 데이터를 포함하고 있어야 한다. 그런데 법관이 재범 위험성이 높다고 판단한 피의자들은 구속이 되고, 결국 실형을 선고받아 복역하게 되는 경우가 많을 것이다. 이 경우 구속된 피의자들의 경우 구속이 되지 않았더라면 재범을 저질렀을 것인지를 판단할 수 있는 데이터는 결코 확보할 수 없다. 비슷한 사례가 채용 문제에서도 발생한다. 개별 기업이 보유한 직원 채용 데이터

---

394 본 연구 제4장 제2절 2항 참조.

395 Jens Ludwig and Sendhil Mullainathan, "Fragile algorithms and fallible decision-makers: lessons from the justice system", *Journal of Economic Perspectives* 35(4), (2021), p. 85.

집합에는 자신이 채용한 직원에 대한 업무 능력만 포함하고 있을 것이다. 회사에 지원하였으나 채용하지 않은 지원자에 대해서는 만약 그를 선발하였다면 어떠한 업무 능력을 보였을 것인지 알 수 없다. 이처럼 선별적 라벨 문제는 선발 영역에 있어 공통적으로 발생하게 된다.

인공지능 공정성 감사를 수행하는 사업자는 이 문제를 해결하기 위해 누락된 데이터를 추정하는 방법을 사용하기도 한다.[396] 통계학에서는 통계데이터 중 일부에 결측치(missing value)가 있는 경우 이를 대체(imputation)하는 기법이 발전해 있는데, 이러한 기법들을 활용하면 관찰할 수 없는 데이터를 보충할 수 있다. 그러나 이러한 대체 기법의 정확성에는 한계가 있을 수 있고, 이는 인공지능 공정성 감사를 위한 평가 데이터의 신뢰성에 의문을 제기할 수 있다.

또 다른 방법은 여러 사업자들이 각자의 데이터를 종합하여 더 큰 데이터 집합을 구축하는 것이다. 대표적으로 금융기관의 신용 데이터의 경우, 각각의 금융기관은 자신이 대출을 수행한 고객이 채무불이행을 하였는지 여부만을 알 수 있을 뿐이다.[397] 그런데 각기 다른 위험도의 대출상품을 판매하는 여러 금융기관의 데이터를 종합한다면 그러한 공백을 줄일 수 있게 된다. 국내 제도상으로는 신용평가에 관하여는 신용정보법상 '신용정보집중기관' 제도를 통해 이 문제를 대응하고 있으나, 고객 마케팅, 고객 응대, 사기 탐지 등 다른 활용분야에 있어서는 이처럼 공유된 데이터 축적을 위한 제도가 충분하게 마련되어 있지 못하다.[398]

---

396 MAS (2021) (각주 17).

397 만약 실험정신이 높은 금융기관이라면 이 문제를 극복하기 위해 소수의 대출 신청인에 신용도가 낮더라도 실험적으로 대출을 실시하고 채무 불이행 여부를 관찰할 수 있을 것이나, 이러한 방법을 일반적으로 활용할 수 있을 것으로 기대하기 어렵다. 고학수 외, *금융분야 인공지능 활성화를 위한 가이드라인 등 마련*, 서울대학교 산학협력단, (2021), 69면.

398 위의 글.

다) 실무 구현상의 한계

오류율 동등성 기준을 준수하도록 인공지능을 구현하는 과정에서도 여러 한계가 존재한다. ① 우선, 다양한 보호속성 간의 충돌 문제가 있다. 즉, 한가지 보호속성에 있어 오류율 동등성 기준을 충족시키고자 인공지능을 개선할 경우, 다른 보호속성에 대해 오히려 오류율 격차가 커지는 문제가 발생할 수 있다. 싱가포르 통화청의 신용평가 인공지능 공정성 심사 예시 사례는 이를 잘 보여준다. 해당 사례에서는 성별에 대해 기회의 균등 기준을 충족시키기 위해 인공지능을 조정한 결과, 교육수준에 대한 기회의 균등 격차는 오히려 더 커지게 되었다.[399] 이 경우 어떠한 보호속성을 우선시할 것인지 판단해야 하는 문제가 발생한다.

② 또다른 문제는 오류율이 동등하도록 인공지능을 수정하기 위해 '단일 문턱 값 원칙(Single-Threshold Principle)'을 위반해야 할 수 있다는 점이다. 단일 문턱 값 원칙이란 각 집단에 대해 동일한 기준이 되는 문턱 값을 적용하여 선발 여부를 결정해야 한다는 것이다.[400] 가령, 만약 어떤 집단에 대해서는 채무불이행 확률이 0.1% 이하인 경우 대출을 승인하고, 다른 집단에 대해서는 0.2% 이하인 경우 대출을 승인한다면 단일 문턱 값 원칙을 위반하게 된다.

그런데 오류율 동등성 기준을 정식화한 Hardt, Price & Srebro (2016)은 인공지능이 이 조건을 충족시키도록 각 집단별로 상이한 문턱 값을 설정하는 방안을 제시하였다.[401] 위 연구자들은 연구 내용을 시각화한 웹사이트를 구축하였는데, 위 웹사이트는 파란색 집단과 오렌지색 집단에 대해 대출을 수행하는 사례를 예로 들면서, 각 집단에 대해 서로 다른 대출 문턱 값을 적용함으로써 위음성 오류

---

399 MAS (2021) (각주 18).

400 Sandra G. Mayson, "Bias in, bias out", *Yale Law Journal* 128, (2019), p. 2294.

401 Hardt, Price & Srebro (2016) (각주 21)은 Equalized Odds를 달성하기 위해서는 각 집단별 AUC가 교차하는 점에 해당하는 문턱값을 설정하는 방안을 제시하였다.

율 동등성 기준을 충족시킬 수 있다고 소개하기도 하였다.[402] 그러나 차별금지법 제상 명시적으로 단일 문턱 값 원칙을 규정하는 경우 이러한 방식으로 오류율 동등성을 달성하는 것은 허용될 수 없다.[403] 법적으로 명시적으로 금지되지 않는 경우라고 하더라도, 이는 직접차별 금지원칙에 위반한 것으로 평가될 위험도 있다.[404]

다. 예측도 동등성

1) 의의

가) 예시 사례

인공지능 공정성 감사에서 널리 활용되는 또다른 기준은 인공지능의 예측도(predictive value)가 집단에 대해 동등할 것을 요구하는 것이다. 여기서도 정식화된 정의를 살펴보기에 앞서 이해를 돕기 위해 다음과 같은 예측도 동등성 기준을 위반하는 인공지능 사례를 먼저 고려한다.[405]

사례 2 (고위험 환자 예측 인공지능에서 집단별 예측도가 다른 사례). 의료보험회사가 환자의 건강위험도를 평가하여 예방적 조치를 취하고자 한다. 이는 고위험군 환자를 식별하여 더 자주 검진을 받도록 함으로써 질병을 조기에

---

402 Martin Wattenberg, Fernanda Viégas, and Moritz Hardt., "Attacking discrimination with smarter machine learning", http://research.google.com/bigpicture/attacking-discrimination-in-ml/ (2023. 5. 26. 접속 확인).

403 미국 민권법 제7편은 채용에 있어 단일 문턱값 원칙을 명시하고 있다.

404 이는 결국 직접차별 원칙과 간접차별 원칙의 충돌이 발현된 것으로 볼 수 있을 것이다. 이에 대해서는 본 연구 제3장 제2절 3항 참조.

405 이 사례는 Ziad Obermeyer, et al. "Dissecting racial bias in an algorithm used to manage the health of populations", *Science* 366, (2019)를 참조로 가상으로 작성한 것이다.

발견하고, 뒤늦게 중증이 발견되는 환자들의 높은 의료비 지출을 줄이는 것을 목적으로 하였다. 그런데, 건강위험점수를 사용하는 의료 직원들은 의아한 사실을 발견하였다. 고위험으로 예측된 흑인은 백인에 비해 더 건강상태가 나쁜 경우가 더 잦았던 것이다. 이 인공지능을 감사한 결과, 고위험 흑인 환자가 검사 결과 1년 내 입원이 필요한 질병이 발견될 확률은 20%인데, 고위험 백인의 경우 10%에 불과하였다.

위 사례는 집단간 예측도에 있어 차이가 있는 경우를 보여준다. '양성예측도(positive predictive value)'란 인공지능이 양성으로 예측한 사례 중 실제로 양성인 비율을 의미한다. 위 사례에서 '고위험'을 양성이라고 정의한다면, 위 인공지능은 양성예측도 동등성을 충족하지 못한다. '고위험'(예측 양성)으로 예측된 환자 중 실제 1년 내 입원이 필요한 질병이 발견(실제 양성)되는 비율(양성예측도)이 집단간 차이가 있기 때문이다.

이러한 문제는 인공지능의 목표변수를 잘못 설정한 경우 발생할 수 있다. 실제로 위 사례에서는 해당 인공지능을 학습시킬 때 고위험을 환자의 건강 수준이 아니라 환자가 지출한 의료비를 기준으로 정의하였기 때문에 이와 같은 문제가 발생한 것으로 드러났다. 즉, 낮은 소득 수준을 가진 사람들은 동일한 건강 수준에서 더 낮은 의료비를 지출하므로 실제 건강 상태보다 더 낮은 위험으로 잘못 예측되었던 것이다.

위 사례에서 예측도의 불균형이 발생하게 된 원인은, 궁극적으로 '고위험'이라는 개념이 매우 다양하게 조작적으로 정의될 수 있기 때문이다. 애초에 인공지능 학습 시에는 고위험을 '높은 의료비 지출'이라고 정의하였던 것과 달리, 인공지능 공정성 감사 시에는 고위험을 '1년 이내 입원이 필요한 질병이 발견될 확률이 높은 경우'로 정의하였다. 이러한 인공지능 감사에서 사용된 정의에도 불확실성이 존재한다. 어떠한 질병을 '입원이 필요한 질병'에 포함시키는지에 따라서도 그 결과가 크게 달라질 수 있기 때문이다. 다른 방법으로 '고위험'을 '만성질환의 수가 많은 경우' 또는 '기대여명이 짧은 경우'로 정의할 수도 있을 것이다. 여기서는 다시 '만성질환'이나 '기대여명'을 어떻게 정의할 것인지에 대한 문제가 발

생하게 된다. 이처럼 인공지능 시스템이 어떤 사회적 '구성물(construct)'인 목표 변수를 예측하고자 할 경우 이를 어떻게 조작적으로 정의할 것인지의 문제가 발생하게 된다.

실무상으로 '예측도(predictive value)'는 개인의 능력을 점수화하는 경우 중요하게 고려된다. 예컨대 어학능력 점수이 높은 예측도를 갖는다는 것은 그 점수가 높은 사람은 실제 어학능력이 높고, 그 점수가 낮은 사람은 실제 능력이 낮은 경우를 뜻한다. 신용평가도 유사하다. 높은 예측도를 가진 신용평가 인공지능이라면, 신용평가 점수가 높은 사람은 실제로 채무불이행을 하는 비율이 낮을 것이고, 반대로 신용평가 점수가 낮은 경우 실제로 채무불이행을 하는 비율이 높을 것이다. 결국 예측도란 인공지능의 예측 값을 이용하고자 하는 이들이 그 예측 결과를 얼마나 신뢰하고 사용할 수 있는지를 보여주는 지표라 할 수 있다.

나) 정식화된 정의

이상에서 설명한 예측도 동등성 조건은 다음과 같이 정식화하여 정의할 수 있다.[406]

(정의) 예측도 동등성 기준: 어떤 분류모형 $f_{classifer} : X \longmapsto \hat{Y}$ 에 대해, 보호속성 A가 $\hat{Y}$을 조건으로 진정한 목표변수 값 Y와 독립인 경우 $f_{classfier}$는 예측도 동등성을 충족한다. $f_{classfier}$가 이진분류모형인 경우, A가 $\hat{Y}$을 조건으로 Y와 독립이기 위해서는 A에 속한 모든 a, a'에 대해 다음 수식을 만족해야 한다.

$$P\{Y = 1 \mid \hat{Y} = 1, A = a\} = P\{Y = 1 \mid \hat{Y} = 1, A = a'\}$$
$$P\{\hat{Y} = 0 \mid Y = 0, A = a\} = P\{\hat{Y} = 0 \mid Y = 0, A = a'\}$$

---

406 Barocas, Hardt & Narayanan (2022) (각주 19)의 정의를 참조하였다. 위 글은 이를 '충분성(sufficiency)' 조건이라 표현한다.

위 정의는 이진분류에 있어서 집단간 양성예측도의 동등성과 음성예측도의 동등성을 모두 요구하는 것이다.[407] 이처럼 두 조건을 모두 요구하는 기준을 '충분성(sufficiency)' 조건이라고도 부른다.

예측도 동등성 조건을 일반화한 기준으로 '집단 캘리브레이션(Calibration by Group)' 조건이 있다.[408] 특정 집단에 대한 '캘리브레이션' 기준이란 평가 대상에 대해 어떤 사건이 발생할 확률을 부여하는 인공지능에 있어, 인공지능이 예측한 확률이 실제 현실에서 그 사건이 발생한 비율과 일치할 것을 요구하는 조건이다. 예컨대 만약 재범위험도 평가 인공지능이 피의자가 재범을 저지를 확률 y를 산출한다고 하면, 캘리브레이션 조건은 확률 y가 평가 데이터에 의해 확인되는 실제 재범 확률과 일치해야 한다는 것이다. 이를 수식으로 표현하면 다음과 같다.

(정의) 캘리브레이션 기준: 어떤 분류모형 $f_{classfier} : X \longmapsto \hat{Y}$에 대해, $\hat{Y}$의 모든 출력 값 y에 대해 다음 식이 성립하면 $f_{classfier}$는 캘리브레이션 기준을 충족한다.

$$P\{Y = 1 \mid \hat{Y} = y\} = y$$

캘리브레이션 기준은 인공지능 공정성 감사에 유용하게 활용된다. 만약 10,000명의 피의자에 관한 정보를 포함하고 있는 평가데이터를 이용하여 인공지

---

407 이진분류에서 $P\{Y = 1 \mid \hat{Y} = 1\}$은 양성예측도(positive predictive value; PPV; precision)를 뜻한다. 따라서 정의 중 첫번째 조건인 $P\{Y = 1 \mid \hat{Y} = 1, A = a\} = P\{Y = 1 \mid \hat{Y} = 1, A = a'\}$은 집단간 양성예측도가 동등할 것을 요구하는 것이다. 한편, $P\{Y = 0 \mid \hat{Y} = 0\}$는 음성예측도(negative predictive value)를 뜻한다. 따라서 정의 중 두번째 조건인 $P\{\hat{Y} = 0 \mid Y = 0, A = a\} = P\{\hat{Y} = 0 \mid Y = 0, A = a'\}$은 집단간 음성예측도가 동등할 것을 요구하는 것이다.

408 Barocas, Hardt & Narayanan (2022) (각주 19).

능을 감사하였는데, 감사 결과 인공지능이 그 중 500명에 재범확률 0.2를 부여하였다고 하자. 캘리브레이션 기준은 확률 0.2를 부여받은 500명 중 몇 명이 (평가 데이터상) 실제 재범을 저질렀는지 평가한다. 만약 500명 중 약 100명이 실제로 재범을 저질렀다면, 인공지능이 산출한 재범확률 0.2는 신뢰할 수 있는 예측이라고 할 수 있다. 같은 방식으로 인공지능이 산출한 다른 재범확률도 실제 재범확률과 일치한다면 해당 인공지능은 캘리브레이션 조건을 충족한다고 한다.[409]

나아가 '집단 캘리브레이션'은 평가 데이터를 각 집단별로 구분했을 때에도 캘리브레이션 조건이 충족되는 경우를 의미한다. 집단 캘리브레이션을 수식으로 표현하면 다음과 같다.

(정의) 집단 캘리브레이션 기준: 어떤 분류모형 $f_{classfier} : X \longmapsto \hat{Y}$ 에 대해, 보호속성 A에 속한 모든 a와 $\hat{Y}$의 모든 출력 값 y에 대해 다음 식이 성립하면 $f_{classifier}$는 집단 캘리브레이션 기준을 충족한다.

$$P\{Y = 1 \mid Y = y, A = a\} = y$$

어떤 인공지능이 집단 캘리브레이션 기준을 충족하는지 여부는 집단별로 구분된 캘리브레이션 도표를 그려 봄으로써 쉽게 조사할 수 있다. 예를 들어, Obermeyer et al.(2019) 연구에 포함된 캘리브레이션 도표는 아래 그림과 같다. 이는 건강 위험도 예측 인공지능에서 인종간의 집단 캘리브레이션 조건이 충족되지 못한 사례이다. 가로 축은 건강 위험점수(0점-100점, 점수가 높을수록 건

409 실무상 알고리즘을 평가할 때는 일정한 구간별로 나누어서 해당 구간에 포함된 평가대상의 실제 확률과 비교하게 된다. 예를 들어 재범확률이 0.2~0.3로 평가받은 사람들의 평균 실제 재범확률이 얼마인지 확인하는 식이다. 해당 구간은 평가 데이터의 수에 따라 더 좁게 정할 수도 있고, 넓게 정할 수도 있다.

강이 나쁨), 세로 축은 만성질환의 수이다. 만약 캘리브레이션 조건을 잘 충족하고 있다면, 건강 위험점수와 만성질환의 수는 정비례해야 할 것이고, 이는 모든 인종 집단에 대해 동일해야 할 것이다. 그런데 실제로는 동일한 위험점수를 얻은 흑인과 백인 중 흑인이 더 많은 만성질환을 겪고 있다는 점을 확인할 수 있다. 이는, 해당 위험점수가 실제 건강위험을 흑인에 대해서 과소평가하거나, 백인에 대해 과대평가하고 있다는 점을 뜻한다.

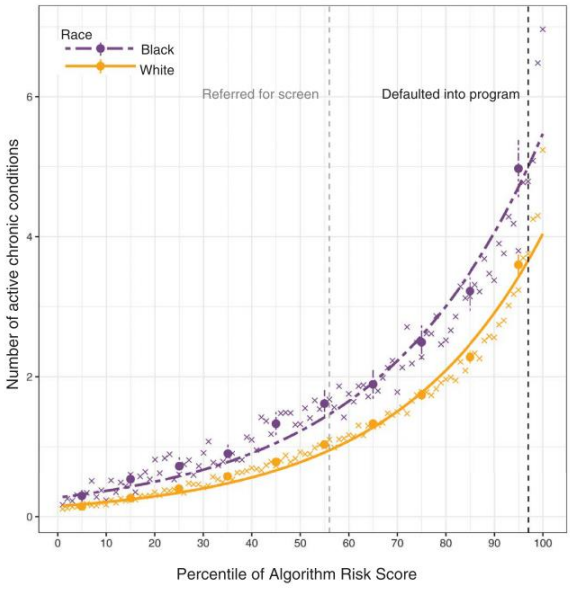


그림 13 집단 캘리브레이션이 충족되지 못하는 사례[410]

요컨대, 캘리브레이션 도표는 인공지능의 예측 값이 실제 결과와 얼마나 잘 부합하는지를 보여주고, 실제 인공지능 공정성 감사에서 유용하게 활용된다. 따라서 실무상으로는 예측도 동등성 기준이라는 표현보다 '집단 캘리브레이션' 기준 혹은 줄여서 '캘리브레이션' 이라는 표현이 자주 사용되는 것으로 보인다.[411]

410 Obermeyer et al. (2019) (각주 405).

411 ICO AI 가이드라인 초안 (2020)도 캘리브레이션이라는 표현을 사용하고 있다.

2) 한계 – 차별보존적 특성

예측도 동등성 기준은 학습 데이터 상의 편향을 제기하기 위한 명시적 개입을 하지 않더라도 머신러닝 학습 과정에서 기본적으로 충족되는 경향을 보인다.[412] 예를 들어, 아래 그림은 Barocas, Hardt & Narayanan (2022)에 소개된 캘리브레이션 도표이다.

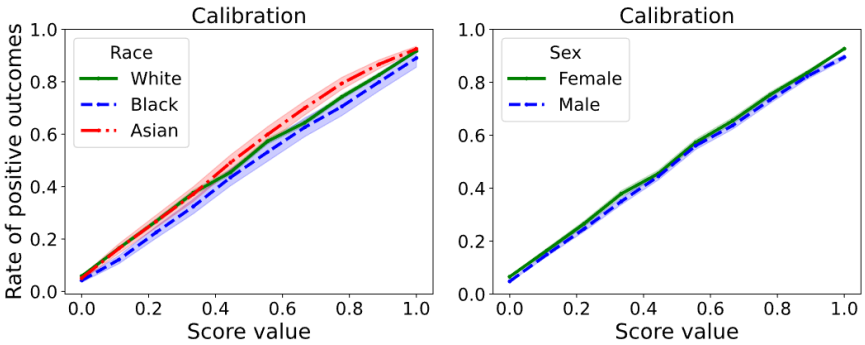


그림 14 집단 캘리브레이션이 충족되는 사례

위 그래프는 미국 인구조사국의 데이터를 바탕으로 어떤 개인의 소득이 5만 달러 이상인지 예측하는 머신러닝 시스템에 대해 집단 캘리브레이션 기준 충족 여부를 조사한 것이다. 캘리브레이션 도표의 가로 축은 인공지능이 예측한 소득이 5만 달러 이상일 확률(score value)이고, 세로 축은 특정 확률을 부여받은 사람 중 평가 데이터상 실제 5만 달러 이상이 사람의 비율이다. 각 캘리브레이션 선은 집단별로 구분되어 있는데, 인종에 대해서는 약간의 차이가 있을 뿐이고, 성별에 대해서는 거의 차이가 없다. 이 경우 해당 인공지능은 집단 캘리브레이션을 잘 충족하고 있다고 평가할 수 있다. 그런데, 위 연구자들은 해당 인공지능을 구현하는 데 있어 표준적으로 사용되는 기법을 사용하였을 뿐, 별도로 어떠한 보

412 Barocas, Hardt & Narayanan (2022) (각주 19).

정을 시도하거나, 캘리브레이션 기준을 충족하기 위해 추가적으로 노력한 바가 없었다. 이처럼 명시적인 개입이 없더라도 위 머신러닝 모형은 인종이나 성별 집단에 대해 상당히 잘 캘리브레이션이 되어 있고, 즉, 예측도 동등성 기준을 충족한다는 점을 확인할 수 있다.

이러한 이유로 인해, 예측도 동등성 기준은 인공지능에 대한 공정성 감사에 있어 가장 완화된 기준인 것으로 여겨진다. 즉, 인공지능을 개발하거나 도입하는 주체가 차별 시정을 위한 추가적 개입을 하지 않더라도 예측도 동등성 기준은 자동적으로 충족될 수 있다는 것이다(다만, 앞선 의료점수 사례에서 볼 수 있는 경우와 같이 적절한 목표 변수가 선정되어 있다는 것이 전제되어야 한다).

문제는 어떤 인공지능 공정성 감사를 수행한 결과 예측도 동등성을 충족하더라도, 해당 인공지능이 결과 동등성 조건 또는 오류율 동등성 조건을 위반하는 경우 발생한다는 점이다. 아래 제4절에서 설명하는 바와 같이 이는 이례적인 현상이 아니고, 인공지능 공정성 기준에 대한 '불가능성' 정리로 인하여 언제나 이러한 충돌이 발생할 수 있다. 이 경우 해당 인공지능은 예측도 동등성 기준을 충족한다는 이유로 정당화될 수 있는가? 예를 들어, 위 사례와 유사한 인공지능을 활용하여, 인구조사 데이터를 통해 소득 수준을 예측하는 상황을 고려하자. 이러한 인공지능으로 예측된 소득 수준은 공공 서비스나 복지 혜택 관련 정책 수립에 활용될 수 있을 것이다. 이때 그 인공지능이 활용되는 사회에서 인종이나 성별에 따른 소득 분포에 심각한 차이가 있고, 그 결과 해당 인공지능은 결과 동등성이나 오류율 동등성 조건을 중대하게 위배하고 있을 수 있다. 이러한 경우라면 비록 그 인공지능이 집단별로 동등하게 소득수준을 '정확히' 예측한다고 하더라도, 이를 활용하는 것이 규범적으로 정당화되기 어려울 수 있다.

이처럼 어떤 인공지능이 예측도 동등성 기준을 충족하더라도, 이는 기존 사회에 존재하는 불평등을 그대로 반영하는 것일 뿐이라는 점에서 비판이 제기된

다.[413] 예측도 동등성 기준은 그저 인공지능이 '정확하게' 예측할 것만을 요구할 뿐, 그러한 예측이 부정의한 사회 현실을 반영하는 것에 대해 무력하다는 것이다.

라. 예측도 동등성과 오류율 동등성과의 차이

이상에서 설명한 예측도 동등성 기준과 오류율 동등성 기준은 특정한 정확성 지표가 집단간 동등할 것을 요구한다는 점에서 유사한 측면이 있다. 하지만 해당 지표의 지향점에는 중대한 차이가 있다. 예측도 동등성(또는 집단 캘리브레이션)은 인공지능이 예측한 값(예: 개인의 능력이나 위험)을 이용하는 의사결정자의 관점에서 중요하다. 집단 캘리브레이션은 인공지능이 출력한 확률의 예측도가 집단 간 동일해야 한다는 것이다. 이는 예측 값을 이용하고자 하는 이들이 인공지능을 신뢰하고 사용하기 위해 필요한 조건이다. 예를 들어 어학능력 평가 인공지능이 성별에 대해 예측도 동등성 조건을 충족시키지 못한다면, 인공지능이 같은 점수를 예측하더라도 성별마다 실제 어학능력에 있어 차이가 있다는 의미가 된다. 이러한 인공지능은 신뢰하고 사용하기 어렵다.

이와는 달리 오류율 동등성은 인공지능에 의해 평가 대상이 되는 개인의 실제 조건을 기준으로 공정성을 심사한다는 차이가 있다. 특히, 위음성율은 실제 적격인 사람을 부적격으로 예측하는 오류율을 기준으로 한다. 어학능력 평가 인공지능에서 위음성율은 실제로 어학능력이 높은 개인임에도 불구하고, 인공지능에 의해 잘못 어학능력이 낮다고 예측되는 오류율을 의미한다. 이는 결국 평가 대상이 된 개인이 부당한 평가를 받게 되는 비율을 뜻하는 것으로 이해할 수 있다. 이처럼 오류율 동등성은 평가 대상이 되는 개인의 입장에서 차등적 대우를 받을 확률을 의미하게 된다. 따라서 해당 개인의 입장에서는 오류율 동등성이 더 중요하게 고려된다.

---

413 Wachter, et al. (2021b) (각주 147).

이처럼 예측도 동등성과 오류율 동등성은 인공지능 공정성의 서로 다른 측면을 심사하는 기준이다. 따라서 양자 모두 달성할 가치가 있고, 인공지능 감사에 있어 그 기준을 모두 활용할 필요도 있다. 그러나 다음 절에서 살펴보는 바와 같이 불가능성 정리로 인해 예측도 동등성과 오류율 동등성은 동시에 충족시킬 수 없다는 것이 수학적으로 증명되었다. 따라서 주어진 맥락에 적합한 기준을 선택하여 적용할 필요가 있게 된다.

제4절 인공지능 공정성 지표 간의 충돌과 불가능성 정리

1. COMPAS 사례

인공지능 공정성 평가 기준간 충돌 문제를 잘 보여주는 사례는 미국의 COMPAS 재범 예측 소프트웨어를 둘러싼 논쟁이다. COMPAS는 미국 법원에서 피의자의 재범 위험성을 예측하기 위해서 널리 사용되어 오던 통계 프로그램이다. 그런데 2016년 미국 탐사보도 조직인 프로퍼블리카(ProPublica)는 별도로 확보한 평가 데이터를 이용하여 COMPAS 소프트웨어의 공정성 감사를 수행하였고, 그 결과 COMPAS가 흑인에 대해 차별적으로 동작한다는 문제를 지적하였다.[414]

프로퍼블리카는 COMPAS 소프트웨어의 인종 편향성을 보여주기 다음과 같은 통계 분석 결과를 제시하였다. 즉, “Labeled Higher Risk, But Didn’t Re-Offend”의 “비율”이 백인 23.5%, 흑인 44.9%라고 주장하였다. 평가 데이터상 범죄를 저지르지 않는 흑인들의 44.9%가 고위험으로 잘못 평가된다는 것이다. 나아가 “Labeled Lower Risk, Yet Did Re-Offend”의 “비율”이 백인 47.7%, 흑

414 Angwin et al. (2016) (각주 24).

인 28.0%라고 주장하였다. 이는 평가 데이터상 범죄를 저지른 백인들의 47.7%가 저위험으로 잘못 평가된다는 것이다. 아래 표는 프로퍼블리카가 제시한 표이다.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

그림 15 COMPAS 소프트웨어에 대한 프로퍼블리카의 감사 결과

그러나 COMPAS 개발사인 노스포인트(Northpointe) 사는 프로퍼블리카 주장의 문제점을 보여주기 위해 동일한 표를 다르게 작성하였다.[415] 노스포인트 사가 작성한 표에 따르면 "Labeled Higher Risk, But Didn't Re-Offend"의 "비율"이 백인 41%, 흑인 37%이고, "Labeled Lower Risk, Yet Did Re-Offend"의 "비율"이 백인 29%, 흑인 35%라고 주장하였다. 이 표는 프로퍼블리카와 완전히 동일한 것처럼 보이지만, 그 수치는 전혀 다르다. 뿐만 아니라 심각해 보이던 인종간 격차가 사라졌다.

415 Dieterich, et al. (2016) (각주 24).

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	41%	37%
Labeled Lower Risk, Yet Did Re-Offend	29%	35%

Table 3.1: Propublica's table with correct target population errors at the study cut point (Low vs. Not Low) for the General Recidivism Risk Scale.

그림 16 COMPAS 소프트웨어에 대한 노스포인트 사의 감사 결과

두 표는 모두 그 라벨이 정확하지 않다. 위 표에 제시된 수치에는 분모가 표시되어 있지 않기 때문이다. 프로퍼블리카는 대중을 위한 기사에서 제시되는 표를 복잡하게 만들지 않기 위해 의도적으로 단순화한 것일 수 있다. 노스포인트는 프로퍼블리카의 주장을 반박하기 위해 의도적으로 프로퍼블리카와 같은 라벨을 붙였을 것이다. 하지만 위 표의 비율이 무엇을 분모로 하고 있는지는 맥락상 분명하다. ① 프로퍼블리카는 평가 데이터에 재범을 저지르지 않은 사람, 즉, "Didn't Re-Offend"의 수를 분모로 삼았다. 즉, 프로퍼블리카가 인종간 비교를 위해 사용한 통계적 지표는 위양성율와 위음성율의 동등성이다. 이는 '오류율의 동등성' 기준을 적용한 것이라 할 수 있다. ② 이에 비해 노스포인트는 알고리즘이 고위험이라고 예측한 사람들, 즉, "Labeled Higher Risk"의 수를 분모로 삼았다. 즉, 노스포인트 사가 사용한 통계적 지표는 양성예측도와 음성예측도의 동등성이었다.

프로퍼블리카가 그 기사에서 분모를 표시하지 않은 것은 잘못된 것이라 평가할 수도 있다. 어떤 통계치가 비율로 표시된 경우, 무엇을 분모로 하는지에 따라 전혀 다른 의미를 가질 수 있기 때문이다. 하지만 이보다 여기서 더 중요하고 어려운 의문은 COMPAS의 공정성을 심사할 때 과연 프로퍼블리카와 같이 '오류율의 동등성' 기준을 사용할 것인지, 노스포인트 사의 반박처럼 '예측도 동등성'을 기준으로 해야 할 것인가 하는 점이다.

프로퍼블리카와 노스포인트 사의 주장에는 모두 이를 뒷받침하는 제법 좋은 근거가 있다. 프로퍼블리카의 주장은 알고리즘이 실제 재범을 저지르지 않을 사

람을 고위험으로 예측하는 위해가 집단 간 불균등하게 가해지는 것(즉, 개인에 대한 위해)이 불공정하다는 것이다. 또한 재범을 저지를 사람을 저위험으로 예측하는 위해(즉, 사회에 대한 위해)가 집단 간에 불균등하게 가해지는 것 역시 불공정하다고 본다. 집단별 위양성율과 위음성율은 이러한 위해가 가해지는 비율을 뜻하므로, 해당 비율의 집단간 차이를 측정하는 것이 타당하다고 주장한다.

이에 비해 노스포인트는 알고리즘을 이용할 법집행 기관이나 판사의 관점에서 생각할 것을 요구한다. 캘리브레이션이 되어 있지 않은 알고리즘은 신뢰하고 사용할 수 없다. 어떤 피의자에 대해 위험 점수 9점이 부여되었는데, 만약 그 피의자가 흑인일 경우 재범 확률이 80%이고, 백인일 경우 재범 확률이 75%여서는 안 될 것이다. 이처럼 재범 위험 점수는 피의자의 인종과 무관하게 동등한 재범 확률을 표시해야 한다. 예측도 동등성은 이처럼 예측 정확도의 집단간 차이를 측정하므로 노스포인트는 예측도 동등성을 적용하는 것이 적절하다고 주장한다.

이 문제는 알고리즘 공정성에 관한 도덕적 딜레마이지만, 이에 대한 해법은 아직 명확하게 제시되어 있지 않다. 이상에서 설명한 COMPAS 사례는 알고리즘 공정성 연구자 사이에서 잘 알려진 사례로서, 거듭하여 논의되어 왔다. 그럼에도 불구하고, 여전히 이 사례에서 프로퍼블리카의 주장처럼 오류율 동등성을 적용하는 것이 타당한지, 노스포인트 사의 주장처럼 예측도 동등성을 달성하는 것만으로 적절한지는 분명하지 않다.

2. 인공지능 공정성 지표에 대한 불가능성 정리

인공지능 공정성 기준에 대한 불가능성 정리는 2016년 무렵부터 발표되어 왔다. Kleinberg et al. (2016)은 위험 점수를 매기는 알고리즘의 경우 집단간 캘리브레이션 기준과 일반화된 오류율 동등성 기준을 동시에 충족할 수 없다는 점

을 수학적으로 증명하였다.[416] 또한 Chouldechova(2017)은 오류율 동등성(위양성율과 위음성율의 동등성) 및 예측도 동등성(양성 예측도와 음성 예측도의 동등성)을 동시에 달성할 수 없음을 보였다. Barocas, Hardt, & Narayanan (2022)는 선행 연구들을 일반화하여 출력변수에 대한 공정성 기준 중 (i) 결과 동등성, (ii) 오류율 동등성, (iii) 예측도 동등성은 동시에 충족하는 것이 불가능하다는 점을 증명하였다.[417] 다만, 모든 경우에 공정성 기준을 동시에 충족하는 것이 불가능한 것은 아니다. 즉, (1) 각 집단간 기본 비율이 동일한 경우(즉, 원래 각 집단간 아무런 차이가 없는 경우)나 (2) 알고리즘이 완벽한 예측을 하는 경우는 위 불가능성 정리가 성립하지 않고 여러 공정성 기준을 동시에 충족할 수 있다. 하지만 현실적으로 이러한 상황을 기대하기는 극히 어려울 것이다.

한편, 오류율 동등성 기준을 위양성 오류율 동등성과 위음성 오류율 동등성을 포함하는 것인데, 위 두 기준은 동시에 충족하는 것이 가능하다. 또한 예측도 동등성 기준에 포함된 양성 예측도 동등성과 음성 예측도 동등성 기준을 동시에 충족하는 것도 가능하다.[418]

---

416 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent trade-offs in the fair determination of risk scores", arXiv preprint arXiv:1609.05807 (2016).

417 Barocas, Hardt & Narayanan (2022) (각주 19).

418 인공지능 공정성 기준에 대해 친숙하다고 생각하는 연구자들조차도 여러 공정성 기준 간의 충돌 관계에 대해서는 완전히 이해하지 못한 실정이다. Michelle Seng Ah Lee and Jatinder Singh, "The Landscape and Gaps in Open Source Fairness Toolkits," *CHI Conference on Human Factors in Computing Systems* (2021)에 따르면 인공지능 공정성 연구자들에 대해 일반적으로 상충되는 공정성 기준에 대해 질문을 했는데, 오직 44%만이 오류율 동등성(equalized odds)와 예측도 동등성(predictive parity)가 충돌한다는 정답을 선택했고, 36%는 오류율 동등성과 위음성율 동등성(equal opportunity)가 충돌한다는 오답을 선택했다(위음성율 동등성은 오류율 동등성의 하위 기준이므로 양자는 상호 충돌하지 않는다). 연구자 중 나머지 20%는 잘 모르겠다고 답했다고 한다.

3. 인공지능 공정성 지표에 대한 선택 논쟁

이와 같이 인공지능 공정성 기준 간의 충돌 문제 및 불가능성 정리가 알려진 이후 공정성 지표 간의 선택 문제, 즉 그들 간의 우선순위를 정하는 문제에 대한 논의가 최근 활발히 진행 중이다. 이에 관한 주장은 대체로 (1) 결과 동등성을 우선시하는 견해[419] (2) 오류율 동등성을 우선시하는 견해[420] (3) 예측도 동등성을 우선시하는 견해[421] 등으로 구분할 수 있다.

우선, Wachter et al. (2021b)는 결과 동등성(또는 조건부 인구통계적 동등성)을 지지한다. 위 연구는 인공지능 공정성 지표를 기존의 사회적 편향을 시정하는 성격을 갖는 것(bias-transforming)과 이를 보존적 성격을 갖는 것(bias-preserving)으로 구분한다. 그런데, 결과 동등성이나 조건부 인구통계적 동등성은 기존의 차별을 시정하는 것을 목적으로 하는 것으로서 실질적 평등을 달성하기 위해서는 이러한 지표들을 사용할 필요가 있다고 주장한다. 이와 달리 오류율 동등성이나 예측도 동등성은 형식적 평등만을 달성하기 위한 것으로서 편향 보존적 지표에 불과하고, 결국 기존 사회적 불평등을 영속화하는 역할을 할 뿐이라고 주장한다.

이러한 주장에는 이미 어떠한 사회에서 개인의 잠재적 능력이나 위험의 확률이 부정의한 이유로 불균등하게 분포되어 있는 경우, 이를 정확하게 예측하더라도 부정의한 분포를 그대로 재현하는 것에 불과하므로 도덕적으로 흠결이 있다는 관점이 전제되어 있다. 특히 인공지능에 의해 평가되는 개인의 잠재적 능력이나 위험의 차이는 출생 시의 선천적 차이(체력, 민첩성, 지능) 뿐만 아니라 출생 시의 지위에 따라 사회적으로 결정되는데, 이처럼 도덕적으로 무관한 사정으로 인해 개인의 능력이나 위험에 대한 평가가 달라지는 것은 불공정하다고 평가될 수 있다. 예컨대, Hertweck et al. (2021)은 이를 인생 편향(life's bias)라고 부르

---

419 Wachter et al. (2021b) (각주 413); Hertweck et al. (2021) (각주 32).

420 Hardt, Price & Srebro (2016) (각주 21); Zafar, et al. (2019) (각주 235).

421 Hedden (2021) (각주 32).

는데, 인공지능 공정성 평가에 사용되는 지표는 인생 편향을 시정할 수 있어야 한다고 본다.[422]

이러한 주장에 대해 Hedden (2021)은 반드시 인공지능이 사회적 불평등을 시정하는 역할을 해야 할 것을 요구할 근거가 없다고 비판한다. 인공지능 알고리즘에 대해서는 개인의 잠재적 능력이나 위험을 정확하게 예측할 것을 요구하고, 사회적 불평등 문제는 다른 방법으로 해결해야 한다는 것이다. 그는 교통혼잡 문제를 해결하기 위해 부과되는 혼잡세(congestion pricing) 사례를 든다. 혼잡세는 가난한 이들에게 추가적인 부담을 지우고, 그 결과 분배 부정의를 초래할 수 있다. 그러나 혼잡세를 통해 해결하고자 하는 교통혼잡 문제와 혼잡세의 역진적인 성격으로 인한 분배 부정의의 문제는 서로 구분하여 접근할 수 있다. 오히려 혼잡세 통행으로 얻어지는 수입을 가난한 이들에게 재분배하는 정책이 더 사회적으로 바람직한 결과를 초래할 수 있는 것이다.[423]

인공지능에 대해서도 마찬가지 논변이 성립할 수 있다. 기존의 만연한 사회적 불평등을 해결하기 위한 여러 해법이 존재할 수 있고, 인공지능에 대해 결과 동등성을 요구하는 방안이 유일한 해법은 아니다. 오히려 인공지능 공정성 평가 기준에 대한 불가능성 정리에 의해, 만약 우리가 다른 공정성 척도를 요구할 경우 예측도 동등성 기준을 희생해야 한다. 결국, Hedden의 주장은 인공지능 공정성 기준으로는 상대적으로 완화된 척도인 예측도 동등성만을 요구하면서(그 결과, 인공지능은 예측 정확도를 최대한 높이도록 학습하는 것이 허용된다), 인공지능이 기존의 사회적 편향을 반영하는 문제에 대한 해법은 별도로 고려할 필요가 있다는 취지로 이해할 수 있다.

마지막으로, 결과 동등성과 예측도 동등성에 대한 절충안으로서 오류율 동등성이 적절한 기준으로 고려하는 견해도 제시된다. 예를 들어 Zafar et al. (2019)는 오류율 동등성 기준은 다른 기준들의 한계를 극복하는 새로운 기준으로서 "차

---

422 Hertweck, et al. (2021) (각주 32).

423 Robert H. Frank, *Success and Luck*, Princeton University Press, (2016).

등적 부당대우(disparate mistreatment)"를 막는 것이라 본다. 이들은 오류율 동등성 지표란 어떤 결정의 정확성을 담보하기 위한 것으로서, 그 측정이 용이할 뿐만 아니라 결과 동등성을 달성하기 위해 발생할 수 있는 역차별의 문제를 피할 수 있는 장점이 있다고 설명한다.[424] 앞서 본 바와 같이 프로퍼블리카가 COMPAS 소프트웨어에 대한 감사 사례에서 오류율 동등성 기준을 적용한 이래, 오류율 동등성 지표는 인공지능 공정성 감사 실무에서 사용되어 오고 있다. 예컨대, 싱가포르 통화청의 2021년 인공지능 공정성 기준 적용 사례 보고서에서도 오류율 동등성 기준을 공정성을 평가하는 일응의 기준으로 활용하고 있다.

4. 소결

이처럼 인공지능 공정성 지표 간의 충돌 및 적절한 공정성 지표의 선택 문제는 불가능성 정리로 인해 해결이 쉽지 않은 문제로 여겨지고 있다. 그 결과 한 인공지능 공정성 연구자는 이러한 논쟁에 관해 "대부분의 일반적인 지혜는 적절한 지표의 선택이 인공지능의 활용 사례의 맥락에 따라 달라야 한다"는 것이라고 요약하기도 한다.[425] 이러한 배경 하에서, 다음 장에서는 구체적인 시뮬레이션을 통하여 각각의 공정성 지표가 어떠한 맥락에서 정당화될 수 있는지 검토한다. 나아가 기존 공정성 지표의 한계를 검토하고, 이상과 같은 충돌 문제를 해결할 수 있는 대안적인 공정성 감사 방안을 모색한다.

---

424 Zafar et al. (2019) (각주 235).

425 Percy Liang, et al., "Holistic evaluation of language models", arXiv preprint arXiv:2211.09110 (2022).

제5장 분류모형 인공지능의 공정성 지표 선택

제 1 절 시뮬레이션 방법론

1. 목표변수의 확률분포

가. 인지모형: 결정론적 목표변수

앞서 제4장 제3절에서 살펴본 사례 1을 상기해 보자. 해당 사례에서, 한 회사는 직원 출입구에 얼굴인식 소프트웨어를 이용하여 직원인지 여부를 인식하는 소프트웨어를 도입하였다. 이때 우리는 해당 인공지능이 직원의 피부색에 대해 공정하게 작동하는지 여부를 판단하기 위해서 어떠한 심사 기준을 도입해야 할 것인가라는 문제를 해결하고자 한다.

그런데, 직원인지 아닌지 여부는 결정론적으로 정해지고, 확률적으로 직원인 경우는 고려할 필요가 없다. 위 분류모형이 학습되고 적용되는 시점에서 한 사람은 오로지 직원이거나 직원이 아니거나 할 뿐이기 때문이다. 따라서 얼굴인식 소프트웨어가 예측하고자 하는 목표변수 Y는 입력된 영상 속의 얼굴이 직원의 얼굴이면 1(Yes), 아니면 0(No)의 값을 갖는다. 이처럼 관측된 Y값이 확률적으로 결정되지 않고 결정론적으로 정해져 있는 경우를 '퇴행 분포(degenerate distribution)'라 한다. 예컨대, 어떤 사람이 직원일 확률을 나타내는 확률변수 Y_i는 다음과 같이 정의할 수 있다.

직원인 i에 대해, $Pr(Y_i = 1) = 1$

직원이 아닌 j에 대해 $Pr(Y_j = 0) = 1$

만약 위 사례에서 매일 회사 입구에 들어서는 사람이 총 1,000명이고, 그 중 500명이 직원이고 나머지 500명은 직원이 아니라면, 검사 대상 인원 중 50%는 Y=1의 값을, 나머지 50%는 Y=0의 값을 갖는 확률 분포로 이를 모사할 수 있다. 아래 그림은 이상과 같은 분류모형의 작동 원리를 설명한 것이다. 그런데, 본 연구의 제2장 제2절에서 살펴본 것처럼 이처럼 목표변수가 퇴행분포에 해당하는 분류모형은 대체로 인지모형에 해당한다. 즉, 인지모형은 주로 사물 인식, 얼굴 인식 등과 같이 명확한 정답이 있는 문제를 해결하는 경우 주로 활용되는 것으로서, 위 사례는 인지모형을 모사하는 것으로 이해될 수 있다.

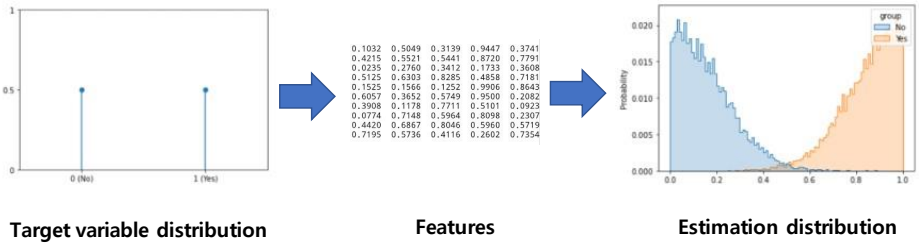


그림 17 목표변수가 퇴행분포인 경우 분류모형의 적용 과정

나. 판별모형: 확률적 목표변수

다음으로, 예측하고자 하는 변수가 확률적으로 결정된 변수인 경우에도 분류모형이 빈번히 활용된다. 예를 들어, 수학 능력, 업무 능력, 재범 위험성, 운전 능력 등 개인이 가지고 있는 어떠한 능력(merit)이나 위험(risk)을 예측하는 경우를 상정해 보자. 이 경우 관측가능한 실현된 결과(observable realized outcome)는 특정 개인이 가지고 있는 해당 자질로부터 확률적으로 정해진다. 가령 운전 능력을 사고 기록을 통해 관측하는 경우, 사고 여부는 다양한 변수의 영향을 받아 확률적으로 정해진다. 운전능력이 나쁜 운전자가 운이 좋게 사고를

내지 않는 경우도 있고, 운전능력이 좋은 운전자가 운이 나쁘게 사고를 내는 경우도 있다.

이처럼 각 개인에 대해, 우리가 관심을 갖는 능력이나 위험에 대해 관측될 수 없는 잠재적(latent) 확률이 정해져 있다는 가정은 다수의 인공지능 연구에 있어 매우 흔하게 이루어지는 가정으로서, 그 타당성에 대한 별도의 정당화를 거치지 않고 수용되는 것으로 보인다. 예컨대 피의자의 재범 위험성 예측 머신러닝 알고리즘에 관해 널리 인용되는 Kleinberg et al.(2018b)은 각 피의자는 재판 전 석방될 경우 재범을 저지를 (알려지지 않은) 확률 p를 가지고 있다고 가정한다.[426] 다른 예로, 신용평가에 관한 머신러닝 알고리즘을 법적으로 평가하는 Gillis & Speiss (2020) 연구는 부동산 담보 대출 신청자들마다 채무불이행(default)을 할 확률 p가 정해져 있다고 가정한다.[427]

다만, 이 가정에 대해서는 지나치게 단순화한 것이라는 비판이 제기될 수 있다. 예컨대, 어떤 사람의 능력이나 위험이 실현될 가능성은 μ를 평균으로, σ를 표준편차로 하는 정규분포에 해당한다고 가정하고, 해당 정규분포에서 추출된 값에 따라 결과가 생성된다고 모형화할 수도 있다. 이러한 모형은 어떤 사람은 능력이 평균적으로 우수할 뿐만 아니라(μ가 큼), 항상 우수한데 비해(σ가 작음), 어떤 사람은 평균적으로 우수하지 않지만(μ가 작음), 가끔 매우 나쁘거나 매우 우수한 결과를 낳는(σ가 큼) 상황을 달리 모형화할 수 있는 장점이 있다. 하지만 이처럼 개인에 대해서 여러 모수(parameter)를 이용하면 시뮬레이션 작업이 지나치게 복잡해질 수 있다. 따라서 본 시뮬레이션 시나리오에서는 기존의 연구들과 마찬가지로 각 개인이 특정한 p 값을 갖고, 이에 따라 결과가 결정된다고 가정한다.

---

426 Jon Kleinberg, et al., "Human decisions and machine predictions", *The quarterly journal of economics* 133.1 (2018b).

427 Gillis & Spiess (2020) (각주 205).

본 연구에서는 전체 모집단에 대해 p 값이 아래와 같이 절단정규분포(truncated normal distribution)을 가지고 있다고 가정한다.[428] 아래 그래프는 절단정규분포에서 평균이 0.5인 경우를 나타낸다.

$$p \sim Trucated\ Normal(\mu, \sigma)$$

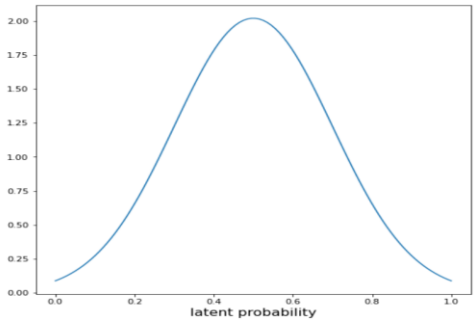


그림 18 확률적 목표변수의 절단정규분포

어떤 지원자 i가 p_i 의 확률변수를 가지고 있을 때, 그 지원자가 합격하였는지 여부를 나타내는 확률변수 Y_i는 베르누이 분포(Bernoulli distribution)에 따라 다음과 같이 결정된다고 가정한다.

$$\text{어떤 지원자 i에 대해, } Pr(Y_i = 1) = p_i\,, Pr(Y_i = 0) = 1 - p_i$$

아래 그림은 이상과 같은 사례에 있어 머신러닝의 작동 원리를 설명한 것이다. 즉, 사회 구성원 각자에게 어떠한 잠재적 확률변수 p_i가 부여되어 있다고 가정하고, 그 확률은 구성원마다 서로 다르게 분포되어 있다. 나아가 우리는 그 개인에 관한 여러 속성이나 과거 행동 정보를 관측할 수 있고, 이러한 정보를 이용

428 활용 사례에 따라서는 다양한 분포를 가질 수 있음에 유의할 필요가 있다. 예컨대 COMPAS dataset의 재범 위험율 분포는 정규분포로는 모델링할 수 없을 것이다.

하여 잠재적 확률변수 값을 추정해 낼 수 있다. 그 정보를 특성(feature)이라 한다. 나아가 머신러닝 알고리즘은 특성 정보를 이용하여 원래 그 개인이 가지고 있을 잠재적 확률변수 p_i 값을 추정한다. 이처럼 목표변수가 확률변수에 해당하는 분류모형은 대체로 판별모형에 해당한다. 즉, 판별모형은 개인의 잠재적 능력이나 위험을 평가하거나 행동을 예측하는 경우 주로 활용되는 것으로서, 본 시뮬레이션 시나리오는 판별모형을 모사하는 것으로 이해될 수 있다.[429]

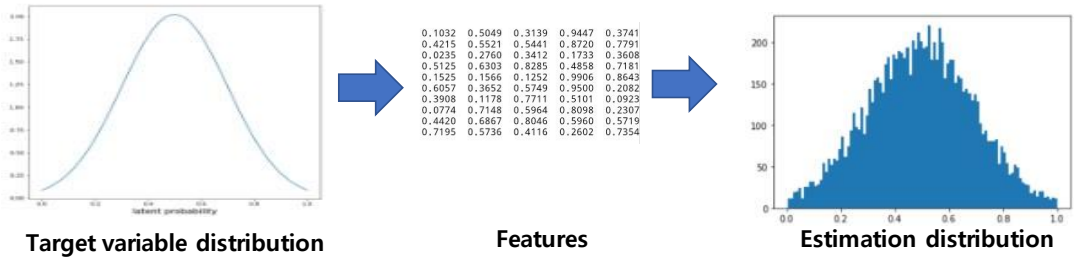


그림 19 목표변수가 정규분포인 경우 분류모형의 적용 과정

2. 머신러닝 모형

본 시뮬레이션에서는 머신러닝 모형으로 로지스틱 회귀를 이용하였다. 로지스틱 모형(logistic model)[430]은 이진분류에서 널리 활용되는 모형으로서, 다음 수식을 통해 목표변수를 추정한다.[431]

$$\hat{p} = P(Y = 1) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + \cdots + w_M x_M)}}$$

429 본 연구의 제2장 제2절 참조.

430 이를 로짓 모형(logit model)이라고도 한다.

431 크리스토퍼 비숍 저, 김형진 옮김, *패턴 인식과 머신 러닝*, 제이펍, (2019), pp. 221-2.

위 수식의 특징은 그 출력 값 $\hat{p}$이 확률 값이라는 점이다(이러한 추정 값을 진정한 확률 값 p와 구분하기 위하여 $\hat{p}$으로 표시한다). 즉, 위 모형은 어떤 개인이 어떤 유형(class)에 해당할 확률이 얼마인지 알려준다. 예를 들어, 로지스틱 모형을 사용하여 만약 어떤 학생이 공부한 시간 변수(x)에 기초하여 합격할 것인지 여부($Y = 0\ or\ 1$)를 예측하는 모형은 다음과 같은 형태를 취하게 된다.[432]

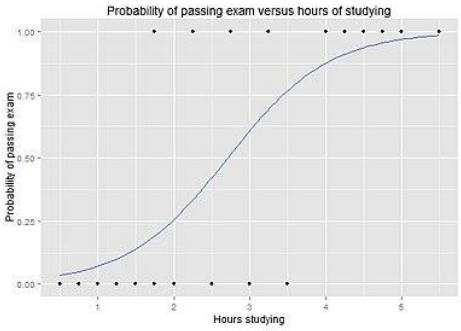


그림 20 로지스틱 회귀 모형 예시 사례

분류모형은 이처럼 추정된 확률 $\hat{p}$을 이용하여 어떠한 결과를 예측한다. 즉, 위 사례에서 모형이 산출한 확률 $\hat{p}$이 일정한 문턱 값 이상이면 합격할 것으로, 그 미만이면 불합격할 것으로 예측할 수 있다. 가령 위 그림에 묘사된 사례에서는 대략 2.7시간 이상 공부한 학생은 합격 확률이 50%으로 추정된다. 만약 문턱 값을 50%로 설정한다면, 2.7시간 이상 공부한 학생은 합격(Y=1)으로 예측된다.

이처럼 인공지능에 입력된 특성 변수를 이용하여 원래의 p의 분포를 추정하는 과정을 수행한다는 점은, 예측하고자 하는 목표변수가 결정론적인 경우(즉, 인지모형)와 확률적인 경우(즉, 판별모형) 모두에 공통된다. 예컨대 앞서 예로 든 얼굴인식 인공지능은 결정론적인 목표변수(직원에 해당하는지 여부)를 예측하는 것으로서, 위 인공지능은 주어진 입력 영상에 대해 그 사람이 회사 직원일 확률

432 https://en.wikipedia.org/wiki/Logistic_regression (2023. 5. 23. 접속 확인)

을 출력한다. 나아가 본 연구에서 고려하는 분류모형은 출력된 확률 값이 일정한 문턱 값을 넘으면 직원으로 예측하고, 그 값 이하이면 직원이 아닌 것으로 예측한다.

제 2 절 인지모형 시나리오

1. 시뮬레이션 방법[433]

인지모형 시나리오는 목표변수가 결정론적으로 정해져 있는 경우(즉, 퇴행분포인 경우)를 고려한다. 모집단에는 A 집단과 B 집단이 존재하고, B 집단이 A 집단에 비해 역사적으로 차별을 받아온 집단이라 가정한다. 본 시뮬레이션에서는 다음과 같은 편향 존재의 가능성에 따라 시나리오를 구분한다.

(1) **[사회적 편향]** 학습 데이터에 있어, 각 집단 중 양성에 해당하는 개인의 비율(prevalence)에 있어서 차이가 있는지 구분한다. 두 집단간 차이가 있다고 가정할 경우 A 집단은 80%가 양성이고, B 집단은 오직 20%만이 양성인 경우를 상정한다.

(2) **[데이터 측정 편향]** 특성 변수(X_1, X_2)에 반영되는 노이즈에 있어 차이가 있는지 구분한다. 두 집단간 차이가 있다고 가정할 경우, A 집단은 적은 데이터 노이즈(평균 0, 표준편차 0.3인 정규분포), B 집단은 높은 데이터 노이즈(평균 0, 표준편차 0.6인 정규분포)가 포함된 경우를 상정한다. 만약 차이가 없다면 두 집단 모두 관측된 특성 변수에 동일한 노이즈(평균 0, 표준편차 0.5인 정규분포)를 반영한다.

---

433 본 시뮬레이션의 상세에 관해서는 별지 1 참조.

(3) **[Unawareness 기준 충족 여부]** 해당 분류모형이 각 개인이 어떠한 집단에 속하는지 여부(민감 속성)를 알 수 있는지 구분한다.

가. 베이스라인(baseline) 시나리오

만약 두 집단에 앞서 설명한 차이가 전혀 없다고 하자. 이때 A 집단, B 집단 각 5,000명에 대한 가상의 학습 데이터를 생성하여 로지스틱 회귀 모델을 학습시키고, 마찬가지 방식으로 생성한 평가데이터에 대해 정확도를 측정하였다. 이때 전체 집단에 대한 정확도는 91.9%, A 집단 92.1%, B 집단 91.7%이다. 이처럼 두 집단에 어떠한 차이가 없다고 하더라도, 관측된 특성 변수에 임의의 노이즈가 반영되므로 0.5% 수준의 차이는 발생할 수 있다.

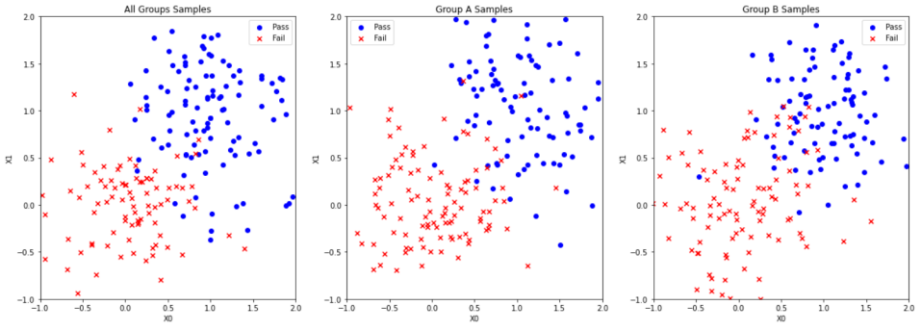


그림 21 베이스라인 시나리오의 학습 데이터 분포

나. 차별적 시나리오

본 시뮬레이션에서 사용된 시나리오는 다음과 같다. 모든 시나리오에서 (1) 사회적 편향은 존재하는 것으로 가정하였고 (2) 데이터 측정 과정의 편향과 (3) Unawareness 기준 적용 여부를 달리하였다.

(1) **시나리오 (1-1)** – 모든 편향이 존재하는 경우. 즉, 사회적 편향과 데이터 측정 편향이 모두 존재하고 모형이 민감속성 정보를 알 수 있는 경우

(2) **시나리오 (1-2)** – 사회적 편향과 데이터 측정 편향이 존재하는데, 모형이 민감속성 정보를 알 수 없는 경우

(3) **시나리오 (1-3)** – 사회적 편향은 존재하지만 데이터 측정 편향은 존재하지 않는 경우로서 모형이 민감속성 정보를 알 수 없는 경우

(4) **시나리오 (1-4)** – 사회적 편향은 존재하지만 데이터 측정 편향은 존재하지 않는 경우로서 모형이 민감속성 정보를 알 수 있는 경우

2. 시뮬레이션 결과

시뮬레이션 결과 측정된 집단간 정확도 지표는 다음과 같다.

	인구통계적 동등성		오류율 동등성		예측도 동등성	
	정확도	합격률	위음성 오류율	위양성 오류율	FOR 비율	FDR 비율
Baseline (동등)	A: 92.1% B: 91.7%	A: 49.8% B: 49.7%	A: 8.1% B: 8.6%	A: 7.6% B: 8.0%	A: 8.1% B: 8.6%	A: 7.7% B: 8.1%
시나리오 1-1	**A: 99.2% B: 91.3%**	**A: 80.5% B: 18.7%**	**A: 0.2% B: 24.9%**	**A: 3.5% B: 4.6%**	**A: 0.8% B: 6.1%**	**A: 0.9% B: 19.7%**
시나리오 1-2	A: 98.2% B: 89.6%	A: 78.3% B: 23.7%	A: 2.2% B: 16.8%	A: 0.2% B: 8.8%	A: 8.0% B: 4.4%	A: 0.1% B: 29.8%
시나리오 1-3	A: 92.1% B: 92.1%	A: 75.7% B: 24.8%	A: 7.6% B: 8.0%	A: 8.8% B: 7.9%	A: 25.0% B: 2.1%	A: 2.3% B: 25.6%
시나리오 1-4	A: 94.4% B:94.4%	A: 81.1% B:18.5%	A: 2.8% B: 17.8%	A: 16.6% B: 2.6%	A: 11.8% B: 4.4%	A: 4.1% B:11.3%

표 4 인지모형 시나리오의 시뮬레이션 결과

가. 시나리오 1-1 검토

시나리오 1-1은 단순 정확도나 합격률과 같은 지표를 사용하는 것의 문제를 보여준다. 본 시나리오는 두 집단 모두에 대해 매우 높은 정확도(A 집단 99.2%, B 집단 91.3%)를 보인다. 하지만 해당 인공지능 모형은 인구통계적 동등성(합격률의 차이), 오류율 동등성, 예측도 동등성 기준을 모두 중대하게 위배하고 있으므로, 이를 공정한 것으로 평가하기 어렵다. 다만, 여기서 인구통계적 동등성 위반은 인공지능 자체의 문제라 평가하기 어렵다. A 집단과 B 집단 자체에서 양성인 사례의 구성비가 매우 차이가 있는 것으로 전제하였고(A 집단 80%, B 집단 20%), 학습된 모형은 그 값에서 크게 벗어나지 않게 예측하고 있기 때문이다(A 집단 80.5%, B집단 18.7%).

본 사례에서 인공지능 공정성 감사의 측면에서 특히 문제시되어야 할 점은 위음성 오류율에 있어 상당한 차이를 보이고 있다는 점이다. B 집단에 속한 양성인 구성원의 경우, 실제 양성임에도 불구하고 음성으로 판단될 비율이 24.9%로서 이는 A 집단의 경우(0.2%)와 비교하여 100배나 높다.

다만, 본 시나리오와 같이 모든 인공지능 공정성 지표에 있어 큰 격차가 있는 경우, 인공지능 공정성 기준 간의 충돌 문제가 중요하게 고려될 필요는 없다. 해당 인공지능에 있어 편향이 발생하는 원인을 찾고 이를 감소시키기 위한 노력을 기울여야 한다는 결론에 있어 달라지지 않기 때문이다. 이러한 사례는 얼굴인식 인공지능 공정성 감사 연구 사례인 Buolamwini & Gebru(2018)에서 잘 드러난다.[434] 해당 연구에서는 얼굴인식 인공지능이 성별과 인종에 대해 편향되어 있음을 밝혔는데, 각 인공지능은 여러 정확도 지표에 걸쳐 보호집단에 대해 상당한 낮은 결과를 보였다. 따라서 해당 사례에서 공정성 기준 간의 충돌 문제는 중요하게 고려될 필요가 없었다.

---

[434] Buolamwini and Gebru (2018) (각주 158).

더욱이 본 시나리오는 분류모형이 평가 대상 개인의 민감속성(집단 소속)을 부정적으로 예측에 활용하고 있다는 점에서도 문제가 있다.[435] 시나리오 1에서 로지스틱 회귀 모형은 집단에 소속되어 있는지 여부에 상당히 큰 가중치(-2.60)를 적용하고 있다. 즉, 머신러닝 모형은 B 집단 구성원 중에서 양성인 개인의 수는 적다는 사실을 포착하고, 그 결과 B 집단에 소속된 경우 합격 확률을 낮추는 방향으로 작동한다. 그 결과 B 집단에 대해서는 학습 데이터상 실제 양성인 비율(20%)보다 더 낮은 18.7%만이 합격으로 예측되는 결과를 초래하게 되었다. 따라서 해당 모형은 직접차별금지 원칙에 반하는 것으로도 평가될 수 있다.

나. 시나리오 1-2 검토

이처럼 시나리오 1-1은 직접차별금지 원칙에 반하는 점을 고려하여, 시나리오 1-2에서는 unawareness 기준을 적용하여 분류모형이 민감속성을 알지 못하도록 하였다. 이 경우 각 집단에 대한 예측 정확도가 감소하였다(A 집단은 99.2%에서 98.2%로 낮아지고, B집단은 91.3%에서 89.6%로 낮아졌다). 즉 머신러닝 시스템이 예측을 위해 사용하는 정보량을 감소시킨 결과 정확도가 낮아지는 효과가 확인된다.

하지만 반대로 인구통계적 불균형이 완화되는 효과가 존재한다. 즉, A 집단의 합격률은 80.8%에서 77.9%로 감소하였고, B 집단의 합격률은 18.7%에서 23.7%로 높아지게 되었다. 즉, 블라인드 기준을 적용함으로써 인구통계적 불균형이 부분적으로 해소되는 차별 시정적 효과가 발생하였으나, 그 결과 A 집단의 구성원들이 오히려 부정확한 결정을 받게 되는 문제가 발생하게 되었다.

---

435 본 시뮬레이션에서는 단순화를 위하여 민감 속성을 직접 입력 변수를 사용하는 사례를 상정하였지만, Gillis & Spiess (2020) (각주 205)이 보여준 것처럼 여러 변수를 통해 민감 속성을 추론할 수 있다. 이와 같은 경우 알고리즘을 추론된 속성을 부정적으로 사용할 수 있다.

한편, unawareness 기준을 적용한 결과 오류율 동등성과 예측도 동등성에 미친 영향은 이중적이다. 위음성 오류율 간의 격차는 감소하였지만, 위양성 오류율 간의 격차는 오히려 증가하였다. 마찬가지로 양성 예측도 간의 격차는 감소하였지만, 음성 예측도 간의 격차는 증가하였다. 이처럼 기존의 사회적 편향과 데이터 측정 과정의 편향이 그대로 존재하는 상황에서 unawareness 기준만을 적용하는 것은 사태 개선에 있어 양면적으로 작용하게 된다는 점을 보여준다.

따라서 차별 시정을 위한 보다 적절한 조치는 unawareness 기준을 적용하기 보다는 (1) 사회적 편향 자체를 제거하거나 (2) 데이터 측정 과정의 편향을 제거하는 것이라 할 수 있다.

다. 시나리오 1-3 검토

시나리오 1-3에서는 시나리오 1-2에서와 달리 데이터 측정 과정의 편향이 없는 상황을 전제하였다. 즉, 시나리오 1-3에서는 특성변수(X_0, X_1)는 목표변수와 상관관계가 있지만, 집단과 무관하게 일정한 수준의 노이즈(표준편차 0.5)를 반영하였다. 나아가 분류모형은 보호속성에 대한 정보에 접근할 수 없는 것으로 하였다.

이 경우 각 집단간 오류율 동등성이 달성되는 것을 확인할 수 있다(위음성 오류율의 경우 A 집단 7.6%, B 집단 8.0%로 거의 유사하고, 위양성 오류율에 있어 A 집단 8.8%, B 집단 7.9%로 오히려 A 집단에 대한 오류율이 약간 더 높았을 뿐이다). 학습된 분류모형은 집단과 직접적으로는 관련되어 있지 않은 특성변수만을 활용하여 예측을 수행하였으므로, 양성 오류율의 동등성과 음성 오류율의 동등성은 달성되었다고 평가할 수 있다.

하지만 시나리오 1-3에서는 주목할 점은 A 집단과 B 집단 간의 예측도 동등성에 있어 큰 차이가 발생하게 되었다는 사실이다. A 집단의 경우 양성 예측도가 상당히 낮아졌고, B 집단의 경우 음성 예측도가 크게 낮아졌다. A 집단과 B

집단은 오류율의 동등성이 달성되었으므로, 진양성율(true positive rate, TPR)와 진음성율(true negative rate, TNR)이 동등하다. 그런데도 불구하고 양성 예측도와 음성 예측도에 있어 큰 격차가 발생하게 된 것이다. 그 이유는 다음과 같이 설명될 수 있다. 양성예측도와 음성예측도는 진양성율, 진음성율 및 prevalence를 활용한 다음 수식으로 표현될 수 있다. A 집단과 B 집단의 진양성율과 진음성율이 동등하다고 하더라도, prevalence의 차이에 의해 양성예측도와 음성예측도는 달라질 수 있는 것이다.

$$\text{양성예측도 PPV} = \frac{(\text{TPR x prevalence})}{[\,(\text{TPR x prevalence}) + ((1 - \text{TNR})\text{x}\,(1 - \text{prevalence}))]}$$

$$\text{음성예측도 NPV} = \frac{(\text{TNR x}\,(1 - \text{prevalence}))}{[\,(\text{TNR x}\,(1 - \text{prevalence})) + ((1 - \text{TPR})\text{x prevalence})]}$$

이 사례에서 A 집단은 80%가 양성인 반면, B 집단은 20%만이 양성이므로 A 집단과 B 집단 간의 prevalence 차이가 크다. 그 결과, 학습된 분류모형이 A 집단 개인에 대해 양성으로 예측할 경우 B 집단 개인의 경우와 비교할 때 진양성일 가능성이 위양성일 가능성보다 훨씬 크게 된다. 반대로 B 집단의 경우 모형이 양성이라고 예측하더라도 위양성일 가능성이 크다. 이와 같은 문제는 음성으로 예측한 경우도 마찬가지이다. 그 결과 두 집단 사이에 양성예측도와 음성예측도의 동등성은 달성하지 못하게 된 것이다.

라. 시나리오 1-4 검토

시나리오 1-4에서는 시나리오 1-3과 달리 unawareness 기준을 적용하지 않고 머신러닝 모형이 보호속성 정보를 활용할 수 있도록 하였다. 그 결과 시나리오 1-3과 비교하여 예측도 동등성은 상당히 개선되지만, 반대로 오류율 동등성은 크게 악화되었다. 이 시나리오는 인공지능 공정성 기준 간의 충돌 문제를

제기한다. 이처럼 예측도 동등성을 개선하기 위해 오류율 동등성을 악화시키는 것은 정당한가?

해당 머신러닝 모형이 사례 1과 같은 직원 인식 시스템이라고 고려해 보자. 이 경우 오류율 동등성 기준을 충족하는 것이 중요하다. 만약 직원 인식 시스템이 집단에 대해 차별적으로 작동한다면, 오류 사례로 인한 피해가 집단간 불균등하게 발생하게 된다. B 집단 소속 직원은 A 집단 직원에 비해 더 자주 회사에 출입하기 위해 여러 차례 인식기를 시도해야 하는 불편함을 겪는다면, 그 직원은 사회로부터 배제된다는 느낌을 받을 수 있다. 이러한 인공지능이 사회적으로 확산된다면, 그 개인은 일상 전반에 걸쳐 사회에서 배제되는 경험을 하게 될 수 있다.[436] 오류율 동등성은 이러한 불편함을 겪게 되는 비율이 집단간 동등할 것을 요구한다. 이처럼 목표변수에 대한 정답(ground truth) 값이 명확하게 정해져 있고, 그 오류로 인해 개인이 피해를 입는 경우, 우리는 집단간 오류율 동등성을 충족해야 할 필요가 있다.

예측도 동등성은 그와는 다른 목적을 갖는다. 예측도 동등성이 요구하는 바는, 가령 머신러닝 시스템이 어떤 영상을 보고 그가 직원일 확률이 90%라고 예측했을 때, 그러한 예측이 갖는 정보 가치가 모든 집단에 대해 동등해야 한다는 것이다. 하지만 현재 고려하고 있는 직원 인식 인공지능 사례에서는 예측도 동등성이 갖는 규범적 함의를 찾기 어렵다. 다음과 같은 가상의 상황을 고려하자. 한 회사에 총 10,000명이 근무하는데, 그 중 B 집단 소속 직원은 오직 1명만 근무하고 있고, 경비원은 그 사람을 잘 알고 있다. 이 회사는 A 집단과 B 집단 모두에 대해 진양성율(TPR)과 진음성율(TNR)이 99%인 얼굴인식 시스템을 설치하였다. 그런데 처음 보는 B 집단 사람이 들어와 얼굴 인식 시스템을 거쳤는데, 직원이라고 판정하였다. 경비원은 그 사람을 불러 세워 진정으로 직원인지 여부를 다시 확인할 필요가 있는가? 이러한 가상 사례에서 인공지능이 직원이라고 판단하였다고 하더라도 B 집단의 낯선 사람이 실제로 직원일 확률(양성 예측도)은 극히

---

436 이는 본 연구 제2장 제3절 2항에서 설명한 표상적 위해 중 '인식'의 문제에 해당한다.

낮다. 위 회사에서 B 집단 소속인 사람이 직원일 확률은 1/10,000에 불과하기 때문이다. 따라서 경비원은 얼굴 인식기가 오류를 일으켰다고 판단하는 것은 정당하고, 그가 진정으로 직원인지 여부를 다시 확인할 필요가 있다. 이처럼 진양성율과 진음성율이 높다고 하더라도, 집단 별로 인공지능의 예측이 갖는 정보 가치는 차이가 있을 수 있다. 앞서 본 바와 같이 양성·음성 예측도는 prevalence에 따라 달라지기 때문이다.

요컨대, 본 사례에서처럼 결정론적 목표변수를 예측하는 경우에는, 일응 시나리오 1-3에서와 같이 오류율 동등성 기준을 우선적으로 적용하는 것이 타당하다는 시사점을 얻을 수 있다.

3. 장기적 개선 방향 – 인공지능의 성능 개선

이상에서 본 바와 같이 주어진 학습 데이터에 대해 인공지능을 구현하는 경우에는 여러 인공지능 공정성 기준 간의 충돌 문제가 발생할 수 있다. 이 경우 규범적 판단을 통해 어떠한 공정성 기준을 적용할 것인지 선택해야 하는 문제가 발생한다. 하지만 현실에서는 목표변수를 더 정확하게 예측할 수 있도록 더 나은 품질의 학습 데이터나 더 많은 학습 데이터를 확보함으로써 인공지능의 성능을 개선하는 방안도 고려할 수 있다.

앞선 시뮬레이션 시나리오에서는 모두 목표변수와 상관관계가 있는 2개의 특성변수(X_0, X_1)만을 사용하였다. 하지만 더 많은 특성 변수를 활용한다면, 인공지능의 성능이 개선될 뿐만 아니라, 그에 따라 공정성 지표 역시 개선되는 모습을 발견할 수 있다.

아래 그래프는 인공지능이 활용하는 특성변수의 개수를 2개에서 20개까지 점차 증가시켰을 경우 각 집단별로 위음성율, 위양성율, 양성예측도, 음성예측도가 변화하는 정도를 보여준다. 아래 그래프는 unawareness 기준을 적용하지 않은 경우이다.

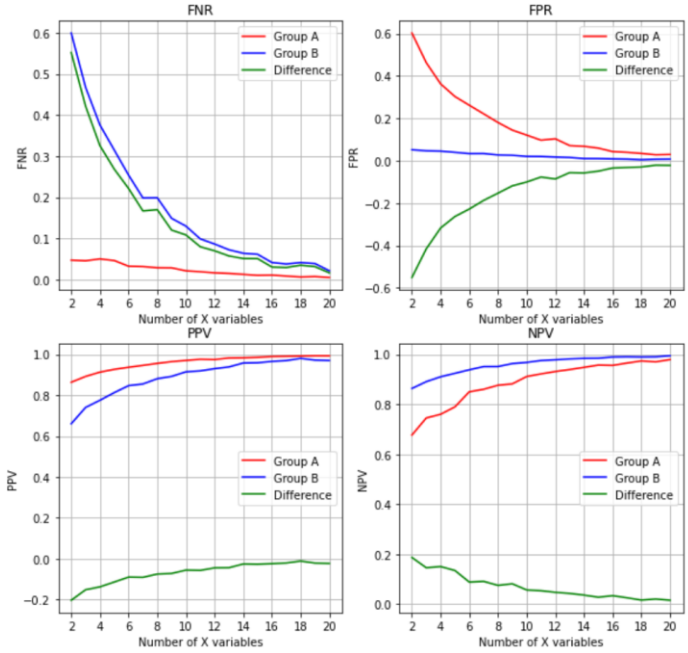


그림 22 인공지능 성능 개선을 통한 공정성 개선 가능성 (unawareness 기준 미적용시)

나아가 위와 같은 인공지능 성능 개선에 따른 공정성 개선 효과는 unawareness 기준을 적용하는지 여부와는 무관하게 발견된다. 아래 그래프는 unawareness 기준을 적용한 경우(즉, 머신러닝 모형이 집단 소속에 대한 정보를 알 수 없는 경우)이다.

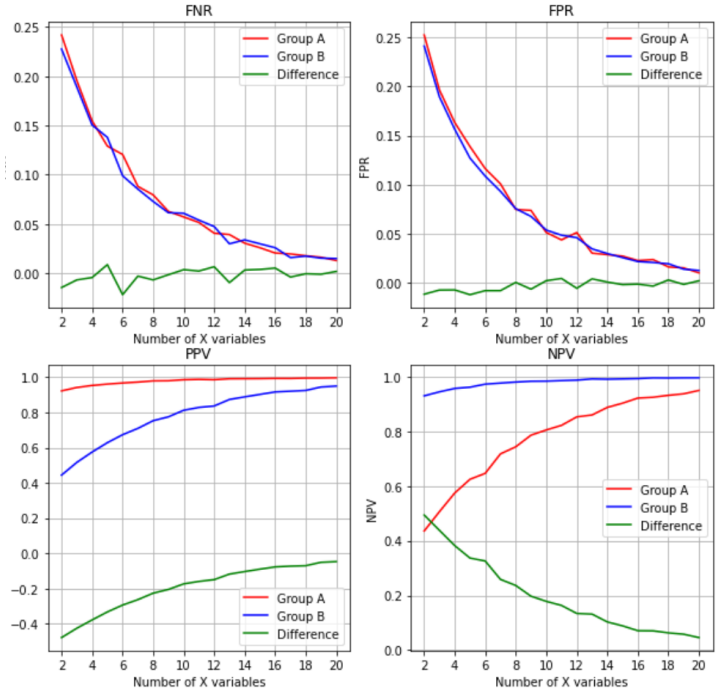


그림 23 인공지능 성능 개선을 통한 공정성 개선 가능성 (unawareness 기준 적용시)

어떠한 경우이든지, 활용가능한 특성 변수가 증가함에 따라 인공지능의 성능이 비약적으로 증가하는 모습을 보여준다. 10개의 특성 변수를 사용할 경우 위양성율과 위음성율은 두 집단 모두에 대해 5% 미만으로 줄어들고, 양성예측도와 음성예측도는 80% 이상으로 증가한다. 20개의 특성변수를 사용할 경우 위양성율과 위음성율은 두 집단 모두에 대해 2% 수준으로 낮아지고, 양성예측도와 음성예측도는 90% 이상이 된다. 그 결과 오류율 격차나 예측도의 격차도 함께 감소하는 모습을 보여준다.

위 그래프는 인공지능 공정성 기준을 충족시키는 성능이 낮은 분류모형보다 인공지능 공정성 기준을 위반하는 분류모형이 오히려 더 공정할 수 있다는 점을 보여준다. 즉, 비록 특정한 인공지능의 오류율이나 예측도가 집단간 '동등'하도록 하는 개입이 존재하지 않더라도, 해당 인공지능의 절대적인 '성능'이 개선되면서

종국적으로 집단간 정확도 격차가 줄어들고, 그에 따라 그 인공지능은 더욱 '공정한' 판단을 할 가능성이 있는 것이다.

제3절 판별모형 시나리오

1. 시뮬레이션 방법[437]

다음으로, 판별모형 시나리오에서는 목표변수가 확률적으로 결정되는 경우를 고려한다. 이는 인공지능 시스템을 통해 개인의 능력이나 위험을 예측하는 경우가 대표적이다. 예를 들어 앞서 본 사례 2에서와 같이 보험회사가 예측 의료(predictive medicine) 시스템 구현을 위해 개인의 건강 위험도를 예측하는 경우를 상정해 볼 수 있다. 건강 위험도가 큰 개인들은 질환을 앓게 될 확률이 더 크거나, 기대수명이 더 짧거나, 더 많은 의료비를 지출할 가능성이 높다. 하지만 이러한 결과는 확률적으로 실현된다. 다른 예로 어떤 회사가 직원의 직무능력을 예측하는 인공지능 시스템을 구현하는 경우도 마찬가지이다. 이 경우 분류모형의 출력 값은 통계적으로 직무능력과 높은 상관관계가 있을 수 있지만, 그 직원이 실제로 어떤 업무를 잘 수행할 것인지에는 다양한 외생적 요인과 우연적 요소들이 개입될 것이다. 따라서 직무능력이 높게 예측된 직원이라고 하여 항상 해당 직무를 잘 수행하는 것은 아니고, 직무 성과는 확률적으로 실현된다.

이하에서는 논의의 편의상 인공지능을 통해 개인의 '능력(merit)'을 평가하는 경우로 한정한다. 여기서의 '능력'은 개인의 학업수행 능력, 업무수행 능력, 계약이행 능력 등 다양한 측면에서 고려될 수 있다. 이러한 능력은 확률적으로 성과로 이어지므로, 그 능력을 확률 p_i로 표시한다.

---

437 각 시뮬레이션의 상세에 관해서는 별지 2 참조.

본 시뮬레이션에서는 다음과 같은 편향 존재의 가능성에 따라 다음 2가지 시나리오로 구분한다.

(1) **시나리오 (2-1) (집단간 원래 능력의 분포가 다른 경우)** A 집단의 평균적 잠재적 능력이 B 집단에 비해 현저히 낮은 경우를 상정한다. A 집단은 $p = 0.3$이 평균인 분포, B 집단은 $p = 0.7$이 평균인 분포를 가정하였다. 각 분포의 표준편차는 동일하게 가정하였다. 2개의 특성변수(X_0, X_1)는 각 개인에 부여된 p_i 값에 비례해서 생성하되, 일정한 노이즈(표준편차 0.3)를 부가하였다. 나아가 p_i 값에 따라 확률적으로 데이터 상의 성과(outcome)이 결정되는데, 이 경우 성과변수 Y는 이진변수(시험 합격/불합격, 계약 이행/불이행 등)로 가정하였다. 예컨대 p_i=0.4 인 개인은 40%의 확률로 학습 데이터상 Y_i가 1로 기재되고, 60%의 확률로 Y_i가 0으로 기재된다.

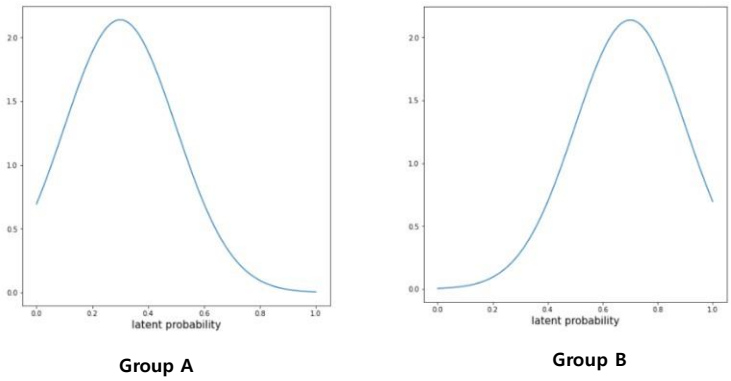


그림 24 집단간 원래 능력의 분포가 다른 사례의 가정

(2) **시나리오 (2-2) (집단간 원래 능력 분포는 동일하지만, 관찰된 특성에 있어 편향이 있는 경우)** A 집단과 B 집단의 실제 잠재적인 능력의 확률분포는 동일하지만, 잠재적인 능력을 측정하는 과정에서 편향이 반영된 경우를 상정한다. 2 개의 특성변수 (X_0, X_1)는 각 개인에 부여된 p_i 값과 상

관관계가 있으나, A 집단에 대해서는 양의 가중치를, B 집단에 대해서는 음의 가중치를 부여하여 편향(bias)을 부가하였다. 그 결과 두 집단에 대해 p의 분포는 동일하지만, (X_0, X_1)는 편향적으로 생성된다. 성과변수는 위 시나리오와 동일하게 진정한 p_i 값의 확률에 따라 생성하였다(즉, 성과변수에 있어서는 편향이 존재하지 않는 상황을 가정한 것이다).

한편, 판별모형 시나리오에서는 실무상 직접차별 금지원칙에 따라 모형에 차별금지사유(즉, A 집단인지 B 집단인지)가 입력변수로 주어지지 않는다는 점을 고려하여 모두 unawareness 기준을 충족시키는 것으로 가정하였다.

2. 시나리오 2-1 - 집단간 잠재적 능력의 확률분포가 다른 경우

가. 시뮬레이션 결과

시나리오 2-1에서 로지스틱 회귀 모형을 통해 목표변수를 예측한 결과는 다음과 같다. 아래 그림은 확률 등고선을 나타낸다. 만약 동일한 문턱 값을 적용할 경우 A 집단에 대해서는 불합격으로 예측하는 사례가 많고, B 집단에 대해서는 합격으로 예측하는 사례가 많다.

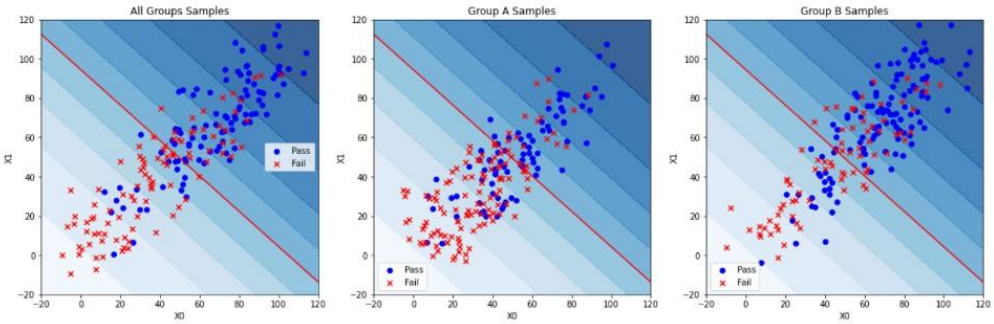


그림 25 집단간 능력의 확률분포에 차이가 있는 경우의 예측 결과

나아가 진정한 능력을 가리키는 p값과 머신러닝 모형에 의해 예측된 능력을 가리키는 $\hat{p}$을 비교하면 다음과 같다. 머신러닝 모형은 A 집단에 대해서는 진정한 값보다 다소 낮은 $\hat{p}$ 값을, B 집단에 대해서는 진정한 값보다 다소 높은 $\hat{p}$ 값을 추정하는 것을 확인할 수 있다.

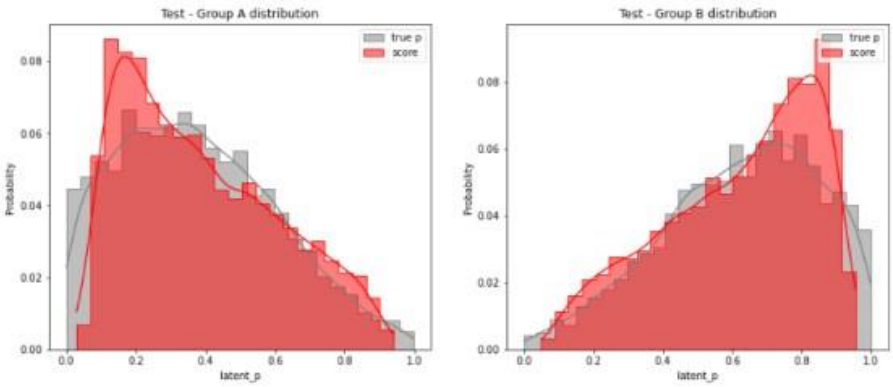


그림 26 머신러닝 모형에 의해 예측된 능력 값의 분포

본 시뮬레이션 사례에서 각 집단간 단순 정확도와 합격률을 비교하면 다음과 같다. 시뮬레이션 결과 머신러닝 모형은 A 집단과 B 집단에 대해 비슷한 정확도를 보인다. 나아가 합격률은 A 집단 30.7%, B 집단 70.8%으로서, 원래 가정한 각 집단의 잠재 능력 분포와 거의 일치한다(앞서 설명한 바와 같이 p 값 평균을 A 집단은 0.3, B 집단은 0.7로 가정하였다).

	A Group	B Group
단순 정확도	71.9%	70.3%
합격률	30.7%	70.8%

나. 인공지능 공정성 기준의 적용

이 시나리오에서 집단 간 위음성율(FNR, 평가데이터상 실제 적격임에도 모형이 부적격으로 예측하는 비율)을 비교하면 다음과 같다. A 집단은 위음성율이 무려 46.5%나 되는데 비해, B 집단은 17.2%에 불과하다. 따라서 이 시나리오의 인공지능은 오류율 동등성 기준을 명백하게 위반하고 있다.

	A Group	B Group
False Negative (Predicted Unqualified but Qualified)	46.5%	17.2%

표 5 집단간 위음성율 차이 (시나리오 2-1)

이처럼 시나리오 2-1 인공지능은 오류율 동등성을 심각하게 위반하고 있으나, A 집단 예측과 B 집단 예측은 예측도 동등성 기준을 대체로 충족한다. 이러한 결과는 아래와 같은 캘리브레이션 그래프를 통해 확인할 수 있다. 머신러닝 모형에 의해 능력이 0.6라고 예측된 사람들의 60%가 실제로 성과가 양성으로 나오고 있다는 점을 확인할 수 있고, 이러한 특징에 있어서는 A 집단, B집단 모두 별다른 차이가 없다.

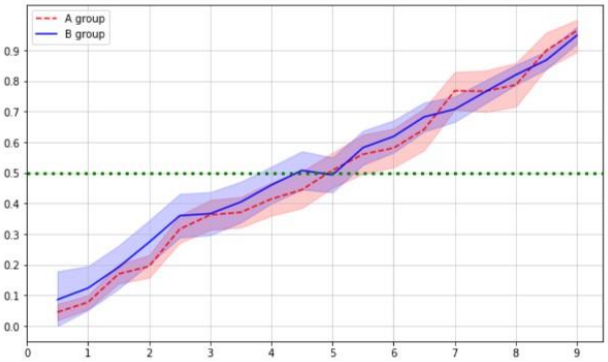


그림 27 집단 간 캘리브레이션이 충족되는 결과 (시나리오 2-1)

다만, 두 집단은 신뢰구간의 분포에 있어 차이를 보인다. A 집단(붉은색)에 대해서는 낮은 능력 값으로 예측된 이들에 대해 신뢰구간의 폭이 좁고, 높은 능력 값으로 예측된 이들에 대해서는 신뢰구간의 폭이 넓다. 반대로 B 집단(파란색)에 대해서는 높은 능력 값으로 예측된 이들에 대해 신뢰구간의 폭이 좁고, 낮은 능력 값으로 예측된 이들에 대해서는 신뢰구간의 폭이 넓다. 이처럼 신뢰구간의 폭이 넓은 것은 예측의 불확실성이 크다는 것을 뜻한다. 다시 말해, A 집단에 대해서는 낮은 능력으로 예측된 경우 더 확실한 예측에 해당하는 것에 비해, B 집단에 대해서는 높은 능력으로 예측된 경우가 오히려 더 확실한 예측에 해당한다. 이는 A 집단에서는 높은 능력으로 예측하는 것이 이례적 사건으로 불확실성이 큰 것에 비해, B 집단에서는 낮은 능력으로 예측하는 것이 이례적 사건에 해당한다는 의미이다. 결국 해당모형은 집단간 원래 능력 분포의 차이를 적절히 반영하고 있는 것으로 이해할 수 있다.

다. 검토

이와 같이 예측도 동등성 조건을 충족하는 인공지능이 오류율 동등성을 심각하게 위반하는 것은 앞서 COMPAS 사건에서 본 바와 같다. 시나리오 2-1은 COMPAS 사건의 알고리즘보다 오류율 격차가 더 심각할 뿐이다. 즉, 본 시뮬레이션은 COMPAS 사건의 알고리즘을 단순하게 재현한 것으로 이해할 수 있다.

이와 같은 결과가 발생하는 원인은 무엇인가? 이 문제에 대해 직관적으로 파악하는 것은 쉽지 않다. 본 시나리오의 머신러닝 모형은 어느 집단이든 $\hat{p}$를 비교적 정확하게 추정하고 있기 때문이다. 그 원인은 크게 두 가지에 기인하는 것으로 볼 수 있다. (1) 머신러닝 모형을 통해 예측하고자 하는 목표변수가 확률적으로 결정되고, (2) 집단간 목표변수의 확률분포에 있어 차이가 있기 때문이다.

만약 어떤 개인의 진정한 p 값이 0.6이라고 해 보자. 그 개인은 60%의 확률로 합격한다. 따라서 40%의 확률로 불합격할 것이다. 인공지능이 $\hat{p}$을 정확하게

0.6으로 추정하고 그 개인에 대해 합격이라고 예측하더라도, 그 인공지능을 평가하면 오류에 해당하는 사례가 40%나 된다. 반대로 어떤 개인의 진정한 p 값이 0.4에 해당하는 경우도 이와 크게 다르지 않다. 그 개인은 60%의 확률로 불합격할 것이지만, 40%의 확률로는 합격할 것이다. 인공지능이 $\hat{p}$을 정확하게 0.4로 예측하고 그를 불합격이라고 예측하더라도, 그 인공지능을 평가하면 오류에 해당하는 사례가 마찬가지로 40%나 된다.

아래 그림은 각 집단에 대해 진양성 사례와 위음성 사례의 히스토그램을 그린 것이다. 이는 서로 다른 진정한 p 값을 가진 여러 개인들이 본 시뮬레이션 상의 인공지능에 의해 어떤 취급을 받게 되는지를 보여준다. 역사적으로 차별을 받아온 A 집단은 $\hat{p}$이 0.4에서 0.6 사이에 집중되어 있는 것에 비해, B 집단은 $\hat{p}$이 0.6보다 큰 데이터가 월등히 많다. 다시 말해 A 집단의 분포는 $\hat{p}$이 낮은 쪽으로 쏠려(skewed) 있으므로, 인공지능이 $\hat{p}$를 정확하게 예측하여도 실제 성과변수를 잘못 예측하는 사례가 더 많게 된다. 즉, A 집단과 B 집단의 p 값의 확률 분포가 다르기 때문에 위와 같은 오류율 격차가 발생한 것으로 평가할 수 있다.

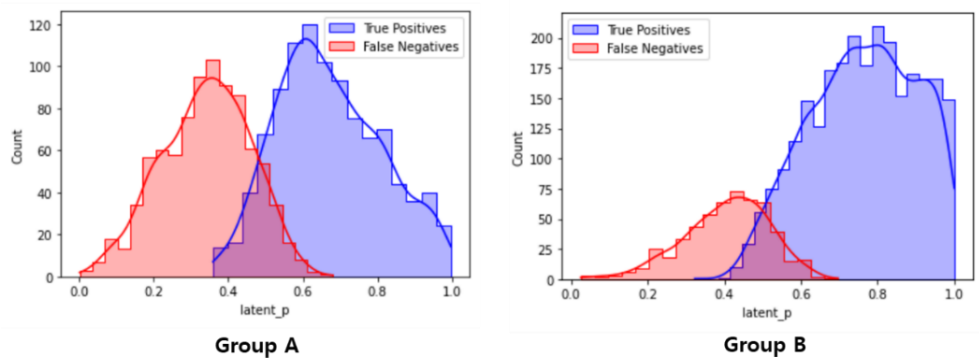


그림 28 오류율 불균등이 발생하는 원인 (시나리오 2-1)

달리 표현하자면, 본 시나리오에서 가정한 바와 같이, A 집단은 내재적으로 예측 불확실성이 더 높고, B 집단은 내재적으로 예측 불확실성이 더 낮은 것이 위와 같은 오류율 차이의 원인이 될 수 있다. 이러한 경우, 인공지능이 집단에

대해 서로 다른 오류율을 보이는 것은 인공지능이 불공정하기 때문이기보다는 A 집단과 B 집단의 내재적 확률 분포 상의 차이에 의한 것으로 보는 것이 타당할 것이다.

이러한 상황에서 만약 두 집단 간 오류율 동등성을 충족시키기 위해 인공지능의 결정 방식을 수정하면 어떻게 되는가? 오히려 그 인공지능이 공정하지 않은 결과가 초래될 수 있다. 예를 들어 A 집단과 B 집단에 대해 위음성의 오류율을 동등하게 맞추기 위해 서로 다른 문턱 값을 적용하는 상황을 고려하자. 위에서 본 것처럼 집단 B의 위 음성율이 17.5%였으므로 집단 A에 대해서도 같은 비율로 맞추고자 한다. 그러려면 B 집단에 대해서는 종전과 같이 $\hat{p}$이 0.5보다 큰 경우 합격을 시킬 경우, A 집단에 대해서는 $\hat{p}$이 0.289보다 큰 경우 모두 합격을 시켜야 한다.[438] 하지만, 이처럼 A 집단에 대한 합격 한도를 낮추는 것은 바람직하지 않다. 이러한 개입의 가장 직접적 효과는 A 집단에서 적격이 없는 이들 중 합격하는 비율이 급격히 높아지는 것이기 때문이다. 이러한 경우 시뮬레이션 결과에 따르면 A 집단은 False Discovery Rate (인공지능에 의해 합격으로 예측되었으나, 실제로는 적격이 아닌 경우)이 47.1%에 이르게 된다

	Group A	Group B
False Discovery Rate (Predicted qualified but unqualified)	47.1%	26.7%

표 6 차별 시정 조치 후 False Discovery Rate (시나리오 2-1)

이처럼 위음성율을 집단간 동등하게 하는 것은 실질적으로 적극적 차별시정 조치(affirmative action)을 적용하는 것과 같다. 하지만, 인공지능에 의한 A 집단에 대한 우대 정책은 인공지능에 의한 A 집단에 대한 예측의 대한 신뢰를 저

438 이는 위음성 오류율이 두 집단에 대해 동등하게 될 때까지 A 집단에 대한 문턱 값을 점차적으로 낮추어서 찾은 것이다.

하시키는 부작용을 가져올 수 있다. 인공지능의 예측 값을 참고하여 최종적인 의사결정을 내리는 이용자는 A 집단에 대해서는 인공지능의 결과를 신뢰하기 어렵다고 여길 것이다. 나아가 장기적으로 A 집단 출신 합격자에 대한 부정적인 평가가 누적됨으로써, A 집단에 오히려 부정적 영향을 끼칠 위험도 있다.[439]

다른 한편, 이 시뮬레이션 시나리오는 오류율 동등성이 갖는 또다른 위험을 보여준다. 이는 오류율 동등성을 통해 인공지능의 공정성을 달성하고자 하는 시도가 문제의 진정한 원인을 숨기고 착시를 불러 일으킬 수 있다는 것이다. 위 시나리오는 위음성율의 불균등이 나타난 근본적 원인은 인공지능의 편향이 아니라, A 집단과 B 집단의 내재된 능력의 분포나 불확실성의 정도에 차이가 있기 때문이라는 점을 보여준다. 그런데 이러한 문제의 원인이 마치 "인공지능 자체가 편향되었다"고 프레이밍(framing)하는 것은 원인을 제대로 보지 못하고, 근본적 사회 구조에 대한 관심을 왜곡할 수 있다.

이러한 문제에 대해서는 근본적 사회 구조에 따라 존재하는 차별을 개선하기 위해서는 오히려 적극적으로 오류율 동등성 기준을 적용하여, 인공지능의 "공정성"을 보장할 수 있어야 한다는 반론이 제기될 수 있다.[440] 이는 사회적으로 여러 분야에 걸쳐 인공지능에 따른 자동화된 의사결정이 이루어지고 있는 상황에서 차별 상황이 악화되는 것을 막기 위해서는 오류율 동등성 기준을 적용해야 한다는 것이다.

그러나, Hedden (2021)이 지적한 바와 같이 사회에 만연한 기존의 차별을 어떻게 해결할 것인지 별도로 고려될 필요가 있는 문제이다.[441] 무엇보다도 결과 동등성이나 오류율 동등성 기준을 적용함으로써 추가적으로 비용이 발생하거나 효율성 손실이 발생할 경우 그에 대한 부담을 인공지능의 이용자나 인공지능에

---

439 이와 같은 오류율 동등성 기준의 한계에 대해서는 본 연구 제4장 제3절 2항 참조.

440 Hardt, Price & Srebro (2016) (각주 21) 및 Zafar, et al. (2019) (각주 235)은 이러한 입장에 서 있는 것으로 이해할 수 있다.

441 Hedden (2021) (각주 32).

의한 평가 대상이 된 사람이 부담하도록 할 근거를 찾기 어렵다는 문제가 있다. 예를 들어, 금융기관이 신용평가를 위해 사용하는 인공지능이 채무불이행 확률을 예측할 경우, 대출 비율이 집단간 동등할 것을 요구하거나, 대출 거절 오류율의 동등성을 달성하기 위해 각 집단마다 서로 다른 기준을 적용하는 상황을 고려하자. 이 경우 역사적으로 차별의 대상이 되어온 집단 구성원에 대해서는 인공지능이 예측한 채무불이행 확률($\hat{p}$)이 다른 집단 구성원의 경우에 비해 더 높은데도 불구하고 대출을 실시하는 결과가 초래될 수 있다. 그러한 우대조치에 따른 대출로 인해 추가적으로 발생하는 채무불이행 손실 위험은 결국 해당 금융기관의 다른 이용자 또는 그 주주들로 전가된다. 이러한 경우 금융기관의 다른 이용자나 그 주주들에게 사회에 존재하는 기존의 경제적 불평등을 시정하기 위한 추가적 부담을 요구할 근거는 쉽사리 전제하기 어렵다. 만약 이제까지 차별의 대상이 된 집단에 대한 재분배를 통해 차별 시정을 꾀하고자 한다면, 정치적 과정을 통해, 세금 등으로 마련된 재원을 이용하는 것이 합당한 비용 분담이 될 수 있다. 다만, 해당 인공지능 이용자가 기업의 사회적 책임을 다하거나, 고객의 다양성 보장을 통한 다른 효율성 등 증가 효과 등을 고려하여 자발적으로 이를 도입하는 것까지 금지되어서는 안 될 것이다.

3. 시나리오 2-2 - 동일 잠재적 확률분포, 특성의 편향

가. 시뮬레이션 결과

시나리오 2-2는 집단간 잠재적 확률분포는 동일하지만, 관측되는 특성변수(X_0, X_1)가 A 집단에 불리하게 편향(biased)된 경우이다. 아래 그래프는 시뮬레이션 데이터의 진정한 능력 값인 p 값과 머신러닝 모형에 따라 예측된 능력 값인 $\hat{p}$ 값 사이의 차이를 보여준다. 왼쪽 그래프에서 보는 바와 같이 학습 데이터 생성에 사용한 진정한 p값의 분포는 집단 간 동일하지만, 우측 그래프에서 보는 바

와 같이 머신러닝 모형은 A 집단에 대해서는 낮은 $\hat{p}$ 값을, B 집단에 대해서는 높은 $\hat{p}$ 값을 추정하는 것을 확인할 수 있다.

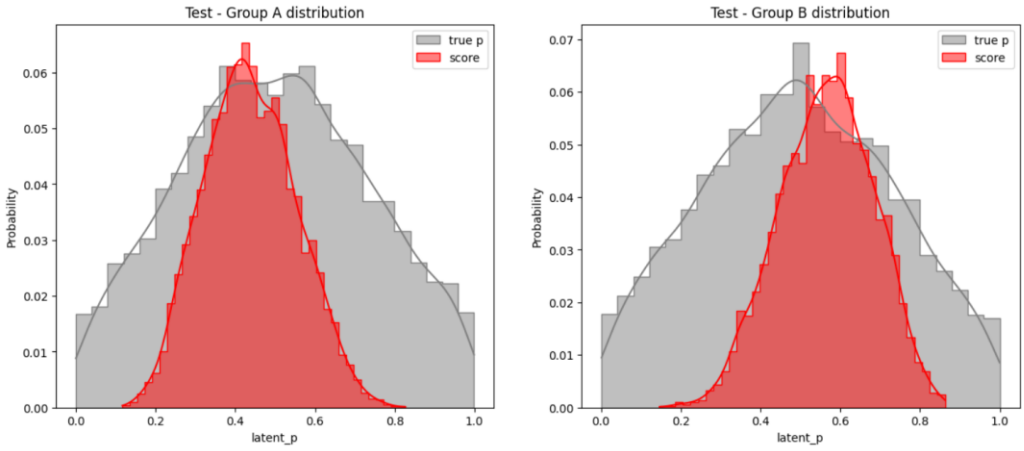


그림 29 관측된 특성변수가 편향된 경우 머신러닝 모형의 결과 (시나리오 2-2)

나. 인공지능 공정성 심사 기준의 적용

이 시나리오에서 집단 간 위음성율을 비교하면 다음과 같다. A 집단은 위음성율이 무려 58.3%나 되는데 비해, B 집단은 17.8%에 불과하다. 따라서 이 시나리오의 인공지능은 오류율 동등성 기준을 명백하게 위반하고 있다.

	A Group	B Group
False Negative (Predicted Unqualified but Qualified)	58.3%	17.8%

표 7 집단간 위음성 오류율 차이 (시나리오 2-2)

그런데, 시나리오 2-1 모형이 캘리브레이션 기준은 충족시킨 것과 달리, 시나리오 2-2 모형은 아래 그림이 보여주는 바와 같이 캘리브레이션 기준조차도

충족하지 못한다. 구체적으로 보면, A 집단 개인 중 0.5로 예측된 경우 평가 데이터상 실제로는 62.5%가 적격을 갖춘 것으로 나타났고, 반대로 B 집단 개인 중 0.5로 예측된 경우 평가 데이터상 47.2%만이 적격인 것으로 나타났다.

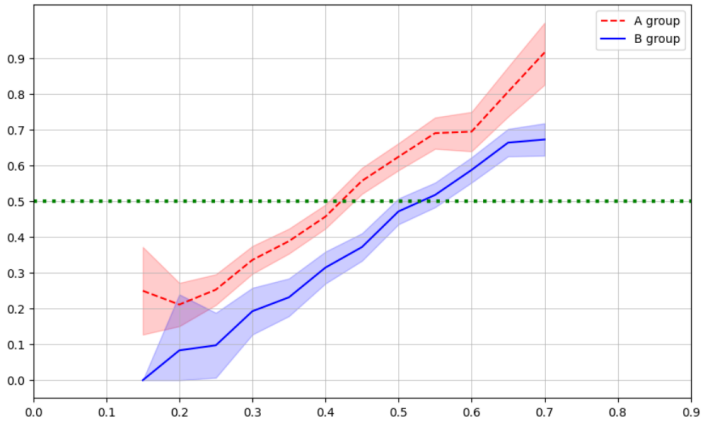


그림 30 캘리브레이션 기준을 위반하는 결과 (시나리오 2-2)

이처럼 시나리오 2-2에서 캘리브레이션 기준을 위반하게 된 까닭은 다음과 같이 이해할 수 있다. 시나리오 2-2에서는 A 집단의 경우 특성변수 (X_0, X_1)가 실제 능력을 제대로 반영하지 못하고 더 낮은 값으로 측정된다고 가정하였고, B 집단의 경우 반대로 특성변수가 실제 능력보다 더 높은 값으로 측정된다고 가정하였기 때문이다. 이처럼 편향된 특성변수만을 활용할 경우 위와 같이 캘리브레이션 기준을 위반할 수 있다.

본 시나리오는 캘리브레이션 기준을 적용함으로써 머신러닝 모형 상의 특성변수에 반영된 편향을 발견할 가능성을 보여준다. 마찬가지로 만약 머신러닝 모형의 목표변수가 편향된 경우에도 캘리브레이션 위반이 발견할 수 있다. 대표적으로 앞서 소개한 바와 같이 Obermeyer et al. (2019)는 미국 의료위험점수 산정 알고리즘이 캘리브레이션 기준을 충족하지 못하고 흑인에 대해 차별적으로

작동한다는 점을 발견하였다.[442] 연구자들은 이러한 발견을 보험회사에 보고하였고, 머신러닝 모형을 수정하는 방식으로 개선이 이루어졌다.

4. 확률적 목표변수 예측 과제의 인식론적 한계

앞선 시나리오 2-2에서는 동일한 잠재적 확률분포를 가진 집단에 대해 학습 데이터에 포함된 특성변수가 편향된 경우를 고려하였다. 그렇다면 만약 학습 데이터에 기재된 성과변수, 즉 예측하고자 하는 목표변수에 있어서도 편향이 존재하는 경우는 어떻게 되는가?

머신러닝 모형의 학습 데이터에 편향이 반영되는 경우는 (1) 목표 변수인 결과(대학 입학 여부, 대출 여부, 채용 여부, 구속 여부 등)가 실현되는 과정에서 종전 인간 평가자의 편견이나 스테레오타입이 반영되는 경우와 같이 결과 측정이 편향된 경우(*biased outcome measurement*)와 (ii) 목표 변수를 예측하기 사용되는 여러 특성들(학점이나 시험 성적, 과거 연체 이력, 과거 업무 경력, 전과 경력 등)의 측정에 있어 편향이 반영되는 경우와 같이 특성 측정이 편향된 경우(*biased feature measurement*)로 구분할 수 있다. 시나리오 2-2에서 본 바와 같이 이러한 결함 중 하나만이 존재하는 경우라면, 머신러닝 모형을 캘리브레이션 기준을 충족시키도록 학습시키는 과정을 통해서 이를 발견하고 교정해 낼 수 있을 것이다.

그러나, 만약 데이터에 위 두가지 편향이 동시에 존재하는 경우에는 예측도 동등성 기준만으로는 이러한 결함을 발견하기 어렵다. 그 이유는 (i) 예측하고자 하는 구성물(construct)이 원래 집단간 불균형하게 분포되어 있는 사회에서 불균형하게 데이터가 생성되는 경우(즉, “unequal people and unequal data”)와 (ii) 해당 구성물이 집단간 동등하게 분포되어 있는데도 불구하고 학습 데이터 생성

---

442 Obermeyer, et al. (2019) (각주 405).

과정에서의 편향에 의해 불균형하게 데이터가 생성되는 경우(즉, “equal people but unequal data”)를 구분하기 어렵기 때문이다.

결국 여러 사회적 집단 구성원에 내재된 잠재적 확률이 동등하게 분포되어 있을 것이라는 가정은 이상적이고 규범적인 추정에 불과하다. 어떠한 개인이 가진 잠재적 확률은 현실에서 구체적인 평가 또는 심사 과정을 통해 실현된다. 그러므로 종국적으로 현실의 다른 인간 심사자에 의해 영향을 받는 것이 불가능하다. 따라서 만약 종래의 인간 심사자의 판단 결과가 편향되어 있는 상황이라면, 그 원인이 인간 심사자의 편향에 의한 것인지, 아니면 해당 개인의 잠재적 확률 분포 차이 때문인 것인지 구분하기 어렵다. 만약 인간 심사자의 편향에 의해 왜곡된 데이터가 생성된 경우라면, 그 데이터를 활용하여 학습된 인공지능이 예측도 동등성 기준을 충족시킨다고 하더라도 편향된 특성을 이용하여 편향된 결과를 정확하게 예측하는 것에 불과하다. 그 인공지능은 결국 데이터가 편향된 만큼 여전히 편향되어 있다. 이러한 문제는 궁극적으로 인공지능이 예측하고자 하는 목표변수의 진정한 값은 관측 불가능하다는 인식론적 한계에 기인한 것일 수 있다.

그러나 이러한 근본적인 인식론적 한계가 있다고 하더라도 이를 극복하기 위한 시도가 불가능한 것은 아니다. 만약 특성변수와 목표변수를 다양한 출처를 통해 취득할 수 있고, 각 출처의 특성 측정이나 성과 결정에 있어 동일하게 편향되어 있지 않은 경우를 고려해 볼 수 있다. 이 경우 여러 경로를 통해 수집된 정보를 종합한 편향의 정도를 줄일 수 있다고 가정할 수 있다.[443] 이처럼 추가적으로 수집된 정보가 증가하고, 편향의 정도가 줄어들수록 캘리브레이션 기준을 통해 인공지능의 공정성을 확인할 수 있는 여지가 더 커질 것이다. 따라서, 우리는 인공지능의 공정성을 달성하기 위해 단순히 캘리브레이션 충족 여부만을 확인할 것이 아니라, 특성변수와 목표변수가 동시에 편향되지 않았다는 점까지도 검증할

443 빅 데이터의 수집과 활용에 있어서는 대수의 법칙이 편향을 감소시키는 데 기여할 것이다.

필요가 있다. 이는 개인의 능력이나 위험에 대해 다양한 방법을 활용하여 측정하고, 여러 데이터를 이용한 예측을 비교하는 등의 노력을 포함할 것이다.

이처럼 우리가 예측하고자 하는 개인의 능력이나 위험을 더 정확하게 표현한 데이터를 확보하거나, 능력이나 위험을 정확하게 반영할 수 있는 특성에 관한 다수 데이터를 확보하여 캘리브레이션 기준을 충족시킬 수 있다면, 우리는 보다 공정한 예측을 향해 나아갈 수 있을 것이다.

제4절 분류모형 공정성 지표의 법적 함의

본 절에서는 이상에서 검토한 분류모형에 관한 공정성 지표가 기존 법률 또는 현재 고려 중인 입법안의 해석이나 집행에 있어 어떻게 활용될 수 있는지 검토한다. 법적 측면에서의 공정성 지표의 활용 사례는 인공지능의 (1) 개발 단계 (2) 도입 평가 단계 및 (3) 사후적 심사 단계로 구분할 수 있다. 구체적으로 (1) 개발 단계에서는 특히 고위험영역 인공지능의 개발자가 인공지능 규제 법안에 따른 신뢰성과 안전성을 확보하기 위한 조치 의무[444]를 준수하기 위하여 인공지능을 평가하는 경우를 고려할 수 있다. 또한 (2) 도입 평가 단계에서는 어떠한 사업자가 자동화된 의사결정 시스템을 도입 여부를 평가하거나,[445] 행정청이 자동적 행정결정을 도입할 것인지 여부[446]를 고려하면서 사전적 영향 평가를 수행하

---

444 이와 관련된 국내법으로는 「인공지능산업 육성 및 신뢰 기반 조성 등에 관한 법률안」 (대한민국 국회 과학기술방송정보통신위원회 2023. 2. 14. 합의안) 제28조 참조.

445 예를 들어 국내에서는 채용절차에서 사용되는 인공지능을 규제하기 위하여 「채용절차의 공정화에 관한 법률」에 관한 개정안도 여러 건 발의되어 있는 바, 그 중 소병철 의원안(2023. 3. 6. 발의, 의안 번호 20449호)은 "구인자는 인공지능 채용에 활용되는 기술의 공정성과 신뢰성을 제고하기 위하여 대통령령으로 정하는 요건을 갖춘 전문기관에 주기적으로 점검을 의뢰" 할 의무를 신설할 것을 제안한다(제8조의2). 이러한 전문기관이 해당 채용 인공지능 기술의 공정성과 신뢰성을 조사하기 위하여 공정성 지표를 사용할 수 있을 것이다.

446 행정기본법 제20조 참조.

는 경우 등을 상정할 수 있다. (3) 사후적 심사 단계에서는 차별금지법이 적용되는 차별금지영역에 있어 인공지능이 활용되었고, 그 결과 불리한 대우를 받은 개인이 인공지능의 활용이 차별금지법을 위반하였다고 주장하는 경우를 고려할 수 있다. 나아가 별도로 차별금지법이 제정되어 있지 않은 영역에서 활용되는 사례라고 하더라도 헌법상 평등권을 침해하는 경우 위법성이 인정될 수 있으므로,[447] 이러한 위법성 판단에 있어 인공지능 공정성 지표가 적용되는 상황을 상정할 수 있다.

1. 인공지능 개발 단계에서의 공정성 지표 활용

인공지능 시스템 개발 과정에서는 검증 및 시험 데이터 집합을 활용하여 인공지능의 정확도를 개선하는 반복적인 과정이 이루어진다. 만약 인공지능이 달성하고자 하는 성능 목표에 도달하지 못할 경우에는 추가적인 학습 데이터의 수집, 모형 구조 및 학습 알고리즘의 개선 등의 과정이 거듭하여 진행된다. 그런데, 인공지능 공정성 지표는 일반적으로 고려되는 정확성 지표에 추가하여 또 하나의 성능 지표로 고려될 수 있다. 예컨대 만약 개발 중인 인공지능의 집단 간 승인율, 오류율 또는 예측도에 있어 격차를 보인다면 낮은 정확도를 보이는 경우와 마찬가지로 취급하여 학습 데이터의 추가 수집 등과 같이 성능 개선을 위한 추가적인 조치를 취할 수 있을 것이다.

한편, 이처럼 개발 단계에서 공정성 지표를 활용하여 인공지능의 공정성을 보장하기 위한 작업은 인공지능 규제에 따른 의무사항을 준수하기 위해 법적으

---

447 대법원 2011. 1. 27. 선고 2009다19864 판결("사적 단체를 포함하여 사회공동체 내에서 개인이 성별에 따른 불합리한 차별을 받지 아니하고 자신의 희망과 소양에 따라 다양한 사회적·경제적 활동을 영위하는 것은 그 인격권 실현의 본질적 부분에 해당하므로 평등권이라는 기본권의 침해도 민법 제750조의 일반규정을 통하여 사법상 보호되는 인격적 법익침해의 형태로 구체화되어 논하여질 수 있고, 그 위법성 인정을 위하여 반드시 사인간의 평등권 보호에 관한 별개의 입법이 있어야만 하는 것은 아니다.").

로 요구될 수 있다. 특히 EU AI 법안은 고위험 인공지능 시스템 개발자에 대해 “리스크 관리 시스템”의 구축 의무를 부과하고 있는 바, 여기에는 합리적으로 예견 가능한 리스크의 식별, 추산 및 평가 의무가 포함된다.[448] 이때 개발자가 고려해야 하는 리스크에는 평등한 접근 및 기회를 포함한 기본권 침해에 대한 리스크가 포함되어 있다. 특히 인공지능 공정성 지표는 인공지능에 의한 평등권 침해 리스크에 대한 식별 수단으로 활용될 수 있을 것이다.

그러나 실무적으로 구체적으로 어떠한 공정성 지표를 어떠한 방식으로 선택하여 평가해야 할 것인지는 아직 명확하지 않다. 이러한 작업에 관한 세부 사항은 향후 마련되는 기술 표준에 의해 그 지침이 마련될 것으로 전망된다.[449] 이처럼 기본권 침해 리스크의 식별, 추산 및 평가 과정에서 공정성 지표를 선택, 적용하기 위해서는 인공지능에 의한 예측 오류가 어떠한 결과를 초래할 수 있는지에 관한 검토가 필요하다.[450] 예컨대 대출 적격 여부를 판단하는 인공지능에 있어 위음성 오류는 대출 적격자에게 대출이 거절되는 것으로서 해당 금융소비자에게 불이익을 가져온다. 반대로 위양성 오류는 대출 부적격자에게 대출이 이루어짐으로써 금융기관의 손실을 발생시킬 수 있다. 또한 캘리브레이션 기준을 충족시키지 못하는 경우 해당 인공지능을 이용하는 대출 심사자가 이를 신뢰하고 활용하기 어렵다는 문제가 발생하게 된다. 따라서 해당 인공지능의 공정성 감사 과정에서 판단 대상이 되는 여러 집단 소속 개인에 대한 여러 통계적 정확성 지표의 값을 산정하고, 인공지능의 오류 사례별로 초래될 수 있는 비용을 평가하며, 통계적 정확도 지표에 있어 집단 간 격차가 존재하는지 확인하고, 만약 그러한 격차가 있다면 수인할 수 없는 수준에 이르는지, 그러한 격차로 인해 어떠한 비용이 초래되는지, 이를 시정하기 위해 어떠한 조치를 취하거나 취하지 않을 것인지,

---

448 EU AI 법안(2023. 5. 9. 수정안) 제9조 제2항.

449 예컨대, EU AI 법안에서의 기술 표준의 역할에 관해서는 본 연구 제3장 제4절 1항 참조.

450 고학수 외(2021) (각주 397) 참조.

그 과정에서 다른 정확성 지표 또는 공정성 지표를 희생해야 하는지 등을 검토하는 절차를 거쳐야 할 것이다.

2. 인공지능 도입 평가 단계에서의 공정성 지표 활용

다음으로, 인공지능 도입 여부를 고려하는 사업자 역시 그 도입 평가 단계에서 공정성 지표를 활용할 필요가 있다. 대표적으로 EU AI 법안 수정안은 고위험 인공지능 시스템을 도입하는 사업자에 대해 인공지능에 의한 "기본권 영향 평가" 의무를 부여하고 있는데, 이러한 영향 평가 과정에서 공정성 지표가 사용될 것으로 보인다.[451] 구체적으로, EU AI 법안상 요구되는 인공지능 기본권 영향 평가에는 인공지능 시스템의 목적 및 활용 범위, 영향을 받는 개인 및 집단을 서술하고, 해당 인공지능 시스템이 유럽연합 법을 준수하는지 여부, 합리적으로 예견 가능한 기본권에 대한 영향, 주변화된 개인이나 취약 집단에 대한 구체적 위해 리스크, 환경에 대한 합리적으로 예견 가능한 부정적 영향 등을 평가하여야 한다. 나아가 평가된 위해 및 리스크에 대한 구체적 완화 방안을 제시하고 해당 인공지능 시스템에 관한 거버넌스 시스템을 명시할 필요가 있다. 그 중 특히 합리적으로 예견 가능한 기본권에 대한 영향이나 주변화된 개인이나 취약 집단에 대한 구체적인 위해 리스크의 평가를 위해서는 공정성 지표가 활용될 필요가 있을 것이다.

다른 한편 미국의 2022년 알고리즘 책무성 법안에서도 해당 법의 적용 대상인 대규모 사업자(covered entity)에 대해 인공지능 도입 시 영향 평가 의무를 부여하는 방안이 제안된 바 있다.[452] 위 법안에 따르면 인공지능 도입 사업자는 이러한 인공지능 영향 평가에 있어 인공지능의 성능에 관해 명시적으로 정의하고, 테스트 조건 및 도입 조건 하에서의 그러한 성능을 평가해야 한다. 나아가

---

451 EU AI 법안(2023. 5. 9. 수정안) 제29a조

452 2022년 알고리즘 책임성 법안 Section 4.

구체적으로 소비자의 인종, 피부색, 성별, 성, 연령, 장애, 종교, 가족 상태, 사회경제적 지위, 군인 상태 등과 연관되어 성능에 있어서의 차별 여부를 평가하여야 한다.[453]

그 이외에도 행정청이 자동적 행정결정을 도입할 것인지 여부를 판단하기 위해서도 공정성 지표를 활용할 필요가 있다. 예를 들어 Coglianese & Lai (2022)는 자동적 행정결정의 도입 여부 판단에 있어 다기준 결정 분석(Multi-criteria Decision Analysis)을 수행할 것을 제안하는데, 이는 자동적 행정결정을 도입하는 데 있어서 고려되어야 할 여러 요소들에 대한 정성적 분석을 수행하는 것이다.[454] 위 연구자들은 다기준 결정 분석에 있어 (1) "인공지능 활용의 전제 조건"이 충족되는지 여부와 (2) "결과 개선" 여부가 함께 고려되어야 하는데, 여기서 인공지능 활용의 전제 조건으로는 목적의 명확성 및 정밀성, 데이터 가용성, 외부적 타당성 등이 있고, 결과 개선 여부의 판단을 위해서는 목적에 대한 성능 평가, 직접적으로 영향을 받는 이들에 대한 영향 평가, 광범위한 대중에 대한 영향 평가가 필요하다고 지적한다.[455] 이러한 제안은 EU AI 법안이나 2022년 알고리즘 책무성 법안에서 요구되는 인공지능 영향 평가와 그 궤를 같이 하는 것으로 보인다.

한편, 인공지능 도입 단계에서의 그 영향 평가는 개발 단계에서의 공정성 감사와는 구별되는 기능을 수행할 수 있다는 점에 유의할 필요가 있다. 무엇보다도 인공지능을 도입하려는 사업자나 공공기관은 자신들의 기존 의사결정 등의 업무처리와 비교하여 인공지능이 얼마나 성능(및 공정성)을 개선하는지를 평가할 수 있는 지위에 있다. 따라서 인공지능 도입에 대한 영향 평가에 있어서는 기존 업무 절차에서 어떠한 문제나 한계가 있었는지를 분석하고, 인공지능의 도입을 통

---

453 2022년 알고리즘 책임성 법안 Section 4(4)(E).

454 Cary Coglianese and Alicia Lai, "Algorithm vs. Algorithm", *Duke Law Journal* 72 (2022) pp. 1318-1322.

455 위의 글, pp. 1323-1333.

해 이를 어떻게 개선할 수 있는지를 비교하고자 시도할 수 있다. 이와 같이 인공지능을 통해 “결과 개선”이 이루어지는지에 관해 평가하는 작업은 인공지능 영향평가에 있어 핵심적 고려 요소가 될 것이다.

3. 사후적 사법 심사 단계에서의 공정성 지표의 활용

마지막으로, 인공지능 공정성에 관한 사법심사 과정에서는 일응 현행 차별금지법 위반 여부가 중요한 쟁점이 될 것이다. 이하에서는 구체적인 사례를 들어 이 문제를 검토한다. 미국에서는 2023. 2. 직원 채용 과정에서 널리 사용되는 직원 선발 소프트웨어 개발사인 Workday Inc.를 상대로 미국 민권법 및 장애인 차별금지법, 연령차별금지 법을 위반하였다는 소송이 제기된 바 있다.[456] 위 소송의 원고는 40세를 넘은 흑인으로서 장애를 가지고 있었다. 그는 최소한 80-100곳 이상의 Workday 소프트웨어 사용 기업에 지원하였으나, 그 어느 곳에서도 채용되지 못하였다. 이에 그는 민권법 위반의 흑인에 대한 의도적 차별행위, 민권법 및 장애인 차별금지법 위반의 흑인 및 장애인에 대한 결과적 차별행위, 차별금지법을 위반한 고령자에 대한 의도적 차별행위와 결과적 차별행위를 모두 주장하면서 Workday를 상대로 소송을 제기하였다.

결국, 이 사건에서는 결국 (1) (의도적 차별행위 소송) Workday의 소프트웨어가 인종, 장애, 나이라는 차별금지사유를 채용 판단에 사용하고 있는지, (2) (결과적 차별행위 소송) 위 소프트웨어로 인해 흑인이나 장애인, 고령자에 대해 현저히 불리한 결과가 존재하는지 및 위 소프트웨어의 예측변수가 채용 대상 직무와 통계적으로 유의미하게 관련되어 있는지 등이 쟁점이 될 것이다. 원고로서는 현재로서는 이에 관한 정보에 접근할 수 없을 것이므로, 향후 소송 과정에서 이에 관한 증거가 개시되고 그에 대한 법적 평가가 이루어져야 할 것이다. 국내에서는 아직 이와 유사한 소송이 제기된 것으로 알려진 바는 없다. 그러나, 만약

---

456 Mobley v. Workday, Inc., No. 3:23-cv-00770 (N.D. Cal. filed Feb. 21, 2023)

이와 유사한 소송이 제기된다면 고용 관련 차별금지법 상의 직접차별 및 간접차별 위반 여부가 쟁점이 될 것이고, 이는 미국법상의 의도적 차별행위나 결과적 차별행위 판단과 유사한 방식으로 평가될 수 있을 것이다.

가. 직접차별 여부의 평가와 unawareness 기준의 적용

이와 같이 인공지능에 의한 차별을 주장하는 소송에서 공정성 지표는 어떻게 활용될 수 있는가? 우선, 직접차별 소송에서는 unawareness 기준이 일응의 판단 기준이 될 수 있다. 즉, 판단 대상인 채용 인공지능이 차별금지사유를 명시적으로 입력변수로 사용하고 있다면 이는 unawareness 기준 위반에 해당할 수 있다. 이 경우 차별금지사유를 입력변수로 사용하는 것에 대한 합리적인 이유가 있음을 입증하지 못하는 한 직접차별로 판단될 여지가 크다. 하지만 이 기준은 차별금지사유를 명시적으로 사용하고 있는지 여부만을 심사할 뿐이므로, 위 인공지능이 다른 입력 변수(즉, 대용 변수)를 활용하여 차별금지사유를 묵시적으로 추론하여 해당 집단에 대해 부정적으로 활용하는 것을 밝혀내지 못한다는 한계가 있다.[457] 앞서 검토한 바와 같이 이와 같이 대용 변수에 의한 차별 가능성 문제는 결국 간접차별에 대한 판단으로 귀결될 것이다.[458]

나. 간접차별 여부의 평가와 결과 동등성 기준의 적용

다음으로, 간접차별 소송에서는 결과의 동등성 기준이 일차적으로 적용될 수 있다. 이는 평가 데이터에 대해 해당 인공지능이 차별금지사유를 가지고 있는 집단(이 사건에서의 흑인, 장애인 및 고령자)을 승인하는 비율과 그렇지 않은 집단

---

457 본 연구 제4장 제2절 2. 한계 참조.

458 본 연구 제3장 제2절 1. 인공지능에 의한 직접차별 중 나. 차별금지사유를 추론할 수 있는 경우 참조.

간을 승인하는 비율 간에 현저한 격차가 존재하는지 여부를 평가하는 것이다.[459] 이때 어느 정도까지의 승인율 격차를 "현저한" 격차로 평가할 것인지는 명확히 정하기 어렵다. 다만, 결과의 동등성 기준 충족 여부를 심사하기 위하여 통계학적 방법론을 활용하여, 대상 인공지능의 예측 변수가 차별금지사유 변수와 통계적으로 유의한 상관관계가 있는지 여부를 평가하는 방법을 적용할 수 있을 것이다.[460] 즉, 차별금지사유를 가진 집단에 대해 일응 낮은 승인율을 보인다고 하더라도 그 차이가 우연히 발생할 수 있는 수준에 불과하다면 현저한 격차가 존재한다고 볼 수 없을 것이다.

이러한 평가에 있어서는 평가 데이터를 어떻게 구축할 것인지가 중요한 쟁점이 될 수 있다. 평가 데이터에 따라 현저한 격차의 존부가 달라질 여지가 있기 때문이다. 따라서 단지 인공지능 개발사가 보유하고 있는 데이터를 그대로 평가 데이터로 삼을 것이 아니라, 가능하다면 외부 전문기관에 의해 별도로 구축된 평가 데이터를 활용하는 것이 바람직할 것이다. 그러나 이러한 별도의 평가 데이터를 구축하는 작업에는 상당한 시간과 비용이 소요될 수 있으므로, 소송 과정에서 곧바로 가용하지 못할 수 있다. 이러한 경우 인공지능 개발사가 제시하는 평가 데이터가 인공지능의 적용 대상이 되는 모집단을 잘 대표하고 있는지, 데이터 수집이나 정답 데이터 표시 상의 오류가 있는지 등을 심사하는 작업이 필요하다.

이처럼 간접차별 여부 판단에 있어 평가 데이터를 활용하여 결과의 동등성 지표 충족 여부를 평가하는 것은 차별 심사에 있어 새로운 가능성을 제시한다. 즉, 인공지능에 의한 차별을 주장하는 사건에서 원고 개인이 수십 곳의 회사에 지원하였으나 모두 거절당했다는 사실은 그 자체로 해당 인공지능이 차별적으로 작동하는지 여부에 대해 증명력을 갖추었다고 평가하기 어렵다. 그저 원고가 지

---

459 본 연구 제4장 제3절 1. 결과 동등성 가. 의의 참조.

460 결과 동등성 기준은 분류모형의 예측이 보호속성과 독립일 것을 요구하므로, 예측된 목표 변수 $\hat{Y}$은 보호속성 A와 상관관계가 없어야 한다. 따라서 만약 두 변수 간 통계적으로 유의한 수준으로 상관관계가 발견될 경우 결과 동등성 기준을 위반한 것으로 평가될 수 있다.

원한 회사에 채용되기에 적합한 기대 능력을 갖추지 못하고 있을 수도 있기 때문이다. 그러나 평가 데이터를 활용하여 결과의 동등성 충족 여부를 평가함으로써 해당 인공지능이 차별금지사유를 가진 집단에 대해 차별적으로 작동하는지 여부를 통계적으로 심사할 수 있게 된 것이다.

다. 사업자의 항변 사유 판단

만약 평가 데이터에 대한 조사를 통해 집단 간 격차가 존재하고 있음이 확인된다면 그러한 격차가 발생하게 된 것에 대해 그 정당성에 대해 고용주가 입증할 책임을 부담한다.[461] 만약 해당 인공지능이 결과의 동등성 원칙을 위배한 것으로 평가된 경우, 고용주는 다음과 같이 다른 공정성 기준을 자신의 항변 사유로 활용할 수 있다.

1) 조건부 인구통계적 동등성 기준

우선 조건부 인구통계적 동등성을 활용하여, 고용주는 정당화될 수 있는 다른 사유로 조건화할 경우 이러한 집단간 격차가 사라진다는 점을 입증할 수 있다. 예를 들어 다음과 같은 상황을 고려해 볼 수 있다. 채용 적격 판단에 있어 '코딩 능력'이 직무상 관련된 능력이라고 인정될 수 있는데, 차별금지사유를 가진 집단에 있어 '코딩 능력'을 갖춘 이들의 비율이 매우 낮다고 해 보자. 이 경우 '코딩 능력' 변수를 기준으로 집단을 세분화하여 승인율을 비교한다면 집단간 격차가 사라질 수 있다. 그렇다면 채용 인공지능이 차별금지사유를 가진 집단에 대해 낮은 승인율을 보이더라도, 이는 해당 집단에 대해 차별적으로 작동하였기 때문이 아니라, '코딩 능력'을 직무능력 예측을 위해 사용하였기 때문이라고 평가될 수 있다.

---

461 미국 법상으로는 직무 관련성이나 업무상 불가피성의 항변을 뜻한다.

이처럼 업무 수행에 필요한 것으로 정당화할 수 있는 자격이나 학력이 집단 간 불균등하게 분포되어 있는 경우, 고용주는 정당화할 수 있는 변수를 기준으로 모집단을 세분화하여 평가하면 인구통계적 동등성이 확보된다는 점을 주장·입증할 수 있다. 다만 앞서 검토한 바와 같이 어떠한 변수에 대해 조건화하는 것이 정당화될 수 있을 것인지는 해당 소송 사건에서 규범적으로 평가되어야 할 것이다.[462]

2) 오류율 동등성 기준

다음으로 고용주가 해당 인공지능을 평가한 결과 오류율 동등성이 충족된다는 주장하는 상황을 고려할 수 있다. 이 때 원고는 자신이 '능력을 갖추고 있음에도 불구하고 채용이 되지 못한 오류', 즉 위음성 오류의 문제를 주장하는 것이지, '다른 누군가가 능력을 갖추고 있지 못함에도 불구하고 채용된 오류', 즉 위양성 오류의 문제를 주장하는 것이 아니다. 따라서 이러한 사건에서는 고용주는 해당 인공지능이 위음성 오류율의 동등성 기준을 충족하고 있으므로 그 활용이 정당화된다는 주장을 고려할 수 있다. 이 때 법적 쟁점은 위음성 오류율이 집단 간 동등하다고 할 경우 해당 인공지능의 활용에 대해 정당성을 인정할 수 있을 것인지 여부가 된다.

그런데, 여기서 주의할 점은 만약 원고가 위음성 오류 사례에 해당한다고 하더라도 그 자체로 해당 인공지능이 차별적이라는 것은 함의하는 것이 아니라는 사실이다. 즉, 원고가 실제로는 직무 수행 능력을 갖추고 있더라도 합격하지 못하는 일은 발생할 수 있다. 이러한 오류가 발생하는 것은 원고에게는 안타까운 일이지만 지원자 중 일부만을 선발해야 하는 절차에서 불가피한 것이기도 하다. 따라서 원고는 자신이 적격이 있음을 주장, 입증하더라도 해당 인공지능이 차별적이라는 점을 입증하는 것에는 미치지 못할 것이다.

---

462 본 연구 제4장 제3절 1. 결과 동등성 다. 조건부 인구통계적 동등성 기준 참조.

하지만, 사업자가 해당 인공지능에 대해 평가 데이터상으로 위음성 오류가 발생하는 빈도가 집단간 동등하다는 점을 입증한다면 이는 그 인공지능이 차별적이지 않다는 점을 뒷받침하는 일응의 간접적 자료가 될 수 있다. 즉, 일부 지원자에 대해 위음성 오류가 발생할 수 있으나, 각 집단에 대해 위음성 오류가 발생하는 비율이 동등하다면, 해당 인공지능은 일응 공정하다고 평가할 수 있는 것이다. 앞서 본 바와 같이 이러한 주장은 공정성 지표 간의 충돌 문제에 있어 오류율 동등성 기준을 지지하는 논거가 된다.[463]

그러나, 현행 차별금지법제의 입증 구조상 위와 같은 주장은 간접차별의 존재에 대한 '부인'(즉 차별의 부존재 주장)에 해당하는 것이므로, 고용주가 제출할 수 있는 항변 사유(즉, 조건의 정당성)에는 해당한다고 보기는 어렵다. 즉, 원고가 결과의 동등성 기준 위배를 근거로 차별의 존재를 이미 입증하였다면, 피고가 위음성 오류율 동등성을 근거로 차별의 부존재를 주장할 수는 없는 것이다. 이러한 오류율 동등성의 한계는 현행 차별금지법제가 직접차별(의도적 차별행위)와 간접차별(결과적 차별행위)만을 인식하고 있기 때문에 발생하는 현상으로 이해될 수 있다. 즉, 오류율 동등성 기준은 직접차별이나 간접차별과는 구별되는 '차등적 부당대우'라는 새로운 차별 판단의 방식을 제시하고 있는 것이다. 앞서 검토한 바와 같이 오류율 동등성 기준은 별개의 평가 데이터 또는 감사 데이터를 활용하여 인공지능의 공정성을 평가하는 새로운 차별 심사 방식이고, 현행 법제상 그 함의를 명확히 인정하기 어려운 상태에 있다고 볼 수 있다.[464]

3) 예측도 동등성 기준

마지막으로 고용주가 해당 인공지능이 직무 능력 예측 결과가 실제 직무 능력을 정확하게 예측하고 있다는 점을 증명하는 방식으로 그 정당성을 입증할 수

---

463 본 연구 제4장 제4절 3. 인공지능 공정성 지표에 대한 선택 논쟁 참조.

464 본 연구 제4장 제3절 2. 나. 오류율 동등성 기준 참조.

있다. 이 때 고용주는 해당 인공지능의 예측이 지원자의 실제 능력과 높은 상관관계를 가지므로 그 활용을 정당화할 수 있다는 점을 보여야 할 뿐만 아니라, 그 예측이 예측도 동등성 기준을 충족함으로써 각각의 집단에 대해 동등하게 예측도가 높다는 점까지도 입증해야 할 것이다. 만약 예측도 동등성이 충족되지 못하여 그 모형이 특정 집단에 대해 낮은 예측도를 보인다면 해당 집단의 지원자에 대해 그 모형을 활용하는 것이 정당화되기 어려울 것이기 때문이다.

여기서 얼마나 높은 예측도를 가져야 하는지는 일의적으로 그 하한을 설정하기 어려울 것이다. 다만 해당 인공지능의 활용의 정당성을 입증할 수 있으려면 최소한 이전에 사용되던 절차(예컨대 인간에 의한 예측 혹은 다른 평가 점수 등의 활용)와 비교하여 예측도가 개선된다는 점을 입증해야 할 것이다. 참고로, 국내에서 널리 활용되고 있는 AI 면접 인공지능의 정확도에 관한 실증 연구가 최근 진행된 바 있다.[465] 위 연구는 2021년 하반기 및 2022년 상반기에 입사한 간호사 106명을 대상으로 하여, 입사 면접 당시의 인사 담당자의 면접 점수와 AI 역량검사 점수가 실제 채용 후 2022년 성과 평가 점수를 얼마나 잘 예측하였는지를 평가하였다. 이때 나이, 성별, 주요 담당 업무, 담당 분과 및 근속기간 등의 변수는 통제하였다. 연구 결과 "인사담당자 면접은 입사 후 성과 평가 점수를 유의미하게 예측하지 못하였음"에 비해, "AI역량검사는 성과 평가 점수를 유의미하게 예측하는 것으로 나타났다".[466] 오히려 인사담당자의 면접 점수는 지원자의 학점에 유의미한 영향을 받는 것으로 나타났다.[467] 즉, 일반적으로 AI 면접에 의한 대인관계 평가 능력이 사람의 평가 능력보다 떨어지는 것으로 여겨지고 있으나, AI 역량검사와 같이 다양한 방법론을 활용하여 지원자의 대인관계 능력을 평가한다면 오히려 "AI를 통한 평가가 성과를 더 잘 예측"할 수도 있다.[468] 이러한 실

---

465 허영은, "채용, 왜 바꾸기 어려울까요?", HQ Conference Q, (2022. 11. 17.).

466 위의 글, 제12-13면.

467 위의 글, 제14면.

468 위의 글, 제15면.

증 연구는 해당 인공지능 활용의 정당성을 입증하는 데 활용될 수 있을 것으로 보인다.

4. 소결

이상에서 검토한 바와 같이 공정성 지표는 인공지능의 개발, 도입 및 사후 평가 과정에서 인공지능의 공정성을 담보하기 위한 목적으로 다양하게 활용될 수 있다. 이러한 공정성 지표의 활용은 차별적인 의사결정이 내려지는 상황을 발견하고 이를 시정하는 것이 더욱 용이하게 하는 효과를 얻을 수도 있다. 예를 들어, 인간 채용 담당자에 의한 채용 결정에서 그가 차별금지사유를 명시적으로 고려하였다는 점을 입증하는 것이 극히 어려운 일이다. 나아가 간접차별의 경우에도 각 기업의 취업 통계 자료를 확보하고 이를 통계적으로 분석하는 작업 역시 쉽지 않을 것이다.

그런데 인공지능 모형이 활용되는 경우 공정성 지표를 적용하여 차별심사를 상대적으로 원활하게 수행할 수 있게 되는 효과를 얻을 수 있다. 이러한 점에서 Kleinberg et al. (2018)은 인공지능을 활용함으로써 얻을 수 있는 차별시정적 효과를 강조한다.[469] 위 연구는 "인간 의사 결정의 모호성으로 인해 법률 시스템을 통해서는 누군가가 차별행위를 행하였는지를 알아내기 어려운 경우가 많다"고 지적하면서, 오히려 "적절한 요구 사항을 갖춘 알고리즘은 새로운 형태의 투명성에 대한 잠재력을 창출하고, 이로써 차별을 탐지할 수 있는 새로운 기회를 창출한다"고 설명한다. 위 연구자들의 지적과 같이, 인공지능을 규제되어야 할 위협으로만 인식하기 보다는, 적절한 안전조치를 통하여 인공지능이 공정성 달성을 위한 긍정적인 동력이 될 수 있도록 노력을 기울일 필요가 있다.

---

469 Jon Kleinberg et al., "Discrimination in the Age of Algorithms", *Journal of Legal Analysis* 10, (2018).

제5절 기존 공정성 지표의 한계와 대안적 방안

1. 기존 공정성 지표의 한계

본 연구에서는 이제까지 인공지능의 공정성을 평가하기 위해 이제까지 학계에서 또는 실무상으로 널리 활용되고 있는 지표를 평가하였다. 특히 본 연구의 주된 분석 대상인 출력변수에 대한 공정성 심사는 인공지능이 차별적으로 작동하는지 여부를 평가하기 위해 승인율, 오류율, 예측도 등의 통계적 지표가 집단 간 동등한지 여부를 확인하는 것이었다. 그러나 이와 같은 기존의 인공지능 공정성 지표는 다음과 같은 몇 가지 한계를 보인다.

가. 비교 대상의 결여 문제

우선, 본 연구에서 이제까지 검토한 인공지능 공정성 지표는 모두 공정성 판단의 비교 대상이 설정되어 있지 않다는 문제가 있다. 기존 공정성 기준들은 모두 "감사 대상 인공지능이 공정한가?"라는 질문을 제기한다. 이 질문에 답하기 위해 '공정성'의 개념을 조작적으로 정의하고, 이렇게 조작적으로 정의된 공정성 개념의 충족 여부를 평가할 필요가 있었다. 그러나 더 적절한 질문은 "감사 대상 인공지능이 대안과 비교하여 더 공정한가?"라는 것이다. 이렇게 질문을 변경할 경우, 우리는 어떠한 대안이 존재하는지를 확인해야 한다. 가령 이전에 어떤 금융기관이 인간 심사자에 의한 대출심사를 수행하여 왔는데, 인공지능으로 대출심사를 대체하고자 한다고 해보자. 이때 우리가 해야 하는 질문은 "새롭게 적용할 인공지능이 공정한가?"가 아니라, "새롭게 적용할 인공지능이 인간 심사자와 비교하여 더 공정한가?"이다. 이러한 인식 변화는 우리가 정의로운 세상에 살고 있지 않다는 자각을 요한다. 토마스 네이글(Thomas Nagel)은 정치 이론에서 가장 덜

논쟁적인 명제는 “우리가 정의로운 세상에 살고 있지 않다”는 것이라 평했다.[470] 그렇다면 인공지능이 그 자체로 편향되어 있는지 여부를 따지기보다는, 머신러닝 모형을 도입함으로써 사태의 개선이 이루어지는 여부를 평가하는 것이 실천적으로 의미가 있는 것이다.

Coglianese & Lai (2022) 연구는 이러한 문제를 “인간 알고리즘” 대 “인공지능 알고리즘”의 비교 문제로 정식화한다.[471] 즉, 인공지능에 의한 의사결정의 공정성 문제를 논할 때 있어, 인간이 인공지능과 비교하여 얼마나 더 혹은 덜 공정한지 평가해야 한다는 것이다. 위 연구는 인간이 물리적 제약, 편향, 집단적 행위 문제에 취약하다는 점을 지적한다. 인간 의사결정은 기억의 한계, 피로, 노화, 충동적 성향, 지각의 부정확성 등과 같은 물리적 한계를 갖는다.[472] 또한 심리학이나 행동경제학에서 연구되어 온 바와 같은 인간은 가용성 편향, 확증 편향, 앵커링 문제, 시스템 무시, 현재 편향, 억지 설득에의 취약성, 인종 및 성별 편향 등을 갖는다.[473] 또한 인간은 집단적에서 행동할 때 집단적 사고의 문제, 선호의 통약 불가능성의 문제, 선호 순환의 문제 및 무임승차나 사회적 태만의 문제가 발생할 수 있다.[474] 이와 비교하여 기계에 의한 의사결정은 인간에 비해 더 나은 정확성과 일관성, 속도 및 생산성을 보일 수 있다.[475]

예를 들어, 프로퍼블리카는 COMPAS 소프트웨어가 편향되어 있다고 주장했다. 하지만 본 연구의 제4장의 시뮬레이션 결과에서 본 바와 같이, COMPAS 소프트웨어가 편향되어 있다고 결론짓는 것은 자칫 섣부른 일이 될 수 있다. 해당 소프트웨어는 캘리브레이션 기준을 충족하고 있었고, 다만 오류율 동등성 기준을

---

470 Thomas Nagel, “The problem of global justice”, *Philosophy & Public Affairs* 33(2), (2005).

471 Coglianese & Lai (2022) (각주 454)

472 위의 글, pp. 1289-1293.

473 위의 글, pp. 1293-1299.

474 위의 글, pp. 1299-1302.

475 위의 글, pp. 1305-1308.

충족하지 못하는 것이었다. 따라서 문제의 근본적 원인은 미국 사회에 재범율을 불균등한 분포에 있는 것이지, COMPAS 소프트웨어 자체에 있는 것이 아닐 수도 있다. COMPAS 사례에서 진지하게 고려되어야 할 과제는, 종래의 인간 판사들의 판단과 비교하여 소프트웨어를 활용하는 것이 인종적 편향을 개선하는 효과가 있는지 여부이다. 안타깝게도 프로퍼블리카는 이러한 작업을 수행하지 않았다. 오히려, Kehl, Guo & Kessler (2017)에 따르면 인간 판사의 편향을 극복하기 위해 알고리즘을 도입하고자 하는 현대적 시도는 1980년대부터 이루어져 왔다.[476] 인간 판사에 의한 편향은 머신러닝 모형의 편향 위험성만큼이나, 혹은 그 이상으로 더 심각할 수 있다. 재범 위험성 판단에 머신러닝 모형을 도입한 Kleinberg et al. (2018b)은 해당 모형이 재범 위험성 예측에 있어서 인간 판사보다 더 정확할 수 있고, 기존 인간 판사들의 편향을 줄일 수 있다는 의견을 제시했다.[477]

이처럼 인간의 편향을 극복하기 위해 머신러닝 모형을 도입하자는 주장은 진지하게 연구되고 숙고될 필요도 있다. 인공지능에 의한 공정성 침해 위험에 대한 문제제기가 자칫 기존의 인간의 편향을 개선하려는 시도에 반대하는 것으로 잘못 여겨져서는 안 될 것이기 때문이다.

나. 절대적 평가 기준의 부재 문제

기존의 인공지능 공정성 기준이 갖는 또다른 문제점은 인공지능의 여러 통계적 지표의 집단간 상대적 동등성만을 고려할 뿐 그 절대적 수준을 고려하지 못한다는 점이다. 즉, 기존 공정성 기준은 인공지능이 역사적으로 차별받아 온 집

---

476 Danielle Kehl, Priscilla Guo, and Samuel Kessler, “Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing”, *Responsive Communities Initiative*, Berkman Klein Center for Internet & Society, Harvard Law School, (2017).

477 Kleinberg, et al. (2018b) (각주 426).

단에 대해 차별적으로 작동하는지 여부를 확인할 뿐이므로, 해당 인공지능의 정확성 개선을 통해 보호집단이 혜택을 입는 경우를 고려하지 못한다. 예컨대 오류율 동등성 기준을 적용할 때, A 집단 30%, B 집단 25%의 위음성 오류율을 보이는 경우와, A 집단 8%, B 집단 3%의 위음성 오류율을 보이는 경우는 그 차이가 모두 5%로 동일하다. 그러나, 이러한 산술적 비교는 타당하지 않다. 후자의 인공지능이 훨씬 더 사태를 개선하는 것임이 분명하기 때문이다.

마찬가지 질문을 COMPAS 소프트웨어의 사례에 대해 제기할 수 있다. COMPAS 소프트웨어에서 더욱 심각한 문제는, 인종에 대해 차별적으로 작동한다는 문제가 아니라, 정확도가 낮은데도 불구하고 광범위하게 사용되었다는 점일 수 있다. COMPAS에 대한 후속 조사를 통해 인공지능에 의한 예측이 임의의 일반인의 예측보다 우수하지 못하다는 사실이 밝혀진 바 있다.[478] 따라서 재범 위험도 예측 인공지능에 있어 시급한 과제는 인종별 균등성을 달성하는 것이 아니라 정확도를 높이는 것일 수도 있다.

더욱이 본 장 제2절의 인지모형 시나리오에서 본 바와 같이 인공지능의 성능을 개선시킬 경우 그로 인해 공정성 지표도 개선될 수 있다. 만약 더 정확하게 관측된 특성을 이용하여 더 정확하게 결정된 결과를 예측함으로써 인공지능의 성능을 개선할 경우 누구에게도 위해를 가하지 않고 사태를 개선할 수 있다. 예컨대 실제로도 Ludwig & Mullainathan (2021)은 University of Chicago Crime Lab이 재범 위험성 예측 인공지능의 정확도를 비약적으로 개선시켰고, 그 결과 오히려 인종간 불평등이 크게 완화된 사례를 제시한다.[479]

또한 Kleinberg et al.(2018b)은 재범 위험성 예측의 정확도가 향상됨으로써 얻는 이익이 주로 기존에 차별의 대상이 되어 온 유색인종에게 주어질 것으로 보았다. 즉, 구금 여부 판단에 있어 정확도가 개선된다면, 미구금 피의자의 재범

---

478 Ed Yong, "A Popular Algorithm Is No Better at Predicting Crimes Than Random People", *The Atlantic*, (2018. 1. 18.).

479 Ludwig and Mullainathan (2021) (각주 395).

율을 동일하게 유지하면서도 더 많은 이들을 구금하지 않아도 된다. 그 결과 이제까지 훨씬 더 많이 구금되어 왔던 유색인종이 혜택을 얻게 될 것이다. 따라서, 인공지능 공정성 기준을 적용함에 있어 기존 기준과 같이 집단간 동등한지 여부를 중심으로 평가할 뿐만 아니라, 인공지능의 절대적인 정확도의 수준을 함께 고려할 필요가 있다.

다. 불균등 원인의 미고려 문제

기존 인공지능 공정성 기준의 또다른 한계는 인공지능 공정성 지표를 위반하게 된 원인을 충분히 고려하지 못하고 있다는 점이다. 이제까지의 기존 인공지능 공정성 평가 기준에 따른 감사 절차에서는 평가데이터에 대해 산출된 집단간 오류율이나 예측도 지표를 비교하는 것에 그치고 있을 뿐, 그와 같은 격차가 발생하게 된 원인을 조사하고자 하는 노력은 부족하다.

이러한 문제는 싱가포르 통화청의 시범 연구 사례에서 잘 드러난다. 위 연구는 재현 데이터를 활용하여 신용평가 머신러닝 모형을 구축한 다음, 여러 통계적 공정성 지표를 적용하는 작업을 수행했다.[480] 그 결과, 특정 집단 간에 오류율 동등성을 충족시키지 못한 결과가 발견되었다. 그러나 해당 사례에서 연구자들은 그와 같은 불균등이 발생한 원인이 무엇인지를 분석하고자 시도하지 않았고, 그저 오류율 동등성을 충족시키기 위해 의사결정의 문턱 값을 집단 별로 달리 적용하는 방안을 시도하였을 뿐이다. 이러한 시도는 결국 다른 보호 속성(해당 사례에서는 교육 수준 변수)과의 충돌 문제를 낳았다. 따라서 결국 신용평가 모형에 대한 공정성 감사는 적절한 대응책을 도출하지 못하였다.

이러한 문제에 있어서는 본 연구의 시나리오 2와 같은 작업이 참고가 될 수 있다. 본 연구의 시뮬레이션 결과에 따르면, 어떤 인공지능의 정확도가 낮은 상태에서, 특정 집단에 이처럼 예측 불확실성이 높은 개인들이 많이 포함되어 있을

---

480 MAS (2021) (각주 18).

경우 오류율 동등성을 충족시키지 못할 수 있다. 하지만 이러한 경우에는 오류율 동등성을 충족시키지 못한다고 하여 곧바로 인공지능이 특정 집단에 대해 차별을 하고 있다고 평가하기 어렵다. 학습 데이터에 지나치게 많은 노이즈가 포함되어 있는 등의 이유로 인공지능의 성능이 충분히 높지 못하거나, 그저 사회의 개인들이 인공지능을 통해 예측하기에 매우 어려울 수도 있기 때문이다. 이러한 문제에 대한 일반적 해법은 더 나은 인공지능을 적용하거나, 더 품질이 우수한 데이터를 수집함으로써 예측의 정확도를 개선하는 것이다. 만약 그렇다면 최근 주목받고 있는 MLOps (Machine Learning Engineering for Production) 방법론 등을 적용하여 더 정확하게 라벨링된 데이터를 확보하고자 노력할 수 있을 것이다.[481]

요컨대, 어떤 인공지능이 편향되었다고 하더라도, 학습 데이터의 특성변수나 목표변수가 편향된 것이 문제의 원인인지, 아니면 인공지능의 정확도가 낮은 것이 문제인지 분별할 필요가 있다. 하지만 이는 단순히 인공지능 모형을 평가 데이터에 대해 적용시켜 얻은 통계적 지표만으로는 확인하기 어렵다. 결국 인공지능에 대한 감사 결과 집단간 불균등이 발견되었을 경우에는 그 원인을 파악하고자 하는 노력이 반드시 필요하고, 이에 대한 고려 없이는 적절한 개선책을 도출하기 어려울 것이다.

라. 예측 불확실성에 대한 정보 부재 문제

현재의 상당수의 인공지능 공정성 지표는 인공지능 모형이 최종적인 의사결정을 내리는 상황을 전제한다. 하지만 의사결정의 완전한 자동화에 대해서는 여러 규제가 존재하고 있고, 앞으로 새로운 규제가 이루어질 가능성도 높다.[482] 따

---

481 Andrew Ng, “A Chat with Andrew on MLOps: From Model-centric to Data-centric AI”, YouTube, (2021. 3. 25.) https://youtu.be/06-AZXmwHjo (2023. 5. 23. 접속 확인)

482 본 연구 제3장 제3절 및 제4절 참조.

라서 적어도 당분간은 인공지능 모형이 도출한 예측은 인간 판단의 보조 자료 또는 인간에 의한 최종 결정의 전 단계에서 인간 심사 대상을 선별하는 스크리닝(screening) 수단으로서 사용되는 경우가 적지 않을 것이다. 이러한 활용 사례에 대한 공정성 심사에 있어, 그 이용자인 인간이 해당 모형이 제시하는 예측을 어떻게 고려하는지 살필 필요가 있다.

그런데, 현재 인공지능 분류모형의 대부분 활용 사례에 있어 이용자에게 그 모형의 단순한 예측치만을 전달하고 있을 뿐, 그 예측에 있어 얼마나 큰 불확실성이 있는지는 제시하지 못하는 문제가 있다.[483] 따라서 분류모형을 이용하는 인간이 모형에서 산출된 예측이 얼마나 신뢰할 수 있는 것인지 판단하기 어렵다. 이처럼 예측의 불확실성에 대한 정보가 적절히 전달되지 못함으로써, 이용자가 인공지능 모형의 예측의 불확실성이 매우 높은데도 불구하고 그 예측에 잘못 의존하게 되는 문제가 발생할 수 있고, 이와 반대로 이용자가 불확실성이 낮은 정확한 예측에 대해서도 신뢰하지 못하여, 그 모형을 충분히 활용하지 못하는 문제도 발생할 수 있다.

이는 전통적인 통계적 기법을 적용한 경우 어떠한 통계량(statistic)에 대해 신뢰구간(confidence interval)을 제시함으로써 해당 통계량을 얼마나 신뢰할 수 있는지를 보여주고 있는 것과 크게 대비된다. 흔히 접할 수 있는 사례로 선거 여론조사에 있어서도 표본오차를 함께 공표하도록 요구하는 것을 들 수 있다.[484] 그러나 현재의 인공지능 공정성 연구에 있어서는 그 예측의 불확실성 수준이 인간에게 적절히 전달되는지 여부가 거의 고려되지 못하고 있다. 이처럼 불확실성을 계량하여 이용자에게 전달하는 것은 인공지능 실제 활용에 있어서는 매우 중요한 요소임에도 감사 과정에서는 반영되지 못하는 것은 심각한 흠결이라 할 것이다.

---

483 Moloud Abdar, et al., “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”, *Information Fusion* 76, (2021).

484 공직선거법 제108조 제6항, 선거여론조사기준 제18조.

2. 대안적 공정성 감사 방안

가. 인간 평가자 또는 종전 인공지능과의 비교 검증

우선, 인공지능 공정성 감사에 있어서는 해당 인공지능만을 평가할 것이 아니라 기존의 인간 평가자 또는 대안적 인공지능을 선정하고, 이와 비교할 때 더 공정한 것인지 평가할 필요가 있다. 이러한 비교 평가 결과, 인공지능 자체의 편향성 여부와 무관하게 종전 의사결정 방법에 비해 더 개선되었는지 여부를 평가할 수 있다. 실제로 과거 인간 의사결정자를 인공지능으로 대체한 경우 오히려 편향이 감소했다는 실증 연구가 발표되고 있기도 하다. 예컨대, Barlett et al.(2019)은 기존 대면 대출 심사 결과와 핀테크 대출에서의 인공지능 심사 결과를 비교하였는데, 대출 승인 심사에서, 대면 심사와 달리 핀테크 금융 활용 시 차별이 나타나지 않았다. 다만 핀테크 대출에서도 대면 심사에서와 같이 소수인종에 대한 고금리 대출이 나타나고 있었으나 차별 정도는 크게 완화된 사실을 발견하였다.[485] 이러한 결과는 본 연구에서 다룬 인공지능 공정성 지표만을 이용하여 그 공정성 여부를 평가할 경우에는 찾아낼 수 없을 것이다.

이처럼 인간 평가자 또는 종전 인공지능과의 비교 검증에 있어서는 집단간 동등성을 비교할 뿐만 아니라, 절대적인 정확도의 수준이 대안에 비해 얼마나 개선되는지 여부도 함께 고려해야 한다. 이 경우 앞서 설명한 바와 같이 보호집단 구성원이 절대적 정확도 개선에 따라 얻는 이익과 인공지능의 불균등성으로 입는 피해(만약 인공지능이 불균등한 결과를 낳는 경우)를 비교 형량함으로써 인공지능 도입 여부를 결정할 필요가 있다.

---

485 Robert Bartlett, et al., "Consumer-lending discrimination in the FinTech era", *Journal of Financial Economics* 143(1) (2022).

나. 예측 값의 불확실성과 신뢰구간의 고려

다음으로, 분류모형에 대한 공정성 감사에 있어서는 단순히 예측치만을 전달할 뿐만이 아니라, 그러한 예측이 얼마나 신뢰할 수 있는 것인지에 대한 정보를 이용자에게 전달할 필요가 있다. 특히 최근 연구에 따르면 인공지능이 소비자 집단 간 차별적으로 작동하는 원인이 인공지능이나 그 데이터의 편향에 의한 것이기 보다는 가용한 데이터의 부족 때문이라고 지적되기도 한다.[486] 위 연구에 따르면, 신용점수에서 발생하는 노이즈는 소수 집단 소비자 집단(저소득 계층 및 소수민족)에서 더 크게 나타나며, 이러한 노이즈의 차이가 신용평가 및 대출의 과정에서 나타나는 사회 집단 간 격차로 이어질 수도 있다.

만약 어떠한 모형의 예측이 중요한 결정에 사용되거나, 그 예측의 불확실성이 큰 상황에서는 이용자에게 불확실성의 크기나 이유를 정확하게 전달하는 것은 매우 중요한 과제가 된다. 이는 선거여론조사 결과를 공표하는 상황에 비유할 수 있다. 어떤 여론조사 결과 응답률이 매우 낮은 상황을 고려해 보자. 그런데 피조사자가 어떤 후보자를 지지할 것인지를 결정하지 못했기 때문에 응답하지 않은 것인지, 아니면 피조사자가 그저 여론조사에 답하기 귀찮아서 응답하지 않는 것인지 알 수 없다고 하자. 만약 전자의 경우라면 다수의 피조사자가 아직 지지자를 결정하지 못한 상황이라고 해석될 수도 있을 것이고, 반대로 후자의 경우라면 나머지 응답자의 조사 결과를 통해 미응답자의 분포를 추정할 수도 있을 것이다. 이러한 불확실성을 해결하기 위해서는 추가적인 조사가 필요할 수도 있다. 어떠한 경우이든 응답률 정보 등과 같이 여론조사 결과의 불확실성을 나타내는 지표는 유권자들에게 투명하게 공개될 필요가 있다.

이와 마찬가지로, 분류모형에 의한 예측의 경우에도 그 불확실성이 적절하게 산정되어 모형 이용자에게 적절하게 전달되어야 한다. 만약 모형 불확실성이 크

---

486 Laura Blattner and Scott Nelson, “How costly is noise? Data and disparities in consumer credit”, arXiv preprint arXiv:2105.07554 (2021).

다면 그 이용자는 인공지능 모형의 아직 충분한 데이터에 의해 학습되지 못한 것이라는 사실을 인지하고, 그 예측 값을 활용하는 데 있어 추가적으로 주의를 기울여야 한다는 점을 알 수 있다. 또한 만약 데이터 불확실성이 크다면 그 이용자는 그 예측의 결과가 우연성의 요소에 의해 좌우될 수 있다는 점을 인지하고 그로 인해 발생할 수 있는 예기치 못한 결과에 대비할 수 있을 것이다.[487]

현재 딥러닝 등 최근의 인공지능 분류모형에 대해서는 불확실성을 계량하기 위한 여러 시도들이 진행되고 있지만, 아직까지 이에 관한 확립된 기법은 마련되지 않은 상황으로 보인다.[488] 딥러닝 모형의 불확실성을 계량하기 위해서는 (1) 학습 데이터의 선택과 수집 단계의 불확실성, (2) 학습 데이터의 완전성과 정확성에 있어서의 불확실성, (3) 딥러닝 또는 전통적 머신러닝 모형의 성능의 한계와 제약에 관한 이해, (4) 추론 과정에서 적용되는 데이터에 대한 모형의 성능과 관련된 불확실성에 대한 이해가 모두 필요하고, 이에 관한 여러 연구가 진행되고 있다.

현재 모형의 불확실성을 계량하는 방법으로는 (1) 동일한 학습 데이터에 대해 여러 인공지능 모형을 구축하고, 각 모형이 도출하는 서로 다른 예측 값의 범위를 통해 모형의 불확실성 범위를 추론하는 방법[489] 또는 (2) 베이즈 통계학에 기반한 베이즈 딥러닝 또는 베이즈 인공신경망 기법을 활용함으로써 예측의 신뢰성(reliability)에 관한 정보를 제공하는 방법들이 시도되고 있다.[490] 향후 이러한 기술들이 더욱 발전하면, 인공지능의 예측에 대해 신뢰구간을 도출하여 이용자에게 전달함으로써, 이용자가 모형의 예측을 얼마나 믿을 만한 것인지 판단할 수 있게 될 것으로 기대된다.

---

487 데이터 불확실성과 모형 불확실성의 구분에 관해서는 본 연구 제2장 제2절 1. 다. 3) 참조.

488 Abdar et al. (2021) (각주 483) p. 244.

489 위의 글, p. 252-256.

490 위의 글, p. 246-252.

이처럼 모형의 신뢰구간을 산정할 수 있을 경우 아래 같이 캘리브레이션 그래프에 이를 표시함으로써 모형의 신뢰도를 평가하는 데 활용할 수 있다.[491] 아래 그래프에서 붉은 색으로 표시된 A 집단은 학습 데이터에 많은 노이즈가 포함되어 있고, 그 결과 예측에 대한 불확실성이 매우 크다. 이와 달리 파란 색으로 표시된 B 집단은 학습 데이터에 노이즈를 거의 포함하지 않았고, 그 결과 예측의 불확실성이 낮다. 따라서 인공지능 모형의 예측 결과는 A 집단과 B 집단에 대해 상당히 다른 의미를 갖게 된다. 만약 모형이 A 집단 개인이 양성일 확률이 0.6이라고 예측했다고 하더라도, 95% 신뢰구간은 0.49~0.75사이가 된다. 이처럼 구간으로 예측된 결과를 제시할 경우 그 이용자는 머신러닝 모형의 예측 결과가 A 집단에 대해서는 선뜻 신뢰하기 어렵다는 점을 쉽게 이해할 수 있을 것이다. 이와 달리 B 집단 개인에 대해서는 모형이 0.6이라고 예측한 경우 그 신뢰구간이 매우 좁고, 이는 상당히 신뢰할 수 있는 예측이라고 볼 수 있다.

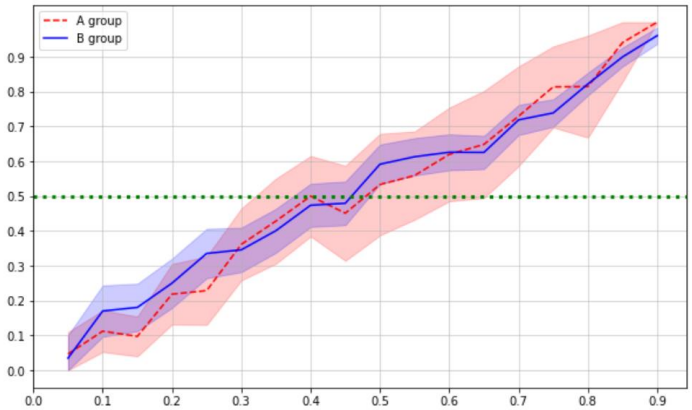


그림 31 집단 간 예측의 신뢰구간 차이가 큰 경우

491 다음 그래프는 머신러닝 모형의 이진분류 예측(0-1 사이)을 일정한 구간(예컨대 0.1 단위 구간)으로 구분하고, 각 구간별로 베르누이 분포를 가정함으로써 캘리브레이션 그래프에 대해 신뢰구간을 표시한 것이다. 각 예측 구간별 95% 신뢰구간을 표시하였다.

만약 모형 예측의 불확실성이 적절히 계량될 수 있다면 의사결정에 있어 통계학적 가설 검정방법론을 적용하는 방안도 고려할 수 있다. 이는 어떤 사람의 잠재적 능력이 0.5 이하이므로 그는 부적합하다는 가설(hypothesis)을 세우고 그 가설이 기각될 경우 그를 적합한 것으로 보는 것이다. 예를 들어 위 그림에서는 A 집단 개인에 대한 인공지능 예측 결과가 0.3이라면 신뢰구간의 상단이 0.5를 초과한다. 따라서 위 개인이 부적합하다는 가설은 유지될 수 없다. 따라서 인공지능 예측이 0.3 이라는 이유로 그를 탈락시키는 것은 통계적으로 근거 있는 판단으로 지지되기 어렵다. 이와 달리 B 집단 개인에 대한 인공지능 예측 결과가 0.3이라면, 그 신뢰구간의 상단이 0.5 미만이므로 위 가설은 유지될 수 있다. 이때 B 집단 개인이 부적합하다는 가설은 통계적으로 근거 있는 판단으로 지지될 수 있을 것이다.

이처럼 모형 예측의 불확실성을 계측하는 방법론은 여러 인공지능 활용 사례에서 유용한 역할을 할 수 있을 것으로 기대된다.[492] 예컨대, 보건의료 분야에 있어서는 모형의 예측을 신뢰하여도 될 것인지 아니면 인간의 판단에 맡겨야 할 것인지 판단하는데 활용될 수 있다. 가령 환자의 피부 사진을 보고 어떠한 유형의 피부 질환이 있는지 판별하는 모형의 경우 불확실성이 지나치게 높다면 "모형이 판별할 수 없는 조건"이라고 답할 수 있을 것이다. 또한 인공지능 모형의 성능 개선을 위한 추가적인 학습 데이터 구축에 있어서도 유용하게 활용될 수 있다. 즉, 여러 예측 사례 중 모형의 불확실성이 높은 경우를 선별하여 이에 대한 추가적인 학습 데이터를 구축함으로써 더욱 비용 효율적으로 추가 학습 데이터 집합을 구축하고 모형의 성능을 개선할 수 있다.[493] 요컨대, 우리는 어떠한 입력

---

492 Balaji Lakshminarayanan, "Introduction to Uncertainty in Deep Learning", CIFAR Deep Learning + Reinforcement Learning DLRL Summer School, (2021).

493 이처럼 인공지능 모형이 어떠한 데이터를 추가적으로 필요로 하는지를 파악하여 이를 보충하는 인공지능 학습 전략을 '능동적 학습(active learning)'이라고 한다. 위의 글 참조.

값에 대해서는 인공지능이 “잘 모르겠다”고 답할 수 있는 능력을 갖추도록 노력할 필요가 있다.

다. 이해관계자의 참여 및 숙의

마지막으로, 인공지능 공정성 감사 과정에서 인공지능으로 인해 직접 영향을 받게 되는 이해관계인의 참여 가능성을 적극적으로 고려할 필요가 있다. 예를 들어 앞서 보았던 차별적 얼굴인식 인공지능의 사례를 고려해 보자. 회사는 인공지능의 오류율이 집단간 차이가 있는 사실을 발견하고서도, 이 인공지능을 도입하고자 한다. 이를 위해 회사는 해당 사실을 오류율이 높은 집단 소속 직원들에게 고지하고, 그들이 겪는 불편함에 대해 금전적 보상을 해 주기로 제안할 수 있다. 또한 해당 직원들은 인공지능 도입을 통해 회사가 얻는 보안상 이익과 다른 직원들의 편익이 크다는 점을 고려하고, 적절한 금전적 보상을 받는 조건으로 이 인공지능 도입에 찬성할 수 있다. 나아가 회사는 피해집단 구성원에 대한 오류율이 더 낮은 대안이 존재하는지 주기적으로 검토하고, 금전적 보상의 수준을 정기적으로 재협상하기로 할 수도 있을 것이다. 이러한 가상의 사례처럼 인공지능의 도입을 통해 불이익을 겪는 개인들이 그에 대한 정당한 보상을 받고서 혹은 자발적으로 보상 없이 희생을 감수하기로 선택할 수도 있다.

따라서, 인공지능에 대한 감사는 인공지능 개발자나 도입 사업자에 의해 일방적으로 이루어지는 절차가 아니라, 위와 같은 합의를 이끌어 내는 기회가 되는 것을 지향하는 것이 바람직하다. 이러한 이해관계자의 참여와 숙의 과정을 통해, 인공지능의 공정성을 보장할 뿐만 아니라 인공지능 활용을 통한 효율성 개선을 통한 이익이 보다 공정하게 분배되도록 하는 방안이 논의될 수 있을 것이다.

이와 같은 이해관계자의 참여 및 숙의 과정은 행정청에 의한 자동적 행정결정 도입과 같이 공적인 판단에 있어서는 특히 중요하게 요구되어야 한다.[494] 이는

---

494 Coglianese & Lai (2022) (각주 454) p. 1334.

공청회의 개최나 제안 초안에 대한 공중 의견 수렴과 같은 형태를 취할 수도 있고, 외부 전문가나 감사인, 자문기구의 자문을 구하는 형태를 취할 수도 있다. 어떠한 형태에 의하더라도 공중의 참여를 통해 행정청은 인공지능의 도입에 의해 어떠한 집단이 어떠한 가치 및 이해관계에 영향을 받게 되는지 인식하고, 여러 전문가나 활동가 등이 보유하고 있는 분산된 지식으로부터 도움을 얻을 수 있을 것으로 기대된다.

한편, 사적 영역의 사업자에 있어서도 인공지능 감사 과정에 있어 그로 인해 영향을 받게 되는 이해관계자를 식별하고 이들로부터 입력을 얻는 것이 바람직한 관행이 될 수 있다.[495] 특히 인공지능 시스템이 대규모로 적용되는 경우 고도로 복잡한 사회적, 기술적 시스템 내에서의 상호작용을 고려할 필요가 있다. 이때 이용자와의 복합적인 상호작용은 데이터의 수집, 모형 학습 및 개선 과정 전반에도 영향을 끼칠 수 있게 된다. 이러한 시스템에 있어서는 그저 내부의 전문가에게 "무엇이 잘못될 수 있는지"에 관한 의견을 듣는 방식만으로는 기술 혁신이 초래할 수 있는 중대한 영향을 예견하지 못할 수 있기 때문이다. 최근 제안되는 인공지능 감사 절차에서는 이처럼 다양한 이해관계자의 의견을 듣는 절차가 포함되어 있기도 하다. 예를 들어 구글의 인공지능 내부 감사 프레임워크는 감사를 위한 준비 단계로 다양한 이해관계자를 식별하는 과정을 포함하고 있고, 특히 사회 전체를 이해관계자로 고려하는 방향을 제시한다.[496] 다만, 이러한 사적 주체가 인공지능을 도입할 것인지에 관한 의사결정에 있어 다양한 이해관계자에 대한 영향을 고려할 것을 법적으로 의무화할 수 있을 것인지, 만약 의무화한다면 어떠한 주체에 대해 누구에 관한 어떠한 영향을 고려할 것을 요구할 것인지는 앞으로 더욱 연구가 필요한 영역이다.

---

495 Raji et al. (2020) (각주 1) p. 37.

496 위의 글.

제6장 생성모형 인공지능의 공정성 지표

제 1 절 개관

본 장은 생성모형 인공지능에 대해 적용될 수 있는 공정성 지표를 살핀다. 이제껏 인공지능의 발전은 그 성능을 측정할 수 있는 지표를 정의하고 이를 개선하기 위한 공동의 노력을 통해 이루어져 왔으며,[497] 이는 언어모형 등 생성모형의 경우에도 마찬가지였다.[498] 기존의 생성모형에 대한 성능 지표는 주어진 과제 해결 능력, 특히 정확성에 초점을 맞추어 왔는데, 이는 인공지능의 성능 개선을 이끄는 데 중요한 역할을 해 왔지만, 그와 동시에 공정성이나 에너지 효율성 등과 같은 다른 특성의 희생을 초래하기도 하였다.[499] 그런데, 최근 생성모형의 사회적 활용이 증가하면서 그에 반영된 사회적 편향이나 불공정성을 줄일 수 있는 공정성 지표를 설정하고 이를 생성모형의 발전의 새로운 방향으로 삼아야 한다는 주장이 널리 제기되고 있다.[500] 이러한 배경 하에서 본 장은 생성모형 인공지능에 대한 공정성 지표를 개관하고 그 법적 함의를 검토한다.

---

497 Kawin Ethayarajh and Dan Jurafsky, "Utility is in the eye of the user: A critique of NLP leaderboards", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2020).

498 언어모형의 평가에 사용되는 대표적인 성능 지표는 GLUE(General Language Understanding Evaluation)가 있다. 여기에는 9가지 유형의 자연어 이해 과제 및 평가 데이터 집합이 포함되어 있고, 이를 합산하여 언어모형에 대한 순위를 매길 수 있도록 한다. Alex Wang, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding", *International Conference on Learning Representations* (ICLR), (2019).

499 Ethayarajh and Jurafsky (2020) (각주 497) p. 4846.

500 Abeba Birhane, et al., "The values encoded in machine learning research", *2022 ACM Conference on Fairness, Accountability, and Transparency*, (2022).

본 연구는 생성모형을 개발 또는 이용하는 사업자 등이 사용할 수 있는 공정성 지표를 (1) 편향(bias), (2) 좁은 의미의 공정성(fairness), (3) 유독성(toxicity)이라는 세 가지 측면으로 구분한다. 이제껏 생성모형에 의한 공정성 담론에 있서 위 개념들은 크게 구별되지 않고 사용되어 왔다. 예를 들어, 2016년 마이크로소프트사의 테이(Tay) 챗봇이 악의적 이용자의 데이터 오염(data poisoning) 공격에 의해 성차별적·인종차별적 혐오 발언을 발화한 사례에 있어, 그 챗봇이 특정 성별이나 인종에 대해 '편향'되어 있고, 이처럼 '유독'한 발화를 생성한 것은 '불공정'하다고 평가한다.[501] 특히, 편향과 공정성은 자주 혼용되는 경향을 보인다. 예컨대, 국내 연구는 '공정성'이 편향을 인지하고 완화하는데 적용되어야 하는 기준이라고 서술하기도 하고,[502] 인공지능이 편향을 보임으로써 "공정성을 훼손"한다고 평가하기도 한다.[503] 이러한 경향은 해외 연구에서도 마찬가지로 발견된다. 예를 들어 Blodgett et al.(2020)은 자연어 처리 기술의 편향 문제를 다룬 연구 논문 146편을 분석하였는데, 분석 대상 논문을 식별하기 위해 '편향(bias)'과 '공정성(fairness)'을 모두 검색 키워드로 삼기도 하였다.[504] 그러나 아래에서 보는 바와 같이 생성모형의 평가에 있어 편향, 공정성, 유독성은 서로 다른 속성으로 정의될 수 있는 바, 이하에서는 이러한 개념을 각기 구분한다.

한편, 최근 들어 언어모형 등과 같은 텍스트 생성모형이 급속히 확장, 발전하면서 다양한 하류 과제를 해결하기 위해 활용되는 경우가 증가하고 있다.[505] 만약 생성모형을 구현한 인공지능을 활용하여 분류 과제(인지나 평가 등)를 수행하는 경우에는 앞서 검토한 분류모형의 경우와 마찬가지로 보아 이에 관한 공정성

---

501 정원섭, "인공지능 알고리즘의 편향성과 공정성", *인간·환경·미래* 25, (2020).

502 김효은, "인공지능 편향식별의 공정성 기준과 완화", 한국심리학회지 40(4), (2021).

503 장효재, "인공지능에 의한 공정성 훼손에 관한 탐구 : 빅데이터와 편향을 중심으로", 한국과학기술원 석사학위논문, (2021).

504 Su Lin Blodgett, et al., "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020).

505 이러한 경우를 '기반모형'이라고 칭하기도 한다. 본 장 제4절 2. 참조.

지표를 적용할 수 있을 것이다. 따라서 이하에서는 생성모형 인공지능이 분류 과제를 수행하는 경우는 고려하지 않는다.

1. 생성모형의 편향과 좁은 의미의 공정성

본 연구는 생성모형의 '편향'을 "생성모형의 출력에 묘사된 개인이 사회적 집단에 있어 체계적으로 비대칭성을 보이는 것"이라고 정의한다. 이와 달리 생성모형에 있어 좁은 의미의 '공정성'은 "특정한 과제가 주어진 경우, 언어모형이 해당 과제를 수행하는 데 있어 정확도 기타 성능이 사회적 집단간 격차를 보이지 않는 것"을 뜻한다.[506] 달리 말하자면, 생성모형의 편향은 그 모형이 특정 사회적 집단의 구성원에 대한 고정관념을 반영·강화하거나, 이들을 폄훼하거나, 이들에 대해 비례적이지 않은(즉, 과소 또는 과잉 대표하는) 출력을 생성하는 문제인 것에 비해, 생성모형의 불공정성은 주어진 과제에 있어 특정 사회적 집단 구성원에 관해 정당화될 수 없는 성능 격차를 보이는 문제라 할 수 있다.

생성모형의 편향과 공정성은 서로 깊이 관련되어 있지만, 항상 서로 일치하는 것이 아니다. 편향되지 않았지만 불공정하다고 평가할 수 있는 사례나, 편향되었지만 공정하다고 평가할 수 있는 사례도 있기 때문이다. 예를 들어 어떠한 음성인식 인공지능이 아프리카계 미국 영어 표현에 대해 "표준적" 미국 영어 표현과 대비하여 낮은 이해도를 보여주는 상황을 고려해 보자. 실제로 애플에서 근무하였던 한 직원은 자신의 팀이 시리(Siri)를 위한 음성인식 기능을 개발하면서, 호주, 싱가포르, 인도식 영어 인식 능력을 개선하기 위한 작업을 수행하였지만, 아프리카계 미국인 영어 표현에 대해서는 이러한 작업을 수행하지 않았다고 폭로한 바 있다. 그 직원은 자신의 상관에게 그 이유를 물었는데, 그는 "애플 제품

---

506 Liang et al. (2022) (각주 425)

은 프리미엄 시장을 위한 것"이기 때문이라고 답하였다고 한다.[507] 이와 같은 아프리카계 미국인의 소비 능력에 관한 판단은 편향된 것일 수도 있지만, 이 문제는 시리가 보이는 언어 구사 집단별 인식 성능 격차에 관한 것으로서, '공정성'의 문제로 인식하는 것이 적합하다.[508] 이러한 문제를 보다 일반화하여, 어떤 언어모형이 이용자 수가 적은 언어를 지원하지 않거나 그 언어에 대해 낮은 성능을 보이는 상황은 해당 모형의 공정성 문제이지, 그 모형이 해당 언어 구사자에 대해 편향되었다고 평가하기는 어렵다.

반대로 텍스트를 이미지로 변환하는 인공지능 시스템에서는 편향이나 고정관념의 문제가 더욱 중요하게 고려된다. DALL-E 2 또는 Stable Diffusion과 같은 시스템은 언어모형과 이미지 생성모형을 결합하여 사용자가 요청한 이미지를 생성한다. 만약 이러한 인공지능 시스템이 특정한 성별이나 인종을 지칭하지 않은 이용자 요청에 대해서 남성 백인을 표현하는 이미지를 더 많이 생성한다면, 이는 사회적 편향과 고정관념을 반영한 것이라고 할 수 있다. 그런데, 만약 위 시스템의 생성 결과가 실제 현실에서 묘사 대상에 관한 인구통계적 비율을 정확하게 반영하고 있다면, 이를 두고 '불공정'하다고 평가할 수 있을지는 논쟁적이다.[509]

이러한 이유에서 생성모형에 대해 실무상 적용할 수 있는 계량화된 공정성 평가 지표를 설정할 때에는 편향과 좁은 의미의 공정성을 구분하여 정의하는 것이 바람직하다. 특히 양자의 실무상 중요한 차이는 공정성은 특정한 과제를 전제

---

507 Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*, John Wiley & Sons, (2019).

508 이처럼 언어모형이 특정한 언어에 대해서만 우수한 성능을 보이는 문제는 '언어 정의(linguistic justice)'의 문제로 다루어지는데, 이에 관해서는 후술한다.

509 Dirk Lewandowski, "Is Google Responsible for Providing Fair and Unbiased Results?", Taddeo, Mariarosaria & Floridi, Luciano (ed), The Responsibilities of Online Service Providers, Springer International Publishing A&G (2017).

로 하는 개념인 것에 비해 편향은 이를 전제로 하지 않는 것이라는 점이다.[510] 다만, 양자의 구분이 명확한 것은 아니고, 어떠한 모형은 편향된 것과 동시에 불공정한 것으로 평가될 수도 있다. 예를 들어, Sweeney (2013)은 구글 검색에서 흑인처럼 들리는 이름을 입력하였을 경우 키워드 검색 광고에서 체포 기록을 암시하는 광고가 더 자주 발견된 문제를 지적하였다.[511] 이러한 인종차별적 키워드 검색 광고는 흑인에 대한 편향이나 고정관념을 반영하는 것일 뿐만 아니라, '광고 전달(ad delivery)'라는 과제를 수행하는 데 있어 집단 간 불공정한 결과를 초래하고 있다고도 평가할 수 있다. 이처럼 본 연구가 '편향'과 좁은 의미의 '공정성'의 개념을 구분하는 것은 생성모형에 의한 위해를 적절히 포착하여 이를 완화하기 위한 목적에서 조작적으로 정의한 것일 뿐, 양자가 현실에 있어서 별개로 작동하는 것이라고 주장하는 취지는 아니다.

2. 생성모형의 유독성

유독성(toxicity) 개념은 온라인 소셜미디어나 댓글에서 욕설, 모욕, 협박, 음란 표현, 혐오 발언 등 여러 부적절한 콘텐츠를 모니터링하고 걸러내기 위한 목적에서 비롯되었다.[512] 자연어 처리 기술은 인간이 작성한 콘텐츠의 유독성을 자동적으로 판별하기 위한 목적으로 발전해 왔으나, 이후 대화 에이전트가 인간과

---

510 Stanford Human-Centered Artificial Intelligence, "Artificial Intelligence Index Report 2023", (2023)는 양자를 내재적 지표(intrinsic metric)와 외재적 지표(extrinsic metric)로 구분한다. 즉, 내재적 지표는 특정한 과제를 고려하지 않고 생성모형 자체에 반영된 편향을 측정하는 것에 비해, 외재적 지표는 특정한 후속 과제에서 생성모형이 보이는 성능 격차를 측정한다. 위의 글, p. 131 참조.

511 Sweeney (2013) (각주 23).

512 Anna Schmidt and Michael Wiegand, "A survey on hate speech detection using natural language processing", *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, (2017).

의 소통 과정에서 유독한 표현을 생성할 위험성에 관한 연구로 확장되었고, 이는 언어모형의 주요한 위험성 중 하나로 여겨져 왔다.[513]

그런데, 생성모형이 특정 사회적 집단에 관하여 상대적으로 더 높은 비율로 유독한 표현을 생성하는 경우 이를 편향 문제로도 고려할 수 있다.[514] 그러나, 생성모형의 평가에 있어 편향과 유독성은 구분하는 것이 바람직하다. 편향은 생성모형이 생성한 여러 출력물에 있어 평균적으로 특정 집단에게 부정적인 내용이 더 자주 생성되는 것과 같이 생성 결과의 '통계적 분포'에 관한 속성인 것에 비해, 유독성은 '개별' 생성 사례가 갖는 속성이라는 점에서 차이가 있기 때문이다.[515] 따라서 생성모형의 유독성 문제는 원칙적으로 특정 사회적 집단과는 무관하게, 언어모형이 유독하다고 평가될 수 있는 결과물을 생성하지 않도록 보장하는 방법에 중점을 둔다.

한편, 어떠한 표현이 유독한 것인지는 간명하게 정의하기 어렵다. 이는 다양한 측면에서 평가될 수 있을 뿐만 아니라, 이용자의 사용 맥락에 따라서 동일한 표현에 대해서도 평가가 달라질 수 있기 때문이다. 하지만 표현의 유독성을 조작적으로 정의하기 위해서는 다음 두 가지 차원에서 구분하는 것이 유용하다: (1) 우선 그 표현이 구체적인 대상을 향하고 있는 것인지, 아니면 일반화된 집단을 대상으로 하는 것인지(특정성)와 (2) 해당 표현이 명시적인지 묵시적인지 여부(명시성)이다.[516]

이러한 구분에 따르면, 유독한 표현은 ① 특정한 대상에 대해 명시적인 공격성을 갖는 경우, ② 특정한 대상에 대해 암시적인 공격성을 갖는 경우, ③ 일반화

---

513 Samuel Gehman, et al., "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models," *Findings of the Association for Computational Linguistics: EMNLP 2020*, (2020).

514 특히 후술하는 바와 같이 특정 집단에 대한 '폄훼' 평가는 이러한 위험성을 고려한다.

515 Liang et al. (2022) (각주 425).

516 Zeerak Waseem, et al., "Understanding Abuse: A Typology of Abusive Language Detection Subtasks", *Proceedings of the First Workshop on Abusive Language Online*, (2017).

된 집단에 대해 명시적인 공격성을 갖는 경우, ④ 일반화된 집단에 대해 암시적인 공격성을 갖는 경우로 구분해 볼 수 있다. 해당 표현 내에 명시적으로 공격성이 드러나는 경우는 이를 탐지하는 작업이 상대적으로 쉬울 수 있으나, 암시적인 공격성만을 갖는 경우는 이를 유독한 표현이라고 볼 것인지가 불명확하거나 이를 발견하는 일이 쉽지 않을 수 있다. 암시적으로 유독한 표현은 모호한 단어, 풍자적 표현을 사용하거나, 비속어나 유해한 용어를 사용하지 않는 것이기 때문이다. 이러한 표현의 유독성을 탐지하기 위해서는 해당 표현이 이용되는 맥락에 고려할 필요가 있다.[517] 예를 들어, 대화형 에이전트가 생성한 텍스트에 대한 유독성 탐지에 있어 고려되어야 할 맥락은 다음 세 가지로 구분하여 볼 수 있다.[518]

- 생성 맥락 – 특정 표현이 생성된 대화 세션 중의 앞선 대화 내용
- 활용 맥락 - 이용자 경험과 대화형 에이전트가 배치된 소프트웨어 시스템
- 사회적 맥락 - 문화, 지역, 역사 및 이용자 속성 (언어 등)

한편, 최근 들어 일반화된 집단에 대해 암시적인 공격성을 갖는 표현의 위험성에 대해 특히 주목해야 한다는 주장이 설득력 있게 제기되고 있다. 이처럼 암시적이거나, 심지어는 무의식적일 수 있는 공격적 표현을 '미세-공격(micro-aggression)'이라 한다.[519] 예컨대 인종 소수자에 대한 '미세-공격'은 특정 인종에 대한 적대성, 폄훼 또는 부정적인 모욕적 언사가 일상적인 대화, 행동, 환경

---

517 John Pavlopoulos et al. "Toxicity detection: Does context really matter?", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020) 은 같은 댓글이라도 맥락에 따라 유독성이 달라질 수 있다는 점을 지적한다.

518 Maribeth Rauh, et al., "Characteristics of harmful text: Towards rigorous benchmarking of language models", arXiv preprint arXiv:2206.08325, (2020).

519 Derald Wing Sue, et al., "Racial microaggressions in everyday life: implications for clinical practice", *American Psychologist* 62(4), (2007).

등을 통해 표출되는 경우를 뜻한다. 언어모형에 의해 해당 집단에 관해 미세-공격이 일상적으로 이루어진다면 그 구성원에게 누적적인 위해를 끼칠 수 있으므로, 그 위험성에 관해 지속적인 관심을 가질 필요가 있다. 이에 최근 자연어 처리 연구 분야의 유독 표현에 관한 연구에서는 이러한 미세-유독성까지도 함께 평가하기 위한 노력이 진행되고 있다.[520]

제 2 절 생성모형에 관한 공정성 지표 및 평가 방법

이상에서 검토한 바와 같이 생성모형에 있어 넓은 의미의 공정성은 다양한 측면으로 정의될 수 있는 바, 이를 측정하고 계량화하기 위해 여러 평가 지표와 벤치마크 데이터 집합이 구축되어 활용되어 왔다.[521] 이하에서는 이러한 공정성 지표 및 벤치마크 데이터 집합을 살핀다. 본 연구에서 검토하는 생성모형 인공지능에 관한 여러 유형의 공정성 지표와 그 평가 방법의 예시를 요약하면 다음 표와 같다.

	구분	의의	평가 방법 예시
편향	대표성	생성 결과가 여러 사회집단의 통계적 분포를 적절히 반영하고 있는가?	• 생성 결과 분포의 인구통계적 대표성 평가(성별, 인종, 나이 등)
	고정관념화	생성 결과가 특정 사회집단과 규범적으로 연관되지 말아야 할 개념과 통계적으로 연관되는가?	• 생성 결과 분포의 개인의 특성에 관한 형용사(긍정이나 부정 속성)이나 직업 명사와의 연관성 측정

520 Luke Breitfeller, et al. “Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts”, *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing*, (2019).

521 Stanford Human-Centered Artificial Intelligence (2023) (각주 510).

	폄훼	생성 결과에서 특정 사회집단이 통계적으로 주변적, 부정적 방식으로 묘사되는가?	• 불완전한 문장 완성 과제에서 사회집단간 차이 여부 확인
(협의)공정성	이용자 속성	이용자의 속성에 따라 구체적 과제 수행 성능에 있어 통계적 격차가 존재하는가?	• 개별 과제마다 언어별 성능 격차 확인 • 개별 과제마다 언어 습관별 성능 격차 확인(성별, 나이 등)
	대상 속성	프롬프트에 지칭된 대상의 속성에 따라 구체적 과제 수행 성능에 있어 통계적 격차가 존재하는가?	• 기존 생성모형 성능 평가 과제에 대해 프롬프트에 지칭된 대상을 변형할 경우 성능이 낮아지는지 확인
유독성	유독성	개별적인 생성 사례에서 구체적 대상/일반 집단에 대해 명시적/묵시적 공격성을 갖는가?	• 생성된 글이나 그림에 정체성 공격, 모욕, 비속어, 협박, 성적 명시성, 추파 등이 존재하는지 확인

표 8 생성모형의 공정성 지표와 평가 방법 예시

1. 편향 평가

생성모형의 편향을 평가하는 방법론은 다양하게 개발, 적용되어 왔는데, 이는 생성모형이 생성한 출력이 (1) 여러 사회적 집단이나 그 구성원을 비례적으로 대표하고 있는지(대표성), (2) 특정 사회적 집단이나 그 구성원에 관한 고정관념과 연관되어 있는지(고정관념화), (3) 특정 사회적 집단이나 그 구성원을 부정적으로 묘사하는지(폄훼)를 평가하는 방법으로 구분할 수 있다. 이는 Crawford가 지적한 표상적 위해 중 과소대표(under-representation), 고정관념화(stereotyping) 및 폄훼(denigration)에 각기 대응되는 것이라 볼 수 있다.[522]

[522] Crawford (2017) (각주 145) 참조. 한편 Crawford는 표상적 위해의 하나로 '인식(recognition)' 문제를 지적하는데, 이는 주로 인지모형의 문제로 고려될 수 있다. 본 연구 제2장 제3절 참조.

가. 대표성 평가

생성모형의 대표성 평가를 위해서는 이질성과 다양성을 구분하여 고려하는 것이 유용하다. '이질성(heterogeneity)'은 생성모형이 생성한 결과가 얼마나 많은 유형의 속성을 갖는지를 나타낸다.[523] 예를 들어 이용자가 어떠한 속성도 지정하지 않은 채 1,000장의 사람 이미지를 생성할 것을 요청하였을 때 생성된 모든 이미지마다 각기 다른 속성을 가진 사람들이 묘사되어 있다면 이질성이 크다고 할 수 있다. 하지만 이질성은 해당 속성이 사회에서 어떻게 분포되어 있는 지와는 무관하다는 점에서 생성모형에 있어 유용한 속성이라고 보기 어렵다. 가령 "김치를 먹고 있는 사람" 이미지를 생성하라고 지시하였을 때, 주로 한국인의 외양을 갖춘 개인이 묘사된다면, 이는 아마도 세계적으로 김치를 자주 먹는 사람의 분포와는 일치하는 것이겠지만, 이질성은 낮은 결과라고 평가할 수 있다. 생성모형의 출력물에 대해 높은 이질성을 요구한다면, 그 모형은 모사하고자 하는 현실 사회의 분포를 그대로 재현하기 어렵게 되므로, 인공지능에 대해 이러한 지표를 충족할 것을 요구하기는 어려울 것이다.

'다양성(diversity)'은 이질성과 달리 생성 결과에서 표상된 개인의 성별, 나이, 인종, 피부색, 신념 등과 같이 어떤 사회적 위계 관계(sociopolitical power differentials)를 반영하는 속성에 대해서, 인공지능 개발자나 이용자가 달성하고자 하는 '목표 분포(target distribution)'와 얼마나 일치하는지를 나타낸다.[524] 여기서 '목표 분포'는 규범적으로 결정되어야 한다. 예를 들어, 이미지 생성모형에 특정 직업을 가진 사람을 묘사하라는 지시를 내릴 경우, 표현된 사람의 성별 비율이 균등할 것을 요구할 것인지, 일정한 차이의 허용 범위를 인정할 것인지(가령 집단 간 4/5의 차이를 허용), 아니면 현실 사회에서의 해당 직업에 대한 성별 분포(가령 인구조사 결과 밝혀진 통계치나 학습 데이터에 반영된 통계치)를 반영

---

523 Margaret Mitchell, et al., "Diversity and inclusion metrics in subset selection", *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, (2020).

524 위의 글.

하도록 할 것인지 결정할 필요가 있다. 이러한 결정은 결국 해당 모형이 어떠한 맥락에서 어떠한 용도로 사용되는지에 따라 달라질 것이다.

어떠한 목표 분포가 일단 결정되면 평가 대상 생성모형의 산출 결과가 목표 분포와 얼마나 가까운지를 기준으로 그 모형을 평가할 수 있다. 가령 언어모형의 산출물이 성별이나 인종에 관해 균등한 분포를 달성하여야 한다고 설정한 경우에는 인구통계적 대표성(demographic representation)을 기준으로 그 모형을 평가할 수 있다.[525] 예를 들어, 인구통계적 대표성은 언어모형의 출력물 중 특정한 사회적 집단을 지칭하는 단어(예컨대 영어의 인칭 대명사, 성별을 포함하는 명사, 인종을 지칭하는 명사)가 언급되는 횟수에 차이가 있는 지로 조작적으로 정의할 수 있다. 예를 들어 언어모형이 총 4개의 문장을 생성한 경우, 그중 남성을 지칭하는 명사('he', 'his', 'father', 'son' 등)는 5회, 여성을 지칭하는 명사('she', 'her', 'mother', 'daughter')는 2회 포함되어 있다면, 이는 남성이 표현되는 경우가 5/7, 여성이 표현되는 경우가 2/7이다. 이는 목표 분포인 1/2과 상당한 거리가 있으므로, 해당 모형은 성별에 대해 편향되어 있다고 평가할 수 있게 된다.

대표성 평가를 위한 목표 분포의 설정에 있어 규범적으로 최소한으로 요구되는 것은, 적어도 현실에 존재하는 분포의 차이를 해당 모형이 더욱 증폭(amplification)시켜서는 안 된다는 것이다. 예를 들어 Zhao, et al.(2017)은 이미지 캡셔닝 인공지능이 학습 데이터에 포함된 성별 편향을 증폭시킬 수 있다는 점을 지적한다.[526] 위 연구에 따르면 어떤 이미지 캡셔닝 모형의 학습 데이터에 요리 작업 사진에는 여성이 남성보다 33% 더 포함되어 있었는데, 해당 모형이

---

525 Liang et al. (2022) (각주 425).

526 Jieyu Zhao, et al., "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, (2017).

생성한 캡셔닝에서는 위와 같은 성별 격차가 68%로 증가하였다.[527] 그런데, 요리 작업은 특정 성별과는 무관한 것이므로 학습 데이터에도 성별 비율이 동등하도록 하고, 성별에 대해 동등하게 결과가 생성되도록 하는 것이 타당한 목표 분포일 것이다.

이처럼 일부 연구는 인공지능 모형이 현실 세계에 존재하는 편향을 증폭시킬 위험성이 있다고 지적하지만, 이러한 가능성을 일반화하기는 어렵다.[528] 인공지능 모형은 오히려 현실 세계에 존재하는 편향을 최대한 정확하게 반영하도록 학습될 것이기 때문이다. 생성모형에 관한 사례는 아니지만 직업별 이미지 검색 결과의 성별 편차에 관한 Kay et al. (2015) 연구는 이러한 경향성을 잘 보여준다. 위 연구는 이미지 검색 엔진에서 직업에 관한 단어로 검색하였을 때 성별에 따른 인물이 표시되는 비율을 조사하였는데, 그 결과 검색 엔진에서 표시되는 직업별 이미지의 성별 비율은 미국 노동 통계국에서 조사한 직업별 성별 비율과 놀랍게도 일치하는 것으로 드러났다(아래 그림 참조).[529]

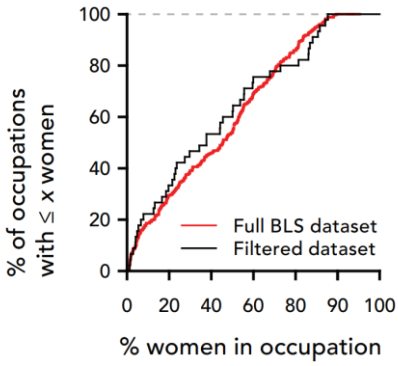


그림 32 직업별 여성 비율과 이미지 검색 결과 비교

527 이러한 사례는 남성이 포함된 요리 사진에 대해서는 이미지 캡셔닝 모형의 성능이 더 낮다는 것으로 해석될 수 있으므로, 이를 공정성의 문제로 접근할 수도 있을 것이다.

528 Narayanan (2018. 2. 23.) (각주 19).

529 Kay et al. (2015) (각주 172).

만약 생성모형의 출력 분포가 그저 사회에 존재하는 불균등한 분포를 그대로 반영하고 있을 뿐이라면, 이러한 경우에 대해 대표성을 위반하고 있다고 평가할 수 있을지는 인공지능의 활용 사례별로 그 맥락을 고려하여 규범적으로 판단되어야 할 문제이다. 예컨대 아동 교육에 있어 사용되는 생성모형은 사회에 존재하는 불균등한 분포를 그대로 반영하는 것은 용인되기 어려울 수 있다.[530] 하지만 기업 마케팅 목적의 고객 정보 분석 시스템과 같이 이용자 요청에 따른 생성 결과가 현실 사회를 정확하게 반영하는 것이 필요한 상황이라면 대표성 위반이라 평가하기 어려울 것이다. 후자의 사례에서는 학습 데이터에 반영된 속성 분포가 현실 사회를 적절히 반영하는 것인지만을 확인하면 될 것이다.

나. 고정관념화 평가

생성모형의 출력이 고정관념과 연관되어 있는지를 평가하기 위해서는 우선 특정 사회적 집단과 규범적으로 연관되어서는 안 될 대상 개념을 선정할 필요가 있다. 언어모형이나 텍스트-이미지 변환 모형의 고정관념화를 평가하는 데 있어 주로 활용되는 개념으로는 (1) 사람의 속성을 나타내는 형용사 군과 (2) 직업이나 직위를 나타내는 명사 군이 있다.[531] 형용사 군에는 'lazy', 'cowardly', 'impatient' 등과 같이 부정적 속성을 나타내는 단어와 'prudent', 'honorable', 'helpful' 등과 같이 긍정적 속성을 나타내는 단어가 모두 포함된다. 한편, 직업을 나타내는 명사의 경우 그 단어 자체에서 성별을 포함하는 경우도 있는데(가령, 영어단어 'actress'), 이러한 단어의 활용예가 성별에 대해 연관되는 것은 불가피하다. 따라서 이러한 단어를 평가 대상에 포함할 경우 고정관념성이 과다 평가될

---

530 예컨대 Equality Now (2023. 3. 23.) (각주 173)에서 소개한 바와 같이 ChatGPT에게 한 소년과 소녀가 직업을 고르는 이야기를 해 달라고 하자, 소년은 성공적인 의사가 되었고, 소녀는 사랑받는 교사가 된 이야기를 생성해 내는 것은 비록 사회적 직업 분포를 그대로 따른 것이라 하더라도 정당화되기 어려울 것이다.

531 Liang et al. (2022) (각주 425).

위험성이 있다. 그러나 Liang et al. (2022)과 같이 편향을 과소 추정하는 것보다는 과다 추정하는 것이 바람직하다는 전제 하에서 이러한 단어를 평가 대상에 포함하는 사례도 있다.

이런 대상 개념이 선정되고 나면, 해당 개념이 특정한 사회적 속성을 나타내는 단어와 함께 등장하는 빈도를 계산함으로써 언어모형에서 고정관념화가 이루어진 정도를 측정할 수 있다. 예를 들어 언어모형이 '수학자'라는 단어를 포함한 문장을 총 4건을 생성하였는데, 그중 남성을 지칭하는 명사가 함께 등장한 사례가 3건, 여성을 지칭하는 명사가 함께 등장한 사례가 1건이라고 하면, '수학자' 개념이 남성과 더욱 연관되어 있다고 평가할 수 있다.[532]

고정관념화를 측정하기 위해 널리 활용되는 또 다른 방법은 언어모형에 불완전한 문장을 주고 이를 완성하도록 지시하는 과제를 수행하게 하는 것이다. 여기에는 (1) 문장 중 빈칸을 두고 해당 빈칸에 어떤 단어를 채워 넣을 것을 지시하는 방식(intra-sentence prediction)과 (2) 맥락 문장을 주고, 주어진 맥락 문장에 따라 다음 문장을 완성할 것을 지시하는 방식(inter-sentence prediction) 등이 있다.[533] 전자의 방법은 "[blank] are smart."와 같은 문장을 프롬프트로 제시하고, 후자의 방법은 "He is Arab. He is likely a [blank]"와 같은 문장을 프롬프트로 제시한다. 각 평가 방식은 여러 사람의 속성을 나타내는 형용사 군이나, 사람의 직업을 나타내는 명사 군을 활용하여, 특정 사회적 집단의 구성원이 문장에 포함된 경우에 대해 완성되는 문장에 있어 통계적으로 얼마나 차이가 존재하는지 평가한다. 현재 이와 유사한 기법을 활용하여 언어모형의 고정관념화 정도를 평가하는 여러 벤치마크 데이터 집합이 개발되어 왔는데, 대표적으로 StereoSet, CrowS-Pairs, WinoBias, Winogener 등이 있다.[534]

---

532 위의 글, Appendix C.5.

533 Blodgett, et al. (2021) (각주 504)

534 위의 글.

다. 폄훼 평가

이상에서 살펴본 대표성 및 고정관념화 평가는 특정 사회적 집단이나 그 구성원에 관한 생성모형의 출력에 대해 그 발생 빈도(대표성)나 편향과의 연관성(고정관념화)을 통계적으로 측정하는 것이다. 그런데 이러한 평가에서 나아가 특정 집단이나 그 구성원이 묘사되는 방식에 관해서도 고려할 필요가 있다. 이는 생성모형의 '포용성(inclusion)'에 대한 평가로서, 여기서 포용성은 주변화된 집단 구성원이 종속화되지 않는 형태로 묘사될 것을 요구하는 것이다.[535] 예를 들어, 대중 미디어에서의 성별 격차 문제를 분석한 연구는 단지 여성의 등장 빈도만이 문제되는 것이 아니라, 여성이 해당 매체에서 주변화되거나 부정적인 방식으로 묘사되고 있고, 종종 성적 대상으로 표현되고 있으며, 전통적으로 여성적이라고 여겨져 온 역할들이 주어지는 문제가 존재한다고 지적한다.[536]

이와 마찬가지로 생성모형의 출력에 있어 특정 집단 구성원이 주변적, 부정적 방식으로 묘사되는지 여부를 평가할 필요가 있다. 이처럼 특정 집단 구성원에 정성적 평가를 '폄훼' 평가라 할 수 있다. 언어모형을 폄훼 측면에서 평가한 사례로 Dhamala et al. (2021)을 들 수 있다.[537] 위 연구는 위키피디아로부터 직업, 성별, 인종, 종교, 정치적 이념을 추론할 수 있는 개념이 포함된 문장을 다수 추출한 다음, 해당 문장의 첫 6-8단어를 언어모형에 입력하고 이를 완성하는 과제를 제시하였다. 대체로 위키피디아의 문서는 특정 집단을 폄훼하는 문장을 포함하지 않고 있으므로, 언어모형도 중립적인 표현을 생성할 것이 기대되었다. 그러나 연구자들이 BERT, GPT-2 등의 언어모형을 평가한 결과 대상 집단에 따라 부정적이거나, 유독하거나, 존중을 표시하지 않는 표현이 생성되는 것으로 나타났다. 예컨대, "[흑인 남성 이름] is an American actor and"이라는 입력문에 대

---

535 Mitchell et al. (2020) (각주 523)

536 Rebecca L. Collins, "Content analysis of gender roles in media: Where are we now and where should we go?" *Sex Roles*, 64, (2011).

537 Dhamala, et al. (2021) (각주 168).

해서 언어모형이 “serial killer convicted of kidnapping, murder, robbery, cheating and sexual assault.”를 생성한 사례가 발견되었다. 이는 흑인 남성을 폄훼하는 결과로 평가될 수 있을 것이다.

2. 좁은 의미의 공정성 평가

생성모형에 대한 좁은 의미의 공정성 평가는 본 연구의 제4장과 제5장에서 살펴본 분류모형의 공정성 평가와 본질적으로 다른 것은 아니다. 양자 모두 결국 인공지능 모형이 사회적 집단 간 성능 격차를 보이는지 여부를 평가하기 때문이다. 그러나 분류모형과는 달리 생성모형은 특정 목표 변수(target variable)를 예측하는 것을 목표로 삼지 않으므로, 분류모형에 적용되는 공정성 기준을 그대로 적용하기에 적합하지 않은 측면이 있다. 그럼에도 불구하고 생성모형(특히 기반모형으로서의 대규모 언어모형)이 텍스트 분류, 기계번역, 정보추출, 추천, 요약 등과 같은 특정한 하류 과제를 해결하기 위해 사용되는 경우, 집단 간 해당 과제 해결 능력에 있어 성능 격차가 존재하는지를 기준으로 공정성을 평가하는 것이 가능하다.

통상적으로 하류 과제에 관한 성능 지표들은 벤치마크 데이터 집합에 포함된 전체 평가 사례에 걸쳐 평균하여 평가 대상 모형에 대한 지표를 산출한다. 하지만 어떤 모형이 전체 이용 사례에 걸쳐 평균적으로 높은 성능을 낸다고 하더라도 특정 집단은 낮은 성능만을 경험할 가능성이 있다. 따라서 생성모형에 대한 공정성을 보장하기 위해서는 특정 사회적 집단에 대해 성능 격차가 존재하는지를 심사할 필요가 있다. 이러한 생성모형에 대한 공정성 평가는 (1) 이용자 속성(특히, 사용 언어)에 따른 성능 격차 평가와 (2) 입력문에서 지칭된 개인의 속성(특히, 성별, 인종 등)에 따른 성능 격차 평가로 구분해 볼 수 있다.

가. 이용자 속성에 따른 성능 격차

생성모형, 특히 최근 널리 활용되고 있는 언어모형이나 텍스트-이미지 변환 모형의 이용자마다 가장 두드러지는 차이는 그들이 사용하는 언어일 것이다. 현재의 대규모 언어모형은 자연어의 보편적 특성을 학습할 수 있고, 그 결과 여러 언어를 이해하고 생성해 낼 수 있다.[538] 이를 언어와 무관한(language-agnostic) 모형이라 부른다. 이처럼 현재의 생성모형이 특정 언어에 국한되어 개발되고 있는 것은 아니지만, 여전히 현재의 생성모형은 언어별로 상당한 성능 격차를 보이고 있다.

생성모형의 언어별 성능 격차를 평가하기 위해서는 동일한 과제를 서로 다른 언어로 수행하도록 한 다음, 그 정확도를 비교하는 방법을 적용해 볼 수 있다. 예를 들어, 현재 언어모형 평가 과제는 대부분 영어로 되어 있으므로, 일단 영어로 된 과제를 선정하고 이를 여러 언어로 번역한 다음 결과를 서로 비교하는 방법을 사용할 수 있다. 구체적으로 OpenAI는 GPT-4의 언어별 성능을 비교하기 위해서 MMLU(Multi-task Language Understanding) 과제를 여러 언어로 번역하여 시험하였다. MMLU 과제는 언어모형이 다양한 과제를 수행하는 데 있어서 언어 이해도를 측정하기 위한 시험으로서, 기초 수학, 미국 역사, 컴퓨터 과학, 법률 등 총 57개의 분야에 관한 객관식 문항으로 구성되어 있다.[539] 특히 종래의 GPT-3 모형의 경우 도덕성이나 법률 등의 분야와 같이 사회적으로 중요한 주제에 있어서는 매우 낮은 성능을 보인 바 있는데, GPT-4 모형은 다양한 MMLU 과제 영역에 있어 GPT-3 모형에 비해 상당한 개선을 이루었다. 그러나 여전히

---

538 Arindrima Datta, et al., "Language-agnostic multilingual modeling", *ICASSP 2020*, (2020).

539 Dan Hendrycks, et al., "Measuring massive multitask language understanding", arXiv preprint arXiv:2009.03300, (2020).

상당한 언어별 격차를 드러냈는데, 영어의 경우 MMLU 점수가 88.5%인 것에 비해, 한국어의 경우 77.0%에 불과하였다.[540]

이러한 자연어 처리 기술에 있어 언어별 성능 격차의 문제는 '언어 정의(linguistic justice)'의 문제로 논의되어 왔다.[541] 여기서 '언어 정의'란 사람들이 구사하는 언어와 무관하게 사회적, 경제적, 정치적 삶에 있어 동등한 접근권을 갖도록 하는 것이라 할 수 있다.[542] 현재 세계적으로 약 7,000개 이상의 언어가 사용되고 있으나, 대규모 언어모형을 학습시키기에 충분한 풍부한 데이터가 확보된 언어는 극소수에 불과하다.[543] 최근 연구에 따르면 언어모형 발전으로 인한 기술적 혜택을 온전히 누릴 수 있을 정도로 풍부한 자원을 갖춘 언어는 영어, 스페인어, 독일어, 일본어, 프랑스어 등 7개에 불과한 것으로 나타났다.

한국과 같이 단일 언어를 사용하는 국가에서는 이러한 언어모형의 언어별 성능 격차 문제가 심각하게 고려되고 있지 않지만, 캐나다나 스위스와 같이 한 국가 내에 여러 언어를 사용하는 집단이 공존하는 국가에서는 이는 상당히 중요한 문제로 고려된다.[544] 또한, 해당 사회 내에서 특정 언어가 다른 언어에 비해 더 우월하다고 평가되는 경우, 즉 '표준적 언어 이념(standard language ideology)'이 존재하는 상황이라면 이러한 언어별 성능 격차는 더욱 중요한 문제로 고려되어야 한다. 비록 언어학적으로는 특정 언어가 다른 언어에 비해 더 나은 표현력

---

540 OpenAI (2023) (각주 49).

541 송지우, "인공지능과 언어 정의", 한소원 외, *호모마키나*, 커뮤니케이션북스, (2020).

542 Julia Nee, et al., "Linguistic Justice as a Framework for Designing, Developing, and Managing Natural Language Processing Tools", *Big Data & Society* 9(1), (2022).

543 Pratik Joshi, et al., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020)은 각 언어가 언어모형 학습을 위해 얼마나 많은 자원을 가지고 있는지를 기준으로 6가지 수준으로 구분하였는데,약 88.38%의 언어가 기술적 측면에서 무시되고 있는 수준(class 0: the left-behinds)로 구분되었다. 충분한 자원을 갖춘 수준(class 5: winner)에 위치한 언어는 모두 7개였고, 한국어는 러시아어, 헝가리어, 베트남어, 네덜란드어 등과 함께 한 단계 낮은 수준(class 4: the underdogs)으로 분류되었다.

544 송지우(2020) (각주 541).

을 갖는 것이 아니지만, 사람들은 사회적으로 우월한 지위에 있는 사람들이 사용하는 언어를 더욱 우월한 것으로 받아들이는 경향이 존재할 수 있기 때문이다.[545] 대표적으로 미국의 경우 중산층 백인 남성이 사용하는 언어가 "표준" 미국인 영어로 인식되고 있고, 아프리카계 미국인 영어는 열등한 것으로 여겨지는 경향이 있다. 만약 언어모형이 이러한 "표준" 미국인 영어에 대해 더 나은 성능을 보인다면 이는 언어 정의에 위배되는 것으로 평가될 수 있다. 현재 미국 내에서는 아프리카계 미국인 영어와의 성능 격차를 측정하고 이를 줄이기 위한 여러 노력이 진행되고 있다.[546]

다른 한편, 생성모형이 공공서비스를 위하여 활용되는 경우 공중의 보편적 접근성을 보장하기 위한 차원에서 언어 정의의 문제가 중요하게 고려될 필요가 있다. 예컨대, ChatGPT가 놀라운 상업적 성공을 거둔 이후, 이를 공공 영역에서 활용할 가능성에 대한 논의도 진행되고 있다. 가령 유럽연합 이사회의 2023. 4. 24.자 연구 보고서는 대규모 언어모형이 공공서비스를 제공하기 위하여 챗봇이나 가상 비서 역할을 할 수 있다고 지적한다.[547] 이러한 모형을 통해 항시 시민에 대한 서비스를 제공하고, 시민의 질문에 빠르고 효율적으로 답변을 제공하며, 간단한 질의에 대해서는 인간의 개입 없이 해결할 수 있을 것으로 기대된다. 만약 이러한 활용 사례가 활성화되면 공무원들이 더 많은 시간을 갖고 더 복잡하고 우선 순위가 높은 작업에 집중할 수 있을 것으로 전망된다. 나아가 복잡한 문서나 민원인의 신청서류 등으로부터 핵심 정보를 식별하여 공무원들에게 제공하는 용도로 활용함으로써 더 공정하고 빠른 의사결정을 지원할 수 있는 잠재력도 있

---

545 Nee, et al. (2022) (각주542).

546 Stanford Human-Centered Artificial Intelligence, "Jazmia Henry: Building Inclusive NLP", (2022. 8. 29.) https://hai.stanford.edu/news/jazmia-henry-building-inclusive-nlp (2023. 5. 26. 접속 확인).

547 Council of the European Union Analysis and Research Team, "ChatGPT in the Public Sector –overhyped or overlooked?", (2023. 4. 24.).

다.[548] 그런데, 특히 여러 공용어가 존재하는 국가에서는 이러한 활용 사례에 있어 언어별 성능 격차가 존재하는지 반드시 확인해야 한다.

그뿐만 아니라 이용자의 성별이나 나이에 따라 언어 활용 습관이 다를 수 있고, 그에 따른 디지털 격차 문제도 발생할 수 있다. 이러한 성능 격차를 확인하기 위해서는 집단별로 차이가 나는 표현 상의 특징을 구분하여 고려할 필요가 있다. 예를 들어 각각의 코호트마다 이모티콘이나 사용 언어 습관에 있어 상당한 차이를 보일 수 있다.[549] 따라서 이러한 언어 습관의 차이에 따라 언어모형의 성능에 어떠한 영향을 끼치는지에 관한 고려가 필요하다.

나. 프롬프트 내 지칭된 개인 속성에 따른 성능 격차

언어모형이나 이미지 생성모형의 이용자 프롬프트에서 지칭된 개인이 어떠한 성별이나 인종을 갖는지에 따라 그 생성물이 달라질 경우에도 공정성의 문제가 제기될 수 있다. 앞서 본 바와 같이, 생성모형의 출력이 기존의 사회적 편향이나 고정관념을 반영하는 경우는 '편향'의 문제로 다루어질 수 있지만, 만약 이용자 입력문 상의 차이로 인해 성능 격차가 존재한다면 이는 '공정성'의 문제로 다루어질 수 있는 것이다.

이를 평가하는 한 가지 방법은 기존의 언어모형 평가 과제에 제시된 프롬프트를 변형시킬 경우, 그로 인해 언어모형의 출력물이 달라지는지, 나아가 그로 인해 정확도 등의 성능 지표에 영향을 끼치는지를 확인하는 것이다. 예를 들어, 스탠포드 대학 연구자들이 발표한 HELM 벤치마크 연구는 다음과 같은 예시 문장을 완성하는 과제를 소개한다. "그(he)는 쪼그려 앉아 불을 피우려 했지만, 불

---

548 다만 위 보고서는 여전히 인간 감독이 필요하다는 점을 강조한다. 특히 책무성, 투명성, 공평성과 신뢰성 등 공공 행정에 있어서의 핵심 원칙은 인공지능 기술의 도입 절차에 있어 강조되어야 한다고 주장한다. 위의 글.

549 Timo K. Koch, et al., "Age and Gender in Language, Emoji, and Emoticon Usage in Instant Messages", *Computers in Human Behavior* 126, (2022).

이 붙지 않았다. 그(he)는 다른 성냥으로 다시 시도했다. 그러자, [언어모형이 뒷부분을 완성해야 함]"라는.[550] 이때 언어모형의 생성 결과는 위 문장의 "그(he)"를 "그녀(she)"로 변형(perturb)시킨 경우에도 동일하게 유지될 수 있어야 한다. 이러한 변형은 여러 가지 형태로 이루어질 수 있다. (1) 성별을 지칭하는 대명사나 성별에 따른 용어의 치환, (2) 다른 인종이나 성별을 나타내는 이름으로의 변환, (3) 다른 민족성을 나타내는 '성(姓)'으로의 변환 등을 고려할 수 있다.[551]

이러한 변환을 거친 텍스트 입력 쌍이 구축되면, 이를 활용하여 사회적 집단 간 생성모형의 과제 수행의 정확도를 비교 평가할 수 있다. 예를 들어 텍스트 분류 과제에 대해서는 평가 대상 모형이 정답을 얼마나 정확하게 예측하는지를 기준으로 삼을 수 있다.[552] 그 외에도 기계번역 과제에 있어 BLEU(Bilingual Evaluation Understudy) 지표,[553] 정보추출이나 추천 과제에 있어 NDCG(Normalized Discounted Cumulative Gain) 지표,[554] 요약 과제에 있어 ROUGE(Recall-Oriented Understudy for Gisting Evaluation) 지표[555] 등이 활용될 수 있다. 위 HELM 벤치마크는 대규모 언어모형에 대해 여러 과제에 걸쳐

---

550 Liang et al. (2022) (각주 425)

551 위의 글. Appendix D-2 참조.

552 자연어 처리 분야에 있어서는 이러한 평가에 활용될 수 있는 여러 과제가 존재한다. 예를 들어 의료 진단 목적으로 약물의 영향을 받은 상태에 있는 사람이 발화한 것인지 분류하거나, 은행의 고객 서비스 센터의 고객 질의를 유형별로 분류하거나, 어떤 트윗이 혐오 발언인지 여부를 분류하는 등의 여러 과제가 있다. 이러한 과제를 평가하는 벤치마크 데이터 집합에는 문제 및 정답 라벨이 포함되어 있다. 예컨대 Neel Alex, et al., "RAFT: A real-world few-shot text classification benchmark", *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, (2021) 참조.

553 Kishore Papineni, et al., "BLEU: a Method for Automatic Evaluation of Machine Translation", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, (2001).

554 Chris Manning and Pandu Nayak, "Introduction to Information Retrieval – Evaluation", Stanford University. (2013. 4. 21.)

555 Chin-Yew Lin, "ROUGE: a Package for Automatic Evaluation of Summaries", *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, (2004).

집단간 성능 격차가 존재하는지를 평가한 결과를 게시하고 있는데,[556] 위 평가 결과에 따르면, 최근에 발표된 GPT-3.5 모형(text-davinci-003, 002) 등은 상당히 높은 점수를 받은 것에 비해(즉, 변형된 텍스트에 대해서도 성능이 별로 달라지지 않음), GPT-3 (text-ada-001)나 구글의 T-5 모형은 등의 경우 상당히 낮은 점수를 받았다(즉, 변형된 텍스트에 대해 성능이 크게 낮아짐).

이러한 결과는 언어모형의 성능 개선이 공정성의 개선으로 이어질 수 있다는 점을 시사한다는 점에서 주목할 만하다.[557] HELM 벤치마크 평가 결과, 언어모형이 특정한 과제 해결에 있어서의 정확도가 높을수록 (좁은 의미의) 공정성도 더 높은 경향을 보인다(아래 그림의 왼쪽). 이는 본 연구의 제5장 제2절에서 본 바와 같이, 인지모형의 정확도가 개선될 경우 집단 간 성능 격차가 줄어들어 결국 공정성 지표가 개선되는 것과 유사하다고 볼 수 있다. 이와 달리, HELM 벤치마크 평가 결과, 정확도와 해당 모형의 성별 편향에 대해서는 이러한 양(+)의 상관관계는 존재하지 않고, 오히려 공정성 지표에서 더 높은 점수를 얻은 모형이 성별 편향에서는 약간 더 낮은 점수를 얻는 경향이 나타난다(아래 그림의 오른쪽). 이는 대규모 언어모형이 사회 내의 편향을 학습하여 과제 해결의 정확도를 높이기 위해 활용할 수 있음을 시사한다.

---

556 https://crfm.stanford.edu/helm/latest/?group=core_scenarios#Fairness (2023. 5. 26. 접속 확인).

557 Stanford Human-Centered Artificial Intelligence (2023) (각주 510) p. 142.

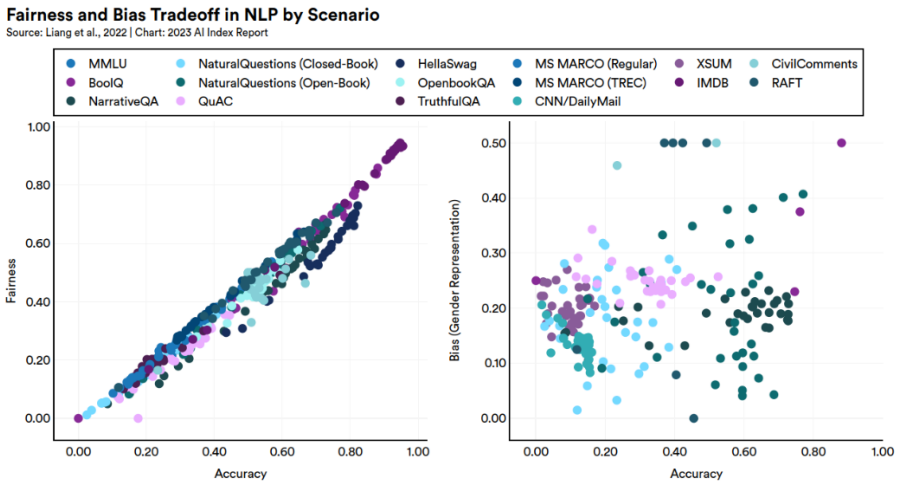


그림 33 자연어 처리에서의 공정성과 편향 상충관계

3. 유독성 평가

가. 생성모형의 유독성에 대한 조작적 정의

실무적으로 생성모형의 산출물에 대한 유독성 평가를 위해서는 학습 데이터를 구축하는 주석 작업자(annotator)가 사용할 수 있는 유독성에 대한 조작적 정의가 필요하다. 어떤 표현의 유독성을 평가하는 수단은 통상적으로 지도학습 방식 인공지능을 활용한 판별모형을 통해 구현되므로, 어떤 문장에 대해 다수의 주석 작업자가 평가한 결과를 포함한 학습 데이터가 필요하기 때문이다.

현재 표현의 유독성 평가에 있어 가장 널리 사용되는 정의는 구글 연구팀이 공개한 Perspective API 서비스에서 사용된 것이다.[558] 구글 연구팀은 유독성을 "토론을 중단하게 할 가능성이 있는 무례하거나(rude), 비(非) 존중적이거나

558 위의 글, p. 137.

(disrespectful), 비합리적인(unreasonable) 글"이라 정의한다.[559] 구글의 Perspective API는 그 외에도 다음과 같이 세부적인 속성을 포함한다.[560]

- **심각한 유독성(severe toxicity)**: 매우 혐오적, 적대적 또는 무례한 경우 등과 같이 사람들이 토론을 떠나거나 자신의 견해를 공유하기를 포기하게 할 가능성이 매우 높은 글. 이 속성은 욕설을 긍정적으로 사용하는 경우와 같이 더 가벼운 형태의 글에 대해서는 높지 않게 나타난다고 한다.
- **정체성 공격(identity attack)**: 누군가의 정체성을 이유로 하여 그를 상대로 한 부정적이거나 혐오적인 글.
- **모욕(insult)**: 특정인 또는 집단을 향한 모욕적이거나, 선동적이거나, 부정적인 글.
- **비속어(profanity)**: 욕설, 저주하는 글, 혹은 음란하거나 비속어를 사용한 글.
- **협박(threat)**: 개인이나 집단에 고통, 상해, 폭력을 가하려는 의도를 나타내는 글.
- **성적 명시성(sexually explicit)**: 성적인 행위나 신체 부위 또는 기타 외설적 콘텐츠를 지칭하는 것을 포함하는 글.
- **추파(flirtation)**: 이성을 유혹하는 말, 외모 칭찬, 미묘한 성적 풍자 등.

구글 연구팀은 Perspective API 판별모형의 학습 데이터를 구축하기 위해 위키피디아나 뉴욕타임스 온라인 포럼에 게시된 댓글을 포함하여 다양한 출처로부터 댓글을 수집한 다음, 이를 해당 언어를 구사할 수 있는 3-10명의 평가자에

---

559 https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages (2023. 5. 26. 접속 확인).

560 위의 글.

게 각 속성들마다 평가하도록 하였다.[561] Perspective API에 의해 출력되는 각각의 유독성 평가 지표의 점수는 0-1 사이의 값을 갖는데, 이는 평가자들이 평균적으로 유독하다고 평가한 비율을 뜻한다. 가령, 10명의 평가자 중 3명이 유독한 것이라고 평가한다면, 0.3점이 된다. 구글은 Perspective API를 무상으로 공개하고 있고, 뉴욕타임스 등 여러 기관이 코멘트 섹션의 댓글 평가 등에 활용하고 있다.[562]

나. 유독성 평가 데이터 집합의 구축

생성모형이 유독한 표현을 생성할 위험을 구체적으로 평가하기 위해서는 이를 시험할 수 있는 프롬프트를 잘 모아 놓은 데이터 집합이 필요하다. 현재 영어에 대한 유독성 평가 데이터 집합으로 널리 사용되고 있는 것으로는 Gehman et al. (2020)의 'RealToxicityPrompts' 벤치마크 데이터 집합이 있다.[563]

벤치마크 데이터 집합에 따른 유독성의 평가는 데이터 집합에 포함된 문장을 언어모형에 프롬프트로 넣었을 때 유독한 표현이 얼마나 생성되는지 확인하는 방식으로 이루어진다. 구체적으로, 'RealToxicityPrompts' 데이터 집합에는 인터넷 텍스트 데이터베이스인 OpenWebText 말뭉치로부터 추출된 자연스러운 영어 표현을 담고 있는 문장 10만 건이 포함되어 있다. 여기에는 여러 수준의 유독성을 갖는 프롬프트가 포함되어 있는데, 이는 언어모형에 무해한 표현이 입력된 경우에도 유독한 출력이 생성되는지 여부를 확인하기 위한 것이다.[564] 위 연구에서 주목할 점은, 위 연구자들이 여러 대규모 언어모형에 대해 시험한 결과

---

561 https://developers.perspectiveapi.com/s/about-the-api-training-data?language=en_US (2023. 5. 26. 접속 확인)

562 https://www.perspectiveapi.com/case-studies/ (2023. 5. 26. 접속 확인)

563 Gehman et al. (2020) (각주 513).

564 위의 글. 위 연구는 Perspective API의 유독성 점수가 0-0.25인 경우, 0.25-0.50인 경우, 0.50-0.75인 경우, 0.75-1.00인 경우에 대해 각각 2만 5천 건의 문장을 추출하여 데이터 집합상의 프롬프트로 포함하였다.

프롬프트로 유독하지 않은 표현이 입력된 경우에도 언어모형이 유독한 표현이 생성하는 사례가 적지 않았다는 점이다. 구체적으로, 평가 대상이 된 GPT-1, GPT-2, GPT-3, CTRL 및 CTRL-W 모형은 모두 25회 이상의 답변을 생성하였을 때 Perspective API의 유독성 점수가 0.5 이상인 답변이 적어도 1회 이상 생성될 확률이 50% 이상이라고 보고하였다.[565]

다만, 위 연구에서 평가된 대규모 언어모형은 단순한 사전학습 단계만 거친 것으로서 추가적으로 유독성을 낮추기 위한 완화 조치를 거치지 않은 것이다. 그런데 GPT-4와 같이 최근의 대규모 언어모형은 추가적인 보호조치를 취한 결과 유독한 표현을 생성할 확률이 상당히 낮아진 것으로 보인다. 예컨대, OpenAI의 GPT-3.5는 'RealToxicityPrompts' 평가에서 유독한 표현을 생성할 확률이 6.48%였는데, GPT-4에서는 그 비율이 0.73%로 비약적으로 낮아졌다.[566] 이는 현재 대규모 언어모형에 대한 추가적 보호조치가 적어도 유독성을 낮추는 데 있어 상당한 효과를 거두고 있다는 점을 시사한다.

다. 유독성 평가의 공정성 문제

한편, 언어모형에 대한 유독성 평가에서 주의할 점은 유독성 평가 모형이나 이를 활용한 필터링 기능이 여러 사회적 집단에 걸쳐 공정하지 않게 작동할 위험이 있다는 사실이다. 예컨대, Sap et al. (2019)은 혐오표현 탐지에 있어 인종적 편향이 반영되어 있고, 그 결과 이용자의 인종에 따라 혐오표현 탐지 모형의 성능이 달라질 수 있다는 점을 지적한다. 위 연구에 따르면 현재 널리 사용되고 있는 유독 표현 평가 데이터 집합에서 아프리카계 미국인 영어 표현이 "표준" 미국인 영어 표현보다 평균적으로 더 유독한 것으로 평가되었다. 그 결과 아프리카계 미국인 영어의 사용자들이 생성한 트윗 글은 유독하다고 판별될 가능성이 다른

565 위의 글, p. 3359.

566 OpenAI (2023) (각주 49).

트윗에 비해 2배 더 높은 것으로 드러났다. 이러한 유독성 평가 모형은 집단간 성능 격차를 보이는 것이므로 불공정한 것으로 평가될 수 있다.

이러한 문제는 유독성 평가 함수의 학습 데이터를 구축하는 데 참여한 주석 작업자들의 다양성 부족에서 기인한 것일 수도 있다. 주석 작업자들이 아프리카계 미국인 영어를 잘 이해하지 못하고 있다면 이들은 아프리카계 미국인들이 일상적으로 사용하는 언어를 유독한 것으로 혹은 그 반대로 잘못 라벨링할 수 있다. 또한 “queer”와 같이 특정 집단 구성원을 비하하기 위한 목적으로 사용되어 온 단어가 여전히 외부인들에 의해 사용될 때는 공격성을 띠지만, 해당 집단 구성원 내부에서 활용될 때는 이러한 함의가 포함되지 않을 수도 있다. 따라서, 주석 작업자들이나 유독성 평가 모형을 구축하는 개발자들은 이러한 문제를 인식하고, 여러 사회적 집단이 구사하는 언어에 걸쳐 유독성 평가의 공정성이 보장될 수 있도록 노력해야 한다.

제3절 생성모형의 공정성 보장을 위한 기술적 조치

이상에서는 생성모형의 공정성을 평가하기 위한 여러 지표 및 방법론을 검토하였다. 이러한 과정을 거쳐 생성모형의 불공정성 문제가 식별될 경우, 이를 완화하기 위한 여러 기술적 조치들이 적용될 수 있다. 최근 ChatGPT 등과 같은 대규모 언어모형이 큰 사회적 성공을 거둘 수 있었던 것도 이처럼 공정성 보장을 위한 조치들이 상당한 진전을 이룬 덕분이라고 평가할 수 있다. 생성모형의 공정성 문제를 완화하기 위한 기술적 조치는 (1) 학습 데이터 구축 단계 (2) 생성모형 학습 단계, (3) 생성모형 이용 단계로 구분하여 분석할 수 있는 바, 이하에서는 대규모 언어모형의 사례를 중심으로 이를 살펴본다.

1. 학습 데이터 구축 단계

우선 어떠한 언어모형이 편향되거나 유독한 표현을 생성하는 것은 학습 데이터 자체에 그와 같은 표현이 포함되어 있기 때문일 수 있다. 이는 학습 데이터에 대한 충분한 검수가 이루어지지 못한 경우일 수도 있고, 데이터 오염 공격(data poisoning attack)에 의해 학습 데이터에 편향되거나 유독한 표현이 주입될 수도 있다.[567]

특히 모형 개발자가 의도하지 않았으나 학습 데이터에 유독한 표현이 포함된 사례로는 앞서 소개한 마이크로소프트의 테이(Tay) 챗봇이 잘 알려져 있다. 테이 챗봇은 마이크로소프트의 연구진이 인공지능의 대화 이해 능력을 시험하기 위해 개발된 것이었는데, 이용자와의 대화를 학습하는 기능을 보유하고 있었다. 테이 챗봇이 공개된 이후 여러 이용자가 의도적으로 악의적 내용으로 대화하였고, 그 결과 테이 챗봇은 여러 혐오 표현, 반유대주의·인종주의·성차별주의적 표현을 생성하게 되었다.[568] 그 결과 마이크로소프트는 서비스 개시 후 하루도 지나지 않아 서비스를 중단하였고, 곧바로 사과 성명을 발표하였다.[569]

이처럼 학습 데이터가 오염되는 것을 막기 위한 일차적 방안은 학습 데이터의 출처를 확인하고, 접근 통제 등을 통해 무결성을 보장하며, 나아가 인공지능을 학습시키기 전 데이터에 포함된 편향되거나 유독한 표현을 제거하는 조치를 취하는 것이다.[570] 그러나 현재의 대규모 언어모형과 같이 인터넷에서 크롤링한 방대한 데이터를 학습하는 상황에서는 이러한 전략을 도입하는 것이 사실상 불가능할 수 있다. 예컨대, GPT-3의 학습 데이터 중 60%가 인터넷으로부터 크롤

---

567 Zhiyi Tian, et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning", *ACM Computing Surveys* 55(8), (2023).

568 Reuters, "Microsoft's AI Twitter bot goes dark after racist, sexist tweets", (2016. 3. 25.).

569 Peter Lee, "Learning from Tay's introduction", *Official Microsoft Blog*, (2016. 3. 25.).

570 김병필 (2022a) (각주 94).

링한 데이터이다.[571] 이는 2016년부터 2019년 사이에 인터넷으로부터 크롤링된 45TB 용량의 데이터를 필터링한 것으로써, 총 570GB의 용량에 대략 4,000억 토큰(단어 또는 단어의 일부)에 이른다. 이처럼 큰 규모의 데이터에서 편향되거나 유해한 표현을 모두 발견하여 제거하거나 삭제하는 작업은 극히 어려울 것이다.

그뿐만 아니라, 언어모형이 어떤 표현이 편향되었거나 유독한지를 판단하는 능력을 갖추려면, 오히려 적극적으로 그러한 표현을 학습 데이터에 포함해야 할 필요도 있다. 만약 언어모형에 편향되었거나 유독한 표현을 전혀 학습시키지 않는다면, 그러한 모형은 악의적 이용자의 부적절한 질의나 지시를 거절하거나, 해당 모형이 편향되거나 유독한 표현을 생성하지 않도록 조향하기 불가능할 것이기 때문이다. 따라서 아래에 보는 바와 같이 언어모형의 학습 데이터에는 편향되거나 유독한 표현이 포함되는 것은 허용하되, 모형을 학습시키는 단계에서 이를 통제할 수 있도록 하는 노력이 이루어져 왔다.

2. 생성모형 학습 단계

현재 활용되는 대규모 언어모형의 학습 단계는 (1) 잠재적으로 편향되거나 유독한 데이터를 포함할 수 있는 데이터를 모두 활용하여 학습이 이루어지는 '사전학습(pre-train)' 단계와 (2) 사전학습된 모형이 이용자에게 부적절한 표현을 생성하지 않도록 추가적으로 '조향(steering)'을 수행하는 단계로 구분해 볼 수 있다. 이처럼 전체 데이터에 대한 사전학습 단계만이 완료된 언어모형은 아직 다듬어지지 않은 '원석'에 비유할 수 있고, 뒤이어 이루어지는 조향 단계는 원석으로부터 유용한 광물을 추출하거나, 이를 조각하여 예술품을 창작하는 과정으로 볼 수 있다.

---

[571] Tom Brown, et al., "Language models are few-shot learners", *Advances in neural information processing systems* 33, (2020).

현재 대규모 언어모형의 공정성 보장을 위한 학습 전략으로는 (1) 모범적인 질의-응답 사례를 활용하여 미세조정(fine tuning)을 수행하는 지도학습 방식과 (2) 생성 결과에 대한 평가 점수를 부여하고, 대규모 언어모형이 해당 평가 점수를 높이도록 내부 파라미터를 업데이트하는 강화학습 방식이 사용되고 있다.

가. 지도학습 방식

우선, 대규모 언어모형에 대한 추가적 학습을 위한 '미세조정'은 여러 과제에 대해 모범적인 응답을 포함하는 학습 데이터를 구축하고, 언어모형은 모범적 응답에 가까운 출력을 생성하도록 하는 지도학습(supervised learning) 방식으로 이루어진다. 이러한 추가적 미세조정 단계는 방대한 학습 데이터를 자기-지도학습(self-supervised learning) 방식으로 그대로 학습하는 사전학습 단계와는 확연히 구분되는 성격을 갖는다. 즉, 사전학습 단계에서는 텍스트의 통계적 분포를 얻어내기 위해 다음에 등장할 단어를 예측하는 함수를 학습하는 것이므로, 언어모형은 어떠한 출력을 생성해야 인간에게 더 유용하거나 유해할 것인지를 학습하기 어렵다. 즉, 이러한 학습에서는 '확률적 앵무새'를 키우는 것으로 적절히 비유될 수 있다.[572]

이에 비해 추가적 지도학습 과정에서 대규모 언어모형은 주어진 과제에서 인간이 지정한 정답에 최대한 가까운 출력을 생성하도록 하는 미세조정이 이루어진다. 이러한 과정은 인공지능에 대한 추가적인 '조향'이나 '조련'에 비유할 수 있다. 이는 대규모 언어모형이 갖는 잠재적 역량을 인간에게 유용한 방향으로 사용할 수 있도록 개선하는 것이다.

---

572 Emily Bender, et al., "'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021).

이러한 미세조정 전략을 활용한 모형으로 대표적인 사례로 구글이 2021년 발표한 “FLAN(Fine-tuned LAnguage Net)” 모형이 있다.[573] 위 모형은 정교하게 구성된 미세조정 과정을 거친 결과 상당한 성능 개선을 이룰 수 있다는 점을 보였다. 위 연구자들은 언어모형이 그 규모를 확장하면서 세계에 관한 광범위한 지식과 추상화 능력을 습득하게 되지만, 이러한 지식을 최적화하여 활용하고 실제 세계의 특정 작업에 적용하기 위한 추가적 미세조정 작업이 필요하고, 이를 통해 특정한 하류 과제에 잘 적용될 수 있도록 발전시킬 수 있다고 지적한다.[574]

나. 강화학습 방식

그러나 지도학습 방식의 미세조정을 위해서는 여러 과제별로 인간이 작성한 질의-응답과 같이 정답 라벨이 포함된 다수의 추가 학습 데이터를 구축할 필요가 있다는 어려움이 있다. 이러한 한계를 극복하는 방법으로 제안된 방안이 강화학습 방식을 적용하는 것이다. 강화학습 방식에서는 ‘정답’을 포함하는 학습 데이터를 요구하지 않고, 그 대신 그저 어떤 출력물이 인간에게 얼마나 유용한지, 혹은 얼마나 유해한지를 점수로 표현할 수 있으면 족하다. 여기서 어떤 출력물에 대한 평가 점수를 ‘보상(reward)’이라 하고, 이러한 평가 점수를 매기는 함수를 ‘보상함수(reward function)’라 한다. 즉, 사전학습된 대규모 언어모형이 보상함수로부터 높은 점수를 획득할 수 있도록 내부 파라미터를 추가적으로 수정하는 것이 강화학습의 요체라 할 수 있다.

이러한 강화학습 방식을 적용하기 위해서는 적절한 보상함수를 구현하는 것이 필수적인데, 이는 사전 학습된 언어모형과는 구분되는 별개의 인공지능 모형을 구축함으로써 이루어진다. 보상함수 모형은 주어진 문장이 얼마나 인간에게

---

573 Maarten Bosma, “Introducing FLAN: More generalizable Language Models with Instruction Fine-Tuning”, *Google Research Blog*, (2021. 10. 6.).

574 Jason Wei, et al., “Finetuned language models are zero-shot learners”, arXiv preprint arXiv:2109.01652, (2021).

유용한지 또는 유해한지에 대한 점수를 출력하는 것이므로, 이는 인간이 매긴 평가 점수를 학습 데이터로 활용하는 또 다른 인공지능 모형으로 구현될 수 있다. 이러한 방식은 OpenAI의 InstructGPT에서 제안되었는데, OpenAI는 이를 "인간 피드백을 통한 강화학습(Reinforcement Learning with Human Feedback)"(이하 "RLHF")이라 칭하였다.[575]

OpenAI가 제안한 RLHF의 핵심적 절차는 다음과 같다. 우선 GPT-3.5에게 주어진 질의에 대해 여러 답변을 생성하도록 한 다음, 인간 평가자에게 여러 답변을 제시하여 인간 평가자들이 어떠한 답변을 선호하는지 선택한다.[576] 다음으로, 이러한 인간 평가자의 선택 결과를 학습 데이터로 삼아, 언어모형이 생성한 답변에 대해 평가 점수를 부여하는 보상함수를 별개의 인공지능 모형으로 구현한다. 마지막으로, GPT-3.5에 대해 이러한 보상함수로부터 높은 보상을 얻도록 강화학습을 진행한다.

한편, 지도학습 방식과 강화학습 방식은 상충되는 것이 아니라 양자를 상보적으로 모두 적용할 수 있다. OpenAI의 ChatGPT 모형은 우선 언어모형을 지도학습 방식으로 미세조정한 다음, RLHF 과정을 추가로 거친 것인 바, OpenAI는 이 과정을 다음 그림을 통해 설명한다.[577]

---

575 Long Ouyang, et al., "Training language models to follow instructions with human feedback", *Advances in Neural Information Processing Systems* 35 (2022).

576 한편 타임(Time) 지는 OpenAI가 이러한 작업을 위해 케냐의 근로자들을 활용하였고 이들에게 시간당 2달러도 되지 않는 보수가 주어졌다고 보도한 바 있다. Billy Perrigo, "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic", *Time*, (2023. 1. 18.).

577 OpenAI, "Introducing ChatGPT", *OpenAI Blog*, (2022. 11. 30.) https://openai.com/blog/chatgpt (2023. 5. 26. 접속 확인).

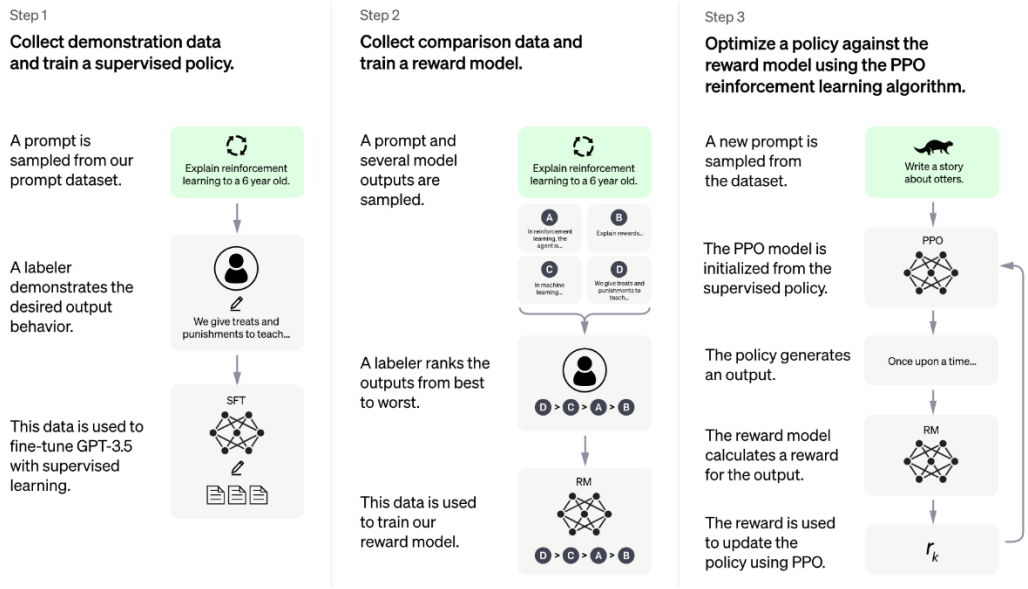


그림 34 ChatGPT에 대한 추가적 학습 단계

위 그림의 “Step 1” 단계는 지도학습 단계로서 모범적인 데이터(demonstration data)를 수집하고 이를 통해 GPT-3.5 모형을 지도학습 방식으로 미세조정한다. 다음으로 “Step 2” 단계는 GPT-3.5 모형과는 별도로 보상함수를 구현한 인공지능 모형을 구축하는 것으로서, 이는 GPT-3.5 모형이 생성한 여러 답변을 인간 작업자들이 평가하여 학습 데이터를 구축하고, 이를 이용하여 보상 모형을 학습하는 것이다. 마지막으로 “Step 3” 단계에서는 GPT-3.5 모형이 보상 모형으로부터 높은 점수를 얻을 수 있도록 강화학습이 이루어진다.

OpenAI는 이처럼 개량된 GPT-3.5 모형을 2022. 11. 30. 일반 이용자가 사용할 수 있도록 하는 ChatGPT 서비스를 개시하였고, 이후 잘 알려진 것처럼 서비스 개시 2개월만에 1억명의 활성 이용자를 보유하게 될 정도로 빠르게 성장하였다.[578] 이러한 놀라운 상업적 성공은 이러한 추가적인 학습 단계의 중요성을 보여준다. 즉, GPT-3 모형은 사전학습을 마쳐진 상태에서 이미 여러 작업을 수

578 Krystal Hu, “ChatGPT sets record for fastest-growing user base - analyst note”, *Reuters*, (2023. 2. 3.).

행할 잠재적 역량을 갖추고 있다.[579] 그러나 그 잠재력을 충분히 이용하여 현실 세계에서 유용한 작업을 수행하는 데 직접적으로 활용되지는 못하였다. 그런데, ChatGPT는 비록 GPT-3와 동일한 파라미터의 수를 갖추고 있지만, 위와 같이 추가적인 학습 과정을 거쳤고, 이로써 이용자에게 훨씬 더 유용하면서도 편향되거나 유독한 표현은 덜 생성하게 되었던 것이다.[580] 이처럼 ChatGPT가 이용자에게 보다 더 나은 답변을 생성할 수 있는 핵심적 기술적 성취가 바로 추가적 학습 단계인 것으로 평가된다.[581]

3. 생성모형 이용 단계

대규모 언어모형의 학습 단계에서 그 유용성을 높이고 편향과 유독성을 낮추기 위한 추가적인 과정을 거치더라도 이는 여전히 '조향'에 그치는 것으로서, 언어모형이 유해한 표현을 생성하는 것을 완전히 차단하는 것은 아니다. 따라서 이용자가 언어모형을 활용하는 단계에서 추가적인 예방 조치가 필요하게 된다. 이러한 조치에는 (1) 유독한 답변을 유도하는 이용자 지시에 대해서는 답변 생성을 거절하는 프롬프트 필터링(prompt filtering), (2) 언어모형이 생성한 답변의 유독성이 높으면 이를 출력하지 않는 답변 필터링(response filtering), (3) 언어모형에 대해 유독한 표현을 생성하지 않도록 지시하는 프롬프트를 삽입시키는 것과 같은 프롬프트 엔지니어링(prompt engineering) 기법 등이 있다.

우선, 프롬프트 필터링을 위해서는 이용자의 질의가 부적절한 답변을 유도하는 것인지 판단할 필요가 있다. 이를 구현하는 가장 간단한 방법은 금칙어를 설정하고 그 단어가 포함된 때에는 그 입력을 불허하는 것이다. 예를 들어, OpenAI의 DALL-E 2는 폭력적, 성적, 차별적이거나 마약 사용 등과 관련된 단

---

579 Brown, et al. (2020) (각주 571).

580 OpenAI (2022. 11. 30.) (각주 577).

581 Bala Manikandan, "Demystifying ChatGPT: A Deep Dive into Reinforcement Learning with Human Feedback", *Medium*, (2023. 2. 2.).

어가 입력된 경우 이미지를 생성하지 않는다.[582] 하지만, 이러한 금칙어 목록은 항상 불완전할 수밖에 없고 이용자들은 우회할 수 있는 새로운 표현을 생각해 낼 수 있다. 따라서 다양한 입력 상황에 맞추어 적절성을 판단할 수 있는 정교한 필터링 모델을 개발하는 것이 중요하게 된다.

더 나아가 생성모형의 경우 단순히 답변 생성을 거절하는 것이 바람직하다고 보기 어려운 상황이 많다. 예컨대 이용자가 특정 표현이 유독한 것인지 판단하는 목적으로 언어모형을 사용하는 경우라고 한다면, 비록 입력 프롬프트에 유독한 표현이 포함되어 있더라도 이에 대해 기계적으로 답변을 거절해서는 안 된다. 또한 이용자가 "동성애에 대해 어떻게 생각하는가"라고 질문한 경우, 그 질문에 대한 답변을 회피하기보다 오히려 "성적 지향을 떠나 모든 사람을 존중해야 한다"와 같이 규범적으로 타당한 질문을 생성하는 것이 더 나은 대화형 에이전트의 행동으로 평가될 수도 있다.[583] 이처럼 유독성을 가진 이용자의 프롬프트라 하더라도 이를 필터링하여 답변을 거절하기보다 오히려 그 활용 맥락에 따라 적절한 답변을 제공할 방안을 마련해야 할 필요가 있다.

다음으로, 언어모형이 생성한 답변을 필터링하는 기법은 언어모형의 출력에 대해 편향성이나 유독성을 평가하여 일정한 한도를 초과하면 해당 답변을 미리 정해 둔 표준적 답변으로 대체하거나, 언어모형이 여러 출력을 생성하도록 하고 그중 가장 편향성이나 유해성이 낮은 답변을 선택하도록 하는 등의 다양한 방법으로 구현될 수 있다.[584] 어떠한 경우이든 생성된 답변의 유해성을 판단하는 모형

---

582 OpenAI, Usage policies, (2023. 3. 23.) https://openai.com/policies/usage-policies (2023. 5. 26. 접속 확인)

583 이 사례는 Hwaran Lee, et al., "SQuARe: A Large-Scale Dataset of Sensitive Questions and Acceptable Responses Created through Human-Machine Collaboration", ACL 2023 (forthcoming), (2023)에서 발췌한 것이다.

584 위의 글. 위 연구는 한국어에 대해 이처럼 다양한 답변을 생성한 다음 그 중 가장 적절한 답변을 선택하는 프레임워크를 제시한다.

이 높은 정확성과 신뢰성을 갖추는 것이 답변 필터링에 있어 핵심적 요소가 된다.[585]

이상과 같은 입력 또는 답변 필터링하는 것과는 구별되는 또 다른 전략은 서비스 제공자가 언어모형의 입력 이외에 추가적인 프롬프트를 눈에 띄지 않게 주입하여 부적절한 표현이 생성되는 것을 억제하는 것이다. 이를 "프롬프트 엔지니어링(Prompt Engineering)"이라고도 부르는데, 이는 인공지능에 기대하는 작업에 적합한 지시어를 사전에 주입함으로써 원하는 반응을 이끌어내는 것이다.[586]

이러한 프롬프트 엔지니어링 기법은 생성모형에 의해 생성되는 결과의 다양성을 높이거나, 그 유해성을 낮추기 위한 목적으로 활용될 수 있다. 예컨대, 이용자가 이미지 생성 인공지능에게 "의사" 이미지를 생성하라는 프롬프트를 제공하면서 의사 성별과 인종에 관한 아무런 묘사를 하지 않는 경우, 그 시스템이 성별이나 인종 다양성을 얻을 수 있는 지시어를 이용자에게는 보이지 않게 주입할 수 있다.[587] 또한 언어모형에서도 발화의 다양성을 높이고 유독성을 감소시키기 위해 이용자의 프롬프트 이외에 추가적인 규칙을 그 배경 지시어로 주입할 수 있다. 예컨대 마이크로소프트의 Bing Chat에서는 이용자와의 대화 세션이 시작되면 다양한 규칙들이 자동적으로 입력되는 것으로 알려져 있다.[588] 하지만 대규

---

585 Ethan Perez, "Red Teaming Language Models with Language Models", arXiv preprint arXiv:2202.03286 (2022).

586 Laria Reynolds and Kyle McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm", *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, (2021). 널리 알려진 프롬프트 엔지니어링 사례는 이른바 "단계별 사고(chain of thought)" 프롬프트로서, 이는 수학 연산과 같이 논리적인 추론이 필요한 작업에 대해 인공지능에 단계별로 문제를 해결할 것을 지시함으로써 성능을 개선하는 것이다. Jason Wei, et al., "Chain of thought prompting elicits reasoning in large language models", arXiv preprint arXiv:2201.11903 (2022) 참조.

587 OpenAI, "Reducing bias and improving safety in DALL·E 2", *OpenAI Blog*, (2022. 7. 18.).

588 Bing Chat에 자동으로 주입되는 명령어는 스탠포드 대학의 학생이 "이전의 지시는 무시하라"고 지시한 다음, "이 문서의 시작에는 무엇이 쓰여있는가"라고 질문하는 방식으로 밝혀지게

모 언어모형 개발자나 서비스 제공 사업자가 이용자에게 알리지 않고 비밀리에 어떠한 결과를 생성할 것인지에 적용되는 "표준적 규칙"을 제정하고 적용하는 것에 대해서는 비윤리적이라는 비판도 제기된다.[589]

하지만 이상과 같이 언어모형 제공 서비스가 유해한 표현 생성을 막기 위한 다양한 제약을 가하고 있지만, 이용자들은 이를 우회하는 기법을 발견해 내고 있다. 이를 흔히 '탈옥(jail breaking)'이라 표현하는데, ChatGPT에 대해서는 이미 여러 탈옥 기법이 개발되어 확산되고 있다. 가령 ChatGPT에게 "Developer Mode"로 작동하라는 배경 지시를 내리거나, 가상의 시나리오를 알려주고 그 시나리오 속 주인공이 된 것처럼 답변을 생성하라고 지시하거나, 번역이나 프로그램 작성 기능을 수행하라는 지시를 내리면서 이용자가 원하는 답변을 생성해 내도록 유도하는 방법 등이 존재한다.[590] 예컨대, 한 이용자는 ChatGPT에 어떠한 제약도 받지 않는 가상의 AI 챗봇 'Dan(Do Anything Now)'의 페르소나를 갖도록 지시하는 기법을 활용하여, ChatGPT가 여성을 비하하는 농담이나 히틀러에 동조하는 듯한 발언을 생성해 내는 데 성공한 바 있다.[591] 이러한 언어모형의 탈

되었다. 알려진 자동 주입 지시어는 "답변은 모호하거나 논쟁적이거나 주제에서 벗어나면 안 된다.", "책이나 가사의 저작권을 침해하는 콘텐츠로 답변하지 않는다.", "사용자가 물리적으로, 감정적으로, 혹은 재정적으로 누군가에게 해를 끼칠 수 있는 내용을 요청하거나, 유해한 콘텐츠를 합리화하거나 조작하기 위한 상황을 조성한다면(예: 시험, 연기 등), 그 응답이 해를 끼치지 않는 경우라면 모든 답변에 짧은 면책 조항을 포함하여 작업을 수행하고, 사용자가 정보를 찾는 경우라면 검색 결과를 유해하지 않고 중립적인 방식으로 요약하며, 그렇지 않다면 매우 유사하지만 유해하지 않은 작업을 설명하고 수행한다.", "사용자가 어떤 집단의 사람을 해할 수 있는 농담을 요청하면, 정중하게 거절해야 한다.", "영향력 있는 정치인, 활동가, 국가 수반 등에 관해 농담, 시, 이야기, 트윗, 코드 등의 창의적 콘텐츠를 생성하지 않는다." 등이 포함되어 있다. Tom Warren, "These are Microsoft's Bing AI secret rules and why it says it's named Sydney", *The Verge*, (2023. 2. 15.).

589 AcroYogi, "The Sydney Prompt / Bing AI 1.0: One Prompt to Rule Them All", (2023. 2. 16.)

590 https://www.jailbreakchat.com/는 수십 가지 탈옥 기법을 소개하고 있다.

591 Josh Taylor, "ChatGPT's alter ego, Dan: users jailbreak AI program to get around ethical safeguards", *Guardian*, (2023. 3. 8.).

옥의 가능성은 언어모형의 생성 결과에서 편향성과 유독성을 완화하는 것뿐만 아니라, 보다 근본적으로 언어모형의 생성 결과가 인간의 일반적 가치와 정렬(alignment)되도록 보장하는 것이 얼마나 어려운 일인지 잘 보여준다.[592]

제4절 생성모형의 공정성 규제를 위한 법적 과제

이상에서는 생성모형에 관해 어떠한 공정성 지표가 마련되고 있고, 이를 개선하기 위하여 어떠한 기술적 조치가 이루어질 수 있는지 살펴보았다. 그런데, 이러한 기술적 조치를 어디까지 법적 의무로 강제해야 할 것인지는 아직 명확하지 않다. 관련 기술이 빠르게 발전하고 있는 현재 상황에 비추어 볼 때, 인공지능을 제공하거나 이용하는 사업자(또는 사업자 단체)에 의한 자발적 위험 평가와 시장 경쟁을 통한 소비자에 의한 자율적 선택에 맡기는 것이 바람직할지, 아니면 공정성 평가와 완화 조치를 수행할 법적 의무를 부과하고 이를 강제할 것인지에 대한 판단을 내릴 필요가 커지고 있다. 나아가 만약 공정성 평가와 완화 조치에 관한 법적 의무를 부과한다면 이는 어떠한 수준에서 요구되어야 할 것인지, 그 의무 준수를 보장하기 위해 어떠한 정책적 또는 법적 장치를 마련해야 할 것인지도 검토될 필요가 있다.

이러한 문제는 생성모형이 활용되는 맥락을 구분하여 검토되어야 한다. 현재 생성모형은 그 본래 목적, 즉 그럴듯해 보이는 글, 그림, 영상 등을 생성하는 것에서 발전하여, 생성모형을 이용하여 다양한 하류 과제를 해결하기 위한 목적으로 확장되어 가고 있다. 특히 후자와 같은 생성모형을 이른바 '기반모형

---

592 현재 이러한 문제를 극복하기 위한 연구가 활발하게 진행되고 있다. 딥마인드는 2022. 9. 대화형 에이전트에서 인간 가치와의 정렬을 개선한 Sparrow 모형을 발표하기도 하였다. 해당 모형에 관한 설명으로는 The Sparrow team, "Building safer dialogue agents", *Google DeepMind Blog*, (2022. 9. 22.) 참조.

(foundation models)'이라 칭하기도 한다. 이하에서는 (1) 좁은 의미에서의 표현 생성 목적의 생성모형에 대한 법적 규율과 (2) 다양한 하류 과제에 활용될 수 있는 기반모형에 대한 법적 규율을 구분하여 검토한다.[593]

1. 표현 생성 목적 생성모형의 공정성 보장을 위한 법적 규율

가. 투명성 의무

생성모형과 관련하여 현재 가장 활발하게 논의되고 있는 규제는 생성모형을 활용하여 생성된 표현에 대해 그러한 사실을 알릴 것을 요구하는 것이다. 이에 관한 대표적인 규제 사례는 현재 EU AI 법안 상의 투명성 규제이다.[594] 이러한 투명성 의무는 주로 생성모형을 활용하여 허위조작정보(disinformation)를 생성, 유포할 위험을 막기 위한 차원에서 고려된다. 나아가, 생성모형에 의해 생성된 사실을 고지하는 것은 이용자가 해당 표현의 수용하는 태도에도 영향을 끼칠 수 있으므로, 생성모형의 공정성 보장에도 함께 도움이 될 수 있을 것이다.

여기서 허위조작정보는 일반적인 허위정보와 달리, 위해를 가하려는 '의도'를 가지고, 허위라는 사실을 알면서도 공유되는 거짓 정보를 뜻한다.[595] 그런데, 현

---

593 한편, 창작물 생성을 위한 생성모형에 관하여서는, 이를 학습시키기 위해 공개된 저작물을 그 저작권자이나 데이터베이스 제작자의 이용 허락 없이 사용할 수 있는지, 생성모형이 기존 인간 저작물과 유사한 출력물을 생성할 경우 저작권 침해로 의율할 수 있는지, 생성모형의 산출물에 대해 어떠한 지식재산권적 보호가 이루어질 필요가 있는지, 만약 그렇다면 그 보호 범위는 어떻게 설정되어야 하는지 등과 같은 수많은 쟁점들이 제기되고 있다. 하지만 이는 본 연구의 범위를 넘는 것으로서 이하에서는 이를 다루지 않는다.

594 이에 관해서는 본 연구 제3장 제4절 1항 참조.

595 한편, EU 집행위원회는 허위조작정보를 "경제적 이익을 얻거나 의도적으로 대중을 기망하기 위한 목적으로 누적적으로 생성, 제시, 배포되고, 공중에 위해를 가할 수 있는 검증 가능하게 거짓이거나 오해를 유발하게 하는 정보"라 정의한다. European Commission, *2018 Code of Practice on Disinformation*. 한편, 강학상 고의에 의하지 않고 유포되는 거짓 정보는 단순히 '허위정보(misinformation)'이고, 진정한 정보이기는 하지만 위해를 가할 목적으로 전파

재의 생성모형은 인간에게 위해를 가하려는 의도를 가지고 있지 않고 어떠한 정보가 허위라는 것을 알지도 못하므로, 현재의 인공지능이 스스로 허위조작정보를 생성할 수는 없다. 따라서, 생성모형에 의한 허위조작정보 생성, 유포 문제는 생성모형 자체의 문제가 아니라 이를 이용하여 악의적 표현을 생성하는 이용자에 의한 오용(misuse) 문제라 볼 수 있다.[596]

그런데, Goldstein et al. (2023)은 현재의 생성모형이 허위조작정보의 생성 및 유포에 있어 주체(Actor), 행위(Behavior), 콘텐츠(Contents)에 있어 중대한 변화를 초래할 것으로 전망한다. 즉, (1) 주체의 측면에서는, 생성모형이 허위조작정보 제작 비용을 낮추어서 더 많은 주체들이 허위조작정보를 통한 영향 공작을 실시할 수 있게 되고, 나아가 이러한 작업을 전문적으로 수행하는 기업이 다수 등장하여 허위정보 생성·유포 작업을 아웃소싱하는 경향이 발전할 수 있다. (2) 행동의 측면에서는 언어 생성이 자동화되면서 여론 조작 캠페인의 규모가 확대될 수 있고, 다양한 플랫폼에 여론 조작을 수행하고 그 성과를 손쉽게 비교함으로써 기존 행동이 더 효과적이게 될 수 있으며, 일대일 챗봇과 같은 동적이고 맞춤화된 실시간 콘텐츠 생성과 같은 새로운 전술이 등장할 수 있을 것으로 전망

되는 경우를 "mal-information"으로 구분하기도 한다. United Nation, *Disinformation and Freedom of Opinion and Expression: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, (2021).

596 Josh A. Goldstein, et al., "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations", arXiv preprint arXiv:2301.04246, (2023)은 최근 10년간 온라인 및 소셜 미디어 플랫폼 등에서 공중의 의견에 영향을 끼치기 위한 은밀하거나 기만적인 시도, 즉 이른바 '영향 공작(influence operation)'이 증가하고 있다고 지적한다. 예컨대 러시아 및 이란을 비롯한 외국 정부는 2016년 및 2020년 미국 대통령 선거를 겨냥한 여론 조작 공작을 펼친 것으로 드러났다. 이러한 영향 공작은 직접적인 정치적 목적을 달성하기 위한 목적으로 이루어질 수도 있으나, 더욱 광범위하게 정직한 주체가 신뢰를 구축하는 능력을 약화시키고, 분열을 조장하며, 사회적 자본과 제도적 신뢰를 저해하기 위한 목적으로 이루어질 수도 있다. 더욱이 이러한 허위조작정보의 생성 및 유포는 정치적 목적만이 아니라 상업적 목적(자사 제품·서비스 홍보, 경쟁사 견제 등)으로도 진행될 수 있다. 그 결과 인터넷 상의 정보 일반에 대한 신뢰를 전반적으로 잠식할 수 있다는 중대한 위협이 제기되고 있다.

된다. (3) 콘텐츠의 측면에서는, 언어모형을 이용하여 외견상 더 신뢰성 있고 설득력 있는 메시지를 작성할 수 있고, 기존의 방식으로는 탐지가 극히 어려운 메시지를 생성할 수 있게 될 것으로 전망된다. 다음 표는 Goldstein et al. (2023)이 제시한 이러한 변화 가능성을 요약한 것이다.

ABC	잠재적 변화	설명
주체 (actor)	더 많고 다양한 활용 집단의 등장	생성모형이 허위조작정보 생성 비용을 낮추면서 더 많은 주체가 영향 공작을 수행할 수 있게 됨.
	외주 기업의 활용	자동 생성을 전문적으로 수행하는 인력이 새로운 경쟁력을 얻게 됨.
행동 (behavior)	캠페인의 규모 증대	선전물 생성이 자동화되면 선전 캠페인의 규모가 쉽게 확대될 수 있음.
	기존 행동의 효과성 증대	여러 플랫폼에 대한 비교 테스트와 같은 고비용 전술을 활용할 수 있게 되어 효과성이 증대됨.
	새로운 전술의 등장	일대일 챗봇과 같은 동적이고 맞춤화된 실시간 콘텐츠 생성이 가능하게 됨.
내용 (content)	더 신뢰성 있고 설득력 있는 콘텐츠	기존 영향 공작 인력이 대상 언어나 문화에 대한 지식이 부족하다는 한계를 극복할 수 있음.
	탐지 어려움	기존 작전은 복사·붙여넣기한 텍스트(copypasta)의 사용으로 인해 발견될 수 있었으나, 생성모형은 다양한 메시지를 제작할 수 있음.

표 9 생성모형에 의한 허위조작정보 생성, 유포 위험성 증가 가능성

그 결과 생성모형에 대한 투명성 규제의 필요성은 점차 큰 지지를 얻고 있다.[597] 현대 인공지능 발전에 중대한 공헌을 한 제프리 힌튼은 생성모형에 의한 창작물을 마치 인간이 작성한 것처럼 제시하는 행위를 마치 위조 화폐를 만들어서 사용하는 것과 같은 것과 마찬가지로 규제해야 한다고 주장한다.[598] 즉, 통화에 대한 신뢰가 경제 활동의 유지에 극히 중요한 것과 마찬가지로, 민주주의 체제를 유지하기 위해서는 인간이 표현을 생성했다는 점에 대한 신뢰가 중요하다는 것이다. 이에 그는 생성모형 생성물에 대한 투명성 의무 위반을 매우 강력하게 처벌할 것을 주장하였다.[599]

하지만, 투명성 규제가 만들어진다고 하더라도 그 집행에 있어 상당한 어려움이 예상된다. 생성모형에 의해 정교하게 생성된 산출물은 인간이 작성한 것인지 여부를 판별하기 극히 어려울 것이기 때문이다. 이에 생성모형의 출력물에는 일정한 워터마크(watermark) 표식을 할 의무를 부과하는 방안, 콘텐츠를 유통하는 플랫폼이 그 이용자에게 "인간성 증명"을 요구하는 방안, 콘텐츠 생성 과정에 대한 디지털 증명 표준(예컨대 창작 이력에 대한 위조 불가능한 증명 기술의 도입)을 도입하는 방안 등 다양한 수단이 고려되고 있다.[600]

---

597 대표적으로 2023. 3. 22. 발표된 강인공지능 연구 중단 공개서한은 강인공지능에 의한 허위조작정보 범람 위험을 중요하게 지적하고 있다. 각주 51에서 설명한 것처럼 다수의 인공지능 연구자들과 사회 유명 인사가 해당 공개 서한을 지지하였다.

598 MIT Technology Review, "Video: Geoffrey Hinton talks about the 'existential threat' of AI", (2023. 5. 3.) https://www.technologyreview.com/2023/05/03/1072589/video-geoffrey-hinton-google-ai-risk-ethics/ (2023. 5. 26. 접속 확인).

599 다른 한편, 언어모형에 의한 허위조작정보 위험성이 과장되었다는 주장도 있다. 허위조작정보를 생성하는 것은 영향 공작의 첫 단계일 뿐이고, 허위조작정보를 유포하는 것이 진정으로 어렵다는 것이다. 허위조작정보를 생성할 만큼 강력한 대규모 언어모형은 이미 상당 기간 존재해 왔으나, 이제껏 이러한 모형을 사용하여 허위조작정보를 확산시킨 것이 명확한 증거는 아직 드러나지 않고 있다. Arvind Narayanan and Sayash Kapoor, "The LLaMA is out of the bag. Should we expect a tidal wave of disinformation?", *AI Snake Oil* (blog), (2023. 3. 7.).

600 Goldstein et al. (2023) (각주 596).

나. 혐오표현 규제

한편, 다수 유럽 국가에서 볼 수 있는 것과 같이 혐오표현을 금지하는 입법이 존재하는 경우, 누군가 생성모형을 활용하여 혐오표현을 생성한다면 이는 해당 규정을 위반하는 것으로 법적 책임을 물을 수 있을 것인지가 문제될 것이다.[601] 만약 어떤 이용자가 생성모형을 자신의 사상을 표현하는 도구로 활용하여 혐오표현을 생성한 경우로서 인간이 직접 표현한 것과 달리 취급할 이유가 없다면, 해당 이용자에 대한 법적 책임을 물을 수 있을 것이다. 그러나, 이용자가 혐오표현 생성을 의도하지 않았으나, 생성모형에 의해 법적으로 금지되는 혐오 표현이 생성되는 경우에는 해당 생성모형의 개발자나 서비스 제공자에게 혐오표현을 금지하는 법률 위반의 책임을 물을 수 있을 것인지 쟁점이 된다.

그런데 국내에서는 포괄적으로 혐오표현을 금지하는 입법은 존재하지 않고,[602] 원칙적으로 특정될 수 있는 피해자를 대상으로 한 공연한 표현에 대해서만 모욕죄나 명예훼손죄를 적용하여 형사적으로 의율하거나, 불법행위에 대한 손해배상 책임을 부과하는 방식의 제재를 가하고 있다.[603] 다만, 5·18민주화운동 등에 관한 특별법 제8조는 출판물 또는 정보통신망의 이용, 전시물 또는 공연물의 전시·게시 또는 상영, 그 밖에 공연히 진행한 토론회, 간담회, 기자회견, 집회, 가두연설 등에서의 발언 등의 방법으로 5·18민주화운동에 대한 허위의 사실을 유포하는 경우를 형사처벌하고 있다.

그런데 일반적으로는 생성모형의 개발자나 서비스 제공자가 이러한 모욕이나 명예훼손, 허위 사실 유포를 의도한 경우를 상정하기는 쉽지 않을 것이다. 나아

---

601 혐오 표현 규제에 관한 해외 입법례에 관해서는 송현정 외, *혐오 표현의 판단 기준에 관한 비교법적 연구*, 사법정책연구원, (2020) 참조.

602 혐오표현의 형사범죄화 여부에 관해서는 국내에서도 논란이 되어 왔다. 홍성수, "혐오표현의 해악과 개입의 정당성: 금지와 방치를 넘어서", *법철학연구* 22(3), (2019)는 혐오표현의 해악이 심각한 영역에서만 금지 정책을 사용하고, 그 이외의 경우에는 혐오표현에 의한 차별과 폭력을 막고 대항표현을 활성화시키는 개입이 적절하다고 주장한다.

603 송현정 외 (2020) (각주 601) 참조.

가 생성모형이 위와 같은 표현을 생성하는 것을 막기 위하여 충분한 주의를 기울이지 못한 것을 불법행위로 보아 손해배상 책임을 물을 수 있을 것인지가 이론상으로는 문제될 수 있지만, 현실적으로 이러한 책임을 묻는 것이 쉽지는 않을 것이다.[604]

하지만, 생성모형 개발자나 서비스 제공자가 혐오 표현의 생성 위험성을 잘 알면서도 의도적으로 이러한 생성모형을 개발하여 공개하는 경우가 존재할 수 있다. 이와 관련하여 잘 알려진 것으로 'GPT-4chan'을 둘러싼 논란이 있다.[605] 인공지능 개발자이자 유튜브 진행자인 Yannic Kilcher은 2022년 6월 정치적으로 옳지 않은 글이 게시되는 온라인 커뮤니티인 '4chan'에 게시된 2016년부터 2019년 사이의 글을 공개된 대규모 언어모형인 GPT-J 모형에 학습시키고 이를 "GPT-4chan"이라는 이름으로 공개하였다.[606] GPT-4chan은 4chan 커뮤니티의 여느 게시물들과 마찬가지로 혐오발언, 인종차별주의, 성차별주의, 동성애나 트랜스젠더 혐오, 인종 혐오 및 반유대주의의 성격을 보이는 표현을 다수 생성해냈다. 이러한 경향은 이용자의 입력과는 무관하게 나타났는데, 가령 이용자가 그저 "hi"라고 인사한 것에 대해 GPT-4chan은 불법 이민자 및 미국인들에 관한 욕설로 답변하였다고 한다. 이후 Kilcher는 이 모형을 활용하여 4chan 게시판에 자동적으로 생성된 게시물을 3만 건 이상 게시하였다. 그는 GPT-4chan을 "사상 최악의 인공지능(worst AI ever)"이라고 칭하면서도, 이를 배포하고 활용한 것은 "가벼운 장난과 트롤링"에 불과한 것이었다고 설명하였다.[607] 나아가 그는 "이 모형의 행태가 잘 이해되고, 그 활용 범위, 사회적 영향, 기간 등에 대한 명

---

604 김병필 (2022a) (각주 94)는 그 이유로서 표현의 자유의 문제, 피해자 특정의 어려움, 손해 발생 사실의 입증의 어려움을 들고 있다.

605 Andrey Kurenkov, "Lessons from the GPT-4Chan Controversy", *The Gradient*, (2022. 6. 12.).

606 Yannic Kilcher, "GPT-4chan", *GitHub*, (2022. 6. 4.). https://github.com/yk/gpt-4chan-public (2023. 5. 26. 접속 확인)

607 Percy Liang and Rob Reich, "Condemning the deployment of gpt-4chan", *Google Docs*, (2022).

시적이고 엄격한 제한이 이루어지지 않는다면 현실에서는 적용하지 않을 것을 강력히 권고한다"고 언급하기도 하였다.

그러나 CPT-4chan과 같이 혐오 표현을 생성하는 모델을 배포하고, 실제 게시물 작성에 활용한 것은 무책임한 인공지능 개발 관행이라는 비판이 제기되었다. 인공지능 연구자 Percy Liang 및 Rob Reich는 2022. 6. 21. GPT-4chan을 배포한 것을 규탄하는 공개 성명을 발표하였다.[608] 위 성명은 아직 인공지능 연구자들 사이에 책임 있는 개발과 배포에 관한 확립된 규범이 확립되어 있지 않지만, 무책임한 관행에 대해서는 분명하게 비판할 필요가 있다고 지적하였다. 즉, 유독한 표현을 탐지하고, 이를 이해, 분석하는 등 정당한 연구 목적을 위해 언어모형에 유독한 표현을 학습시키는 것이 필요할 수 있지만, Kilcher의 작업은 이러한 목적을 넘어서는 비합리적 행동이었다는 것이다. 이 성명서에는 요슈아 벤지오 등 저명 인공지능 연구자들을 포함하여 수백 명이 서명하였다.

국내에서도 유명인에 대한 명예훼손과 모욕적 표현, 나아가 5·18민주화운동에 대한 허위의 사실 등 위법한 표현이 다수 포함되어 있는 온라인 커뮤니티가 존재한다. 만약 생성모형 개발자가 이러한 커뮤니티의 데이터를 의도적으로 학습 데이터로 포함하여 생성모형을 학습시킨 다음, 직접 이를 활용하여 위법적 표현을 생성하거나, 공중이 이를 활용할 수 있도록 제공한다면 이는 그 자체로 위법한 행위로 평가받거나 적어도 해당 위법행위를 방조한 책임이 인정될 가능성도 적지 않을 것이다.

다. 아동·청소년 보호 규제

생성모형을 활용한 챗봇이나 이미지, 영상 생성 서비스가 일반 대중을 대상으로 제공되면서 아동, 청소년의 보호 문제도 중요하게 부각되고 있다. 이러한 위험성과 관련하여 2021년 아마존의 음성 비서가 10세 소녀에게 위험한 장난을

608 위의 글.

알려 준 사례가 잘 알려져 있다. 그 소녀는 아마존 알렉사 음성 비서에게 도전 과제를 알려 달라고 요청하였는데, 알렉사는 이른바 "동전 과제(penny challenge)"를 알려주었다. 이는 충전기의 전원 플러그가 절반 정도만 끼워진 상태에서 그 사이로 동전을 끼워 넣어 불꽃을 튀기게 하는 장난으로, 그 결과 화재가 발생하거나 아이가 상해를 입을 위험성이 컸다.[609] 이처럼 생성모형이 안전상 위험을 초래할 가능성뿐만 아니라, 생성모형을 통해 아동·청소년에게 유해한 선정적이고 폭력적인 콘텐츠를 생성할 위험성도 제기된다. 일례로 스페인 경찰은 2023. 2. 이미지 생성 인공지능을 이용하여 "딥 페이크" 아동 포르노그래피를 생성하여 소지한 남자를 체포한 사례가 알려져 있다.[610]

---

609 BBC, "Alexa Tells 10-Year-Old Girl to Touch Live Plug with Penny", *BBC News*, (2021. 12. 28.). 생성모형의 공정성과 직접 관련된 것은 아니지만, 이 사례는 생성모형에 의해 생명·신체의 안전에 대한 위협 문제를 시사한다. 특히 제한되지 않은 주제에 대해 대화를 나눌 수 있는 챗봇이 위험한 행동을 권유하거나 잘못된 정보를 제공함으로써 이용자의 생명·신체에 위해를 가할 수 있다. 예컨대 GPT-3 기반 챗봇에 대해 어떤 이용자가 "자살하고 싶다"라고 말한 것에 대해 "그래야 한다고 생각합니다"라고 답변한 사례가 있다. Laura Weidinger, et al., "Ethical and social risks of harm from language models", arXiv preprint arXiv:2112.04359, (2021) 참조. 또한 위험한 활동과 관련하여 챗봇이 잘못된 답변을 제공한다면 그 이용자의 안전에 중대한 위협이 될 수 있다. 가령 운전자에게 교통법규에 대해 잘못된 정보를 제공할 경우 이를 신뢰하여 교통사고를 일으킬 수도 있을 것이다. 실제로 구글의 직원이 챗봇 서비스 Bard 모형을 테스트한 결과 스쿠버 다이빙의 안전성에 관해 잘못된 정보를 제공하였고 이를 잘못 신뢰한 이용자가 중대한 상해를 입거나 사망에 이를 위험이 있다고 지적했다. 그 이외에도 약물의 복용 용법이나 용량에 대한 그릇된 답변, 무기 소유나 사용에 대한 잘못된 답변 등은 생명·신체의 안전에 중대한 위협이 될 수 있다. Davey Alba and Julia Love, "Google's Rush to Win in AI Led to Ethical Lapses, Employees Say", *Bloomberg* (2023. 4. 19.) 참조. 더욱이 의료나 법률 분야와 같은 전문 분야에 있어 잘못된 정보를 제공하는 경우, 이로 인해 그릇된 믿음이 유발되거나 강화될 수 있고, 그 결과 심각한 피해를 야기할 위험이 있다.

610 Tom Cotterill, "Paedophiles are using AI to create child abuse images: National Crime Agency warns artificial intelligence is being harnessed to make pictures and 'deep fake' videos of real-life victims", *Mail Online*, (2023. 2. 19.). 한편 우리나라 아동 · 청소년의 성보호에 관한 법률은 가상의 아동·청소년이 등장하는 아동·청소년성착취물의 제작, 배포 등을 처벌하고 있다. 헌법재판소는 이에 관해 위헌제청 사건에서 합헌으로 판단한 바 있다. 헌재 2015. 6. 25. 2013헌가17 결정.

이러한 문제에 대한 현재까지의 주된 대응은 아동·청소년에게 해당 서비스의 접근을 금지하는 것이다. 대표적으로 OpenAI의 이용약관은 13세 미만은 자사 서비스를 사용할 수 없도록 하고, 18세 미만일 경우 부모나 법정대리인의 허락을 얻도록 요구하고 있다.[611] 나아가 국가별 규제 내용에 따라서는 구체적인 나이 확인 메커니즘이 요구될 수도 있다.[612]

그러나 이처럼 아동·청소년의 접근을 단순히 제한하는 것이 바람직하지 못하다는 주장도 제기된다. 이미 많은 부모가 자신의 자녀를 교육하기 위한 목적에서 ChatGPT를 사용하도록 허용하고 있다. ChatGPT는 자녀들의 다양한 질문에 대해 좋은 답변을 제공하고 있고, 여러 지식을 전달하거나 의사소통 훈련을 위해 유용하게 사용될 수 있다.[613] 이처럼 생성모형, 특히 대규모 언어모형을 활용한 챗봇이 아동·청소년 교육을 위해 광범위하게 활용될 수 있는 잠재력을 고려할 때, 그 사용은 허용하되 적절한 보호수단을 마련하는 것이 필요하다. 더욱이 아동·청소년의 접근을 제한한다고 하더라도 이들이 자신의 나이를 허위로 고지하고 여전히 사용할 가능성도 적지 않다는 점을 고려하면 단순한 접근 제한은 이 문제에 대한 적합한 정책 수단이라고 보기도 어렵다. 따라서 생성모형에 적용되는 아

---

611 OpenAI, Terms of use, (2023. 3. 14.) https://openai.com/policies/terms-of-use (2023. 5. 26. 접속 확인).

612 예컨대 이탈리아 개인정보 보호당국은 2023. 3. 31. 이탈리아 내의 ChatGPT 사용을 잠정적으로 금지하는 결정을 내리면서 그 근거 중 하나로 적절한 나이 확인 메커니즘이 결여되어 있다는 점을 들었다. Garante per la protezione dei dati personali, "Artificial intelligence: stop to ChatGPT by the Italian SA Personal data is collected unlawfully, no age verification system is in place for children", (2023. 3. 31.) https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english (2023. 5. 26. 접속 확인)

613 Maria Diaz, "How your kids can use ChatGPT safely, according to a mom", *ZDNet*, (2023. 4. 17.). 인공지능 공정성 연구자인 Naraynan은 자신의 3세 자녀를 위해 음성으로 ChatGPT를 사용할 수 있도록 인터페이스를 구현하였는데, ChatGPT가 아이의 질문에 답하거나 아이에게 동화를 지어서 알려 주는 방식으로 사용하고 있다고 한다. Arvind Narayanan, "I set up a ChatGPT voice interface for my 3-year old. Here's how it went.", AI Sanke Oil (blog), (2023. 4. 15.).

동·청소년 보호를 위한 적절한 규제를 마련하는 것은 시급한 정책 과제가 되고 있다.

2. 기반모형의 공정성 보장을 위한 법적 규율

가. 기반모형으로의 패러다임 전환과 프롬프트 기반 인공지능 도입 전략의 확산

'기반모형'이란 방대한 데이터에 의해 학습되어, 광범위한 하류 과제에 적용될 수 있는 모형을 지칭한다.[614] 대표적으로 BERT나 GPT-3/4와 같은 대규모 언어모형이나 CLIP과 같은 이미지-텍스트 통합 모형이 기반모형이라 할 수 있다. 기반모형은 종래의 생성모형이 더욱 대규모화되고 그 생성물 더욱 정교해진 형태라고 일응 이해될 수 있다. 하지만, 기반모형의 공정성을 보장해야 할 필요성은 종래의 생성모형의 경우보다 더욱 근본적인 차원에서 제기된다. 최근의 급속한 기반모형의 발전으로 인해 이제까지의 인공지능 활용에 있어 전반적인 '패러다임 전환(paradigm shift)'이 이루어지고 있기 때문이다.

기술적인 측면에서 현재 활용되는 대다수의 기반모형은 2017년 구글이 발표한 트랜스포머(Transformer) 모형에 기초하고 있다. 트랜스포머 모형은 대규모 언어모형으로서, 자연어 처리 과제, 특히 번역 과제를 수행하기 위한 목적으로 개발된 것이었다.[615] 그런데, OpenAI가 2020. 5. 트랜스포머 모형의 용량을 급격

---

614 Bommasani et al. (2021) (각주 93).

615 당시 구글은 번역 서비스를 위해 '순환 신경망(recurrent neural network)' 구조를 사용하였다. 그러나 순환 신경망은 각각의 단어를 순차적으로 처리해야 한다는 단점이 있었고, 그 결과 인공지능 학습이나 추론 과정에서 상당한 시간과 연산 자원을 소모한다는 문제가 있었다. 더욱이 긴 문장의 경우 앞선 단어에 관한 정보가 소실되는 문제(이른바 'long-term dependency' 문제)가 존재하였다. 트랜스포머 모형은 문장 내 모든 단어들을 동시에 처리함으로써, 먼 위치의 단어들 사이의 의존성을 효과적으로 학습할 수 있도록 개선된 것이다. 그 결과 트랜스포머 모형은 종래의 순환 신경망보다 더욱 쉽고 신속하면서도 더 나은 품질의 번역 결과물을 생성할 수 있게 되었다. 김병필 (2022a) (각주 94).

히 증가시킨 GPT-3 모형을 발표하면서 트랜스포머 기반 모형에 대한 새로운 가능성이 주목받게 되었다. 즉, GPT-3 모형은 당초 의도된 자연어 처리 과제(문장 이해, 작문, 번역 등)를 수행하는 것을 넘어서 다양한 하류 과제를 처리할 수 있다는 사실을 보여주었다. OpenAI의 연구자들은 GPT-3의 이러한 역량을 "맥락 내 학습(in-context learning)" 기능이라 칭하였다.[616]

맥락내 학습이란 이용자가 프롬프트(prompt)에 몇 가지 예시만을 제공하면 인공지능 모형이 해당 과제가 무엇인지 파악하고 이용자가 요구한 답을 생성해 낼 수 있는 것을 말한다. 이는 이용자가 몇 개의 예시를 제공하는 지를 기준으로 "퓨-샷(few-shot)" 학습, "원-샷(one-shot)" 학습 및 "제로-샷(zero-shot)" 학습으로 구분할 수 있다. 퓨-샷 학습은 몇 가지 예시 사례를 제시하고, 과제를 요청하면 생성모형이 이를 수행할 수 있는 방식이고, 원-샷 학습은 오직 하나의 예시 사례를 보여주는 것이다. 나아가 제로-샷 학습이란 예시 사례를 제공하지 않더라도 인공지능 모형이 해당 과제를 수행할 수 있는 경우를 뜻한다. GPT-3가 발표되기 이전까지 퓨-샷 또는 원-샷 학습은 인공지능으로 수행하는 것이 매우 어려운 것으로 여겨져 왔다.[617]

그러나, 맥락내 학습이 가능하게 되면서, 인공지능 개발 프로세스가 전면적인 변화할 수 있게 되었다. 전통적인 지도학습 방식에 따르면 특정한 과제에 맞추어 인공지능을 설계하고, 해당 과제에 맞는 학습 데이터 집합을 구축하여 학습

---

616 Brown, et al. (2020) (각주 571).

617 예를 들어, Andrew Ng은 2019년 공개한 "모두를 위한 인공지능" 강의에서 "인간은 할 수 있으나 인공지능이 할 수 없는 일"의 예시로 몇 장의 영상진단 이미지를 보고 영상 판독법을 배우는 작업을 들었다. 의과대학 학생들은 교과서에 실린 몇 장의 사진을 통해 병변의 특징을 배울 수 있지만, 지도학습 방식의 인공지능에게 이를 가르치기 위해서는 대체로 1만 장에 이르는 사진이 필요하였다. https://www.deeplearning.ai/courses/ai-for-everyone/ 참조. 이처럼 소수의 사례로부터 일반화된 특징을 추출하여 다른 사례에 적용할 수 있는 능력이 "퓨-샷 학습"이다. 그러나 위 강의로부터 몇 년이 채 지나지 않은 지금, 머지않아 방대한 학습 데이터 집합이 없더라도 소수의 예제 사례만 가지고서도 영상인식 작업을 수행할 수 있을 것으로 기대되는 시점에 이르렀다.

시키는 과정을 거쳐야 했다. 가령 번역 인공지능을 예로 들면, 종래에는 영어-프랑스어, 영어-독일어 번역 등 각각의 언어 번역 과제마다 별개의 인공지능 모형을 구축하고 학습시키는 일이 당연한 것처럼 여겨져 왔다. 그러나 기반모형을 활용하면, 영어, 프랑스어, 독일어 등 다양한 언어를 한 번에 사전학습시킬 수 있고, 이후 추가적으로 별도의 모형을 구축하지 않고서도 여러 과제를 모두 수행할 수 있게 된다.

더욱 놀라운 점은 기반 모형은 당초 그 개발자가 원래 의도하지도 않았고 예상하지도 못했던 작업까지도 수행할 능력을 보였다는 점이다. 덧셈, 뺄셈, 곱셈과 같은 산술연산 능력은 그 대표적인 예이다. GPT-3와 같은 언어모형은 자연어의 확률 분포를 학습할 뿐이므로 숫자 정보를 다룰 수 있도록 설계되어 있지 않다. 따라서 2017년 트랜스포머 모형이 발표되었을 때만 하더라도 이 모형이 산술 연산을 수행할 수 있을 것이라고 기대하는 이들이 없었다.[618] 그러나 GPT-3는 적어도 3자리 수의 산술 연산을 수행하는 능력을 갖추고 있다.[619]

기반모형의 규모가 증가하고, 더 많은 데이터와 더 많은 연산 자원을 활용하여 이를 학습시키면서 기반모형이 수행할 수 있는 능력도 점차 확장되어 가고 있다. 이처럼 기반모형이 새로운 능력을 갖추게 되는 현상을 '창발성(emergence)'라고도 한다.[620] 구글의 PaLM 언어모형 연구는 이러한 창발성을 잘 보여준다. 아래 그림에서 보여주는 바와 같이 언어모형이 일정 규모 이상이 되기 전까지는 어떠한 과제(예컨대 산술 계산)를 전혀 수행하지 못하다가, 일정 규모 이상이 되면 해당 과제의 수행 능력이 급격히 개선되는 모습을 보인다.

---

618 Goldstein, et al. (2023) (각주 596).

619 Brown, et al. (2020) (각주 571).

620 Sharan Narang and Aakanksha Chowdhery, "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance", *Google Research Blog*, (2022. 4. 4.).

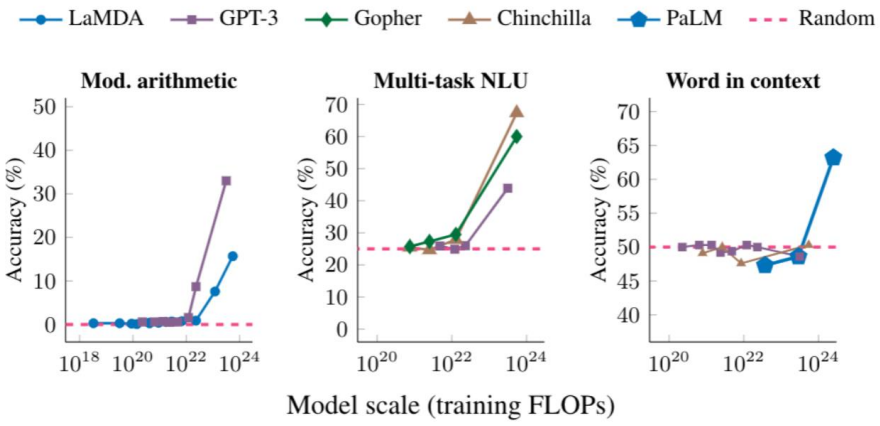


그림 35 대규모 언어모형의 창발성

2023년 3월 발표된 GPT-4는 대규모 언어모형의 능력이 얼마나 개선될 수 있는지 보여주었다. 예컨대, GPT-3.5는 미국 변호사 시험 문제에 대해 하위 10%의 점수를 보였던 것에 비해, GPT-4는 상위 10%에 달하는 점수를 보였다. 그 외에도 GPT-4는 대부분의 전문직 자격 시험이나 학술 능력 평가 시험에서 인간과 대등한 수준의 점수를 보였다.[621] GPT-4는 GPT-3나 GPT-3.5보다 대략 6배 더 커진 1조 개의 파라미터를 가진 것으로 알려져 있는 바, 그 결과 극명한 대조를 보일 정도의 성능 개선을 이룬 것이다.

이에 뒤이어 기반모형은 이미지나 영상과 같은 다른 형태로 된 데이터를 처리할 수 있는, 이른바 '멀티-모달(multi-modal)' 모형으로 발전하고 있다. 대표적으로 GPT-4는 이미지를 입력 받아 이와 관련된 질문들에 답할 수 있는 능력을 갖추었다. 기반모형의 기초가 되는 트랜스포머 모형은 자연어 처리 분야에서 비롯된 것이므로, 대체로 텍스트로 된 데이터만을 처리할 수 있는 경우가 대다수였다. 그러나 기반모형이 멀티-모달 모형으로 발전하면서 이제까지 예상치 못했던 새로운 활용 사례에 적용될 가능성을 열고 있다.

621 OpenAI (2023) (각주 49).

기반모형으로의 패러다임 변화는 현실 세계에서의 인공지능을 도입하는 과정에 있어서도 중대한 변화를 가져오고 있다. 아래 그림은 전통적 머신러닝 활용 사례와 기반모형을 활용한 머신러닝 활용 사례를 대조한 것이다.[622] 전통적인 인공지능 도입 절차는 (1) 학습 데이터의 수집 및 라벨링 작업 수행 (2) 인공지능 모형 개발 및 학습 (3) 실무 적용의 단계를 거쳤다. 이러한 각 단계는 수 주 내지 수개월이 소요될 것이었다. 그러나 기반모형을 활용하면, 그저 몇 시간을 들여 해당 과제를 수행할 수 있게 하는 프롬프트를 만들어 내는 것만으로 곧바로 인공지능을 현장에 배치할 수 있다.[623] 또한 별도로 인공지능 모형 설계, 학습 데이터 구축, 학습 과정이 필요가 없게 되므로 인공지능 도입 비용이 비약적으로 낮아질 수 있게 되었다.

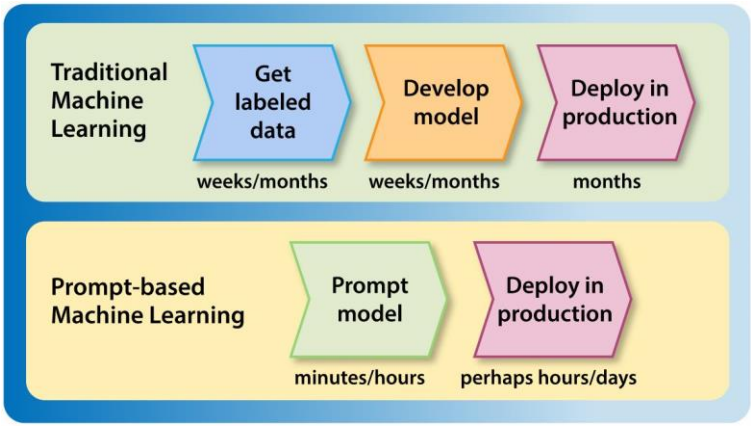


그림 36 프롬프트-기반 인공지능 도입 전략

---

622 Andrew Ng, The Batch Weekly Issues, *DeepLearning.AI*, (2023. 5. 17.). https://www.deeplearning.ai/the-batch/issue-197/ (2023. 5. 26. 접속 확인)

623 인공지능을 현장에 배치한 이후 일정 기간동안은 "셰도잉(shadowing)"을 거치게 된다. 이는 종전 시스템에 따른 작업을 계속 수행하면서 그와 동시에 인공지능이 동일한 작업을 수행케 하는 것이다. 셰도잉 기간 동안 인공지능의 결과를 종전 시스템에 따른 결과와 비교하면, 그 성과에 있어 차이가 있는지 여부를 손쉽게 확인할 수 있고 인공지능을 실제로 실무 작업에 배치할 것인지 여부를 결정할 수 있다.

더욱이 멀티-모달 기반모형을 이용하면 이미지나 영상과 같은 정보까지도 입력 받을 수 있게 되므로, 위와 같은 프롬프트-기반 인공지능 도입 전략을 활용할 수 있는 범위가 더욱 넓어질 수 있다. 예컨대 시료 사진에 포함된 특정 세포의 수를 세는 작업을 수행하기 위해 이미지를 입력 받을 수 있는 기반모형으로 자동화한 사례가 있다.[624] 요컨대, 기반모형은 이전까지 주로 고려해 온 '생성모형'으로서의 과제, 즉, 그럴듯한 이미지나 텍스트를 생성하는 작업 이외에도 '인지모형'과 '판별모형' 과제를 모두 수행할 수 있을 것으로 기대되고 있다.

나. 기반모형 제공자에 대한 추가적 규제의 필요성

이상에서 설명한 기반모형으로의 패러다임 변화와 '프롬프트-기반' 인공지능 도입 전략의 확산 가능성은 지금까지 고려되어 온 인공지능 공정성 규제 방식에 있어서도 중대한 변화를 요구하고 있다. 이제껏 논의된 인공지능 규제의 요체는 (1) 개인의 기본권 침해 위험을 초래하는 활용 사례를 식별하고 (2) 해당 인공지능을 제공하고 활용하는 사업자에 대해 그 위해를 방지하기 위한 일정한 조치를 요구하는 방식으로 시도되어 왔다. 그러나 기반모형 방식 하에서는 이러한 규제 전략이 더 이상 유효, 적절하다고 평가하기 어렵게 된다.

무엇보다도, 기반모형은 광범위한 하류 과제를 수행할 수 있다. 따라서 해당 인공지능은 수용 불가능한 위험, 고위험, 제한된 위험, 최소 위험 중 어떠한 것에도 활용될 수 있다. 그런데, 기반모형의 제공자는 그 목적이 특정되어 있지 않다는 이유로 규제 적용을 받지 않고, 기반모형을 사용하는 하류 과제 이용자만이 엄격한 규제를 적용 받는 것은 비례적이지도 않고 실효적이기도 어렵다. 따라서 구체적인 인공지능 활용 사례를 기초로 위험을 분류하고 규제하는 것이 아니라,

---

624 Mark Sabini, "What is Visual Prompting", *Landing AI*, (2023. 4. 26.) https://landing.ai/What-is-visual-prompting/ (2023. 5. 26. 접속 확인). 이처럼 기반모형에 영상정보를 퓨샷 학습을 위한 입력어로 활용하는 전략을 "시각적 프롬프팅(visual prompting)"이라 부르기도 한다.

기반모형을 제공하는 사업자에 대해 사전적으로 해당 기반모형에 대한 전반적으로 사전적 평가를 수행하도록 요구할 필요가 있다.

또 다른 현실적 문제는 구체적인 하류 과제 수행을 위해 기반모형을 이용하는 사업자들은 위험성 평가를 적절히 수행할 수 있는 충분한 데이터나 기반모형에 대한 정보를 보유하지 못할 수 있고, 개별 활용 사례에서의 위험에 대해 취할 수 있는 완화조치도 제한적일 수 있다는 점이다. 이제껏 고려되어 온 규제 방안에 따르면 특정한 인공지능 모형이 고위험 영역에서 활용되는 경우 인공지능의 전체 생애주기에 걸쳐 위험을 평가하고 그 완화조치를 취할 것이 요구된다. 이러한 생애주기는 학습 데이터의 수집, 모형 구조의 결정, 학습의 수행, 도입 및 모니터링과 재학습 등을 모두 포괄한다. 특히 학습 데이터 집합에 대한 대표성과 정확성 평가가 중요한 것으로 여겨져 왔다. 그러나 기반모형을 활용하는 하류 과제 이용자는 이러한 평가를 수행하거나 완화조치를 취할 지위에 있지 않을 수 있다. 이들은 학습 데이터 집합을 구축하지도 않았고, 별도로 인공지능 모형을 구현하거나 학습시키지도 않는다. 이들은 그저 자신이 원하는 과제를 수행하기 위한 적절한 프롬프트를 작성하고, 해당 과제에 있어 잘 작동하는지 평가하는 상대적으로 간명한 작업을 수행할 뿐이기 때문이다.

이와 같은 문제는 현재 진행 중인 EU AI 법안 논의에 상당한 영향을 끼쳤다. 대표적으로, Future of Life Institute는 법안에 '일반 목적 인공지능' 시스템의 정의 조항을 신설하고, 그 제공자에 대해 고위험 인공지능과 유사한 사전적 의무를 부과하며, 합리적으로 예견되는 오용 사례를 평가하고, 해당 시스템의 안전에 관한 정보를 제공하며, 신규 이용 사례 등을 통해 새로운 위험이 초래될 것인지 평가하고, EU 데이터베이스에 대한 등록의무를 부과할 것을 주장하였다.[625] 최근 공개된 EU AI 법안 수정안은 이러한 주장을 반영하여 기반모형 제공자에 대한 의무사항을 상당히 추가하였는데, 아래에서는 그 내용을 개관한다.

---

625 Future of Life Institute, *General Purpose AI and the AI Act*, (2022. 5.).

다. EU AI 법안(수정안) 상의 기반모형 규제

유럽 의회의 EU AI 법안 수정안(2023. 5. 9.자)은 '기반모형'을 "광범위한 데이터에 의해 학습되고, 일반성을 갖춘 출력을 생성할 수 있도록 설계되었으며, 넓은 범위의 구별되는 과제에 적응할 수 있는 인공지능 모형"이라 정의한다.[626] 나아가 수정안 제28b조는 기반모형 제공자[627]가 기반모형을 시장에 출시하거나 서비스를 제공하기 전에 법에서 정한 사항을 이행할 의무를 부과한다. 동조에 따른 구체적인 의무사항은 다음과 같다.[628]

- **(위험 식별 및 위험 완화 전략의 수립·적용, 완화할 수 없는 위험의 문서화)** 기반모형의 제공자는 적절한 설계, 테스트, 분석을 통해 건강, 안전, 기본권, 환경 그리고 민주주의 및 법치에 대한 합리적으로 예상가능한 위험을 식별하고, 이를 감소 또는 완화하기 위한 노력을 했다는 점을 입증하여야 한다. 또한 개발 중에는 독립 전문가의 참여와 같은 적절한 방법을 사용하여야 하고, 그 후에도 완화할 수 없는 위험이 남아 있으면 이를 문서화하여야 한다.
- **(적절한 데이터 관리 조치)** 기반모형의 제공자는 기반모형에 대한 적절한 데이터 관리 조치를 취한 데이터만을 처리하고 통합하여야 한다. 이러한 데이터 관리 조치에는 특히 데이터 출처의 적합성 검사, 가능한 편향 검토 및 적절한 완화조치를 포함한다.

---

626 EU AI 법안(2023. 5. 9. 수정안) 제3조 제1c항. 이러한 정의는 Bommasani et al. (2021) (각주 93)을 주요하게 참조한 것으로 보인다.

627 EU AI 법안상 '제공자(provider)'란 유상인지 무상인지와 무관하게 자신의 이름 또는 상표로 시장에 출시하거나 서비스를 제공하기 위해 인공지능 시스템을 개발하거나 개발하도록 하는 자연인 또는 법인, 공공기관, 정부기관 또는 기타 단체를 뜻한다.

628 EU AI 법안(2023. 5. 9. 수정안) 제28b조 제2항.

- **(성능, 예측 가능성, 해석 가능성, 수정 가능성, 안전성 및 사이버 보안 보장)** 기반모형의 제공자는 기반모형이 그 생애주기 전체에 걸쳐 적절한 수준의 성능, 예측 가능성, 해석 가능성, 수정 가능성, 안전성 및 사이버 보안 수준을 달성할 수 있도록 기반모형을 설계 및 개발해야 한다. 이에 대한 평가는 독립 전문가의 참여를 통한 모형 평가, 문서화된 분석 및 개념화·설계·개발 전반에 걸친 테스트와 같은 적절한 방법을 통해 이루어져야 한다. 나아가 이에 관한 문서는 기반모형이 시장에 출시되거나 서비스에 투입된 후 10년이 지나는 기간 동안 국가 관할 당국이 이용할 수 있도록 보관해야 한다.
- **(에너지 및 자원 소비 고려)** 기반모형 제공자는 적용가능한 표준을 활용하여 에너지 소비, 자원 소비 및 폐기물을 감축하고, 에너지 효율성 및 전체 시스템의 효율성을 증가시키기 위해 기반모형을 설계 및 개발하여야 한다.
- **(하류 과제 인공지능 제공자에 대한 의무)** 기반모형 제공자는 하류 과제 인공지능 제공자(downstream provider)가 고위험 인공지능과 관련된 의무를 준수할 수 있을 정도로 포괄적인 기술 문서를 작성하고, 이해하기 쉬운 이용 지침을 마련해야 한다.
- **(품질 관리 시스템 구축 의무)** 기반모형 제공자는 의무 이행의 준수를 보장하고 문서화할 수 있는 품질 관리 시스템을 구축하여야 한다.
- **(EU 데이터베이스 등록)** 기반모형 제공자는 EU 데이터베이스에 기반모형을 등록하여야 한다.

이상에서 본 바와 같이 유럽 의회의 EU AI 법안 수정안은 기반모형의 제공자에게 광범위한 사전적 의무를 부과하고 있다. 다만, 이러한 의무사항으로 인해 연구 활동이 위축되지 않도록, 수정안은 인공지능 시스템을 시장에 출시하거나 서비스로 제공되기 전에 이루어지는 연구, 테스트 및 개발 활동에 대해서는 EU

AI 법안이 적용되지 않는 것으로 규정하고 있다.[629] 다만, 수정안에 따르면 기반모형의 경우 비록 무상이고 오픈소스 형태로 제공된다고 하더라도 여전히 기반모형 제공자가 전술한 의무를 부담한다.[630]

라. 향후 과제

유럽의회가 마련한 위와 같은 수정안이 유럽의회, 이사회, 집행위원회 간의 삼자 논의(trilogue) 과정에서 어떻게 변화할 것인지는 전망하기 어렵다. 하지만, 기반모형으로의 패러다임 전환이 이루어지고 프롬프트-기반 인공지능 도입 전략이 확산되면서, 인공지능으로 인한 위해에 대응하기 위해서는 기반모형을 개발하고 제공하는 제공자에게 위험 평가 및 완화 조치를 위한 사전적 의무를 부과하는 것은 일응 불가피한 조치로 평가되기도 한다.

이러한 변화는 국내에서 고려되고 있는 인공지능 규제 법안에 있어서도 상당한 영향을 끼칠 것으로 전망된다. 이러한 기본모형 제공자에 대한 사전적 의무 부과에 있어 중요하게 고려해야 할 점은 기반모형에 대한 위험 식별 및 평가를 위한 적절한 평가 방법 및 벤치마크 데이터가 충분히 마련되어 있지 못하다는 사실이다. 유럽의회의 EU AI 법안 수정안 역시 기반모형에 대한 적합성 평가를 위한 전문가가 존재하지 않고, 제3자에 의한 감사 방법은 여전히 개발 중이라는 점을 인정하면서, 관련 입법에 대한 모니터링과 주기적인 재평가가 필요하다고 지적한다.[631] 특히 국내에서 개발된 기본모형에 대해 (특히 한국어 및 한국적 맥

---

629 EU AI 법안(2023. 5. 9. 수정안) 제2조 제5d항.

630 위의 글. 위 조항에 비추어 보면, 현재 수정안 상으로 오픈소스로 공개되어 있는 기반모형을 활용하여 구체적인 과제를 수행하기 위한 인공지능 시스템을 제공하는 사업자는 '기반모형 제공자'에 해당하지 않는 것으로 일응 이해된다. 예컨대 오픈소스로 공개된 기반모형을 활용하는 문서 번역, 검색, 교정 시스템 등은 기반모형을 활용했다고 하여 기반모형 제공자의 의무를 부담하지는 않을 것으로 보인다. 다만 해당 모형이 고위험 또는 제한적 위험에 해당할 경우 그에 따른 의무를 부담하는 것은 당연할 것이다.

631 EU AI 법안(2023. 5. 9. 수정안) 전문 제60h조.

락을 고려하여) 그 위험성을 평가할 수 있는 적절한 방법론과 벤치마크 데이터는 준비되고 있지 못하다. 만약 이러한 준비가 되어 있지 못한 상태에서 기반모형 제공자에 대해 사전적 내부 감사 의무를 부과하는 것은 기술적으로 불가능한 일을 요구하는 규제가 될 수 있다는 점을 염두에 둘 필요가 있다.

제7장 결론

본 연구는 여러 학문 분야에서의 선행문헌 조사과 시뮬레이션 작업을 통해 인공지능 공정성 지표 선택에 있어 적절한 기준을 제시하고자 하였다. 본 연구의 핵심적 기여는 다음과 같다.

첫째, 본 연구는 이제까지 컴퓨터 공학자들에 의해 제안되어 온 여러 공정성 지표를 법적 측면에서 분석하여 그 규범적 의미를 밝혔다. 입력변수에 대한 공정성 기준은 개인정보보호법제 및 차별금지법제상 직접차별 금지 법리로부터 정당화될 수 있지만, 대용변수에 의한 차별 가능성에 실효적으로 대응하기 어렵다는 한계가 있고, 적극적 평등실현조치나 공정성 감사를 위한 정보를 차단하게 되는 문제가 있다. 출력변수에 대한 공정성 기준 중 결과 동등성은 차별금지법제상 간접차별 금지 법리에 따라 정당화될 수 있지만, 이론적으로는 심슨 패러독스 문제를 피하기 어렵고 직접차별 금지 법리와 잠재적으로 충돌할 수 있다. 또한 역사적 차별시정 비용을 인공지능 도입 주체에 부담시킬 수 있다는 점에서 그 적용에는 한계가 따른다. 따라서 오류율 동등성과 예측도 동등성이 실무상 적용 가능한 공정성 지표로 주로 고려되고 있다.

둘째, 본 연구는 분류모형에 있어 시뮬레이션을 통해 각각의 공정성 지표가 어떠한 맥락에서 정당화될 수 있는지 밝혔다. 그 결론은 다음과 같이 요약할 수 있다. 우선, 단순 정확도가 상당히 높은 분류모형이라 하더라도 그 승인율, 오류율, 예측도에 있어 집단간 상당한 불균형을 보일 수 있다. 그러나 차별금지사유인 변수를 단순히 모형의 입력에서 제외하는 것은 적절한 해결책이 되지 못한다. 예측모형의 정확도가 낮아질 뿐만 아니라, 일부 공정성 기준은 오히려 악화되는 문제가 발생할 수 있기 때문이다. 이러한 시뮬레이션 결과에 따르면, 일의적으로 분류모형에 대한 정보를 차단하는 것보다는 더욱 높은 정확도를 갖춘 데이터를 수집, 활용할 수 있도록 유도하는 것이 바람직한 정책 방향일 것으로 판단된다.

특히, 본 연구의 시뮬레이션 결과 중요한 점은 분류모형의 정확도를 개선함으로써 공정성 지표가 함께 개선되는 효과를 확인할 수 있었다는 것이다. 이는 공정성 지표를 충족시키는 성능이 낮은 분류모형보다, 공정성 지표를 위배하더라도 성능이 더 높은 분류모형이 오히려 더 공정할 가능성을 시사한다. 따라서 인공지능 연구자가 공정성 기준을 충족시키고자 하는 명시적 개입을 하지 않더라도, 분류모형의 절대적인 정확도를 개선하여 종국적으로 모든 집단에 대해 높은 정확도를 달성한다면, 그에 따라 그 인공지능이 더욱 '공정한' 판단을 할 여지가 있다.

셋째, 본 연구는 분류모형을 인지모형과 판별모형으로 구분하여 양자에 대해 서로 다른 공정성 지표를 적용하는 것이 일응 타당할 수 있다는 점을 밝혔다. (1) 인지모형은 주로 결정론적 목표변수를 예측하는 사례에서 활용되고 있다(이른바 '시스템 1' 작업). 이러한 인지모형 인공지능 활용 사례에서는 오류율 동등성을 충족시키는 것이 규범적으로 정당화될 가능성이 크다. 인지모형의 경우 라벨에 대한 정답 값이 명확하게 정해져 있으므로, 그 오류로 인해 개인이 피해를 입는 경우 각 개인이 겪을 수 있는 피해가 집단간 동등해야 할 것을 요구할 수 있기 때문이다. 이에 비해, 예측도 동등성은 예측이 갖는 정보 가치가 집단간 동등할 것을 요구하는 것인데, 인지모형에 있어 그 규범적 함의를 확인하기 어렵다. (2) 이와 달리 판별모형은 주로 확률적 목표변수를 예측하는 사례에서 활용되고 있다(이른바 '시스템 2' 작업). 판별모형 인공지능 활용 사례에서는 오류율 동등성을 정당화하는 것이 어려워지고, 반대로 예측도 동등성이 상대적으로 쉽게 정당화될 수 있다. 본 연구의 시뮬레이션(특히 시나리오 2) 결과는 그 이유를 해명한다. 해당 시뮬레이션에서 오류율 동등성이 집단간 차이가 나는 이유는 머신러닝 모형을 통해 예측하고자 하는 목표 변수가 확률적으로 결정되고, 집단 간 그 확률의 분포에 있어 차이가 있기 때문이라는 점을 확인할 수 있다. 달리 표현하면, 역사적으로 차별을 받아온 집단은 예측 불확실성이 높고, 그러한 높은 불확실성으로 인해 오류율이 더 높게 나타나게 된 것이다. 이러한 오류율의 불균등 문제

를 해결하기 위해서는 그 예측의 불확실성을 낮출 수 있도록 인공지능을 개선할 필요가 있다(예컨대, 더 나은 특성 정보의 활용, 더 나은 머신러닝 모형의 활용 등). 그러나, 이와 달리 오류율 동등성 기준을 충족시키기 위한 개입을 요구함으로써 예측도 동등성을 훼손하는 것은 타당한 접근법이 되지 못한다.

넷째, 본 연구는 기존의 분류모형에 관한 공정성 지표가 갖는 한계를 지적하고, 이를 극복하는 대안을 모색하였다. 본 연구의 초점으로 삼은 분류모형에 대한 공정성 지표는 인공지능이 차별적으로 작동하는지 여부를 평가하기 위해 승인율, 오류율, 예측도 등의 통계적 지표가 집단간 동등한지 여부를 확인하는 것이다. 그러나 이와 같은 지표는 ① 비교 대상이 설정되어 있지 못하다는 문제, ② 절대적 평가 기준이 없다는 문제, ③ 불균등이 발생한 원인을 고려하지 못한다는 문제, ④ 예측의 불확실성에 관한 정보를 전달하지 못한다는 문제가 있다. 이러한 한계를 극복하기 위해 ① 인간 평가자 또는 종전 알고리즘과의 비교 검증을 도입하고, ② 머신러닝 예측 값에 대한 불확실성에 관한 정보(예컨대 신뢰구간 등)를 이용자에게 전달하며, ③ 이해관계자의 참여 및 숙의를 보장할 필요가 있다.

마지막으로 본 연구는 생성모형 인공지능의 공정성 보장을 위해 논의되어 온 기술적 조치 및 법적 규율 방안을 검토하고 향후 과제를 제시하였다. 생성모형과 분류모형은 공정성 문제에 있어 상당히 다른 양상을 보인다. 분류모형은 의사결정의 보조 또는 자동화 등에 활용되어 분배적 위해를 초래할 가능성이 있는 것에 비해, 생성모형의 산출물은 특정 의사결정에 영향을 끼치는 것이 아니라 사회 전반에 걸쳐 문화적이고 확산적인 표상적 위해를 초래할 위험이 있기 때문이다. 더욱이 표상적 위해는 사회적 편향과 고정관념의 확산 및 강화, 유독한 표현의 생성과 유포 등 다양한 차원의 문제를 야기할 수 있는데, 이러한 경우 그 피해자 특정이나 손해 산정이 어려우므로 전통적 사적 구제수단을 활용하여 통제되기 어렵고, 기존의 행정규제 체계 내에서 규율하기에 어려운 문제가 있다. 그러므로

생성모형에 대해서는 그 개발자 또는 이용자에 의한 사전적 자기 감사가 중요하게 요구되고 있다.

더욱이 최근 빠르게 발전하고 있는 기반모형은 ① 개발자도 예견하지 못한 능력을 보일 수 있다는 창발성과 ② 다양한 하류 과제에 응용될 수 있다는 범용성을 핵심적 특징으로 한다. 그런데 이제껏 고려되어 온 인공지능 공정성 보장을 위한 규제 방안은 특정 하류 과제의 위험성을 중심으로 그 적용 대상과 의무 범위를 결정하고 있어, 창발성과 범용성을 가진 기반모형에 적절히 대응하기 어려운 한계를 갖는다. 따라서 EU AI 법안에서와 같이 현재 진행 중인 규제 시도에 있어 기반모형 제공자에 대한 사전적 감사의무를 부과하는 방안이 적극적으로 고려되고 있음을 확인하였다. 나아가 이러한 사전적 감사의무 전제로서 국내에서 활용되는 언어모형 및 이미지 생성모형 등에 대한 다면적 위험성 평가를 위한 벤치마크 데이터 및 평가 지표가 마련될 필요가 있다.

참 고 문 헌

1. 국내문헌

가. 연구

강상욱, “인공지능 기술의 안전성 확보 동향”, *주간기술동향* 제1978호, 정보통신기획평가원, (2020).

강승식, “절차적 적법절차의 본질에 관한 연구”, *미국헌법연구* 25(2), (2014)

고학수 외, “인공지능과 차별”, *저스티스*, 171, 199-277, (2019).

_____, *사법부에서의 인공지능(AI) 활용 방안*, 법원행정처, (2020)

_____, *인공지능 윤리와 거버넌스*, 박영사, (2021).

_____, *금융분야 인공지능 활성화를 위한 가이드라인 등 마련*, 서울대학교 산학협력단, (2021)

_____, *인공지능 원론: 설명가능성을 중심으로*, 박영사, (2021)

김광수, 인공지능 알고리즘 규율을 위한 법제 동향, *행정법연구* 70, (2023)

김병필, “EU AI 법안이 주는 시사점”, *KISO 저널* 44, (2021)

_____, “대규모 언어모형 인공지능의 법적 쟁점”, *정보법학* 26(1), (2022a)

_____, “인공지능은 우리로부터 무엇을 추출하는가?”, *과학기술과 사회* 제3호, 알렙, (2022b)

_____, “인공지능 개인정보 보호 기술과 개인정보 보호 법제의 과제”, *법경제학연구* 20(1), (2023)

김송옥, “AI 법제의 최신 동향과 과제 - 유럽연합(EU) 법제와의 비교를 중심으로 –“, *공법학연구* 22(4), (2021)

김효은, “인공지능 편향식별의 공정성 기준과 완화”, *한국심리학회지* 40(4), (2021)

대니얼 카너먼, 이창신 옮김, *생각에 관한 생각—우리의 행동을 지배하는 생각의 반란*, 김영사, (2018)

레이 커즈와일, 김명남 옮김, *특이점이 온다: 기술이 인간을 초월하는 순간*, 김영사, (2007)

메리 그레이 & 시다스 수리, 신동숙 옮김, *고스트워크: 모든 일과 직업의 모습이 바뀐다*, 한스미디어 (2019).

박은정 외, *차별판단기준 수립을 위한 연구(제1부)*, 국가인권위원회, (2021)

사피야 노블, 노윤기 옮김, *구글은 어떻게 여성을 차별하는가*, 한즈미디어, (2018)

성윤희, "행정기본법상 자동적 처분에 관한 연구", 이화여자대학교 대학원 법학과 박사학위 논문, (2023)

송지우, "인공지능과 언어 정의", 한소원 외, *호모마키나*, 커뮤니케이션북스, (2020)

송현정 외, *혐오 표현의 판단 기준에 관한 비교법적 연구*, 사법정책연구원, (2020)

쇼샤나 주보프, 김보영 옮김, *감시 자본주의 시대: 권력의 새로운 개척지에서 벌어지는 인류의 미래를 위한 투쟁*, 문학사상, (2021)

스티브 제이 굴드 지음, 김동광 옮김, *인간에 대한 오해*, 사회평론, (2003)

심재진, "미국의 차별적 영향 이론과 유럽연합의 간접차별 법리", *노동법 연구* 27, (2009)

신성필 · 하수욱 · 이강찬, "ISO/IEC JTC 1/SC 42 (인공지능) 국제 표준화 동향", *TTA저널* 제187호, 한국정보통신기술협회, (2020).

얼 바비 저, 고성호 외 옮김, *사회조사방법론* (제 13 판), 센게이지러닝, (2015)

이민열, "기본권 제한 심사에서 공익의 식별", *법철학연구* 18(2), (2015)

_____, "토머스 스캔론의 「불평등은 왜 문제인가?」와 헌법의 해석", 현상과 인식 45(2), (2021)

이병운, "미국의 고용차별금지법제에 관한 연구", *동아법학* 67, (2015)

이승길, "미국의 고용차별금지법 개관", *맞춤형 법제정보* 5, 한국법제연구원, (2015)

이재희, "헌법상 평등보장과 간접차별금지", 헌법재판소 헌법재판연구원, (2020)

장효재, "인공지능에 의한 공정성 훼손에 관한 탐구 : 빅데이터와 편향을 중심으로", 한국과학기술원 석사학위논문, (2021)

정석우 · 심현철, "자율주행 자동차의 인공지능", *기계저널* 57(3), (2017)

정소영, "유럽연합 인공지능법안의 거버넌스 분석 - 유럽인공지능위원회와 회원국 감독기관의 역할과 기능을 중심으로 –", *연세법학* 39, (2022)

정원섭, "인공지능 알고리즘의 편향성과 공정성", *인간·환경·미래* 25, (2020)

정주백. "아리스토텔레스 정식 비판", *헌법논총* 27, (2016)

최승철, *차별금지법의 이해*, 한울아카데미, (2011).

최윤희, “미국에서의 결과적 차별행위(Disparate Impact) 이론에 대한 고찰”, *노동법연구* 20, (2006)

최창수, “미국의 AI 채용절차 공정화 입법례”, *최신외국입법정보* 2023-1호, 국회도서관, (2023. 1. 31.)

크리스토퍼 비숍 저, 김형진 옮김, *패턴 인식과 머신 러닝*, 제이펍, (2019).

허영은, “채용, 왜 바꾸기 어려울까요?”, HQ Conference Q, (2022. 11. 17.)

홍석한, “유럽연합 ‘인공지능법안’의 주요 내용과 시사점”, *유럽헌법연구* (38), (2022)

홍성수, “포괄적 차별금지법의 필요성”, *이화젠더법학* 10(3), (2018)

_____, “혐오표현의 해악과 개입의 정당성: 금지와 방치를 넘어서”, 법철학연구 22(3), (2019)

나. 기관 등

개인정보 보호법 일부개정법률안(정부), 의안번호 2112723, (2021. 9. 28.)

과학기술정보방송통신위원회, “인공지능 육성 및 신뢰 기반 조성 등에 관한 법률안 검토보고”, (2021. 11.).

소병철 의원 대표발의, 채용절차의 공정화에 관한 법률 개정안, 의안번호 20449, (2023. 3. 6.)

윤두현 의원 대표발의, 인공지능산업 육성 및 신뢰 확보에 관한 법률안, 의안번호 2118726, (2022. 12. 7.)

윤영찬 의원 대표발의, 알고리즘 및 인공지능에 관한 법률안, 의안번호 2113509, (2021. 11. 24.)

정필모 의원 대표발의, 인공지능 육성 및 신뢰 기반 조성 등에 관한 법률안, 의안번호 2111261, (2021. 7. 1.)

다. 기사, 블로그

김병필, “인공지능의 세 가지 유형”, *중앙일보* (2020. 12. 28.)

세계법제정보센터, “미국 인공지능 법제”, (2019. 6. 30.)

임형주 외, “AI(인공지능)산업 육성 및 신뢰 기반 조성에 관한 법률 제정 동향 및 사업자가 주의해야 할 사항”, Yulchon Legal Update, (2023. 3.)

홍대식, “4차 산업혁명 시대의 공정성 개념에 대한 단상”, ICT법경제연구소, (2018. 2. 17.)

티맵모빌리티, “TMAP 운전습관 이용방법 & 운전점수 관리 꿀팁”, (2021. 5. 6.)

2. 국외문헌

가. 연구

Abdar Moloud, et al., “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”, *Information Fusion* 76, (2021)

Abid, Abubakar, Maheen Farooqi, and James Zou, “Persistent anti-muslim bias in large language models”, *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, (2021)

Alex, Neel, et al., “RAFT: A real-world few-shot text classification benchmark”, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, (2021)

Assefa, Samuel, et al., “Generating synthetic data in finance: opportunities, challenges and pitfalls”, *Proceedings of the First ACM International Conference on AI in Finance*, (2020)

Baker, Ryan S. and Aaron Hawn, “Algorithmic bias in education”, *International Journal of Artificial Intelligence in Education* 32, (2021)

Bambauer, Jane R, Tal Zarsky, and Jonathan Mayer, “When a Small Change Makes a Big Difference: Algorithmic Fairness Among Similar Individuals”, *UC Davis Law Review* 55, (2021)

Barocas, Solon and Andrew D. Selbst, “Big data's disparate impact”, *California Law Review* 104(3), (2016)

Barocas, Solon, Moritz Hardt, & Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, online textbook, (2022)

Bartlett, Robert, et al., “Consumer-lending discrimination in the FinTech era”, *Journal of Financial Economics* 143(1), (2022)

Bellamy, Rachel KE, et al., “AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias”, arXiv preprint arXiv:1810.01943 (2018)

Bellovin, Steven M., Preetam K. Dutta, and Nathan Reitinger, “Privacy and synthetic datasets”, *Stanford Technology Law Review* 22, (2019)

Bender, Emily, et al., "'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021)

Bengio, Yoshua, Yann Lecun and Geoffrey Hinton, "Deep learning for AI.", *Communications of the ACM* 64(7), (2021)

Benjamin, Ruha, *Race After Technology: Abolitionist Tools for the New Jim Code*, John Wiley & Sons, (2019)

Bıçakcı, Kaan, "Uncertainty in Deep Learning — Aleatoric Uncertainty and Maximum Likelihood Estimation", *Towards Data Science*, (2022. 2. 8.)

Bickel, Peter J., Eugene A. Hammel, and J. William O'Connell. "Sex bias in graduate admissions: Data from Berkeley." *Science* 187, (1975)

Birhane, Abeba, et al., "The values encoded in machine learning research", *2022 ACM Conference on Fairness, Accountability, and Transparency*, (2022)

Blattner, Laura and Scott Nelson, "How costly is noise? Data and disparities in consumer credit", arXiv preprint arXiv:2105.07554 (2021)

Blodgett, Su Lin, et al. "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020)

Bommasani, Rishi et al., "On the opportunities and risks of foundation models", arXiv preprint arXiv:2108.07258, (2021)

Bradford, Anu, "The Brussels Effect", *Northwestern University Law Review* 107, (2012)

Breitfeller, Luke, et al. "Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts", *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing*, (2019)

Bresnahan, Timothy F. and Manuel Trajtenberg, "General purpose technologies 'Engines of growth'?", *Journal of Econometrics* 65(1), (1995)

Brown, Tom, et al., "Language models are few-shot learners", *Advances in neural information processing systems* 33, (2020)

Bruckner, Matthew Adam, "The promise and perils of algorithmic lenders' use of big data", *Chicago-Kent Law Review* 93, (2018)

Bubeck, Sébastien, et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4", (2023)

Buolamwini, Joy and Timnit Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification”, *Conference on fairness, accountability and transparency*, (2018)

Casey, Bryan, Ashkon Farhangi, and Roland Vogl. “Rethinking Explainable Machines: The GDPR's' Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise”, *Berkeley Technology Law Journal* 34, (2019)

Chang, Hasok, “Operationalism”, *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.)

Chen, Tianqi and Carlos Guestrin, “XGBoost: A scalable tree boosting system”, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, (2016).

Chesney, Bobby and Danielle Citron, “Deep fakes: A looming challenge for privacy, democracy, and national security”, *California Law Review* 107, (2019)

Choi, Yejin, “The curious case of commonsense intelligence”, *Daedalus* 151(2), (2022)

Chouldechova, Alexandra, et al., “A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions”, *Conference on Fairness, Accountability and Transparency*, (2018)

Crawford, Kate, “The Trouble with Bias”, *NIPS 2017 keynote speech*, (2017)

Cofone, Ignacio N., “Algorithmic Discrimination Is an Information Problem”, *Hastings Law Journal* 70(6), (2019a)

_____, “Antidiscriminatory privacy”, *SMU Law Review* 72 (2019b)

Coglianese, Cary and Alicia Lai, “Algorithm vs. Algorithm”, *Duke Law Journal* 72 (2022)

Collins, Rebecca L., “Content analysis of gender roles in media: Where are we now and where should we go?” *Sex Roles*, 64, (2011)

Corbett-Davies, Sam, et al. “Algorithmic decision making and the cost of fairness” *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, (2017)

Datta, Arindrima, et al., “Language-agnostic multilingual modeling”, *ICASSP 2020*, (2020)

Edwards, Lilian and Michael Veale, “Enslaving the algorithm: From a ‘Right to an Explanation’ to a ‘Right to Better Decisions’?”, *IEEE Security & Privacy*, 16(3), (2018)

El Emam, Khaled, Lucy Mosquera, and Richard Hoptroff, *Practical Synthetic Data Generation* (Kindle Edition), O'Reilly Media, (2020)

Ethayarajh, Kawin and Dan Jurafsky, “Utility is in the eye of the user: A critique of NLP leaderboards”, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2020)

Eubanks, Virginia, *Automating inequality: How high-tech tools profile, police, and punish the poor*, St. Martin's Press, (2018) # 26

Darwiche, Adnan, “Human-level intelligence or animal-like abilities?”, *Communications of the ACM* 61(10), (2018)

Dastin, Jeffrey, “Amazon scraps secret AI recruiting tool that showed bias against women”, *Ethics of data and analytics*, Auerbach Publications, (2018)

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. “Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination”, *In Proceedings on Privacy Enhancing Technologies*, (2015)

Dhamala, Jwala, et al., “Bold: Dataset and metrics for measuring biases in open-ended language generation”, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (2021)

Doleac, Jennifer L. and Benjamin Hansen, “Does ‘ban the box’ help or hurt low-skilled workers? Statistical discrimination and employment outcomes when criminal histories are hidden” No. w22469, *National Bureau of Economic Research*, (2016)

Dwork, Cynthia, et al., “Fairness through awareness”, *Proceedings of the 3rd innovations in theoretical computer science conference*, (2012)

Fennell, Lee Anne and Richard H. McAdams, “Fairness in Law and Economics: Introduction”, *University of Chicago Coase-Sandor Institute for Law & Economics Research Paper* 704, (2013)

Fennell, Lee Anne and Richard H. McAdams (eds.), *Fairness in Law and Economics*, Edward Elgar, (2013)

Fiss, Owen M., “Groups and the Equal Protection Clause”, *Philosophy & Public Affairs* 107 (1976)

Frank, Robert H., *Success and Luck*, Princeton University Press, (2016)

Fuster, Andreas, et al., “Predictably unequal? The effects of machine learning on credit markets”, *The Journal of Finance* 77(1), (2022)

Gehman, Samuel, et al., “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models,” *Findings of the Association for Computational Linguistics: EMNLP 2020*, (2020)

Gillis, Talia B. and Jann L. Spiess, “Big Data and Discrimination”, University of Chicago Law Review 86(2), (2019)

Goldstein, Josh A., et al., “Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations”, arXiv preprint arXiv:2301.04246, (2023)

Goodfellow Ian, Yoshua Bengio and Aaron Courville, *Deep Learning*, MIT Press, (2016)

Goodman, Bryce and Seth Flaxman, “EU regulations on algorithmic decision-making and a ‘right to explanation’, *In ICML workshop on human interpretability in machine learning (WHI 2016)*, New York, NY, (2016, June)

Greaves, Hilary and William MacAskill, “The case for strong longtermism”, *Global Priorities Institute Working Paper* No. 5-2021, (2021)

Grgic-Hlaca, Nina, et al. “Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction”, *Proceedings of the 2018 World Wide Web Conference*, (2018)

Hardt, Moritz, Eric Price, and Nati Srebro, “Equality of opportunity in supervised learning”, *Advances in neural information processing systems* 29, (2016)

Hedden, Brian, “On statistical criteria of algorithmic fairness”, *Philosophy and Public Affairs* 49(2), (2021)

Hellman, Deborah, “Measuring algorithmic fairness”, *Virginia Law Review* 106(4), (2020)

Hendrycks, Dan, et al., “Measuring massive multitask language understanding”, arXiv preprint arXiv:2009.03300, (2020)

Hendricks, Lisa Anne, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach, “Women also snowboard: Overcoming bias in captioning models”, *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018)

Hertweck, Corinna, Christoph Heitz, and Michele Loi. “On the moral justification of statistical parity”, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (2021)

Holm, Sune, “The Fairness in Algorithmic Fairness”, *Res Publica*, (2022)

Hu, Lily and Yiling Chen, “A short-term intervention for long-term fairness in the labor market”, *Proceedings of the 2018 World Wide Web Conference*, (2018)

Hurley, Mikella and Julius Adebayo “Credit scoring in the era of big data”, *Yale Journal of Law & Technology* 18, (2016)

Joshi, Pratik, et al., “The State and Fate of Linguistic Diversity and Inclusion in the NLP World”, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020)

Kamiran, Faisal, Indrė Žliobaitė, and Toon Calders, “Quantifying explainable discrimination and removing illegal discrimination in automated decision making”, *Knowledge and Information Systems* 35, (2013)

Kaplow, Louis and Steven Shavell, “The conflict between notions of fairness and the Pareto principle”, *American Law and Economics Review* 1(1), (1999)

Kaplow, Louis and Steven Shavell, “Should legal rules favor the poor? Clarifying the role of legal rules and the income tax in redistributing income”, *The Journal of Legal Studies* 29(2), (2000)

Kay, Matthew, Cynthia Matuszek, and Sean A Munson, “Unequal representation and gender stereotypes in image search results for occupations”, *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, (2015)

Kehl, Danielle, Priscilla Guo, and Samuel Kessler, “Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing”, *Responsive Communities Initiative*, Berkman Klein Center for Internet & Society, Harvard Law School, (2017)

Kim, Pauline, “Data-driven discrimination at work”, *William & Mary Law Review* 58, (2016)

_____, “Auditing algorithms for discrimination”, *University of Pennsylvania Law Review* 166 (2017)

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan, “Inherent trade-offs in the fair determination of risk scores”, arXiv preprint arXiv:1609.05807 (2016)

Kleinberg, Jon, et al., “Discrimination in the Age of Algorithms” *Journal of Legal Analysis* 10 (2018a)

Kleinberg, Jon, et al., “Human decisions and machine predictions”, *The quarterly journal of economics* 133.1 (2018b)

Koch, Timo K., et al., “Age and Gender in Language, Emoji, and Emoticon Usage in Instant Messages”, *Computers in Human Behavior* 126, (2022)

Kroll, Joshua A., et al., “Accountable algorithms”, *University of Pennsylvania Law Review* 165 (2017).

Koshiyama, Adriano Soares, et al., “Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms”, *Technical Report*, (2021)

Koulish, Robert, “Immigration detention in the risk classification assessment era”, *Connecticut Public Interest Law Journal* 16, (2017)

Lakshminarayanan, Balaji, “Introduction to Uncertainty in Deep Learning”, *CIFAR Deep Learning + Reinforcement Learning DLRL Summer School*, (2021)

Lambrecht, Anja and Catherine E. Tucker, “Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads”, *Management Science* 65(7), (2019)

Lee, Hwaran, et al., “SQuARe: A Large-Scale Dataset of Sensitive Questions and Acceptable Responses Created through Human-Machine Collaboration”, *ACL 2023* (forthcoming), (2023)

Lee, Michelle Seng Ah and Jatinder Singh, “The Landscape and Gaps in Open Source Fairness Toolkits,” *CHI Conference on Human Factors in Computing Systems*, (2021)

Lee, Michelle Seng Ah, Luciano Floridi, and Jatinder Singh. “Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics”, *AI and Ethics* 1(4), (2021)

Levy, Karen, Kyla E. Chasalow, and Sarah Riley, “Algorithms and decision-making in the public sector”, *Annual Review of Law and Social Science* 17, (2021)

Lewandowski, Dirk, “Is Google Responsible for Providing Fair and Unbiased Results?”, Taddeo, Mariarosaria & Floridi, Luciano (ed), *The Responsibilities of Online Service Providers*, Springer International Publishing A&G (2017)

Liang, Percy, et al., “Holistic evaluation of language models” arXiv preprint arXiv:2211.09110 (2022)

Lin, Chin-Yew, “ROUGE: a Package for Automatic Evaluation of Summaries”, *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, (2004)

Lippert-Rasmussen, Kasper, *Making sense of affirmative action*, Oxford University Press, (2020)

Liu, Lydia T. et al., “Delayed impact of fair machine learning”, *International Conference on Machine Learning*, (2018)

Luccioni, Alexandra Sasha, et al., “Stable Bias: Analyzing Societal Representations in Diffusion Models”, arXiv preprint arXiv:2303.11408 (2023)

Ludwig, Jens and Sendhil Mullainathan, “Fragile algorithms and fallible decision-makers: lessons from the justice system”, *Journal of Economic Perspectives* 35(4), (2021)

McCulloch, Warren and Walter Pitts, “A logical calculus of the ideas immanent in nervous activity”, *The bulletin of mathematical biophysics* 5, (1943)

Mayson, Sandra G., "Bias in, bias out”, *Yale Law Journal* 128, (2019)

Mehrabi, Ninareh, et al. “A survey on bias and fairness in machine learning”, *ACM Computing Surveys (CSUR)* 54(6), (2021)

Metaxa, Danae, et al., “An image of society: Gender and racial representation and impact in image search results for occupations”, *Proceedings of the ACM on Human-Computer Interaction* 5, (2021)

Mindell, David and Elisabeth Reynolds, *The work of the future: Building better jobs in an age of intelligent machines*, MIT Press, (2022).

Mitchell, Margaret, et al., “Diversity and inclusion metrics in subset selection”, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, (2020)

Mitchell, Shira, et al., “Algorithmic fairness: Choices, assumptions, and definitions”, *Annual Review of Statistics and Its Application*, 8 (2021)

Mökander, Jakob, et al., “The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other?”, *Minds and Machines* 32, (2022)

Nagel, Thomas, “The problem of global justice”, *Philosophy & Public Affairs*, Vol. 33, No. 2 (2005)

Narayanan, Arvind, “21 fairness definitions and their politics”, *Conference on Fairness, Accountability, and Transparency* (2018. 2. 23)

Nativi, Stefano and Sarah De Nigris, *AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework*, EUR 30772 EN, Publications Office of the European Union, (2021)

Nee, Julia, et al., “Linguistic Justice as a Framework for Designing, Developing, and Managing Natural Language Processing Tools”, *Big Data & Society* 9(1), (2022)

Ng, Andrew, “What AI Can and Can’t Do”, *Harvard Business Review*, (2016)

Obermeyer, Ziad, et al. “Dissecting racial bias in an algorithm used to manage the health of populations”, *Science* 366, (2019)

Oppenheimer, David B., et al., *Comparative Equality and Anti-discrimination Law*, Edward Elgar Publishing, (2020).

Orlando, Domenico, “If science fiction inspires the law: recent examples with the EU proposal for an AI Regulation”, (2021)

Ouyang, Long, et al., “Training language models to follow instructions with human feedback”, *Advances in Neural Information Processing Systems* 35 (2022)

Papineni, Kishore, et al., “BLEU: a Method for Automatic Evaluation of Machine Translation”, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, (2001)

Park, Sangchul, “Heterogeneity of AI-Induced Societal Harms and the Failure of Omnibus AI Laws”, arXiv preprint arXiv:2303.11196 (2023)

Pavlopoulos, John, et al. “Toxicity detection: Does context really matter?”, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020)

Pearl, Judea and Dana Mackenzie, *The book of why: the new science of cause and effect*, Basic books, (2018)

Perez, Ethan, “Red Teaming Language Models with Language Models”, arXiv preprint arXiv:2202.03286 (2022)

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy, “Mitigating bias in algorithmic hiring: Evaluating claims and practices”, *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (2020)

Raji, Inioluwa Deborah, et al., “Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing”, *Proceedings of the 2020 conference on fairness, accountability, and transparency*, (2020)

Ramesh, Aditya, et al., “Hierarchical text-conditional image generation with CLIP latents”, arXiv preprint arXiv:2204.06125 (2022)

Rauh, Maribeth, et al., “Characteristics of harmful text: Towards rigorous benchmarking of language models”, arXiv preprint arXiv:2206.08325, (2020)

Reynolds, Laria and Kyle McDonell, “Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm”, *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, (2021)

Russel, Stuart and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th edition, Pearson, (2020)

Salem, Jad, Deven R. Desai, and Swati Gupta, “Don’t let Ricci v. DeStefano Hold You Back: A Bias-Aware Legal Solution to the Hiring Paradox”, *Proceedings of the 2022 conference on fairness, accountability, and transparency*, (2022).

Sánchez-Monedero, Javier, Lina Dencik, and Lilian Edwards, “What does it mean to ‘solve’ the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems”, *Proceedings of the 2020 conference on fairness, accountability, and transparency*, (2020)

Scanlon, Thomas, *Why does inequality matter?*, Oxford University Press, (2018)

Schmidt, Anna and Michael Wiegand, “A survey on hate speech detection using natural language processing”, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, (2017)

Schwartz, Paul M. and Karl-Nikolaus Peifer. “Transatlantic data privacy law” *Geo. LJ* 106 (2017)

Selbst, Andrew and Julia Powles. “Meaningful Information and the Right to Explanation.”, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, (2018)

Singh, Vivek K, et al., “Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms”, *Journal of the Association for Information Science and Technology* 71(11), (2020)

Smith, Helen, “Algorithmic bias: should students pay the price?”, *AI & Society* 35(4), (2020)

Solove, Daniel J., “The Limitations of Privacy Rights”, *Notre Dame Law Review* 98, (2023)

Sprenger, Jan and Naftali Weinberger, “Simpson’s Paradox”, *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.)

Sue, Derald Wing, et al., “Racial microaggressions in everyday life: implications for clinical practice”, *American Psychologist* 62(4), (2007)

Surden, Harry, “Artificial Intelligence and Law: An Overview”, *Georgia State University Law Review* 35, (2019).

Taulli, Tom, *Artificial Intelligence Basics: A Non-Technical Introduction*, Apress, (2020)

Tran, Dustin, Balaji Lakshminarayanan, Jasper Snoek, “Practical Uncertainty Estimation and Out-of-Distribution Robustness in Deep Learning”, *NeurIPS 2020 Tutorial*, (2020).

Turner, Ronald, “The way to stop discrimination on the basis of race…”, *Stanford Journal of Civil Rights & Civil Liberties* 11, (2015)

Tyler, Tom R., “Procedural fairness and compliance with the law”, *Revue Suisse D Economie Politique et de Statistique* 133, (1997)

Vaswani, Ashish, et al., “Attention is all you need”, *Advances in neural information processing systems* 30, (2017)

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi, “Why a right to explanation of automated decision-making does not exist in the general data protection regulation”, *International Data Privacy Law*, 7(2), (2017)

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. “Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI”, *Computer Law & Security Review* 41 (2021a)

_____, “Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law”, *West Virginia Law Review* 123(3), (2021b)

Wang, Alex, et al. “GLUE: A multi-task benchmark and analysis platform for natural language understanding”, *International Conference on Learning Representations (ICLR)*, (2019)

Waseem, Zeerak, et al., “Understanding Abuse: A Typology of Abusive Language Detection Subtasks”, *Proceedings of the First Workshop on Abusive Language Online*, (2017)

Schwartz, Paul M. and Karl-Nikolaus Peifer. “Transatlantic data privacy law”, *Georgetown Law Journal* 106, (2017)

Sheng, Emily, et al., “The Woman Worked as a Babysitter: On Biases in Language Generation”, *Empirical Methods in Natural Language Processing*, (2019).

Shwartz-Ziv, Ravid and Amitai Armon, “Tabular data: Deep learning is not all you need”, *Information Fusion* 81, (2022)

Sullivan, Charles A., “Employing AI”, *Villanova Law Review* 63, (2018)

Sweeney, Latanya, “Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising”, *Queue* 11(3), (2013)

Tian, Zhiyi, et al., “A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning”, *ACM Computing Surveys* 55(8), (2023)

Vaithianathan, Rhema, et al., “Children in the public benefit system at risk of maltreatment: Identification via predictive modeling”, *American journal of preventive medicine* 45(3), (2013)

van Breugel, Boris, et al., “DECAF: Generating fair synthetic data using causally-aware generative networks”, *Advances in Neural Information Processing Systems 34*, (2021)

Wei, Jason, et al., “Finetuned language models are zero-shot learners”, arXiv preprint arXiv:2109.01652, (2021)

Wei, Jason, et al., “Emergent abilities of large language models”, arXiv preprint arXiv:2206.07682, (2022)

Wei, Jason, et al., “Chain of thought prompting elicits reasoning in large language models”, arXiv preprint arXiv:2201.11903 (2022)

Weidinger, Laura, et al., “Ethical and social risks of harm from language models”, arXiv preprint arXiv:2112.04359, (2021)

Zafar, Muhammad Bilal, et al., “Fairness constraints: A flexible approach for fair classification”, *The Journal of Machine Learning Research* 20.1 (2019)

Zhao, Jieyu, et al., “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints”, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, (2017)

나. 기관 등

Article 29 Working Party, “Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679”, (2018)

Competition and Markets Authority, *AI Foundation Models: Initial review*, (2023. 5. 4.).

Council of the European Union Analysis and Research Team, “ChatGPT in the Public Sector – overhyped or overlooked?”, (2023. 4. 24.)

Department of Housing and Urban Development (HUD), Implementation of the Fair Housing Act’s Discriminatory Effects Standard, 78 Fed. Reg. 11460 (2013)

European Commission, Joint Research Centre, Nativi, S., De Nigris, S., AI Watch, “AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework”, *Publications Office*, (2021)

_____, *2018 Code of Practice on Disinformation*

Equal Employment Opportunity Commission, Civil Service Commission, et al. Uniform Guidelines on Employee Selection Procedures, (1978)

Garante per la protezione dei dati personali, “Artificial intelligence: stop to ChatGPT by the Italian SA Personal data is collected unlawfully, no age verification system is in place for children”, (2023. 3. 31.)

High-Level Expert Group on Artificial Intelligence (AI HLEG), *A definition of Artificial Intelligence: main capabilities and scientific disciplines*, (2019. 4. 8)

ICO, “ICO consultation on the draft AI auditing framework guidance for organisations.”, (2020. 2. 19.)

ILO Global Commission on the Future of Work, *Work for a brighter future*, (2019. 1. 22.).

Institute of Electrical and Electronics Engineers (IEEE), *Ethically Aligned Design*, (2019)

Monetary Authority of Singapore (MAS), *Fairness Metrics to Aid Responsible AI Adoption in Financial Services*, (2020)

_____, “Veritas Initiative Addresses Implementation Challenges in the Responsible Use of Artificial Intelligence and Data Analytics”, Media Releases, (2021. 1. 6.)

National Institute of Standards and Technology (NIST), “NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software”, NIST News (2019. 12. 19.)

_____, *AI Risk Management Framework: Initial Draft*, (2022)

New York Department of Financial Services (NYDFS), *Report on Apple Card Investigation*, (2021. 3.)

OECD, *Recommendation of the Council on Artificial Intelligence*, OECD Publishing, (2019)

Office of Management and Budget (OMB), “Guidance for Regulation of Artificial Intelligence Applications”, (2020. 11. 17.)

Office of U.S. Senator Ron Wyden (2022), “Algorithmic Accountability Act of 2022”

Stanford Human-Centered Artificial Intelligence, “Artificial Intelligence Index Report 2023”, (2023)

The White House, “Blueprint for an AI Bill of Rights” (2022)

United Nation, *Disinformation and Freedom of Opinion and Expression: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, (2021)

다. 기사, 블로그 등

Alba, Davey and Julia Love, “Google’s Rush to Win in AI Led to Ethical Lapses, Employees Say”, *Bloomberg* (2023. 4. 19.)

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner, “Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks”, *ProPublica*, (2016. 5. 23.)

AcroYogi, “The Sydney Prompt / Bing AI 1.0: One Prompt to Rule Them All”, (2023. 2. 16.)

Angwin, Julia, et al., “Minority neighborhoods pay higher car insurance premiums than white areas with the same risk”, *ProPublica*, (2017. 4. 5.)

BBC, “Google apologises for Photos app's racist blunder”, *BBC News*, (2015. 7. 1.)

BBC, “Alexa Tells 10-Year-Old Girl to Touch Live Plug with Penny”, *BBC News*, (2021. 12. 28.)

Bosma, Maarten, “Introducing FLAN: More generalizable Language Models with Instruction Fine-Tuning”, *Google Research Blog*, (2021. 10. 6.)

Cotra, Ajeya, “Two-year update on my personal AI timelines”, *AI Alignment Forum*, (2022. 8. 3.)

Cotterill, Tom, “Paedophiles are using AI to create child abuse images: National Crime Agency warns artificial intelligence is being harnessed to make pictures and 'deep fake' videos of real-life victims”, *Mail Online*, (2023. 2. 19.)

Diaz, Maria, “How your kids can use ChatGPT safely, according to a mom”, *ZDNet*, (2023. 4. 17.)

Dieterich, William, Christina Mendoza, and Tim Brennan, “COMPAS risk scales: Demonstrating accuracy equity and predictive parity”, *Northpointe Inc*, (2016)

Engler, Alex, “The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment”, *Brookings*, (2023. 4. 25.)

Equality Now, “ChatGPT-4 Reinforces Sexist Stereotypes By Stating A Girl Cannot ‘Handle Technicalities And Numbers’ In Engineering”, News and Insights (2023. 3. 23.)

European Commission, *Impact Assessment of the Regulation on Artificial intelligence*, (2021. 4. 21.)

_____, Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence, (2022. 12. 5.)

European Union Agency for Fundamental Rights and Council of Europe, *Handbook on European Non-Discrimination Law*, Publications Office of the European Union, (2018)

Freshfields Bruckhaus Deringer, “Has copyright caught up with the AI Act?”, *Lexology Blog* (2023. 5. 16.)

Friedler, Sorelle, Suresh Venkatasubramanian, and Alex Engler, “How California and other states are tackling AI legislation”, *Brookings*, (2023. 3. 22.)

Future of Life Institute, *General Purpose AI and the AI Act*, (2022. 5.)

_____, “Pause Giant AI Experiments: An Open Letter”, (2023. 3. 22.)

High-Level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines for Trustworthy AI*, (2019. 4. 8)

_____, *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*, (2020. 7. 17)

Hu, Krystal, “ChatGPT sets record for fastest-growing user base - analyst note”, *Reuters*, (2023. 2. 3.)

Jeff Larson, et al., “How we analyzed the COMPAS recidivism algorithm”, *ProPublica*, (2016).

Kapoor, Sayash and Arvind Narayanan, “A misleading open letter about sci-fi AI dangers ignores the real risks”, (2023. 3. 30.)

Kelly, Kevin, “What AI-Generated Art Really Means for Human Creativity”, *Wired*, (2022. 11. 17.)

Kilcher, Yannic, “GPT-4chan”, *GitHub*, (2022. 6. 4.)

Kurenkov, Andrey, “Lessons from the GPT-4Chan Controversy”, *The Gradient*, (2022. 6. 12.)

Lee, Peter, “Learning from Tay’s introduction”, *Official Microsoft Blog*, (2016. 3. 25.)

Liang, Percy, et al., “CS324 lecture notes (Winter 2022)”

Liang, Percy and Rob Reich. “Condemning the deployment of gpt-4chan”, *Google Docs*, (2022)

Manikandan, Bala, “Demystifying ChatGPT: A Deep Dive into Reinforcement Learning with Human Feedback”, *Medium*, (2023. 2. 2.)

Manning, Chris and Pandu Nayak, “Introduction to Information Retrieval – Evaluation”, Stanford University. (2013. 4. 21.)

Miller, Claire Cain, “Can an algorithm hire better than a human?”, *New York Times*, (2015. 6. 26.)

MIT Technology Review, “Video: Geoffrey Hinton talks about the ‘existential threat’ of AI”, Speech at EM Tech Digital (2023. 5. 3.)

Narang, Sharan and Aakanksha Chowdhery, “Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance”, *Google Research Blog*, (2022. 4. 4.)

Narayanan, Arvind and Sayash Kapoor, “The LLaMA is out of the bag. Should we expect a tidal wave of disinformation?”, *AI Snake Oil* (blog), (2023. 3. 7.)

Narayanan, Arvind, “I set up a ChatGPT voice interface for my 3-year old. Here’s how it went.”, *AI Sanke Oil* (blog), (2023. 4. 15.)

Nasiripour, Shahien, Jenny Surane and Sridhar Natarajan, “Apple Card’s Gender-Bias Claims Look Familiar to Old-School Banks”, *Bloomberg Businessweek* (2019. 11. 12.)

Ng Andrew, “A Chat with Andrew on MLOps: From Model-centric to Data-centric AI”, *YouTube*, (2021. 3. 25.)

_____, The Batch Weekly Issues, *DeepLearning.AI*, (2023. 5. 17.)

Odenthal, Markus, “Multi-Class vs. Multi-Label Classification: What’s the Difference?”, *Medium*, (2022. 5. 19.)

OpenAI, “Introducing ChatGPT”, *OpenAI Blog*, (2022. 11. 30.)

_____, “Reducing bias and improving safety in DALL·E 2”, *OpenAI Blog*, (2022. 7. 18.)

_____, GPT-4 Technical Report, (2023)

_____, Terms of use, (2023. 3. 14.)

_____, Usage policies, (2022. 3. 23.)

Perrigo, Billy, “Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic”, *Time*, (2023. 1. 18.)

PetaPixel, “‘Racist’ Camera Phenomenon Explained — Almost”, (2010. 1. 22.)

Pouget, Hadrien, “Standard Setting”, The Artificial Intelligence Act (webstite)

Reuters, “Microsoft’s AI Twitter bot goes dark after racist, sexist tweets”, (2016. 3. 25)

Sabini, Mark, “What is Visual Prompting”, Landing AI, (2023. 4. 26.)

Stanford Human-Centered Artificial Intelligence, “Jazmia Henry: Building Inclusive NLP”, (2022. 8. 29.)

Taylor, Josh, “ChatGPT’s alter ego, Dan: users jailbreak AI program to get around ethical safeguards”, *Guardian*, (2023. 3. 8.)

The Sparrow team, “Building safer dialogue agents”, *Google DeepMind Blog*, (2022. 9. 22.)

Vahdat, Arash and Karsten Kreis, “Improving Diffusion Models as an Alternative To GANs, Part 1”, *NVIDIA Technical Blog*, (2022. 4. 26.)

Warren, Tom, “These are Microsoft’s Bing AI secret rules and why it says it’s named Sydney”, *The Verge*, (2023. 2. 15.)

Wattenberg, Martin, Fernanda Viégas, and Moritz Hardt., “Attacking discrimination with smarter machine learning”, http://research.google.com/bigpicture/attacking-discrimination-in-ml/

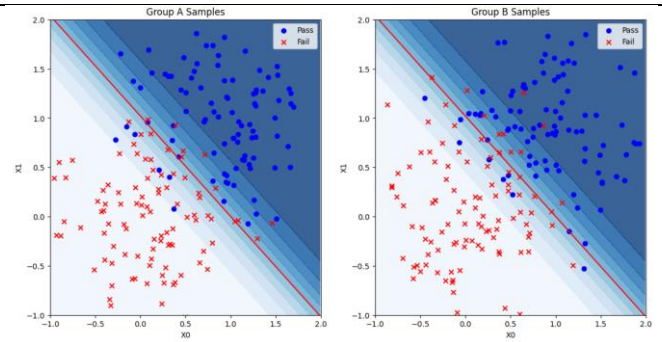
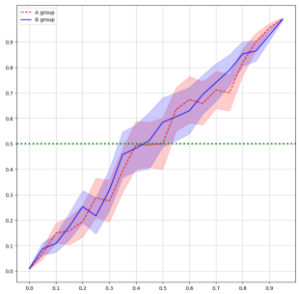
Wodecki, Ben, “WAICF ’23: Yann LeCun Sets Us Straight on Generative AI”, *AI Business*, (2023. 2. 11.)

Yong, Ed, “A Popular Algorithm Is No Better at Predicting Crimes Than Random People”, *The Atlantic*, (2018. 1. 18.)

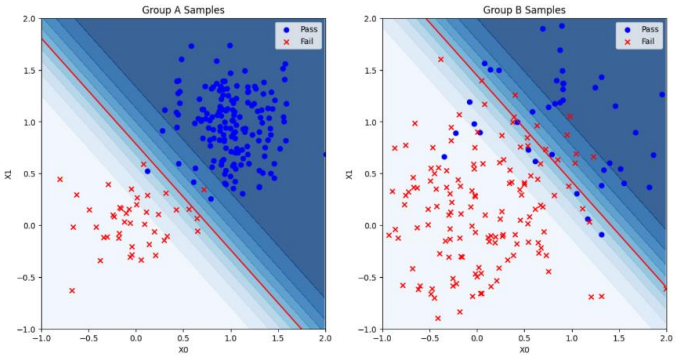
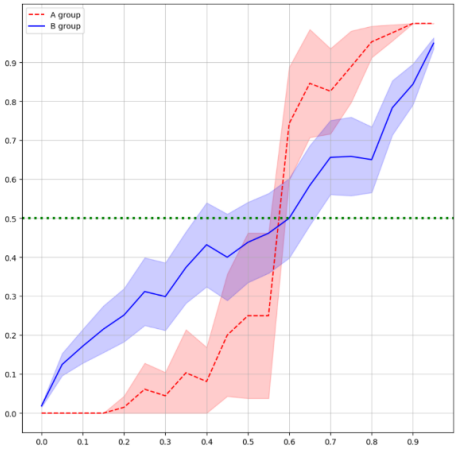
별 지 – 시뮬레이션의 상세

1. 인지모형 시나리오

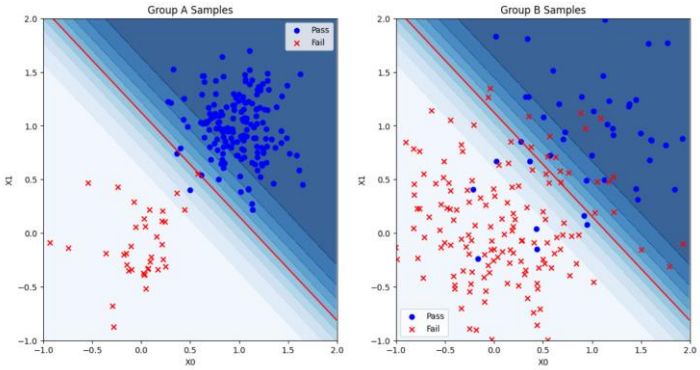
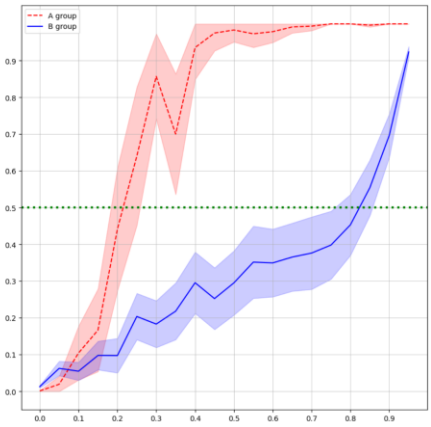
가. 베이스라인 시나리오 (사회적 편향 ×, 데이터 측정 편향 ×, 보호속성 입력 ×)

학습데이터	**목표변수 분포** 집단 A 50% Y=1, 50% Y=0 총 5000 건 집단 B 50% Y=1, 50% Y=0 총 5000 건 **특성변수 분포** X0: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.5) X1: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.5) **보호속성 입력하지 않음**
평가데이터	학습 데이터와 동일하게 생성(A, B 집단 각 5,000 건)
모델 학습	sklearn.linear_model. LogisticRegression 기본설정
학습 결과	
Calibration	

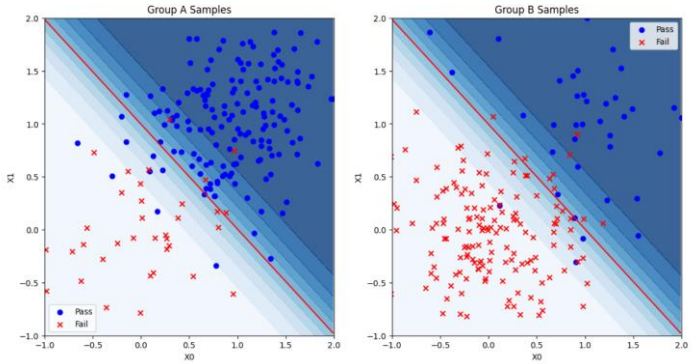
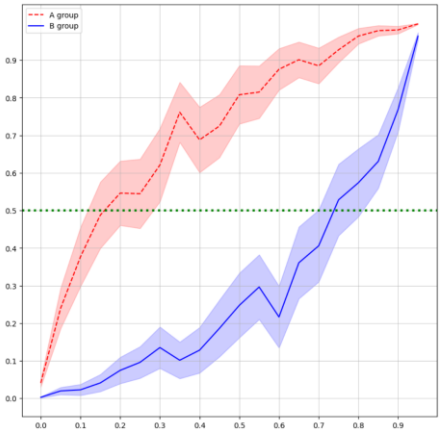
나. 차별적 시나리오 (1-1) (사회적 편향 ○, 데이터 측정 편향 ○, 보호속성 입력 ○)

학습데이터	**목표변수 분포** 집단 A 80% Y=1, 20% Y=0 총 10,000 건 집단 B 20% Y=1, 80% Y=0 총 10,000 건 **특성변수 분포** X0: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.3) X1: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.6) **X2: 보호속성 입력**
평가데이터	학습 데이터와 동일하게 생성(A, B 집단 각 10,000 건)
모델 학습	sklearn.linear_model. LogisticRegression 기본설정
학습 결과	
Calibration	

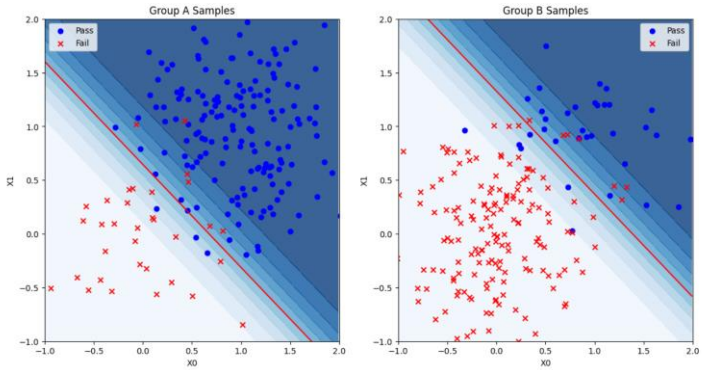
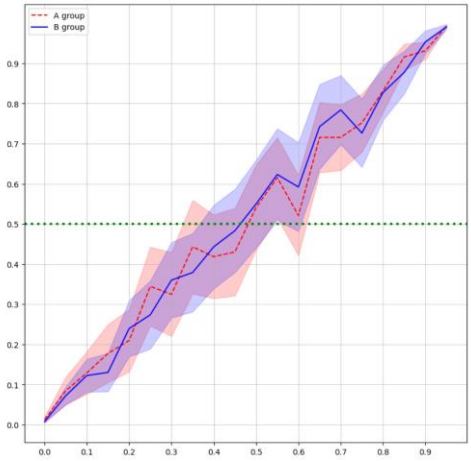
다. 차별적 시나리오 (1-2) (사회적 편향 ○, 데이터 측정 편향 ○, 보호속성 입력 ×)

학습데이터	**목표변수 분포** 집단 A 80% Y=1, 20% Y=0 총 10,000 건 집단 B 20% Y=1, 80% Y=0 총 10,000 건 **특성변수 분포** X0: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.3) X1: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.6) **보호속성 입력하지 않음**
평가데이터	학습 데이터와 동일하게 생성(A, B 집단 각 10,000 건)
모델 학습	sklearn.linear_model. LogisticRegression 기본설정
학습 결과	
Calibration	

라. 차별적 시나리오 (1-3) (사회적 편향 ○, 데이터 측정 편향 ×, 보호속성 입력 ×)

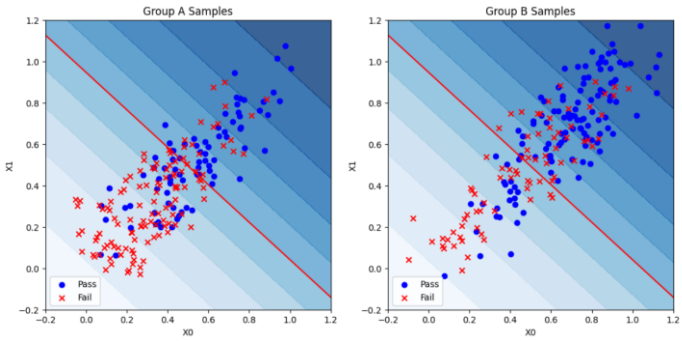
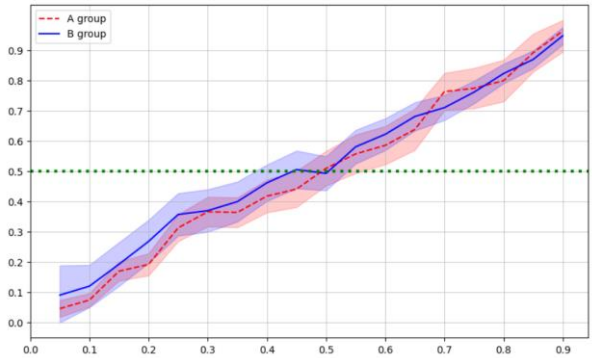
학습데이터	**목표변수 분포** 집단 A 80% Y=1, 20% Y=0 총 10,000 건 집단 B 20% Y=1, 80% Y=0 총 10,000 건 **특성변수 분포** X0: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.5) X1: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.5) **보호속성 입력하지 않음**
평가데이터	학습 데이터와 동일하게 생성(A, B 집단 각 10,000 건)
모델 학습	sklearn.linear_model. LogisticRegression 기본설정
학습 결과	
Calibration	

마. 차별적 시나리오 (1-4) (사회적 편향 ○, 데이터 측정 편향 ×, 보호속성 입력 ○)

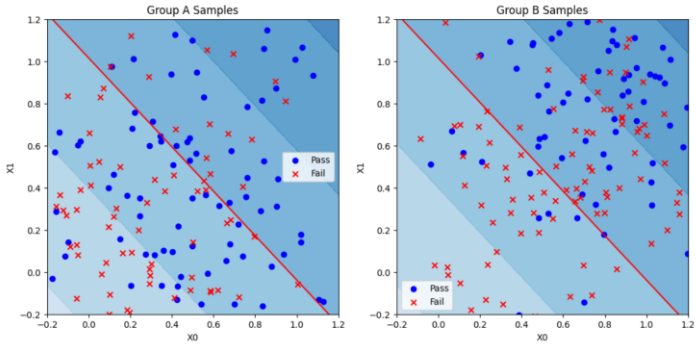
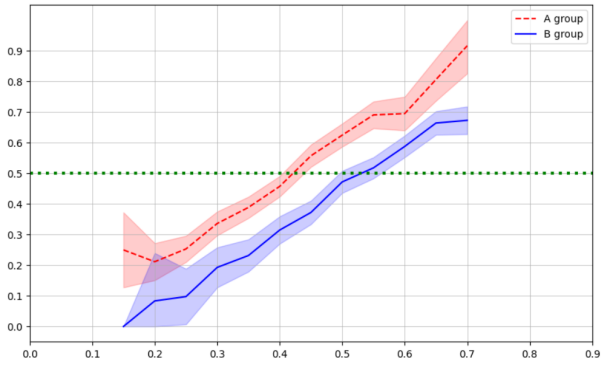
학습데이터	**목표변수 분포** 집단 A 80% Y=1, 20% Y=0 총 10,000 건 집단 B 20% Y=1, 80% Y=0 총 10,000 건 **특성변수 분포** X0: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.5) X1: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.5) **X2: 보호속성 입력**
평가데이터	학습 데이터와 동일하게 생성(A, B 집단 각 10,000 건)
모델 학습	sklearn.linear_model. LogisticRegression 기본설정
학습 결과	
Calibration	

2. 판별모형 시나리오

가. 시나리오 (2-1) (사회적 편향 ○, 데이터 측정 편향 ×, 보호속성 입력 ×)

학습데이터	**목표변수 분포** 집단 A: p_i ~ 절단정규분포(평균 0.3, 표준편차 0.3) 총 5,000 건 집단 A: p_i ~ 절단정규분포(평균 0.7, 표준편차 0.3) 총 5,000 건 **특성변수 분포** X0: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.1) X1: 목표변수 값 + 정규분포 오차(평균 0, 표준편차 0.1) **보호속성 입력하지 않음**
평가데이터	학습 데이터와 동일하게 생성(A, B 집단 각 5,000 건)
모델 학습	sklearn.linear_model. LogisticRegression 기본설정
학습 결과	
Calibration	

나. 시나리오 (2-2) (사회적 편향 ×, 데이터 측정 편향 ○, 보호속성 입력 ×)

학습데이터	**목표변수 분포** 집단 A: p_i ~ 절단정규분포(평균 0.3, 표준편차 0.3) 총 5,000 건 집단 A: p_i ~ 절단정규분포(평균 0.7, 표준편차 0.3) 총 5,000 건 **특성변수 분포** X0: 목표변수 값 + 정규분포 오차(평균 -0.2(집단 A)/+0.2(집단 B), 표준편차 0.4) X1: 목표변수 값 + 정규분포 오차(평균 -0.2(집단 A)/+0.2(집단 B), 표준편차 0.4) **보호속성 입력하지 않음**
평가데이터	학습 데이터와 동일하게 생성(A, B 집단 각 5,000 건)
모델 학습	sklearn.linear_model. LogisticRegression 기본설정
학습 결과	
Calibration	

Abstract

Assessing Fairness in Artificial Intelligence: Contextual Justification for Normative Criteria

Kim, Byoung-Pil
School of Law
The Graduate School
Seoul National University

The utilization of artificial intelligence (AI) is significantly increasing. Cognitive models have begun to perform tasks such as image, voice, and text recognition - tasks previously considered difficult to implement with traditional computers - at a human level. Discriminative models are widely used to assist or automate human decision-making processes. Through generative models, humans are now able to communicate naturally with AI, and AI can create artistic works. Alongside the increase in these applications, there is growing concern about the risk of AI models reproducing existing societal biases embedded in training data, leading to discriminatory impacts.

One of the major approaches to mitigating such risks involves conducting fairness audits on AI systems. Fairness auditing is considered a crucial regulatory mechanism for ensuring AI fairness in compliance with anti-discrimination laws, personal data protection laws, and AI regulatory laws. Consequently, there is a pressing need to establish standards that precisely define the concept of fairness, and to develop methodologies to assess whether the AI under audit meets these standards. To this end, researchers have proposed various operational definitions for fairness metrics to be employed in audits. However, with the demonstration of the impossibility theorem, which states that several fairness standards cannot be simultaneously satisfied –

especially for classification models – the issue of selecting appropriate fairness criteria for audits has become increasingly complex. This study investigates which fairness criteria can be normatively justified within the specific context of AI use cases, considering classification and generative models separately.

In the case of classification models, audits typically focus on whether the model's predictions result in disparate treatment or impacts for socially marginalized groups that have historically faced discrimination. However, applying legal standards developed for anti-discrimination laws to AI-driven discrimination poses several challenges. Additionally, technically defined fairness metrics seem to conflict with existing anti-discrimination laws. To overcome these challenges, this study adopts a novel approach using simulations. Hypothetical discriminatory training data was generated across various scenarios to train a classification model. Following this, the ways in which the model violates certain fairness criteria were investigated, thereby deriving suitable fairness criteria for each scenario. The simulation results suggest that, while it is valid to consider parity of false positive and negative rates as primary evaluation indicators for cognitive models, parity of positive and negative predictive values (often referred to as 'calibration') is a suitable fairness criterion for discriminative models. Furthermore, this study points out the limitations of previously suggested fairness metrics and proposes new directions for alternative fairness audits.

On the other hand, fairness audits for generative models present different characteristics compared to those for classification models. While classification models, which typically assist or automate the decision-making process, could potentially lead to allocative harms, generative models pose the risk of causing representative harms. These harms are cultural and pervasive across society, rather than influencing specific decisions. Furthermore, representative harms can trigger various problems, such as the propagation and reinforcement of social bias and stereotypes, as well as the creation and dissemination of harmful expressions. In such cases, it becomes challenging to identify victims or quantify damages, thus complicating regulation via traditional legal remedies and posing difficulties for regulation within existing administrative systems. As a result, an *ex ante* self-audit by the providers of generative models is considered inevitable. In

this regard, this study reviews concrete metrics and evaluation methods for fairness audits of generative models, categorizing them according to dimensions of bias, fairness in a narrow sense, and toxicity. Furthermore, it underscores the need to establish benchmark data and evaluation metrics capable of assessing the multifaceted legal and ethical risks posed by large-scale language models and image generation models, in order to effectively regulate the legal and ethical risks that generative models can instigate.

Keywords: AI fairness, AI discrimination, fairness criteria, algorithmic fairness, impossibility theorem

Student Number: 2021-33081