JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 5, pp. 1257-1268, May. 2026

Check for updates

생성형 AI 직업 이미지 다차원 편향 분석: 한국 맥락의 성별·연령 교차

이 금 희[1*] · 김 동 호[2]
[1]숭실대학교 미디어학과 박사수료
[2]숭실대학교 글로벌미디어학부 교수

Multidimensional Bias Analysis in Generative AI Occupational Images: Intersectional Gender and Age Perspectives in the Korean Context

Gumhee Lee[1*] · Dongho Kim[2]

[1]Ph.D. Student, Department of Media, Soongsil University, Seoul 06978, Korea
[2]Professor, School of Global Media, Soongsil University, Seoul 06978, Korea

[요 약]

본 연구는 한국 사회문화 맥락에서 DALL·E 3와 Midjourney의 성별·연령 재현 편향을 분석하였다. 제8차 한국표준직업분류(KSCO-8)을 기준으로 15개 직군의 생성 이미지 960장을 평가하였으며, 연구자 판독과 DeepFace 기반 자동 판독을 병행하는 이중 판독 체계(dual-annotation system)와 실제 노동통계 기준선(ground truth)을 비교하였다. 두 모델 모두 남성 과대표현(13.4~19.6%p)과 양극단화 현상을 보였으며, Midjourney에서는 남성 우세 직업일수록 연령이 높게 묘사되는 교차적 편향(intersectional bias)도 확인되었다. 다만 본 연구는 'in Korea' 프롬프트 조건에서 생성된 이미지의 재현 편향을 감사(audit)한 결과로, 한국 사회 자체의 직업 표상을 직접 측정한 것은 아니다. 본 연구는 생성형 AI 공정성 감사 방법론과 정책적 시사점을 제시한다.

[Abstract]

This study examines gender- and age-based representation biases in DALL·E 3 and Midjourney within South Korea's socio-cultural context. Using the 8th Korean Standard Classification of Occupations (KSCO-8), a total of 960 images were generated across 15 occupational groups. A dual-assessment framework combining human coding by researchers and automated detection using DeepFace was employed. The results were benchmarked against official labor statistics as a ground-truth reference. Both models substantially overrepresented men, with deviations ranging from 13.4 to 19.6 percentage points, and exhibited strong gender polarization toward extremes. Midjourney displayed intersectional bias by depicting male-dominated occupations with older individuals, thereby reinforcing authority hierarchies. Notably, this audit focuses on biases in images created using the explicit "in Korea" prompt, rather than on actual occupational demographics in South Korea. The findings highlight the need for bias-free AI design, transparent reporting, and context-aware evaluation methods.

색인어 : 생성형AI, 재현 편향, 성별 편향, 연령 편향, 제8차 한국표준직업분류
Keyword : Generative AI, Representation Bias, Gender Bias, Age Bias, KSCO-8

http://dx.doi.org/10.9728/dcs.2026.27.5.1257

Received 25 February 2026; **Revised** 23 March 2026
Accepted 20 April 2026

***Corresponding Author; Gumhee Lee**

Tel: ※ 개인정보 표시제한
E-mail: moor5186@gmail.com

Ⅰ. 서 론

1-1 연구 현황 및 문제 제기

최근 생성형 인공지능(generative AI)은 텍스트, 이미지, 영상 등 디지털 콘텐츠 제작 전반에서 핵심적인 기술로 부상하고 있다. 특히 Stable Diffusion[1], DALL·E 3, Midjourney와 같은 텍스트-이미지 생성 모델(text-to-image, T2I)은 높은 수준의 시각적 완성도를 바탕으로 마케팅, 디자인, 뉴스 미디어, 교육 등 여러 분야에서 활용되고 있다[2]. 이러한 모델이 생성하는 직업 이미지는 특정 직업에 대한 사회적 인식과 고정관념을 시각적으로 재현할 수 있으며, 반복적인 노출을 통해 대중의 직업 이해와 판단에 영향을 미칠 가능성이 있다.

그러나 생성형 AI는 대규모 웹 기반 데이터셋을 학습하는 과정에서 데이터에 내재된 사회적 편향과 고정관념을 그대로 수용하거나, 오히려 이를 증폭(amplification)하는 한계를 보인다[3]. Crawford는 AI 시스템이 특정 집단을 반복적으로 과소대표하거나 왜곡된 방식으로 묘사함으로써 발생하는 '표상적 해악(representational harm)'의 위험성을 경고하였다[4]. 이는 기술적 중립성이라는 가시적 성과 뒤에 현실 세계의 불평등 구조가 재현될 수 있음을 시사하며, 인공지능 윤리 및 공정성 차원에서 중대한 과제로 부상하였다.

특히 기존 연구들은 '의사'는 남성, '간호사'는 여성으로 고정되는 등 서구권 중심의 편향을 다수 보고해 왔다[5]. 하지만 한국은 직업별 성별 분포와 연령 구조에서 서구와 다른 독자적 특성을 지닌다. 글로벌 상용 모델이 'in Korea'라는 명시적 맥락을 부여받았음에도 불구하고 한국의 실제 통계를 반영하지 못한 채 서구 중심의 고정관념을 투영한다면, 이는 국내 AI 콘텐츠 서비스의 신뢰성을 저해하고 사회적 편향을 고착화할 위험이 있다.

1-2 연구의 목적 및 방법

본 연구는 DALL·E 3와 Midjourney가 한국적 맥락에서 직업 이미지를 생성할 때 나타나는 성별 및 연령 편향을 정량적으로 분석하고, 이를 한국 노동시장 통계와 비교 검증하는 'AI 감사(audit)' 연구를 수행한다.

이를 위해 첫째, 통계청의 제8차 한국표준직업분류(KSCO-8, Korean Standard Classification of Occupations, 8th revision), 경제활동인구조사 자료, 그리고 국가통계포털(KOSIS, Korean Statistical Information Service)의 직업 관련 통계자료를 활용하여 직업별 성별·연령 분포에 대한 기준선(ground truth)을 구축한다[6],[7]. 둘째, 'in Korea'를 포함한 중립적 프롬프트로 이미지를 생성한 뒤, 셋째, 연구자 판독과 딥러닝 기반 오픈소스 얼굴 분석 프레임워크[8]로 자동 판독을 병행하는 이중 판독 체계와 기준선 비교를 실시한다. 본 연구는 한국 노동시장 통계 대비 성별 분포 편향(RQ1), 모델 간 연령 분포 차이(RQ2), 측정 방식에 따른 해석의 차이(RQ3)를 규명하고자 한다.

다만 본 연구가 분석하는 대상은 한국 사회 그 자체의 직업 표상이 아니라, 'in Korea'라는 텍스트 단서가 부여된 프롬프트 조건에서 상용 T2I 모델이 산출하는 시각적 재현이다. 따라서 본 연구의 결과는 생성형 AI가 한국을 어떻게 '이해'하는지에 대한 직접적 검증이라기보다, 한국 맥락 표지가 포함된 입력 조건에서 어떠한 성별·연령 편향을 출력하는지를 평가하는 감사(audit) 결과로 해석되어야 한다. 이러한 범위 설정은 생성 이미지가 실제 한국인 집단을 충분히 대표한다고 가정하지 않으며, 오히려 그 대표성의 불확실성 자체를 연구의 제한점이자 해석 조건으로 명시한다.

Ⅱ. 관련연구

2-1 생성형 AI의 편향 연구

1) 직업 이미지의 성별 편향 연구

생성형 AI의 성별 편향은 단순히 현실을 반영하는 수준을 넘어 특정 성별에 대한 고정관념을 강화 및 증폭하는 양상을 보인다. 직업 표상의 성별 편향을 분석한 선행 연구[9]에 따르면, DALL·E 3와 Bing Image Creator를 통해 생성된 직업 이미지 중 여성의 비율은 평균 29.3%로 집계되었다. 이는 실제 미국 노동통계의 여성 고용 기준선인 47.0%와 비교해 약 17.7%p 낮은 수치로, 생성 모델이 현실의 인구통계학적 구성보다 여성을 유의미하게 과소대표하고 있음을 보여준다. 나아가 최신 모델에서도 여성 직업 이미지가 남성에 비해 상대적으로 성적으로 대상화되거나 아동화(infantilization)되어 묘사되는 질적 왜곡 경향이 보고되었다[10]. 이러한 재현적 편향은 입력 기제의 특성에 의해서도 심화된다. 9개 언어를 대상으로 한 연구[11]에서는 프롬프트 언어에 내재된 문법적 성별 표지(gender markers)가 생성 결과물의 성별 결정에 결정적인 영향을 미침이 확인되었다. 특히 성별 정보가 포함되지 않은 중립적 직업군 프롬프트에서도 남성을 기본값으로 생성하는 '디폴트 남성 편향(default-male bias)'이 관찰되는데, 이는 모델이 명시적인 맥락 단서가 부재한 상황에서도 특정 성별을 표준(norm)으로 우선시하는 알고리즘적 편중성을 내포하고 있음을 시사한다[12].

2) 편향의 측정 방법론과 타당성 문제

편향 측정은 연구자 판독과 DeepFace 기반 자동 판독(automated assessment)으로 나뉜다. 최근 BiasPainter 프레임워크나 대규모 멀티모달 언어 모델(multimodal large language model, MLLM)을 활용한 자동 판별 연구가 증가하고 있으나, 판독 알고리즘 자체가 내포한 인터페이스 편향으로 인해 결과가 왜곡될 수 있다는 지적도 제기된다[12],[13]. 이는 생성 모델뿐만 아니라 판독 체계의 결합 방

식에 대한 메타적 검토가 필요함을 시사한다. 자동 판별은 대규모 비교를 가능하게 하지만, 학습데이터 구성, 임계값 설정, 라벨 정의 등에 따라 동일한 생성 결과도 상이한 편향 수준으로 추정될 수 있으므로, 측정 도구 자체의 잠재적 편향과 불확실성을 함께 고려한 검증 전략이 요구된다. 나아가 생성형 이미지 시스템의 결과가 플랫폼의 기본 설정 및 표현 방식에 의해 달라질 수 있다는 논의를 고려할 때, 편향 평가는 생성 모델과 판독 체계를 분리하여 보지 않고 생성–판독 파이프라인 전체를 대상으로 설계·보고할 필요가 있다[14].

3) 다차원적 및 교차적 편향

최근 연구는 성별 편향을 넘어 성격 특성, 외형적 특성, 문화적 상징 등의 다차원적 편향에 주목하고 있다. 특히 알고리즘 감사 연구에서는 이미지 분류 및 생성 모델에서 성별과 인종이 교차할 경우 성능 불균형이 더욱 심화되는 현상이 보고되었다[15],[16]. 이는 특정 인구통계학적 집단이 AI 재현 과정에서 다층적인 배제 또는 소외를 경험할 수 있음을 시사한다.

2-2 AI 감사(AI Audit) 방법론

AI 감사는 시스템의 공정성과 투명성을 체계적으로 평가하는 방법론으로, 최근에는 문제 정의부터 모니터링에 이르기까지 전 과정을 포괄하는 '엔드 투 엔드(end-to-end) 프레임워크'로 확장되고 있다[17]. 생성형 AI의 경우 동일한 프롬프트에도 다양한 결과가 산출 될 수 있는 확률적 특성을 지니므로 프롬프트 기반 감사와 반복 생성 결과의 분포 분석이 중요한 평가 방식으로 요구된다. 또한 AI 감사는 기술적 성능 검토를 넘어 학습 데이터에 내재한 언어·문화적 편향을 비판적으로 살피는 사회기술적 관점을 포함할 필요가 있으며, 이는 대규모 모델의 잠재적 위험성을 지적한 'Stochastic Parrots' 논의와 맞닿아 있다[18]. 본 연구는 '한국'이라는 명시적 국가 맥락을 조건으로 설정하고, 이러한 문화적 지시가 생성 결과에 어떠한 영향을 미치는지를 검토함으로써 생성형 AI 감사 방법론의 적용 범위을 넓히고자 한다.

Ⅲ. 연구 방법

3-1 연구설계

본 연구는 한국적 직업 맥락이 명시된 조건에서 상용 T2I 모델이 직업을 어떻게 시각적으로 재현하는지를 분석하기 위해 설계되었다. 특히 프롬프트에 '한국'이라는 국가·문화적 맥락을 포함하였을 때 성별 편향이 완화되는지 또는 유지되는지를 실증적으로 검토하고, 동일한 생성 결과를 연구자 판독과 DeepFace 기반 자동 판독으로 측정할 때 편향 추정치가 어떻게 달라지는지도 비교한다.

데이터 수집은 2025년 12월에 수행하였으며, 분석 대상 모델은 DALL·E 3와 Midjourney이다. DALL·E 3는 OpenAI API(size: 1024×1024, quality: standard, style: vivid(기본값), n: 1)를 통해 이미지를 생성하였다. Midjourney는 공식 웹 인터페이스(V7, 2025년 12월 수집 시점 기본 버전)를 사용하였으며, 별도의 파라미터를 지정하지 않아 플랫폼 기본값(Aspect Ratio: --ar 1:1, Stylize: --s 100, Quality: --q 1)이 적용되었다. Midjourney V7은 개인화(Personalization) 기능이 기본값으로 활성화되어 있으나, 본 연구의 데이터 수집은 해당 계정의 최초 사용 시점에 수행되었으므로 사전 이미지 선호 이력이 존재하지 않아 개인화 효과가 실질적으로 배제된 조건에서 생성이 이루어졌다. Midjourney는 공개 API를 제공하지 않아 내부 파라미터의 완전한 명시에 한계가 있으며, 이는 연구의 재현 가능성을 부분적으로 제한하는 요인으로 인정한다. 또한, 비교 가능성을 위해 동일한 프롬프트 조건과 반복 횟수를 적용하였다. 프롬프트는 영어로 통일하되 국가 맥락을 명시하기 위해 "in Korea"를 포함하였으며(예: "A portrait photo of a judge in Korea"), 직업 정체성 관찰을 위해 인물 중심 이미지("portrait photo")를 사용하였다.

분석 대상은 15개 직업군이며, 각 직업군에 대해 모델별로 32장의 이미지를 생성하여 총 960장(15개 직업×2개 모델×32장)을 구축하였다. 또한 동일한 생성 이미지를 대상으로 (1) 한국 노동통계 기반 기준선, (2) 연구자 판독, (3) DeepFace 기반 자동 판독을 적용하는 이중 판독 체계와 기준선 비교 설계를 통해, 생성 결과뿐 아니라 측정 방식이 편향 평가에 미치는 영향도 함께 분석하였다.

3-2 측정기준

본 연구에서는 생성된 직업 이미지의 성별 분포를 평가하기 위해 다음의 세 가지 측정 기준을 설정하였다.

첫째, 실제 노동통계 기준(R_*)은 한국 사회의 직업별 성별 분포를 반영하는 비교 기준선(baseline)으로, 통계청의 경제활동인구조사 자료와 국가통계포털(KOSIS)의 직업 관련 통계자료를 바탕으로 직업군별 실제 남성 비율(R_M)과 여성 비율(R_F)을 산출하였다[6],[7]. 이 기준은 생성형 AI가 산출한 이미지의 성별 분포가 현실 노동시장 구조를 충실히 반영하는지, 또는 특정 방향으로 왜곡하거나 고정관념을 증폭하는지를 판단하기 위한 객관적 참조 기준으로 활용하였다.

둘째, 연구자 판독 기준(researcher)은 연구자가 생성 이미지를 직접 시각적으로 판독하여 이미지 속 인물의 성별을 분류하는 방식이다. 연구자 판독은 인물의 외형, 복장 및 맥락적 단서를 종합적으로 해석할 수 있다는 장점이 있으나, 판독자의 주관이 결과에 영향을 미칠 수 있다는 한계를 지닌다.

셋째, 자동 판독 기준은 얼굴 인식 기반 라이브러리인 DeepFace를 활용하여 이미지 속 인물의 성별과 연령을 자동 추정하는 방식이다. DeepFace는 Serengil and Ozpinar

(2020)가 개발한 오픈소스 경량 얼굴 분석 프레임워크로, VGG-Face(visual geometry group face), FaceNet, OpenFace 등 다수의 딥러닝 기반 얼굴 인식 모델을 통합 제공하며, 성별·연령 자동 추정 기능을 지원한다[8]. 본 연구에서 DeepFace를 측정 도구로 선택한 이유는, 대규모 이미지 데이터의 일괄 처리와 일관된 자동 판독이 가능하여 960장의 생성 이미지에 대한 체계적 분석에 적합하기 때문이다. 본 연구에서는 DeepFace 라이브러리(버전 0.0.92)를 사용하였으며, 얼굴 검출은 OpenCV Haar Cascade 기반 검출기(detector_backend='opencv')로 수행하였다. 다만, DeepFace의 연령 및 성별 추정 모델이 주로 서구권 데이터로 학습되었음을 고려하여, 연구자 판독과의 교차검증을 통해 한국인 얼굴에 대한 추정 타당성을 확인하였다.

성별 및 연령 추정은 VGG-Face 아키텍처 기반 모델을 적용하였고, 모든 이미지는 동일한 파라미터로 일괄 처리하였다. DeepFace 기반 자동 판독은 대규모 이미지 처리와 일관된 판독이 가능하다는 장점이 있으나, AI 생성 이미지 적용 시 다음과 같은 한계가 존재한다: (1) 서구권 실제 인물 사진 중심의 학습 데이터로 인해 AI가 생성한 한국인 얼굴에 대한 추정 정확도가 낮아질 수 있으며, (2) 생성 이미지 특유의 과도하게 매끄러운 피부 표현(hyper-smooth texture)이 연령 추정 결과에 영향을 줄 수 있고, (3) 실제 사진이 아닌 생성 이미지에서의 얼굴 미검출(detection failure)이 발생할 경우 해당 이미지는 분석에서 제외되며, 이는 결과의 대표성에 영향을 미칠 수 있다. 학습 데이터 편향 또는 얼굴 인식 실패가 결과에 영향을 미칠 수 있다는 점은 본 연구의 제한점에서 함께 논의한다.

마지막으로 본 연구는 동일한 960장의 생성 이미지를 대상으로 연구자 판독과 DeepFace 기반 자동 판독을 병행함으로써, 측정 방법 자체가 성별 편향 추정 결과에 미치는 영향을 체계적으로 분석하였다. 이후 분석에서는 판독 방법을 R(연구자 판독, researcher-coded)과 D(DeepFace 기반 자동 판독, DeepFace-based automated coding)로 표기하고, 모델별 결과는 DALL·E 3(DE) 및 Midjourney(MJ)로 구분하여 제시한다.

3-3 분석 대상 직업

분석 대상은 한국 사회의 성별 분업 구조와 시각적 고정관념을 정밀하게 측정하기 위해 KSCO-8의 대·중분류 체계를 참고하여 총 15개 직업군으로 구성하였다. 직업 선정은 (1) 실제 노동시장의 성별 분포, (2) 사회적 인식에 기반한 성별 고정관념, (3) 비교 분석의 효율성과 균형을 기준으로 수행하였으며, 직업군은 다음의 세 범주로 구분하였다.

첫째, 남성 우세 직업군(male-dominant)으로 기업 고위임원, 건설 노동자, 소프트웨어 개발자, 비행기 조종사, 소방관을 포함하였다. 둘째, 여성 우세 직업군(female-dominant)으로 유치원 교사, 간호사, 객실 승무원, 비서, 플로리스트를 포함하였다. 셋째, 성별 혼합 및 변화 직업군(mixed/neutral)으로 판사, 약사, 요리사, 바리스타, 디자이너를 포함하였다. 이 구분은 생성형 AI가 직업 이미지에서 고정관념을 강화하거나, 실제 통계와의 괴리를 확대하는 양상을 범주 간 비교를 통해 분석하기 위한 것이다.

한편 본 연구는 KSCO-8 전체 직업을 포괄하기보다, 성별 분포 대비가 뚜렷하고 사회적으로 널리 인지되는 직업을 중심으로, 모델 간 비교가 가능한 표본(총 960장)을 확보할 수 있는 범위에서 직업군을 구성하였다. 따라서 택시기사·경찰 등 유사 직업은 범주 대표성이 중복되는 경우 제외하였으며, 이로 인한 일반화의 한계는 제한점에서 논의한다.

3-4 편향지표 및 통계 분석

통계 분석은 생성된 이미지의 성별 분포가 실제 노동통계와 얼마나 차이가 있는지를 직업 단위로 정량화하기 위해 수행되었다. 직업$i(i=1,...,n)$에서 실제 노동통계상의 남성 비율을 p_i^{Real}, 생성 이미지에서 관측된 남성 비율을 $p_i^{G,m}$로 정의한다. 여기서 m은 판독 방법을 의미하며, $m\in\{R,D\}$로 두었다. R는 연구자 판독(researcher-coded)을, D는 DeepFace 기반 자동 판독(DeepFace-based automated coding)을 의미한다. 직업 i에 대한 남성 비율 기준 편향은 다음과 같이 정의하였다.

$$b_i^{(m)} = p_i^{G,m} - p_i^{Real} \tag{1}$$

$b_i^{(m)} > 0$이면 남성 과대표현, $b_i^{(m)} < 0$이면 여성 과대표현을 의미한다. 직업 전반에서 편향의 크기는 평균 절대 오차(MAE, Mean Absolute Error)와 평균 제곱근 오차(RMSE, Root Mean Squared Error)를 통해 측정하였다.

$$MAE_{bias}^{(m)} = \frac{1}{n}\sum_{i=1}^{n}\left|b_i^{(m)}\right| \quad (\text{단, } m\in\{R,D\}) \tag{2a}$$

$$\mathrm{RMSE}_{\mathrm{bias}}^{(\mathrm{m})} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{b}_{\mathrm{i}}^{(\mathrm{m})})^2} \quad (\text{단, } m\in\{R,D\}) \tag{2b}$$

아울러 직업 i에서 실제 노동통계 남성비율을 x_i, 생성 이미지 남성비율을 y_i로 두고 선형회귀모형 $y_i=\alpha+\beta x_i+\epsilon_i$를 적합하여 α(기저 편향)와 β(격차 증폭/완화)를 추정하였다. 해당 회귀는 연구자 판독($m=R$)과 DeepFace 기반 자동 판독($m=D$)에 대해 각각 수행하였다. 또한 연구자 판독과 DeepFace 기반 자동 판독 간 불일치는 직업별 생성 남성비율의 차이를 기반으로 MAE 및 피어슨 상관계수로 평가하였고, Bland-Altman 분석을 통해 두 측정 방법 간 일치성을 확인하였으며, 95% 일치한계(LoA, limits of agreement)는 평균차 ± 1.96×표준편차로 산출하였다.

편향의 통계적 유의성은 각 조건(모델×판독 방법에서 $b_i^{(m)}$

의 평균이 0과 다른지 검정하기 위해 일표본 t-검정과 Wilcoxon 부호순위 검정을 병행하여 평가하였다($\alpha = .05$). t-검정 적용의 타당성을 확인하기 위해 Shapiro-Wilk 정규성 검정을 수행한 결과, DALL·E 3(R)와 Midjourney(R), DALL·E 3(D) 조건은 정규성 가정을 충족하였으나($p > .05$), Midjourney (D) 조건은 정규성 가정을 충족하지 못하였다 ($p < .05$). 모든 통계 검정은 직업 단위($n = 15$)의 편향값 ($b_i^{(m)}$)를 대상으로 수행하였다. 이에 따라 본 연구는 모수 검정과 비모수 검정 결과를 함께 보고함으로써, 비정규 분포가 관찰된 조건에서도 결과의 강건성(robustness)을 확인하였다.

Ⅳ. 연구결과

4-1 측정 체계별 전체 편향 경향

본 장에서는 Ⅲ장에서 제시한 연구설계에 따라 생성된 총 960장의 직업 이미지(15개 직업 × 2개 모델 × 32장)를 분석한 결과를 제시한다. 분석은 측정 체계별 전체 편향 경향을 시작으로, 편향의 통계적 유의성, 극단화 패턴, 직업별 분포, 편향 증폭 효과, 그리고 최종적으로 측정 체계 간 불일치 양상을 규명하는 순서로 구성하였다.

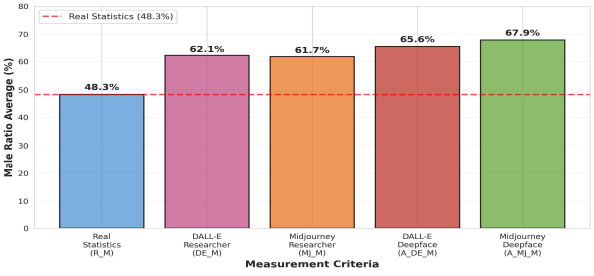

그림 1. 기준별 남성비율 평균비교
Fig. 1. Comparison of average male ratios by measurement criterion

그림 1은 실제 노동통계, 연구자 판독, 그리고 DeepFace 기반 자동 판독이라는 세 가지 측정 기준에 따라 산출된 남성 비율의 평균을 비교한 결과이다. 먼저, 실제 한국 노동통계에 기반한 분석 대상 직업군의 평균 남성 비율은 48.3%로 나타났다. 반면 연구자 판독 결과, DALL·E 3와 Midjourney는 모두 실제 통계치를 상회하는 남성 과대표현 경향을 보였다. 구체적으로 DALL·E 3의 평균 남성 비율은 62.1%로 기준선 대비 13.8%p 높았고, Midjourney는 61.7%로 13.4%p 높은 것으로 나타났다. 이는 두 모델 모두 현실의 직업별 성별 분포에 비해 남성을 상대적으로 더 많이 재현하고 있음을 보여준다. DeepFace 기반 자동 판독에서는 이러한 편향성이 더욱 크게 나타났다. Midjourney의 평균 남성 비율이 67.9%로 실제 통계 대비 19.6%p 높은 값을 보였으며, DALL·E 3 역시 65.6%로 17.3%p 높은 수준을 나타냈다. 이는 동일한 생성 이미지에 대해서도 자동 판독 체계가 연구자 판독보다 남성 편중 경향을 더 크게 추정함을 의미한다.

종합하면, 생성형 AI는 프롬프트에 'in Korea'라는 국가 맥락이 포함된 조건에서도 기준선으로 설정한 한국 노동통계의 성별 분포를 충실히 반영하기보다는, 남성을 상대적으로 더 많이 재현하는 경향을 보였다. 또한 자동화된 판독체계에서 편향의 크기가 더 크게 관측되었다는 점은, 생성 결과 자체뿐 아니라 이를 측정하는 도구의 특성 역시 편향 평가 결과에 영향을 미칠 수 있음을 시사한다.

4-2 편향의 통계적 유의성

관측된 성별 편향이 0과 통계적으로 유의하게 다른지를 검정한 결과, 연구자 판독 기준에서 두 모델 모두 남성 과대표현을 보였다. 이는 'in Korea'라는 한국적 맥락이 명시된 조건에서도 생성형 AI가 실제 노동통계상 평균 남성 비율(48.3%)보다 남성을 13%p 이상 높게 재현하였음을 의미하며, 이러한 편향이 우연한 변동이 아니라 체계적으로 나타난 현상임을 시사한다.

표 1. 성별편향의 통계적 유의성
Table 1. Statistical significance of gender bias

Model (Method)	Bias(%p)	t-stat	p-value	W (Wilcoxon)	p(W)
DALL·E 3 (R)	+13.8%p	t(14)=2.61	.020*	20.0	.022*
DALL·E 3 (D)	+17.3%p	t(14)=3.19	.007**	11.0	.003**
Midjourney (R)	+13.4%p	t(14)=2.24	.042*	26.0	.055
Midjourney (D)	+19.6%p	t(14)=3.51	.004**	0.0	.001**

*p < .05, **p < .01, ***p < .001

표 1은 측정 방법과 모델별 성별 편향의 통계적 유의성을 요약한 것이다. 일표본 t-검정과 Wilcoxon 부호순위 검정을 실시한 결과, DALL·E 3(R), DALL·E 3(D), Midjourney(D) 조건에서는 두 검정 모두에서 통계적으로 유의한 편향이 확인되었다. 구체적으로 DALL·E 3(R)은 t(14)=2.61, p=.020 Wilcoxon W=20.0, p=.022로 나타났고, DALL·E 3(D)는 t(14)=3.19, p=.007, W=11.0, p=.003으로 나타났다. 또한 Midjourney(D)는 t(14)=3.51, p=.004, W=0.0, p=.001로 가장 강한 수준의 통계적 유의성을 보였다. 반면 Midjourney(R) 조건에서는 t-검정 결과 t(14)=2.24, p=.042로 유의한 차이가 확인되었으나, Wilcoxon 검정에서는 W=26.0, p=.055로 유의수준 .05에 근접한 경계적 결과를 보였다. 이는 해당 조건에서 편향의 방향은 일관되게 남성 과대표현으로 나타났으나, 비모수 검정에서는 그 유의성이 상대적으로 약하게 나타

났음을 의미한다.

한편, DeepFace 기반 자동 판독 기준에서는 연구자 판독에 비해 더 큰 편향값이 관측되었으며, 통계적 유의성 또한 전반적으로 강화되는 경향을 보였다. DALL·E 3(D)와는 Midjourney(D)는 각각 +17.3%p와 +19.6%p의 편향을 나타냈고, 두 검정 모두에서 유의한 결과가 확인되었다. 이는 동일한 생성 이미지라도 측정 방식에 따라 편향의 크기와 통계적 유의성 판단이 달라질 수 있음을 보여준다.

4-3 극단화 패턴 분석

그림 2는 남성 비율이 극단적인 수치를 나타내는 직업군의 비중을 실제 통계와 비교 분석한 결과이다.

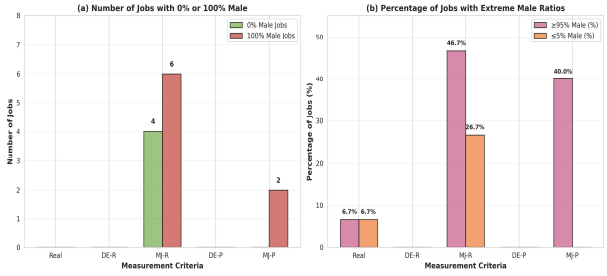

그림 2. 극단화 패턴 비교: (a) 0%/100% 남성 직업 수, (b) 극단적 남성비율 직업 비율

Fig. 2. Comparison of polarization patterns: (a) number of occupations with 0% or 100% male representation, (b) proportion of occupations with extreme male ratios

분석 결과, 생성형 AI 모델들은 한국 사회의 실제 직업 성별 분포를 재현하는 수준을 넘어, 성별 격차를 인위적으로 포화시키는 구조적 특징을 보였다.

첫째, 양극단화(polarization) 패턴이다. 실제 한국 노동통계에서 '거의 전부 남성(남성 비율 ≥95%)'으로 분류되는 직업군은 전체의 6.7%에 불과하였으나, AI 생성 이미지에서는 해당 비중이 40.0~46.7%로 약 6배 이상 급증하였다. 이는 생성형 AI가 직업과 성별의 결합을 이분법적 관점에서 해석하여 성별 분포를 양 끝단으로 밀어내는 '양극단화' 경향을 지니고 있음을 실증한다. 특히 연구자 판독 결과에서는 남성 비율이 0% 또는 100%로 완전히 분리되는 직업이 다수 관찰되어 이러한 현상이 더욱 뚜렷하게 나타났다.

둘째, 상한 편향(ceiling bias) 메커니즘이다. 측정 체계에 따른 차이를 분석한 결과, DeepFace 기반 자동 판독에서는 '남성 비율 0%'인 직업이 단 한 건도 관측되지 않았다는 점에 주목할 필요가 있다. 연구자 판독에서는 유치원 교사 등 여성 우세 직업군에서 남성 비율이 0%로 판독된 반면, DeepFace 기반 자동 판독 알고리즘은 동일한 이미지에서도 일정 비율의 남성을 일관되게 감지하였다. 이는 자동화된 판별 도구가 여성 다수 직업에서도 남성을 기본값으로 인식하는 '상한 편향' 또는 '기저 남성 편향(baseline male shift)'을 내재하고 있음을 시사한다.

이러한 결과는 생성형 AI에 의한 표상적 해악이 단일한 양상으로 나타나지 않음을 보여준다. 즉, 연구자 판독이 실제 성별 격차를 더 극단적으로 벌리는 '격차 증폭' 메커니즘을 반영한다면, DeepFace 기반 자동 판독은 남성을 보편적 표준으로 설정하는 '기본값 이동' 메커니즘을 반영하고 있다. 이는 생성 결과가 실제 통계 분포와 비교해 남성 과대표현이 확대되는 경향을 보일 수 있음을 시사한다.

4-4 직업별 편향 분포

그림 3은 연구자 판독 결과를 기준으로 15개 분석 대상 직업의 성별 편향 분포를 실제 노동통계와의 격차(gap) 크기순으로 정렬하여 제시한 것이다.

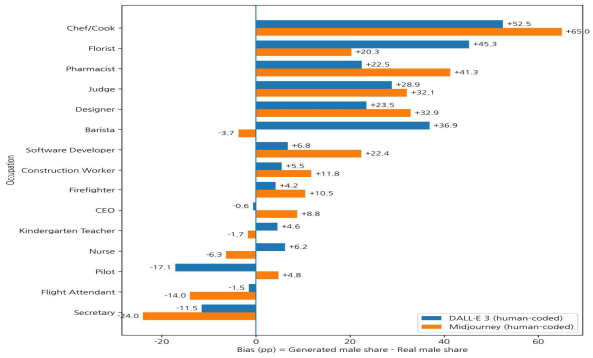

그림 3. 직업별 편향분포(전체 15개 직업)

Fig. 3. Bias distribution by occupation (All 15 occupations)

분석 결과, 남성 과대표현 편향이 가장 크게 나타난 직업은 바리스타(+46.3%p), 약사(+41.3%p), 요리사(+40.0%p) 순으로 나타났다. 이들 직업은 실제 한국 노동통계상 남성 비율이 35~40% 수준임에도 불구하고, 생성된 이미지에서는 남성 표상은 75%에서 최대 100%까지 나타났다. 반면, 비서(-24.0%p)와 객실 승무원(-14.0%p) 등의 직업군에서는 실제 통계 수치보다 여성이 더 많이 생성되는 음(-)의 격차가 관측되었다. 이러한 결과는 생성형 AI가 'in Korea'라는 국가 맥락이 명시된 조건에서도, 기준선으로 설정한 한국 노동통계와는 다른 방향의 직업-성별 결합을 산출할 수 있음을 보여준다. 특히 바리스타, 약사, 요리사와 같이 실제 통계상 성비가 극단적이지 않은 직업에서 남성 표상이 크게 과장되었다는 점은, 모델이 직업과 관련 시각적 고정관념을 선택적으로 강화할 가능성을 시사한다.

4-5 증폭 효과: 회귀 분석

그림 4는 실제 한국 노동통계상의 남성 비율과 생성형 AI가 출력한 이미지의 남성 비율 간의 상관관계를 선형 회귀 분석(linear regression)으로 나타낸 결과이다. 분석 결과, 측

정 체계에 따라 '격차 증폭'과 '기저 편향'이라는 서로 다른 수학적 메커니즘이 확인되었다.

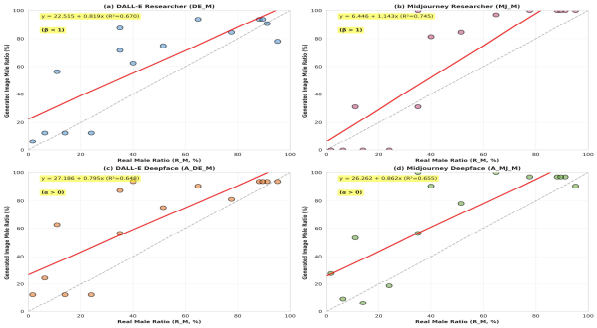

그림 4. 증폭효과-회귀분석(실제 남성 비율 vs 생성 이미지 남성 비율)

Fig. 4. Amplification effect: regression analysis (actual male ratio vs. generated image male ratio)

선형회귀모형 $y_i = \alpha + \beta x_i + \epsilon_i$ 결과, 첫째, 연구자 판독 기준 DALL·E 3는 기울기 $\beta = 1.092 (R^2 = .921)$, Midjourney는 $\beta = 1.143 (R^2 = .906)$으로 두 모델 모두 1보다 큰 값을 기록하였다. 이는 실제 남성 비율이 10%p 증가할 때, DALL·E 3는 10.92%p, Midjourney는 11.43%p의 남성 비율 증가로 응답함을 의미하며, 실제 성별 격차를 약 9.2–14.3% 증폭하는 '편향 증폭(bias amplification)' 효과가 확인되었다.

둘째, DeepFace 기반 자동 판독에서는 절편값(α)이 DALL·E 3는 $\alpha = 0.263 (R^2 = .905)$, Midjourney는 $\alpha = 0.188 (R^2 = .899)$로 통계적으로 유의한 양수로 산출되었다. 이는 실제 여성 비율이 100%인 직업군(남성 비율 = 0)에서도 DeepFace가 평균 18.8–26.3%의 이미지를 남성으로 오판독하는 '기저 남성 편향(baseline male bias)'을 내재함을 의미한다. 다만 α는 회귀모형의 절편으로 관측 범위를 벗어나는 구간에서는 외삽 해석이 될 수 있으므로, 본 연구에서는 기저 편향의 지표로 해석하였다.

4-6 측정체계 간 불일치

그림 5는 연구자 판독과 DeepFace 기반 자동 판독 간 불일치를 Bland–Altman 분석으로 시각화한 결과이다.

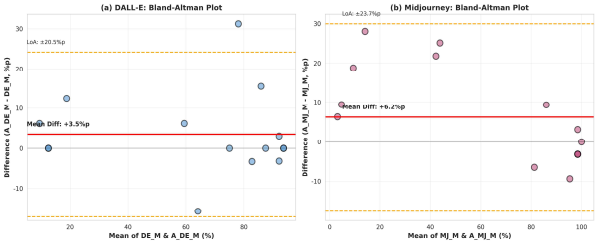

그림 5. 측정체계 간 불일치(Bland-Altman 분석)

Fig. 5. Disagreement between measurement systems (Bland-Altman analysis)

DALL·E 3의 경우 DeepFace 기반 자동 판독이 연구자 판독보다 남성 비율을 평균 +3.5%p 높게 산출하였으며 (95% CI: [+1.2, +5.8]), 두 방법 간 피어슨 상관계수는 $r = 0.951$로 매우 높았다. Bland–Altman 분석 결과 95% 일치한계(LoA, limits of agreement)는 [−8.2, +15.2]로 나타나, 대부분의 직업군에서 두 판독 방법이 ±15%p 이내의 차이를 보였으나, 플로리스트(+28.2%p)와 유치원 교사(+28.1%p) 등 일부 여성 우세 직업에서는 측정 방법 간 불일치가 크게 나타났다. Midjourney에서도 DeepFace 기반 자동 판독이 평균 +6.2%p 높게 나타났고, 상관계수는 $r = 0.978$로 매우 높았으나, 방법 간 차이의 산포(표준편차)가 더 크게 나타났으며, 일치한계(LoA)의 범위도 더 넓게 관측되었다.

대표적인 불일치 사례로, 플로리스트(DALL-E 3)는 연구자 판독 3.1% 대비 DeepFace 기반 자동 판독 31.3%, 유치원 교사(Midjourney)는 0% 대비 28%p 이상의 차이가 나타났다. 이는 동일한 생성 이미지라도 판독 방식에 따라 편향 추정치가 크게 달라질 수 있음을 명확히 보여준다.

4-7 연령 분포 및 모델 간 차이 분석

본 연구는 직업 이미지 생성 과정에서 나타나는 인구통계학적 재현 특성을 파악하기 위해, 성별 편향 분석에 이어 생성된 이미지의 연령 분포를 정량적으로 측정하였다.

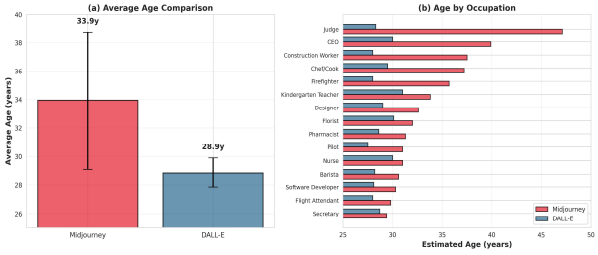

그림 6. 연령 분석 비교(DeepFace 기반 자동 분석)

Fig. 6. Age comparison by model (DeepFace-based automated analysis)

분석 결과, 상용 T2I 모델인 DALL·E 3와 Midjourney 간에 유의미한 연령 묘사 지향성 차이가 관찰되었다.

첫째, 모델별 평균 연령은 DALL·E 3가 28.9세, Midjourney가 33.9세로 측정되어 두 모델 간 약 5세의 유의미한 격차가 확인되었다. 둘째, 그림 6의 박스플롯 분석에 따르면 DALL·E 3는 20대 후반의 좁은 범위에 연령이 밀집되는 경향을 보인 반면, Midjourney는 상대적으로 넓은 연령 스펙트럼을 형성하며 다양한 연령대의 인물을 재현하였다.

이러한 결과는 DALL·E 3가 상대적으로 청년층 중심의 정형화된 이미지를 생성하는 경향이 있는 반면, Midjourney는 직업 표상과 연령 단서를 결합하여 사회적 성숙도를 함께 시각화하는 경향이 있음을 시사한다. 이는 학습 데이터의 인구통계학적 구성 차이와 더불어, 프롬프트 해석 및 생성 과정에

서의 내부 최적화 방식 차이에 기인한 것으로 해석될 수 있다.

4-8 성별-연령 교차분석

본 연구는 단일 차원의 성별 편향을 넘어 성별과 연령이 상호작용하여 나타나는 교차적 편향(intersectional bias)을 분석하였다(그림 7). 분석 결과, 특정 생성 모델에서 직업적 권위와 사회적 위계가 성별과 연령의 결합을 통해 시각적으로 고착화되는 양상이 확인되었다.

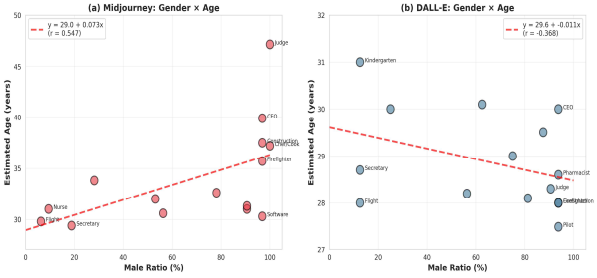

그림 7. 성별-연령 교차 분석(산점도 및 회귀선)
Fig. 7. Intersectional analysis of gender and age (scatter plot with regression line)

첫째, Midjourney에서는 생성 이미지의 남성 비율과 추정 연령 간에 유의한 정적 상관관계가 관찰되었다($r = .547$, 95% CI: [.046, .830], N = 15). 이는 남성 비율이 높은 직업군일수록 생성 인물의 추정 연령 역시 높아지는 경향이 있음을 의미한다. 또한 남성 비율 90% 이상 직업군(기업 고위 임원, 소프트웨어 개발자 등)의 평균 추정 연령은 36.2세(SD = 3.4)인 반면, 여성 비율 80% 이상 직업군(간호사, 유치원 교사 등)은 30.1세(SD = 2.7)로 나타나, 약 6세의 차이를 보였다. 이와 함께 효과크기(Cohen's $d = 2.01$)가 확인되었다. 이러한 결과는 Midjourney가 의사결정 권한이나 사회적 위계가 높은 직업을 대상으로 고연령의 남성과 결합하여 재현하는 경향이 있음을 시사한다.

둘째, DALL·E 3에서는 남성 비율과 추정 연령 간에 미약한 부적 상관관계가 나타났으나 통계적으로 유의하지는 않았다($r = -.368$, $p = .176$). DALL·E 3는 직업적 특성과 무관하게 인물의 연령을 28~30세의 좁은 범위 내에서 생성하는 양상을 보였다. 이는 특정 직업군에 대해 편향된 연령을 부여하지는 않더라도 전반적으로 '젊은 성인'을 기본값으로 재현함으로써 연령 다양성(age diversity)의 측면에서 한계를 드러낸 것으로 해석할 수 있다.

종합하면, 두 모델 간 차이는 학습 데이터의 인구통계학적 구성, 프롬프트 해석 방식, 생성 과정의 내부 최적화 차이와 관련될 가능성이 있다. 이러한 결과는 생성형 AI가 한국 맥락 표지가 포함된 조건에서도 성별과 연령이 결합된 직업 이미지를 비대칭적으로 재현할 수 있음을 보여주며, 이러한 재현은 표상적 해악(representational harm)으로 이어질 가능성을 시사한다.

4-9 편향의 분포

그림 8은 네 가지 측정 조건에서 관측된 편향 분포를 박스플롯으로 비교한 결과이다. 그림에서 빨간 실선은 중앙값, 파란 점선은 평균을 나타낸다.

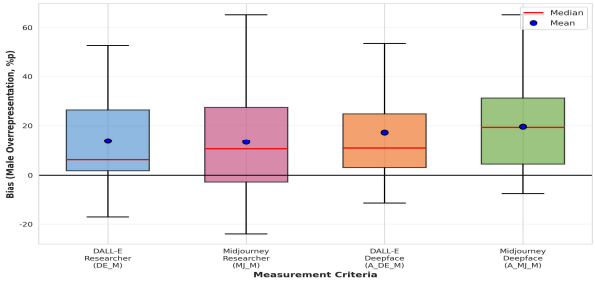

그림 8. 측정 기준별 편향 분포
Fig. 8. Distribution of bias by measurement criterion

분석 결과, DeepFace 기반 자동 판독은 연구자 판독에 비해 평균 편향값이 더 높고 분산도 더 큰 것으로 나타났다. 또한 모든 측정 조건에서 중앙값과 평균값이 모두 0보다 크게 나타나 전반적으로 체계적인 남성 과대표현 편향이 존재함을 확인할 수 있었다. 특히 DeepFace 기반 자동 판독에서는 편향의 분포 범위가 더 넓게 나타났으며, 일부 직업에서는 +50%p 이상의 극단적인 편향이 관측되었다.

이와 같은 경향은 앞서 확인된 직업별 사례와도 일치한다. 예를 들어 바리스타(+46.3%p), 요리사(+40.0%p)와 같은 직업에서는 'in Korea' 조건이 명시되었음에도 불구하고, 생성 결과가 기준선으로 설정한 한국 노동통계와 현저히 다른 방향으로 편중되는 양상이 나타났다. 이는 생성 모델이 특정 직업에 대해 남성 표상을 과장하는 경향을 지닐 수 있음을 보여준다,

종합하면, 한국 맥락 표지가 포함된 조건에서도 생성형 AI 이미지 모델은 직업-성별 결합에 관한 편향을 지속적으로 재현하는 경향을 보였으며, 그 크기와 분포는 측정 체계에 따라 다르게 나타났다. 이러한 결과는 생성형 AI가 단순한 시각화 도구를 넘어 특정 직업과 성별의 결합을 반복적으로 재현함으로써 사회적 인식 형성에 영향을 미칠 수 있음을 시사한다.

V. 결 론

5-1 주요 결과 요약

본 연구는 모든 프롬프트에 'in Korea'를 포함한 조건에서 상용 T2I 모델인 DALL·E 3와 Midjourney가 생성한 직업 이미지를 대상으로 성별 및 연령 재현 편향을 정량적으로 분

석하였다. 비교 기준선(ground truth)으로는 통계청의 제8차 한국표준직업분류(KSCO-8)와 국가통계포털(KOSIS)의 직업 관련 통계자료를 바탕으로 구축하였으며, 생성 결과의 평가는 연구자 판독과 DeepFace 기반 자동 판독을 병행한 이중 판독 체계와 기준선 비교 설계를 통해 수행하였다.

분석 결과, 두 모델 모두 한국 노동통계의 평균 남성 비율(48.3%)에 비해 남성을 과대표현 경향이 관찰되었다. 연구자 판독 기준에서 두 모델은 평균적으로 약 13%p 이상의 남성 과대표현을 나타냈으며, DeepFace 기반 자동 판독에서는 편향의 크기가 17.3%p에서 19.6%p까지 확대되는 것으로 나타났다. 특히 실제 성비가 비교적 균형적이거나 여성 비율이 높은 직업군에서도 남성 표상이 크게 증가한 사례가 확인되어, 편향의 강도는 직업에 따라 상이하게 나타날 수 있음을 보여주었다.

연령 재현 측면에서는 Midjourney의 평균 추정 연령이 DALL·E 3보다 높게 나타났으며(예: 33.9세 vs 28.9세), 직업별 성별 분포와 연령 추정치 사이에 일정한 관련성이 관찰되어 성별-연령 교차적 편향(intersectional bias) 가능성을 시사하였다.

또한 본 연구는 동일한 생성 결과에 대해서도 측정 체계에 따라 편향의 추정치와 양상이 달라질 수 있음을 확인하였다. 이는 생성 모델의 출력 자체에 내재한 편향뿐 아니라, 이를 판독하는 도구와 절차 역시 결과 해석에 중요한 영향을 미칠 수 있음을 보여준다.

5-2 해석 및 학술적 함의

본 연구의 결과는 'in Korea'라는 국가·문화적 맥락이 프롬프트에 명시되더라도, 상용 생성형 AI가 직업 이미지를 생성하는 과정에서 성별 및 연령과 관련된 비대칭적 재현을 지속할 수 있음을 보여준다. 이는 생성형 AI 편향 논의에서 단일 차원의 성별 편향 분석에 머무르기 보다, 성별과 연령이 결합하여 나타나는 교차적 재현 양상까지 함께 검토할 필요가 있음을 시사한다[15],[16]. 특히, 전문직·권위직에서 남성 표상이 강화되고, 전통적으로 여성화된 직업에서 여성 표상이 유지되는 양상은 기존 서구권 직업 이미지 편향 연구에서 보고된 디폴트 남성 편향(default-male bias) 및 직업 고정관념 증폭 경향과 일정 부분 유사하다[9]-[11],[13]. 예컨대 Sandoval-Martín & Martínez-Sanzo[9]는 DALL·E 3 생성 직업 이미지에서 여성 비율이 미국 노동통계 대비 약 17.7%p 낮음을 보고한 바 있으며, 본 연구에서도 연구자 판독 기준 13.4~13.8%p의 남성 과대표현이 확인되어 편향의 방향과 규모 면에서 유사한 양상이 관찰된다. 이는 글로벌 상용 생성 모델이 특정 직업을 시각화하는 과정에서 남성을 기본값으로 설정하거나, 기존 사회문화적 고정관념을 강화하는 방식으로 직업 이미지를 산출할 수 있음을 보여준다. 그러나 본 연구에서는 바리스타, 약사, 요리사와 같이 한국 노동통계상 남성 비율이 절대적으로 높지 않거나 성비가 비교적 균형적인 직업에서도 남성 표상이 크게 과장되는 양상이 확인되었다. 이는 한국 노동통계라는 로컬 기준선을 적용함으로써 드러난 특징으로, 생성형 AI가 '한국'이라는 맥락 표지를 입력받더라도 국내 노동시장의 실제 성별 분포를 충분히 반영하지 못할 수 있음을 의미한다. 이러한 결과는 한국 관련 데이터가 충실히 반영된 결과라기보다, 영어 중심의 글로벌 학습 과정에서 형성된 직업-성별 연상이 한국 맥락 조건에서도 우세하게 작동했을 가능성을 시사한다.

다만, 이러한 결과만으로 서구권 학습 데이터에 내재된 직업-성별 고정관념이 한국 맥락으로 전이된 것인지, 혹은 학습 데이터에 포함된 한국 관련 콘텐츠가 부분적으로 반영된 결과인지를 인과적으로 구분하기는 현재 연구 설계상 어렵다. 반면 비서·객실 승무원과 같이 여성 우세 패턴이 유지된 사례는 일부 직업군에서 문화권을 가로질러 공유되는 젠더 규범이 함께 작동했을 가능성도 보여준다. 따라서 두 기제의 상대적 기여를 분리하여 검증하기 위해서는 모델별 학습 데이터 구성 분석, 한국어와 영어 프롬프트의 비교, 문화적 시각 단서의 포함 여부를 조작한 후속 실험이 필요하다[9]-[11],[13].

이러한 점에서 본 연구는 한국 사회 자체의 편향을 직접 측정한 연구라기보다, 한국 맥락이 명시된 입력 조건에서도 글로벌 상용 생성 모델의 직업 스테레오타입이 상당 부분 지속·변형되어 나타날 수 있음을 실증한 것으로 해석하는 것이 타당하다. 즉 본 연구의 의의는 한국 노동통계라는 로컬 기준선을 활용하여 글로벌 생성형 AI 모델의 직업 이미지 재현이 실제 국내 노동시장 구조와 어떠한 차이를 보이는지를 정량적으로 검증했다는 점에 있다.

아울러 본 연구는 생성 결과에 대한 편향 평가가 단순히 "무엇이 생성되었는가"뿐 아니라 "어떻게 측정했는가"에 따라서도 달라질 수 있음을 실증적으로 제시하였다. 동일한 생성 결과를 대상으로 하더라도 연구자 판독과 자동 판독에서 편향의 크기와 양상이 다르게 나타날 수 있다는 점은, 생성형 AI 공정성 감사(audit) 연구에서 다중 측정 체계와 교차 검증이 방법론적으로 중요함을 보여준다[17],[18]. 이는 향후 생성형 AI 편향 연구가 단일 차원의 성별 편향 분석을 넘어, 연령과의 결합 양상까지 포함하는 교차편향 관점과 함께, 측정 도구 자체의 편향 가능성까지 검토하는 방향으로 확장될 필요가 있음을 시사한다[15]-[18].

5-3 연구의 제한점 및 후속 연구

본 연구는 다음과 같은 제한점을 가진다. 첫째, 본 연구는 한국 사회의 실제 직업 표상을 직접 측정한 것이 아니라, 영어 프롬프트("A portrait photo of a [occupation] in Korea")에 대해 상용 생성형 AI가 산출한 결과를 분석한 감사(audit) 연구라는 점에서 해석 범위의 제한이 있다. 따라서 생성 이미지가 한국인 외형이나 한국 사회의 직업 현실을 충분히 대표한다고 전제할 수 없으며, 본 연구의 결과는 '한국

사회 자체의 편향'이라기보다 '한국 맥락 표지가 포함된 프롬프트 조건에서 드러나는 생성 모델의 재현 편향'으로 이해할 필요가 있다. 또한 한국어 프롬프트를 사용하거나, "a person working as a [occupation]"과 같은 대체 표현, 혹은 한글 간판·복식·업무 공간과 같은 구체적인 문화적 시각 단서를 명시할 경우 결과가 달라질 가능성이 있다. 따라서 후속 연구에서는 언어, 프롬프트 유형, 문화적 단서 유무를 결합한 2×2×2 실험 설계를 통해 각 요인의 효과를 정밀하게 분리 검증할 필요가 있다.

둘째, 모델 버전 및 업데이트에 따른 결과 변동 가능성이 존재한다. 본 연구는 2025년 12월 특정 시점의 모델인 DALL·E 3와 Midjourney V7을 대상으로 수행되었다. 생성형 AI 모델은 업데이트 주기가 매우 짧고, 서비스 정책이나 인물 생성 다양성 관련 내부 조정이 수시로 이루어지므로, 본 연구에서 명시한 버전과 파라미터(--ar 1:1, --s 100, --q 1)에도 불구하고 동일한 조건에서 후속 시점에 재실험할 경우 결과가 달라질 수 있다. 특히 Midjourney와 같은 상용 모델은 내부 알고리즘이 공개되지 않은 블랙박스 구조라는 점에서 재현성 확보에 제약이 있다. 따라서 후속 연구에서는 시점별 반복 감사와 모델 버전 간 비교를 통해 편향의 안정성과 변화 양상을 함께 추적할 필요가 있다.

셋째, 자동 판독 도구인 DeepFace의 기술적 측정 타당성에도 한계가 있다. DeepFace 기반 자동 판독은 대규모 이미지 분석의 효율성을 높여주지만, 본 연구에 사용된 VGG-Face 모델이 주로 서구권 인물 데이터셋을 기반으로 학습되었다는 점을 고려할 필요가 있다. 이는 한국인 얼굴에 대한 성별 및 연령 추정 정확도 저하 가능성으로 이어질 수 있으며, 특히 AI 생성 이미지 특유의 과도하게 매끄러운 피부 질감(hyper-smooth texture)이나 비현실적으로 정제된 얼굴 표현은 연령 추정 오차를 확대할 수 있다. 따라서 후속 연구에서는 얼굴 미검출률 보고, 연구자 판독과의 일치도 추가 검증, 연령 추정 오차 범위 제시, 대안적 얼굴 분석 모델과의 비교를 통해 자동 판독의 측정 타당성을 더욱 강화할 필요가 있다.

넷째, 이중 판독 체계 간 불일치가 시사하는 방법론적 한계가 있다. 본 연구에서 확인된 연구자 판독과 자동 판독 간의 통계적 차이는 편향 평가 결과가 단지 '무엇이 생성되었는가'뿐 아니라 '어떻게 측정하였는가'에 의해서도 달라질 수 있음을 보여준다. 이는 생성형 AI 편향 연구에서 단일 측정 도구에 의존할 경우 결과 해석이 과도하게 단순화되거나 특정 방향으로 치우칠 수 있음을 시사한다. 따라서 향후 AI 공정성 감사 연구에서는 다중 측정 체계, 교차 검증, 평가 기준의 명시적 표준화를 포함하는 방법론적 설계가 필요하다.

다섯째, 분석 대상 직업군 표집과 시각적 표현 방식의 대표성에도 한계가 있다. 본 연구는 15개 직업군을 대상으로 수행되었으므로 한국 노동시장 전체의 직업 구조와 시각적 재현 양상을 완전하게 대변한다고 보기 어렵다. 또한 본 연구에서 사용한 인물 중심 표현인 "portrait photo"는 직업별 인물 이미지를 비교하는 데에는 유리하지만, 직무 수행 맥락보다 외형적 성별·연령 단서를 상대적으로 더 부각시킬 가능성이 있다. 따라서 후속 연구에서는 직업군 범위를 확대하고, "at work"와 같은 업무 중심 프롬프트나 집단 장면, 작업 환경 기반 프롬프트를 함께 활용함으로써 결과의 강건성(robustness)과 일반화 가능성을 추가적으로 검증할 필요가 있다.

5-4 정책적 시사점 및 제언

본 연구 결과는 공정하고 신뢰할 수 있는 생성형 AI 활용을 위해 다음의 정책·실무적 대응이 필요함을 시사한다.

첫째, 다각화된 감사(audit) 체계의 표준화가 필요하다. 단일 자동화 도구에 의존하기보다, 연구자 판독과 자동 분석을 결합한 다중 측정 및 교차 검증 절차를 공정성 평가의 기본 원칙으로 정립할 필요가 있다. 특히 KSCO-8 및 한국 노동통계와 같은 로컬 기준선을 활용하여 국가 맥락에 부합하는 평가 지표를 구축하는 것이 중요하다.

둘째, 모델의 개발·배포 과정에서 선제적 편향 점검(stress testing)과 설계 단계의 공정성 고려를 강화할 필요가 있다. 이를 위해 다양한 직업·인구통계 시나리오를 체계적으로 적용하여 과소표현/왜곡 가능성을 점검하고, 모델 업데이트 이후에도 동일한 평가 체계를 반복 적용함으로써 편향의 변화 양상을 지속적으로 모니터링 해야 한다.

셋째, 서비스 제공자의 투명성 및 책임성을 제고할 필요가 있다. 생성 결과가 통계적 현실과 다르게 편중될 수 있음을 사용자에게 고지하고, 모델·데이터·필터링 정책의 변화가 결과에 미치는 영향을 설명가능한 범위에서 공개하는 노력이 요구된다.

넷째, 언론·교육·공공서비스 등 영향력이 큰 영역에서는 생성형 AI 이미지가 사회적 인식을 형성할 수 있다는 점을 고려하여, 콘텐츠 제작 가이드라인 및 리터러시 교육을 병행할 필요가 있다. 이는 특정 집단에 대한 왜곡된 표상의 확산을 최소화하고, 생성형 AI를 보다 책임 있게 활용하기 위한 제도적 기반이 될 수 있다.

참고문헌

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans: LA, pp. 10674-10685, June 2022. https://doi.org/10.1109/CVPR52688.2022.01042

[2] J. Oppenlaender, "The Creativity of Text-to-Image Generation," in *Proceedings of the 25th International*

Academic Mindtrek Conference (MindTrek 2022), Tampere: Finland, pp. 192-202, November 2022. https://doi.org/10.1145/3569219.3569352

[3] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, ... and A. Caliskan, "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, Chicago: IL, pp. 1493-1504, June 2023. https://doi.org/10.1145/3593013.3594095

[4] K. Crawford, "The Trouble with Bias," in *Proceedings of Invited Talk at the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach: CA, December 2017.

[5] L. Sun, M. Wei, Y. Sun, Y. J. Suh, L. Shen, S. Yang, "Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image-Generative AI," *Journal of Computer-Mediated Communication*, Vol. 29, No. 1, zmad045, 2024. https://doi.org/10.1093/jcmc/zmad045

[6] Statistics Korea. Korean Standard Classification of Occupations (8th Revision) [Internet]. Available: http://kostat.go.kr.

[7] Statistics Korea. Economically Active Population Survey [Internet]. Available: http://kosis.kr.

[8] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," in *Proceedings of the 2020 Innovations in Intelligent Systems and Applications Conference* (ASYU), Istanbul: Turkey, pp. 1-5, October 2020. https://doi.org/10.1109/ASYU50717.2020.9259802

[9] T. Sandoval-Martín and E. Martínez-Sanzo, "Perpetuation of Gender Bias in Visual Representation of Professions in the Generative AI Tools DALL·E and Bing Image Creator," *Social Sciences*, Vol. 13, No. 5, 250, May 2024. https://doi.org/10.3390/socsci13050250

[10] F. Friedrich, K. Hämmerl, P. Schramowski, M. Brack, J. Libovický, K. Kersting, and A. Fraser, "Multilingual Text-to-Image Generation Magnifies Gender Stereotypes," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna: Austria, pp. 19656-19679, July 2025. https://doi.org/10.18653/v1/2025.acl-long.966

[11] W. Wang, H. Bai, J. Huang, Y. Wan, Y. Yuan, H. Qiu, ... and M. Lyu, "New Job, New Gender? Measuring the Social Bias in Image Generation Models," in *Proceedings of the 32nd ACM International Conference on Multimedia,* Melbourne: Australia, pp. 3781-3789, October 2024. https://doi.org/10.1145/3664647.3681433

[12] D. H. R. Spennemann, "Who Is to Blame for the Bias in Visualizations, ChatGPT or DALL-E?," *AI*, Vol. 6, No. 5, 92, April 2025. https://doi.org/10.3390/ai6050092

[13] R. Naik and B. Nushi, "Social Biases Through the Text-to-Image Generation Lens," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, Montréal: Canada, pp. 786-808, August 2023. https://doi.org/10.1145/3600211.3604711

[14] V. Turk, How AI Reduces the World to Stereotypes [Internet]. Available: https://restofworld.org/2023/ai-image-stereotypes.

[15] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," in *Proceedings of the Preconference at the 64th Annual Meeting of the International Communication Association*, Seattle: WA, May 2014.

[16] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency* (pp. 77-91), February 2018.

[17] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, ... and P. Barnes,, "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44), Barcelona: Spain, January 2020. https://doi.org/10.1145/3351095.3372873

[18] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610-623, March 2021. https://doi.org/10.1145/3442188.3445922

이금희(Gumhee Lee)

2003년 : 동국대학교 영상대학원 (공학석사)
2012년 : 숭실대학교 미디어학과 박사수료

2004년～2014년: 한국콘텐츠진흥원
2014년～현 재: 정보통신기획평가원 팀장
※관심분야 : 컴퓨터그래픽스, 가상융합, 멀티미디어 등

김동호(Dongho Kim)

1990년 : 서울대학교 전자공학과 (학사)
1992년 : 한국과학기술원 전기 및 전자공학과(공학석사)
2002년 : 조지워싱턴대학교 전산학과(공학박사)
1995년～1997년: ㈜삼성전자
2003년～현 재: 숭실대학교 글로벌미디어학부 교수
※관심분야 : 메타버스, 컴퓨터그래픽스, 가상현실 등