Multimedia-Assisted Language Learning
Vol. 27, No. 4, 2024
ISSN 1229-8107
e-ISSN: 2982-7302

https://kamall.or.kr
https://kamall.joss.co.kr/
DOI: https://doi.org/10.15702/mall.2024.27.4.70

인공지능 기반 학습자 맞춤형 교육을 위한 형평성과 편향성 연구*

Addressing Bias for Equity in AI-driven Customized Learning Systems

방준성 [1], 이상민 [2]
Junseong Bang, Sangmin-Michelle Lee

Bang, Junseong & Lee, Sangmin-Michelle. (2024). Addressing bias for equity in AI-driven customized learning systems. *Multimedia-Assisted Language Learning, 27*(4), 70-86.

The integration of AI into education has catalyzed transformative changes, particularly in enabling customized and personalized learning experiences that were previously impractical in traditional classroom settings. While earlier AI automation systems focused on educational equality by providing uniform resources and opportunities, more recent AI-driven customized learning systems aim to achieve educational equity by providing differentiated support tailored to individual needs, ensuring comparable learning outcomes across diverse student populations. However, these AI-driven customized learning systems can inadvertently introduce bias in both data collection and algorithmic processing, potentially compromising educational equity. This risk is particularly pronounced in foreign language education, where learner populations exhibit significant demographic and cultural diversity, increasing the potential for data bias. With the planned introduction of AI English digital textbooks in 2025, addressing AI bias and educational inequality becomes critical, particularly in English education. This study systematically examines the sources of AI bias that occur at multiple stages: data collection, analysis, classification, and algorithmic processing. It identifies the manifestations and implications of these biases in educational contexts, and proposes comprehensive solutions that include both technical and non-technical approaches, drawing on existing literature. It also outlines recommendations for key stakeholders -developers, educators, and policymakers- to ensure equity in AI-driven educational systems.

Key words: artificial intelligence, customized learning system, bias, equity, data
Applicable level: all levels

* This study was supported in part through Institute of Information & Communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2024-RS-2024-00425383) grant funded by the Korea government (MSIT) and in part through NRF-2023S1A5A2A21083590.
[1] Junseong Bang (CEO, 1st author, Ymatics Corp., Dogok branch office, Gangnam-gu, Seoul, Korea; E-mail: hjbang21pp@gmail.com)
[2] Sangmin-Michelle Lee (Professor, Corresponding author, Kyung Hee University, 1732, Deogyeong-daero, Yongin, Seoul, Korea; E-mail: sangminlee@khu.ac.kr)

Received: October 31, 2024; Reviewed: November 20, 2024; Accepted: December 15, 2024

I. 서론

디지털 전환(digital transformation)은 사회 전반에 변화와 혁신을 이끌며 교육 분야에도 지대한 영향을 미치고 있다(Kim et al., 2018). 인공지능(AI: Artificial Intelligence)에 의한 학습 환경의 변화는 교육 분야 핵심 관심사들 중에 하나이며, 특히, 텍스트, 이미지, 사운드, 비디오 등 새로운 콘텐츠를 생성할 수 있는 생성형 인공지능(Generative AI)은 기술의 그 활용 범위를 더 확장시켰다. OECD는 디지털 혁신이 교육 분야에 미치는 영향과 각국의 동향을 분석하여 그 활용 방안을 모색하기 위해 2019년부터 매년 「디지털 교육 개관(Digital Education Outlook)」을 발간해 오고 있다. 또한, OECD는 2021년 「국가 AI 전략 및 정책 개요(An Overview Of National AI Strategies and Policies)」 보고서에서 AI가 빠르게 도입되는 분야에 교육을 포함하였다. 특히, OECD는 2024년 8월에 교육에서 형평성(equity)과 포용성(inclusion)을 증진하기 위한 AI의 가능성을 탐색하고 AI 도구의 유형별 기회와 과제를 제시한 보고서 발간하였다(Varsik & Vosberg, 2024).

대한민국 교육부(2023)는 AI 디지털 교과서(AI Digital Textbook, AIDT)를 2025년부터 초등학교 3~4학년, 중학교 1학년, 고등학교 1학년을 대상으로 하여 영어, 수학, 정보 과목부터 단계적으로 도입할 계획이라고 공고했다. 교육 분야 AI 활용 관점에서 세계적으로 선제적이라고 할 수 있으나(Varsik & Vosberg, 2024), 그만큼 교육 환경에 안착을 위해서는 다양한 도전적인 이슈가 존재할 것으로 본다. 기존의 종이 교과서는 다수의 학생들에게 동일한 학습 콘텐츠를 전달할 수 밖에 없는 한계로 인해 교수자가 학습자 역량에 따라 차별화된 콘텐츠를 동시에 다수의 학생들에게 제공하기 어려웠다. 교육 분야 AI 활용의 가장 큰 장점은 개별 학습자에게 맞춤형(customized), 개인화(personalized), 적응형(adaptive) 학습 환경을 제공할 수 있다는 점이다. 이는 “500만 개의 AI 디지털 교과서”를 표방하는 교육부의 목표이기도 하다(Ministry of Education of Korea, 2023). 지식 수준, 흥미, 진로와 같은 개별 학습자의 다양한 변인들을 고려한 맞춤형 학습 제공은 교육의 오랜 숙원이었으나, 교실 현장에서는 여러 제약으로 인해 사실상 실현이 거의 불가능했다(Hong et al., 2024). 그러나 AI 기술의 발전은 학습자 프로파일링 데이터를 이용하여 지식 추적(knowledge tracing) 방식의 학습자 분석을 기반으로, 개별 학습자에게 최적화된 학습 콘텐츠와 학습 속도를 고려한 학습자 맞춤형의 개인화된 학습 환경을 제공할 수 있는 길을 열어주었다(Park et al., 2024). AIDT에서는 교사가 담당 교과의 내용을 학습 목표에 따라 재구성하여 활용할 수 있으며, 학생들은 각자의 수준에 맞는 별도의 활동이나 추가적인 문제 풀이를 통해 개별화된 학습이 가능할 것으로 기대된다(Hong et al., 2024; Ministry of Education of Korea, 2023).

AI 교육시스템에서 작동하는 AI 모델의 출력 결과는 학습에 사용된 데이터에 의해 영향을 받는다. 이러한 특성을 교육에 적용할 때, 데이터 기반 맞춤형 교육이 형평성 문제를 야기할 수 있다는 점에 주목해야 한다. 교육에서의 형평성은 학습자가 자신의 배경이나 환경에 관계없이 공정한 교육 기회와 자원에 접근할 수 있는 것을 의미한다. 이는 모든 학생에게 동일한 기회와 자원을 제공하는 평등성(equality)과는 다른 개념이다. 평등성이 기회와 자원의 동일한 분배에 초점을

맞추는 반면, 형평성은 모든 학습자가 공정한 학습 결과를 달성하는 것을 궁극적 목표로 한다(Lee & Singh, 2021). 따라서 형평성의 관점에서 교육은 각 학습자의 개별적 요구와 상황을 고려하여 성공적인 학습 성과를 달성할 수 있도록 학생 개개인의 학습 능력, 환경, 배경에 따른 차별화된 지원을 제공하고, 이를 통해 다양한 교육적 필요와 학습 잠재력을 가진 학생들이 합리적 기준에서 공정한 결과를 얻을 수 있도록 돕는다(Varsik & Vosberg, 2024). 이 때 공정성(fairness)은 모든 학생이 같은 규칙과 절차에 따라 평가받고 편견 없이 동등하게 대우받는 것을 의미한다(Smith et al., 2017; Wong et al., 2022). 학교 교실의 예를 들면, 시험 시간에 학생 모두가 같은 시간과 규칙 아래 시험을 치르도록 하여 동일한 조건에서 평가받을 수 있도록 하는 것이 공정성이고, 교실에서 수업을 받는 개별 학생들이 같은 교육 자료를 제공받음으로써 동등한 학습 기회를 제공받는 것이 평등성이며, 학습장애가 있는 학생이 다른 학생들과 동일한 학습 성과를 낼 수 있도록 추가 자원을 지원하거나 맞춤형 학습 자료를 제공하는 것이 형평성이라 할 수 있다.

교육 분야 AI 활용의 긍정적 측면으로는 AI가 개별 학습자에게 맞춤형 교육 환경을 제공하고 교육의 접근성과 공정성을 확대함으로써, 상대적 교육 격차와 불평등을 해소하고 교육의 형평성을 달성하는데 기여할 수 있다(Kim & Bang, 2019; Reich & Ito, 2017). 그러나 잠재적인 우려로, AI 활용에 따라 교육 환경에서 기존의 불평등이 강화되거나 새로운 형태의 불평등이 발생할 수 있음이 제기되기도 한다. 예를 들어, 과소대표되는 인종과 언어, 기존에 불평등한 대우를 받던 그룹에 대해서는 AI의 데이터에서도 이러한 편향성이 녹아 존재하게 된다(Holstein & Doroudi, 2021). 편향된 AI가 사회적 불평등을 악화시킬 수 있다는 점이 지적된 것을 상기할 때(Lee & Singh, 2021; Varsik & Vosberg, 2024), AI의 교육 분야 활용에 있어서 공정성·형평성을 확보하지 못하면 기존의 사회적, 경제적 불평등이 심화될 수 있을 것이다. 다시 말하면, AI 도구나 AI 기반 교육 자원 접근성이 AI의 편향성(bias)으로 인해 학습자의 사회·경제적 배경이나 기타 여건과 조건에 따라 달라지게 된다면 교육에서의 불평등을 심화하는 결과를 초래하게 된다(Ibrahim et al., 2020; Varsik & Vosberg, 2024). AI의 편향성이 학습자 대상의 인구학적, 사회·문화적 배경에 의해 영향을 받을 수 있음을 고려할 때, AI의 편향성은 모국어, 제2외국어, 성별, 국적, 인종, 지역, 경제적 배경 등에서 광범위한 차이가 나타나는 언어교육 분야에 더 큰 영향을 미치게 될 것이다(Baker & Hawn, 2022).

형평성은 교육의 핵심 가치를 실현하는 필수 요건이다. AI 기반 학습자 맞춤형 교육이 데이터와 모델 편향성에 의해 형평성을 위배한다면 이는 교육의 가장 기본적인 가치에 반하는 상황을 존속시켜 잠재적으로 사회적, 문화적, 경제적 문제를 양산할 수 있다. 그렇기 때문에, AIDT의 성능적 측면과 더불어, 교육의 기본적 가치 실현에도 관심을 둘 필요가 있다. 맞춤형 교육의 과정에서 AI 편향성 이슈가 존재함에도 불구하고 아직까지 교육 분야에서 이러한 논의가 거의 이루어지지 않고 있다. 교육 분야 AI 활용과 관련하여 국내에 출간된 논문과 보고서의 대부분이 AI 활용의 효과성 탐색이나 AI 윤리(AI ethics)에 집중되어 있다. AIDT가 전국적으로 도입되면 대규모 학습 데이터가 수집될 것으로 예상되는데, 이 과정에서 지역 규모, 학습 성취도, 한국어 능력 등 다양한 요인에 따른 데이터 편향성이 발생할 수 있다. 특히 인구가 적은 지역의 학습자들의 경우, 데이터 표본이

충분히 확보되지 못해 AI 시스템의 정확도가 저하될 수 있으며, 이는 해당 지역 학습자들에 대한 교육적 불이익으로 이어질 수 있다(Lee & Singh, 2021; Varsik & Vosberg, 2024). 따라서 본 연구에서는 기존 문헌을 바탕으로 AI 기반 학습자 맞춤형 교육시스템에서 AI의 편향성이 어떻게 발생하며, 이로 인해 어떤 문제가 생길 수 있는지를 살펴보고, AI의 편향성을 줄일 수 있는 기술적 대안의 방향성에 대해 논하고자 한다. 본 연구는 다음과 같은 연구 질문에 답하고자 한다.

1. AI의 편향성이 발생되는 원인은 무엇이며, 교육 환경에서 어떠한 문제를 야기하는가?

2. AI기반 학습자 맞춤형 교육에서 형평성 향상을 위해 편향성을 줄일 수 있는 기술적 해결 방안은 무엇인가?

II. AI 기반 학습자 맞춤형 교육에서 형평성 이슈와 문제 정의

교육의 형평성은 수평적 형평성과 수직적 형평성으로 구별하여 생각해 볼 수 있다(Varsik &Vosberg, 2024). 수평적 형평성은 교육 시스템의 각 부분에서 공정하게 자원을 제공하고 유사한 상황에 있는 대상에게 비슷한 자원을 제공하는 것으로, 평등성에 유사한 개념이다. 수직적 형평성은 불리한 그룹이나 학교에 그들의 필요에 따라 추가적인 자원을 제공하는 것을 의미하는 것으로, 형평성의 특징적인 개념이다. 기존의 AI 기반 교육은 동일한 콘텐츠를 이용할 수 있는 기회를 모든 학생들에게 제공하고자 하는 수평적 형평성에 더 초점을 맞추었으며, 그 과정에서 평등성의 실현을 위해 교육 분야 디지털 전환을 위한 인터넷 인프라와 스마트 기기들과 같은 환경 구축에 많은 힘을 쏟아 왔다(Lee & Singh, 2021). 그러나 <Figure 1>에서 보이듯이, 디지털 교육 환경이 구성되고 AI 기술이 발전하게 됨에 따라, 최근에 콘텐츠 개인화가 가능한 학습자 맞춤형 교육 실현의 시도에서 수직적 형평성 이슈가 부각되기 시작하였다(Varsik & Vosberg, 2024). 즉, 개별 학습자의 배경과 변인에 관계없이 교육 환경과 학습 과정이 균등하게 제공되었던 이전과 다르게 AI 기반 학습자 맞춤형 교육에서는 개별 학습자의 요구에 맞춰 맞춤형 콘텐츠를 제공받을 수 있다는 장점이 있는 반면, 개인의 학습 이력, 성향, 국적, 거주지, 성별 등에 따라 상대적 불평등(disparity)이 발생할 수 있는 위험성을 내포하고 있다(Kasneci et al., 2023; Kwak & Pardos, 2024). 즉, 맞춤형 콘텐츠를 제공하는 과정에서 전체 학생들을 구분 짓는 여러 속성들(attributes)에 따라 몇 개의 집단(group)으로 혹은 개인(person)으로 세분화하여 학습 콘텐츠를 제공할 때 형평성 이슈가 발생하게 된다.

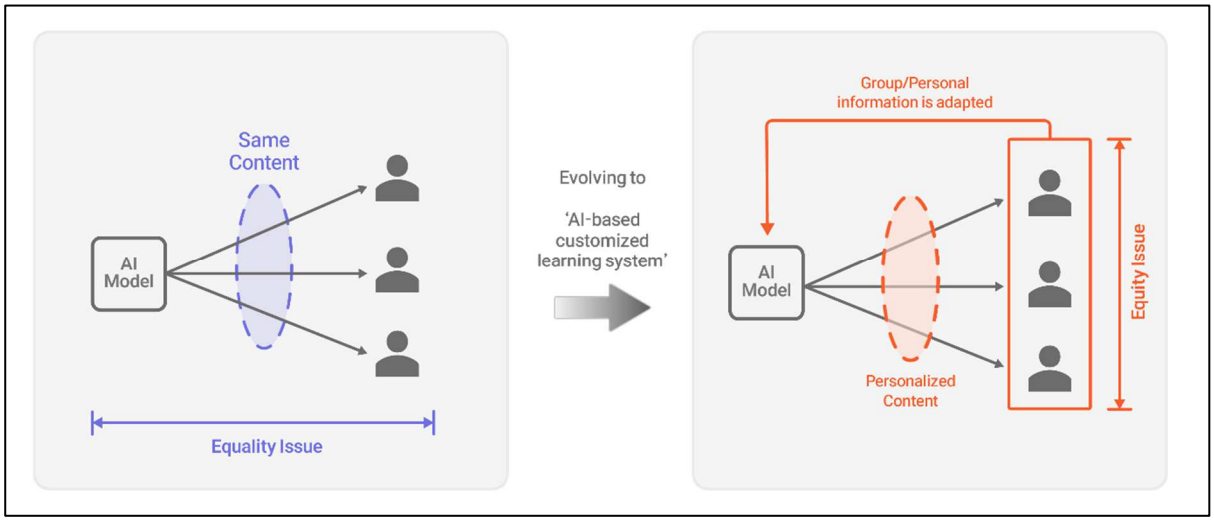

Figure 1

Transition from Equality to Equity in the AI-driven Learning Systems

AI 기반 학습자 맞춤형 교육에서 콘텐츠의 개인화를 위해서는 학생 개별 정보와 학습 분석이 필요하다(Chen et al., 2020). 이 때, 학습 과정 및 결과 데이터 자체에 편향성이 포함되어 있거나 데이터를 처리하는 알고리즘에 편향성이 있는 경우에 형평성 문제, 특히 수직적 형평성 문제를 일으킬 수 있다(Suresh & Guttag, 2021). AI 시스템에서 편향성은 측정과 오류의 통계적 편향부터 모델이 각 그룹에 대해 수행 시 발생하는 불균형, 결과의 체계적인 편향, 그리고 모델 결과가 해석되고 적용되는 과정에서 발생하는 불균형적 영향과 차별에 이르기까지 다양하게 나타난다(Baker & Hawn, 2022, p. 1055). Holstein과 Doroudi(2021)는 AI(AIEd)가 교육 불평등을 야기하는 이유를 1)전반적인 사회-기술 시스템 설계에 내재된 요인들, 2)역사적 불평등을 반영하는 데이터셋의 사용, 3)기계 학습과 자동화된 의사결정을 주도하는 기본 알고리즘에 내재된 요인들, 4)자동화된 의사결정과 인간의 의사결정 사이의 복잡한 상호작용을 통해 나타나는 요인들로 분석하였다.

대표적인 AI의 편향성으로는 과거의 차별적 관행이 데이터에 반영되는 역사적 편견 편향, 특정 그룹이 과대 또는 과소 대표되어 발생하는 표본 편향, 데이터 수집방법이나 도구가 특정 그룹에게 불리하게 작용하는 측정 편향, 데이터를 결합·요약할 때 차이점이 무시되는 집계편향, AI모델 자체가 특정 패턴을 강화하는 알고리즘 편향 등이 있다(Chouldechova & Roth, 2020; Suresh & Guttag, 2021; Zou & Schiebinger, 2018). Sha 외(2022)에 따르면, 교육에서 수집된 학습자 데이터는 성별, 인종, 지역 등에서 실제 사회 데이터와 차이가 나는 "클래스 불균형(class imbalance)" 문제를 발생시키고, 이 문제는 다시 알고리즘의 편향성 문제로 이어지며, 결국에서는 교육에서의 형평성과 공정성을 해치는 결과를 초래한다(Idowu et al., 2024; Plaza et al., 2020). 즉, 실제 인간사회에서는 모든 그룹이 같은 비율로 구성되어 있지 않기 때문에(예: 인종, 지역), 어떤 그룹의 데이터는 다른 그룹에 비해 적거나(분포 편향, distribution bias), 데이터셋 내에서 특정그룹의 예측이 더 어렵거나 쉬워지는 문제(난이도 편향, hardness bias) 등이 발생할 수 있다. 그 결과, 교육 시스템에서 학습 데이터를 바탕으로 한 예측 모델이 학생들의 학습 경로를 결정하는 데 사용될 때, 특정 그룹(예: 저소득층

학생이나 비영어권 학생)에 대한 학습 지원이 부족하게 된다. 즉, AI 시스템의 편향성으로 인해 이 그룹에 대한 예측 정확도가 떨어지게 되고, 교육현장에서 이들을 위한 교사의 적절한 개입이나 처치가 적시에 이루어지지 못하여 불이익을 당할 수 있다(Barbosa et al., 2020; Cavalcanti et al., 2020).

학습 과정 및 결과 데이터 자체에 편향성이 있는 경우 뿐만 아니라, AI 시스템 설계나 구현 과정, 의사결정을 포함하는 알고리즘에서도 편향성이 발생할 수 있다(Baker & Hawn, 2022). 이러한 가능성의 대표적 예로는 과적합(overfitting), 최적화(optimization), 피드백 루프(feedback loop) 등이 있다. 과적합은 모델이 훈련데이터에 너무 잘 맞춰져 있거나 지나치게 특화되어 새로운 데이터에 대한 일반화 능력이 저하되는 경우를 말한다(Salman & Liu, 2019). 최적화는 모델의 성능을 최대화하고 설정한 목표를 극대화하기 위해 매개변수를 조정하는 과정으로 목적 함수에 따라 일부의 데이터 특성이나 문제에 과도하게 집중되어 발생한다. 피드백 루프란 머신러닝 시스템이 편향된 데이터를 학습할 때 발생하는데, 편향된 입력 데이터가 모델의 학습에 영향을 미쳐서 편향된 결과가 나타나고 그 결과가 다시 시스템에 입력되면서 편향이 지속적으로 강화되는 현상을 말한다(Stray, 2023).

AI 기반 학습자 맞춤형 교육을 제공하기 위해서는 학습자 데이터의 수집, 분류, 비교·분석과 이에 근거한 의사결정까지 다단계의 복잡한 과정을 거치게 되는데, 이때, 각각의 과정에서 데이터나 알고리즘의 편향이 발생할 위험이 점차 증가하게 된다(Suresh & Guttag, 2021). AI 기반의 자동화된 학습 콘텐츠 제공 기능은 동등한 교육 기회와 자원 제공의 차원에서 평등성에 위배되지 않았지만, AI 기반 학습자 맞춤형 교육에서는 의도하지 않게 내재된 데이터나 알고리즘 편향에 의해 형평성을 저해할 위험성이 증가된다(Lee et al., 2024). 즉, AI 기반 학습자 맞춤형 교육 시스템의 개발과 도입에는 성능 평가와 함께 다양한 관점에서의 형평성 연구가 필요하며 형평성 향상을 위해 기본적으로 편향성 완화나 설명가능성 등을 실현할 수 있는 기술적 논의가 있어야 함을 알 수 있다.

III. 편향성과 형평성 이슈 사례 분석

AI 기반 학습자 맞춤형 교육은 개별 학생의 학습이력 데이터를 실시간으로 분석하고 그에 맞는 학습 자료를 제공하는 방식으로 작동된다. 이 과정에서 학생들의 다양한 배경과 학습 능력에 맞는 공정한 기회와 자원이 제공되지 못할 때 형평성 문제가 발생한다. 앞에서 설명한 바와 같이 AI 기반 교육 시스템을 사용하는 과정에서 형평성 문제가 데이터의 편향성과 알고리즘의 편향성에서 촉발될 수 있음을 인지할 때, 우리나라의 교육 현장의 시나리오나 사례 연구를 통해 기술적으로 실현해야 하는 형평성 보장의 범위와 함께 이에 영향을 주는 편향성에 대해 살펴볼 필요가 있다. 교육에서의 데이터 편향성은 데이터 수집에서 학습자 분석, 개별 맞춤형 학습을 결정하는 전 과정에서 발생할 수 있는데(Suresh & Guttag, 2021), 대표적인 사례는 다음과 같다. 첫째, 교육용 AI 시스템이 훈련되는 데이터가 특정 집단을 불균형적으로 반영할 경우에 데이터 편향이 발생한다. 예를 들어, AI가 주로 고소득층이나 특정 인종이나 지역의

데이터를 기반으로 학습하면, 다양한 학생 집단을 충분히 대표하지 못한다. 이로 인해 AI가 소수자 그룹을 정확하게 평가하지 못하게 되는 문제가 생긴다. 고소득층 학교는 디지털 학습 도구에 대한 접근성이 더 확보되기 때문에 이들의 학습 데이터가 많이 수집되는 반면, 저소득층 학교는 데이터가 부족할 수 있다. 또한, 인구밀집도가 큰 대도시 지역의 학생들에 대한 데이터는 많이 수집되고, 농촌 지역의 학생 데이터는 부족할 수 있다. 이런 경우에 저소득층 학생들이나 농촌 학생들의 성적과 학습 패턴에 대한 데이터 부족으로 인해 고소득층 지역이나 대도시 학생들에 비해 이들에 대한 예측 정확도가 떨어지게 된다. 실제로, 교사 인터뷰를 통해 조사한 Holstein과 McLaren외 (2019)의 연구에 의하면 AI시스템이 특정학생 그룹에 대해 부정확하거나 불공정한 평가를 내릴 수 있는 우려가 있는 것으로 나타났다.

둘째, AI 시스템이 편향이 존재하는 기존의 데이터만을 학습할 경우, 데이터에 녹아 있는 역사적 편향을 그대로 학습하게 된다(Suresh & Guttag, 2021). 그러므로, 편향성이 있는 과거의 학습 기록이나 성적 데이터를 사용하면 이 데이터는 기존의 사회적 불평등을 반영하게 될 가능성이 커진다(Lee & Singh, 2021). Alesina 외(2024)의 연구에 의하면, 교사들의 고정관념이 학생들의 평가에 영향을 미쳐서 이민자 학생들이 대해서는 더 부정적인 평가를 하는 것으로 나타났다. 이와 같이 인간이 수행하는 평가에 편견이나 고정관념이 개입되는 것은 드물지 않는 일인데, 성별이나 인종에 따라 평가 기준이 달랐던 이러한 과거 데이터가 AI에 그대로 사용될 경우에 AI는 이 차별적 패턴을 반복하게 된다. 이런 맥락에서 Perry(2019)의 연구에서도 AI의 활용이 교육에서의 형평성을 높이기 보다는 오히려 흑인과 라틴계 학생들과 같은 특정 인종에 대한 기존의 편견을 심화시킬 수 있음을 지적하고 있다.

셋째, 훈련 데이터의 편향성 때문에 실제 학습자 데이터 간 불일치가 발생하여 특정 그룹의 학습자들에게 불리한 결과를 초래할 수 있다. Meaney와 Fikes(2019)는 MOOC와 같은 가상학습공간(Virtual Learning Environment, VLE)의 학습분석 설계를 조사해본 결과, VLE을 활용하는 학습 플랫폼이 학습자들의 초기 패턴을 기반으로 학습코스를 설계하였다는 것을 발견하였다. VLE 시스템은 초기에 학습 준비도와 동기가 높은 선도적 학습자들의 행동패턴을 기반으로 설계되었다. 그러나 이후 유입된 학습자들은 상대적으로 학습자원과 준비도가 부족했음에도 불구하고, 시스템은 이들의 특성을 반영하지 못했다. 결과적으로 이 시스템은 고학력자와 학습 환경에 익숙한 학생들에게 유리하게 작용한 반면, 추가적인 학습 지원이 필요한 저학력 학습자들에게는 적절한 학습 경험을 제공하지 못하는 한계를 보였다.

마지막으로, 데이터 자체 또는 그 처리 과정에서의 편향 이외에 알고리즘을 최적화하는 과정에서도 편향이 발생할 수 있는데, 이 또한 학습자 간 불평등을 초래하게 된다. 그 예로, 표준화 시험을 최적화하는 과정에서 창의성이나 비판적 사고 같은 인지능력이 간과되어 불리한 결과를 받는 학습자가 발생할 수 있다(Baker & Hawn, 2022; Holstein, Wortman, et al., 2019). 다른 사례로 AI 피드백 루프를 들 수 있다. AI 시스템에서의 피드백 루프는 AI의 예측이나 추천이 학생의 행동을 변화시키고 이 변화된 행동이 다시 AI의 판단에 영향을 미치는 순환적 과정으로

확대되는데, 교육 사이클에서 이 과정은 환류효과(washback effect)와 유사한 형태의 결과를 가져올 수 있다. 즉, AI가 특정 학생을 저성과자로 분류하면 교사는 그 학생에게 덜 도전적인 과제를 제공하게 되고 이로 인해 그 학생의 성과가 실제로 낮아져서 AI의 초기 분류가 강화되어 결국 그 학생에게 불리한 결과를 초래하게 된다(Yu et al., 2017).

교육 분야에서 AI는 언어교육에도 큰 영향을 주어 언어교육의 방식과 학습 시스템 구현에도 변화를 이끌어내고 있다. AI 번역 및 문법교정 도구, 대화형 챗봇 등의 활용은 영어가 모국어가 아닌 학습자들이 정보와 지식을 이해하고 전달함에 있어서 접근성을 크게 높이는 긍정적인 결과를 가져왔으며, 영어 뿐만 아니라 다양한 언어 간 장벽을 낮추는데 기여하고 있다(Warschauer et al., 2023). 그러나 언어교육의 특성상 학습 대상의 배경, 인종, 학교급, 지역, 언어 등과 같은 다양한 변인이 내재되어 데이터와 알고리즘에서의 편향에 대한 가능성이 더 높아질 수 있음을 짐작할 수 있고, 실제로 Baker와 Hawn(2022)는 AI 시스템에 특정 언어에 대한 편견과 편향이 존재한다고 보고하였다. Sha 외(2021)의 연구에서는 학습자가 게시판에 올린 글의 언어가 영어인지의 여부에 따라 AI(Automatic Classifiers)가 학습자의 게시물을 불공정하게 자동분류를 하는 것으로 드러났다. Naismith 외(2018)의 연구에서는 어휘 수준을 측정하는 일반적인 지표가 모국어 화자에게 더 적합한 단어 목록을 바탕으로 개발되었기 때문에 제2언어 학습자에게는 적합하지 않거나 불리하게 작용하는 것으로 나타났다.

언어교육에서 관심을 받고 있는 AI 기반 학습과 평가도구에 대한 연구에서도 AI의 이점과 한계에 대한 논의가 이어지고 있다. Holmes(2023)는 AI의 활용이 더 객관적이고 일관적인 평가결과를 제시하여 평가의 공정성을 높일 수 있다고 주장하였다. 그러나 다른 연구에서는 AI가 학습자의 언어능력을 평가할 때 학습자의 인종, 문화적 배경, 작문 스타일 등에 따라 편향성이 나타나는 것으로 조사되었다. Bridgeman 외(2012)의 초기 연구에 의하면, ETS에서 도입한 자동 쓰기평가 시스템인 E-Rater를 조사한 결과, 인간 평가자(human rater)의 평가에 비해 중국인과 한국인 피평가자들에게 더 높은 점수를 부여하고 아랍과 힌두 학습자에게는 상대적으로 낮게 편향되어 채점한 것으로 나타났다. Ramineni와 Williamson(2018)의 연구에서도 E-Rater가 인간 평가자의 평가에 비해 아프리카 미국인 학습자들에게 더 낮은 점수를 부여한 것으로 조사되었다. 또 다른 연구에서 SpeechRater 시스템이 외국어 의사소통능력을 평가할 때 특정 모국어 그룹의 피평가자들을 더 높거나 낮게 평가하는 것으로 나타났다(Wang et al., 2018). 외국어 학습에서 널리 사용되고 있는 AI 챗봇에서도 대화 과정에서 다양한 편견이 나타났는데, 이는 문화적 다양성에 대한 인식과 이에 대한 특정 그룹에 대한 데이터 부족에 의한 것으로 볼 수 있다(Holmes, 2023).

IV. 기술적 해결 방안에 대한 탐구

AI 기반 학습자 맞춤형 교육 시스템을 개발하는 과정에서 데이터와 알고리즘의 편향성을

완화하고 형평성을 실현하기 위한 다양한 방법이 존재할 수 있다. 앞에서 언급한 내용들을 바탕으로, 훈련 데이터가 특정 집단을 불균형적으로 반영한 경우, 평가 기준이 달랐던 과거의 학습 이력에 대한 데이터만 사용한 경우, 개발에 사용된 데이터와 실제 학습자의 데이터 간 불일치가 큰 경우, 알고리즘 최적화나 AI 시스템에 대한 피드백 루프에 의한 편향이 발생하는 경우에 대해 고민해 볼 수 있다.

훈련 데이터가 특정 집단을 불균형적으로 반영한 경우는 데이터 다양성을 충분히 확보하여 AI가 특정 집단을 과소평가하거나 과대평가하지 않도록 할 수 있다. 데이터 다양성을 확보하기 위해서는 데이터 수집 단계에서 다양한 인종, 성별, 나이, 사회경제적 배경, 지역적 특성을 가진 인구 집단을 포함해야 한다(Baker & Hawn, 2022). 또한, AI 모델이 특정 집단을 중심으로 최적화되지 않고 전체 집단에 공정하게 동작할 수 있도록 보장해야 한다. 이를 위해 무작위 샘플링을 통한 데이터 수집보다는 인구통계 데이터를 기반으로 대표성이 부족한 그룹을 인위적으로 더 포함할 수 있도록 하는 가중치 샘플링(weighted sampling)을 사용하는 것도 한 방법이다. 이와 함께 편향된 데이터를 교정하기 위해 소수 집단의 데이터를 인위적으로 생성하여 증강하는 방법도 활용할 수 있다(Jiang & Pardos, 2021; Zanna & Sano, 2024). 예를 들어, 한부모가정, 탈북자가정 등의 소외계층에 대한 데이터를 의도적으로 더 많이 수집하거나 데이터를 증강시킬 수 있다. 이때 인구통계학적 데이터 수집과 처리가 맞춤형 교육에 대한 전체 변인들을 대변하는 것이 아니기 때문에, AI 기반 학습자 맞춤형 교육을 위해 콘텐츠를 개인화하는 과정에서 사용된 파라미터(parameters)들을 포함한 속성들(attributes)로 다양성에 대한 형평성 이슈를 시뮬레이션 하는 것이 도움이 될 수 있다(Jiang & Pardos, 2021). 간단하게는 데이터셋이 속성들에 의해 그룹화된 특정 집단을 과소대표하거나 과대대표하고 있는지 통계적으로 데이터 불균형 상태를 살펴볼 수 있기도 하다. 단, 효과적인 분석을 위해서는 형평성 평가를 위한 지표(metric)에 대한 연구가 선행될 필요가 있다. 현재까지 알려진 교육에서의 AI 데이터 편향성을 줄이기 위한 지표로는 대표적으로 Area Between Receiver Operating Characteristic Curves (ABROCA; 서로 다른 인구집단 간 모델 성능 차이를 평가하는 지표), Area Under the Curve(AUC) Gap (다중 하위 그룹 간의 성능 격차를 포착하여 비이진적 형평성을 평가하는 지표), Model Absolute Density Distance (MADD; 예측 성능과 독립적으로 모델의 차별적 행동을 분석할 때 사용되는 지표) 등이 있다.

AI 모델 학습을 위해 충분히 크고 편향되지 않은 데이터셋을 얻는 것은 어렵다. 최근에는 작은 데이터 세트와 잠재적으로 편향된 사전 학습된 모델만 사용하여 AI 모델의 편향을 제거하는 여러 방법을 사용하는 포괄적인 접근 방식이 연구되었는데, 이는 데이터 분할, 로컬 학습 및 정규화된 미세 조정을 통해 사전 학습된 모델의 역편향으로 여러 모델을 학습하여 잠재적으로 역편향된 모델을 얻고 모든 모델에 앙상블 학습을 사용하여 편향되지 않은 예측을 하는 방법이다(Radwan et al., 2024). 평가 기준이 달랐던 과거의 학습 이력에 대한 데이터만 사용한 경우는 최신의 데이터로 AI 모델을 재학습시킴으로써 그 편향성을 일부 완화할 수 있다. 그러나, 동일한 입력에 대한 AI 모델의 출력 결과는 그 AI 모델이 학습한 데이터에 의해 영향을 받기

때문에, 의도한 것과 다르게 한 그룹의 편향성을 완화하기 위해 최신의 데이터를 사용하는 것이 다른 그룹의 편향성을 악화시키는 상황을 초래할 수 있다. 그렇기 때문에 최신의 데이터로 AI 모델을 학습하기 전에 다양한 집단들 사이의 형평성을 사전에 평가해 볼 필요가 있으며 이를 근거로 하여 데이터셋 조정이 필요할 수 있다(Yang, et al., 2022; Zanna & Sano, 2024).

앞에서 설명한 것과 같이 데이터 다양성 확보와 최신화를 통해 편향성을 완화하고 형평성을 향상시킬 수 있지만, 특정 입력에 대한 시스템 출력의 결과를 비교·분석하여 형평성을 평가하고 AI 시스템을 모니터링 해볼 수 있다. 이는 개발에 사용된 데이터와 실제 학습자의 데이터 간 불일치가 큰 경우에도 활용이 가능할 것이라 본다. 개념적으로 AI 모델이 인종, 성별, 지역 등의 속성에 따라 분류된 각 그룹에 대해 동등한 예측 성능을 제공하는지를 평가해볼 수 있다. 입력 데이터 θ에 대한 AI 모델의 예상 출력 결과를 $\hat{R}(\theta)$, 속성에 따라 분류된 N개 그룹의 출력 결과를 $G_k(\theta)\ (k = 1,..,N)$라고 할 때, 예를 들어, $\left|\hat{R}(\theta) - G_k(\theta)\right|$과 같이 참조값과 각 그룹의 출력값을 비교할 수 있고 $\left|G_i(\theta) - G_j(\theta)\right|\ (i \neq j,\ i = 1,..,N, j = 1,..,N)$와 같이 서로 다른 각 그룹의 출력값을 비교하는 방식을 고려해볼 수 있다. 참조값과 각 그룹의 출력값을 비교하는 방식은 $\sum_k \left|\hat{R}(\theta) - G_k(\theta)\right|$와 같이 전체 그룹에 대해 순차적으로 상대적 비교가 가능하나 $\hat{R}(\theta)$의 추정이 중요하다. 서로 다른 각 그룹의 출력값을 비교하는 방식은 단순하나 속성에 따라 분류된 그룹의 수(N)가 증가할수록 그 복잡성이 증가하게 된다. 알고리즘에 손실 함수(Loss Function)을 추가하여 두 그룹간 차이를 줄이는 방식으로 AI 모델의 성능 평가를 추가할 수 있다. 형평성 평가 시뮬레이션에서 결과 비교값의 차이가 기준치(Φ)를 넘길 때, 즉, $\left|\hat{R}(\theta) - G_k(\theta)\right| > \Phi$나 $\left|G_i(\theta) - G_j(\theta)\right| > \Phi$일 때 알람이 있도록 설정할 수 있다. 입력값 사이의 벡터 거리와 출력값 사이의 벡터 거리를 측정하여 비교하는 방법으로 비슷한 특성과 학습 이력을 가진 학생들이 유사한 결과를 받는지 살펴보는 방법으로 확장하여 구성해볼 수 있다.

AI 시스템은 형평성을 고려하여 맞춤형 콘텐츠 제공하는 기능을 갖겠지만, 그룹 간 상황에 따라 불평등에 따른 그 영향 정도가 다르게 된다. 추가적인 연구를 통해 불평등 영향(Inequality Impact)을 추정하여 형평성 기준에 반영해볼 수 있다. 평등한 기회 제공과 함께, 예측 형평성(predictive equity) 측면에서 각각 그룹에 소속된 학생들의 예측된 학업 성취도가 실제 성취도와 얼마나 일치하는지 살펴봄으로써 형평성 평가 성능 개선이 가능하다.

V. AI 활용에 따른 형평성에 대한 논의

AI는 교육 현장의 기존 질서를 새롭게 하는 혁신 도구가 될 것이라는 예측에서 우리는 기대와 우려의 양면을 본다. 최근 우리나라에서는 AIDT 도입을 앞두고 AI의 교육적 활용에 대한 효과성 연구가 활발히 진행되고 있다. 그러나 이러한 연구들이 주로 AI 기술 활용을 통한 학업 성과 향상에만 초점을 맞추고 있어서 AI가 교실 환경과 전반적인 교육 생태계에 미치는 영향에

대해서는 간과하고 있다(Varsik & Vosberg, 2024). AI 기술의 발전으로 AI는 단순한 교수-학습 보조 도구를 넘어서 데이터를 기반으로 학생들의 학습과 평가를 지원하며 인간의 교육 활동을 지원할 수 있다(Kim & Bang, 2020). 그 과정에서 AI 시스템 도입에 교육의 형평성과 공정성을 고려하여 학생들 사이의 불평등을 줄이고 교사들의 무의식적 편견이 발생하지 않도록 함으로써(Holstein & Doroudi, 2021), 교육 분야 AI 활용에 따른 잠재적인 문제들을 사전에 방지하고 교육 현장을 더 민주적인 환경으로 개선할 수 있을 것으로 기대된다.

교육에서의 디지털 형평성(digital equity)은 학생들이 디지털 기술, 기술 습득, 활용 및 태도에 있어 공정성과 형평성을 촉진하는 것을 의미한다(Varsik & Vosberg, 2024). 과거에 디지털 격차(digital divide)에 의해 디지털 형평성이 훼손될 수 있음에 대한 우려가 있었던 것처럼, 최근에는 AI 격차(AI divide)가 새로운 교육 불평등의 원인으로 떠오르게 되었다. 인터넷 접근성이 낮거나 저소득층에 속한 학생들, 또는 AI 리터러시(AI literacy)가 부족한 학생들은 AI 기술을 충분히 활용하지 못해 교육 기회에서 소외될 위험이 있으며 이는 교육 불평등을 더욱 증대시킬 수 있다(Bulathwela et al., 2024; Reich & Ito, 2017). AI 시스템에 사용되는 데이터와 알고리즘은 다양한 이유로 편향이 발생할 수 있으며, 이는 특정 학습자들에게 불평등과 불공정을 야기할 수 있다(Chouldechova & Roth, 2020; Suresh & Guttag, 2021). 편향을 갖는 AI 모델에 의해 학습자들은 학업 성적 뿐만 아니라 진로, 학습 동기, 자기효능감 등 사회·정서적 측면에서 부정적 영향을 받게 된다. AI 기술 개발이 주로 선진국과 중국 등 일부 국가에 집중되고 있는 현실은 개인 간 격차를 넘어 국가 간 경제적 불균형을 심화시킬 가능성 또한 크다(Varsik & Vosberg, 2024). 더욱이, AI 격차는 기존의 디지털 격차와는 본질적으로 다른 양상을 보일 것으로 예측된다. 디지털 격차가 주로 인터넷 접근성이나 하드웨어 기기 보유 여부와 관련된다면, AI 격차는 데이터의 대표성, 알고리즘의 공정성, AI 리터러시 등 훨씬 더 복잡하고 다층적인 요인들로 구성되기 때문이다. 이러한 AI 격차의 복잡성은 그 해결 방안 또한 단순하지 않을 것을 시사한다. 따라서 AI 격차 해소를 위해서는 기술적 접근 뿐만 아니라 정책적, 교육적, 그리고 윤리적 차원에서의 종합적인 노력이 요구된다.

모든 학생들이 동일한 콘텐츠를 이용하는 상황에서는 불공정한 상황이 발생하더라도 모두가 함께 적응하거나 문제 발견시 의견 합치에 의해 빠르게 해결하려는 노력을 하기 때문에 긍정적이던 부정적이던 평등하다. 그러나, 형평성의 관점에서 학습자 맞춤형 콘텐츠를 이용하는 학생들은 상대적 불평등 상황을 인식하지 못할 가능성이 크며, 누군가 이러한 문제를 인식하였다고 해도 이를 시정하기에 교육 기관의 조사와 해결 방안 강구에 많은 노력과 시간이 필요하다. 특히, 불평등의 입장에 놓인 대상이 소외계층일 경우에 이를 개선하기는 쉽지 않다. 데이터의 검증이 어려울 뿐 아니라, 해당 소외계층의 데이터를 교체나 증식하는 과정에서 AI 모델 전체에 영향을 주게 되어 또다른 소외계층 집단의 불평등이 발생시킬 수 있기 때문이다(Knox et al., 2019; Suresh & Guttag, 2021). AI 기반 학습자 맞춤형 교육 프레임 내에서는 데이터와 알고리즘을 다루는 과정 중 편향성 발생은 피하기 어렵고 따라서 형평성 문제는

필연적으로 -보이거나 보이지 않거나- 발생하게 된다.

AI 기반 학습자 맞춤형 교육에서 형평성을 향상하기 위해서 우선적으로 데이터 구축 단계에서 데이터의 품질과 다양성을 확보해야 한다. 이를 위해서는 다양한 인구통계학적 특성을 반영한 데이터셋을 구축하고, 학습 데이터의 대표성과 포괄성을 확보하여 특정 집단이 과대 또는 과소대표되지 않도록 해야 한다(Eden et al., 2024; Holstein & Doroudi, 2021). AI가 학습하는 데이터는 시간이 지남에 따라 현재의 그것과 다를 수 있다. 따라서 데이터 수집 및 모델 학습 과정에서 지속적인 모니터링과 업데이트를 통해 데이터의 다양성과 공정성을 유지해야 할 필요가 있다. 데이터 사용 후 그 성과와 결과를 분석하여, 부족한 부분을 다시 보완할 수 있는 피드백 루프를 구축해야 한다. 특정 시점에서 수집한 데이터만을 사용하는 것이 아니라, 주기적으로 데이터를 갱신하여 최신 데이터를 반영하고, 사회적 변화나 인구학적 변동을 추적하여 모델의 성능과 공정성과 형평성을 개선해야 한다(Knox et al., 2019). AI 알고리즘은 통상 복잡하고 불투명하여 블랙박스로 여겨지는데, 교육의 형평성을 향상시키기 위해서는 알고리즘 설계 및 개발 단계에서부터 공정한 매트릭스(예: fairness toolkits)를 도입해야 한다. 이때 다양한 이해관계자(교사, 학생, 학부모, 교육전문가)를 참여시켜 편향 여부를 지속적으로 점검하고, 이해관계자들이 알고리즘을 이해할 수 있도록 투명하고(transparent) 설명가능(eXplainable)하도록 설계하여 한다(Lee & Singh, 2021). AI 기반 학습자 맞춤형 교육에서 편향성으로 인하여 형평성이 크게 훼손되는 경우 일정 수준까지만 맞춤형 서비스를 제공하는, 즉 개인화 맞춤형 서비스의 범위를 다소간 제한하여 부정적 효과를 낮추는 방법도 고려할 수 있다.

그러나 기술적 해결책만으로는 교육에서의 편향성, 형평성, 공정성 문제를 완전히 해결할 수 없다. AI 도구의 효과는 사용자의 사용 방식에 따라 크게 달라질 수 있다. 따라서 AI 도구의 이점을 최대화하기 위해서는 교사의 역할이 중요하다. AI는 교육에 큰 잠재력을 가지고 있지만, 그 잠재력을 실현하기 위해서는 신중한 접근과 지속적인 연구, 그리고 교사들의 적극적인 노력이 필요하다. AI가 이미 깊숙이 교육에 들어온 현재, 실제 현장에서 학습자와 직접 상호작용을 하고 많은 교수적 결정을 내려야 하는 교사가 AI 격차나 불형평성에서 발생할 수 있는 차별과 부정적인 영향을 최소화하고 공정성과 형평성을 유지해야 할 책임을 맡게 된다. 그러므로 교사는 AI 시스템에서 발생할 수 있는 편견과 편향성에 대해 인식하고 대처할 수 있는 역량을 강화해야 한다. 이에 Deng과 Zhang(2023)은 21세기 교육에서 필요한 다양한 교사 역량을 강조하는 기존의 TPACK(Technological Pedagogical Content Knowledge) 모형에서 교사의 윤리 역량까지 포함한 TPCEK(Technological Pedagogical Content Ethical Knowledge) 모형을 제안하였다. AI 시대의 교육에서 TPCEK 영역의 교사 역량은 필수적이며, 이를 개발하기 위해 포괄적이고 지속적인 교사 연수가 요구된다.

Unterhalter(2009)는 교육의 형평성을 사회 정의 실현, 경제적 효율성 향상, 사회통합 촉진, 개인의 삶의 질 향상, 그리고 지속가능한 발전을 위한 필수 요소로 보았다. Unterhalter에 따르면 교육의 형평성은 단순히 교육 분야의 문제가 아닌 사회 전체의 발전과 정의를 위한 핵심적인

요소인 것이다. 더 나아가 Ladson-Billings(2006)는 교육의 형평성 달성 실패를 후세대에 대한 부채로 규정하고, 이를 역사적, 경제적, 사회정치적, 도덕적 부채로 분류하였다. Ladson-Billings(2006)는 이러한 다양한 형태의 부채가 누적되어 현재의 교육 불평등을 야기했다고 주장하였다. 교육에서의 불평등은 복합적인 원인으로 발생하고, 본 연구에서 설명한 바와 같이 AI 기반 학습자 맞춤형 교육에서는 그 원인이 더욱 복잡하기 때문에, 교육의 형평성 달성을 위해서는 문제에 대한 근본적인 접근과 함께 교육의 접근성, 참여, 결과 등에서 다차원적인 노력이 요구된다(Darling-Hammond, 2001; Unterhalter, 2009).

VI. 결론

AI 기술 발전에 의해 학생들의 소수 집단 또는 개인 맞춤형 서비스가 가능해짐에 따라 교육의 평등성 뿐 아니라 교육의 형평성 문제는 시급하게 짚어보아야 할 주제가 되었다. 이에 본 연구는 기존 문헌연구를 바탕으로 AI 기반 학습자 맞춤형 교육에서의 형평성 문제를 살펴보고 AI 편향성이 발생할 수 있는 원인과 결과, 교육현장에서의 사례를 조사하였다. 연구 결과, AI 기반 학습자 맞춤형 교육에서는 여러 단계에서 다양한 원인으로 데이터와 알고리즘 편향성이 발생할 가능성이 존재하였고, 실제로 이런 사례들이 교육현장에서 나타나고 있음을 발견하였다. 본 연구에서는 AI 기반 학습자 맞춤형 교육을 제공 시 발생할 수 있는 AI 편향성을 완화시키거나 예방할 수 있는 기술적, 기술외적 해결방안을 제시하였다. 자동화와 교육의 평등성을 강조했던 기존의 AI 서비스에서 발생하는 편향은 전체 대상이 모두 영향을 받았지만, 특정 학습자나 그룹에 맞춤형 학습을 목적으로 한 AI기반 교육시스템에서는 소수집단에 발생하는 형평성 문제를 발견하기 어려우며 이를 개선하기도 쉽지 않다. 학습 효과 향상을 위해 맞춤형 학습을 제공하는 것은 교육적 이점이 있으나, AI 기반 교육시스템에서의 상대적 불평등을 줄이기 위해 형평성에 대한 모니터링이 필요하다.

AI는 교육 현장에서 맞춤형 학습을 통해 교육의 형평성을 높일 수 있는 잠재력을 지닌 동시에, 데이터와 알고리즘의 편향성으로 인해 오히려 불평등을 심화시킬 수 있는 양면성을 가지고 있다. AIDT 도입을 앞둔 우리나라 교육계는 이러한 AI의 특성을 인지하고, 데이터 축적 과정에서 발생할 수 있는 편향성을 지속적으로 검증하고 최소화하기 위한 체계적인 연구와 대응이 필요하다. 또한 교육 구성원 모두가 AI의 객관성에 대한 맹신을 경계하고, 그 한계를 명확히 인식해야 한다. AIDT의 도입을 앞두고 현재는 AIDT의 교실현장 적용과 학습효과성에 주의가 집중되어 있으나, 편향성 문제는 언제든지 일어날 수 있는 일이므로, 도입 초기부터 기술적으로, 그리고 정책적으로 이를 막을 수 있는 방법을 모색해야 한다.

본 연구는 AIDT가 2025년에 도입되어 현장에서 활용될 때 발생할 수 있는 중요한 주제인 AI 기반 학습자 맞춤형 교육에서의 형평성과 편향성에 대해 화두를 던졌다는 점에서 그 의의가 있다. 연구의 제한점으로는 아직 학습자 데이터가 존재하지 않아서 실제 데이터를 기반으로 AI

편향성을 조사하지 못했다는 점이다. AI 기반 교육에서의 교육 형평성은 중요한 주제이므로, 향후 과제에서는 실제 학습자 데이터가 확보된 이후에 데이터와 알고리즘을 분석하고 현장 조사를 해서 편향성을 줄이고 형평성을 향상시키는 실질적인 방안을 마련해야 한다. 또한 AI 기반 학습자 맞춤형 교육이 시스템적으로 제공될 수 있도록 형평성을 평가할 수 있는 지표와 알고리즘 개발이 필요하다. 지표 개발을 위해서는 다양한 속성에서의 학습자 집단 연구가 수행되어야 할 것이다.

REFERENCES

Alesina, A., Carlana, M., LaFerrara, E., & Pinotti, P. (2024). Revealing stereotypes: Evidence from immigrants in schools. *American Economic Review, 114* (7), 1916-1948. https://doi.org/10.1257/aer.20191184

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32,* 1052-1092. https://doi.org/10.1007/s40593-021-00285-9

Barbosa, G., Camelo, R., Cavalcanti, A., Miranda, P., Mello, R., Kovanović, V., & Gašević, D. (2020). Towards automatic cross-language classification of cognitive presence in online discussions. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge,* 605- 614. https://doi.org/10.1145/3375462.3375496

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27-40. https://doi.org/10.1080/08957 347.2012.635502

Bulathwela, S., Pérez-Ortiz, M., Holloway, C., Cukurova, M., & Shawe-Taylor, J. (2024). Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools. *Sustainability, 16*(2), 781-801. https://doi.org/10.3390/su16020781

Cavalcanti, A. Diego, R. F. Mello, K. Mangaroska, A. Nascimento, F. Freitas, & Gašević, D. (2020). How good is my feedback? A content analysis of written feedback. *Proceedings of the tenth international conference on learning analytics & knowledge,* 428-437. https://doi.org/10.1145/3375462.337547

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264-75278. http://doi: 10.1109/ACCESS.2020.2988510.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM, 63*(5), 82-89. https://doi.org/10.1145/3376898

Darling-Hammond, L. (2001). Inequality in teaching and schooling: How opportunity is rationed to students of color in America. In B. D. Smedley, A. Y. Stith, & A. R. Nelson (Eds.), *The right thing to do, the smart thing to do: Enhancing diversity in the health professions* (pp. 208-233). National Academy Press. https://www.ncbi.nlm.nih.gov/books/NBK223640/

Deng, G., & Zhang, J. (2023). Technological pedagogical content ethical knowledge (TPCEK): The development of an assessment instrument for pre-service teachers. *Computers & Education, 197*, Article, 10740. https://doi.org/10.1016/j.compedu.2023.104740

Eden, C-A., Chisom, O. N., & Adeniyi, I. S. (2024). Integrating AI in education: Opportunities, challenges, and ethical considerations. *Magna Scientia Advanced Research and Review. 10*(2), 6-13. https://doi.org/10.30574/msarr.2024.10.2.0039

Holmes, W. (2023). The unintended consequences of artificial intelligence and education. *Education International.* Retrieved October 10, 2024 from https://discovery.ucl.ac.uk/id/eprint/10179267

Holstein, K., & Doroudi, S. (2021). Equity and artificial intelligence in education: Will "AIEd" amplify or alleviate inequities in education? *arXiv,* 2104.12920. https://doi.org/10.48550/arXiv.2104.12920

Holstein, K., McLaren, B., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher-AI complementarity. *Journal of Learning Analytics, 6*(2), 27-52. https://doi.org/10.18608/jla.2019.62.3

Holstein, K., Wortman, V., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 600. https://doi.org/10.1145/3290605.3300830

Hong, S., Hwang, Y., Park, Y., & Lee, S. (2024). Expectations and concerns about adopting AI digital textbooks: Based on investigation of teachers' use of AI and digital tools. *The Journal of Studies in Language, 40*(1), 7-20. https://doi.org/10.18627/jslg.40.1.202405.7

Ibrahim, A., Thiruvady, D., Schneider, J-G., & Abdelrazek, M. (2020). The challenges of leveraging threat intelligence to stop data breaches. *Frontier in Computer Science, 28*, 1-11. https://doi.org/10.3389/fcomp.2020.00036

Idowu, J., Koshiyama, A. S., & Treleaven, P. (2024). Investigating algorithmic bias in student progress monitoring. *Computers and Education: Artificial Intelligence, 7,* 100267. https://doi.org/10.1016/j.caeai.2024.100267

Jiang, W., & Pardos, Z. (2021). Towards equity and algorithmic fairness in student grade prediction. *arXiv,* 2105.06604. https://arxiv.org/abs/2105.06604?utm_source=chatgpt.com

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., & Krusche, S. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, S., & Bang, J. (2019). One more way of understanding the education in the era of AI. *The Journal of Educational Principles, 24*(1), 83-105. http://dx.doi.org/10.19118/edp.2019.24.1.83

Kim, S., & Bang, J. (2020). The introduction of the concept of 'Education AI': The challenge of sustainable education by the cooperation between human and AI. *The Journal of Educational Principles, 25*(1), 1-21. http://dx.doi.org/10.19118/edp.2020.25.1.1

Kim, S., Bang, J., & Kwon, H. (2018). A discussion on the planning of national digital transformation in the education sector. *The Journal of Korea Education, 45(*4), 173-200. https://doi.org/10.22804/jke.2018.45.4.007

Knox, J., Williamson, B., & Bayne, S. (2019). Machine behaviourism: Future visions of 'learnification' and 'datafication' across humans and digital technologies. *Learning, Media and Technology, 45*(1), 31-45. https://doi.org/10.1080/17439884.2019.1623251

Kwak, Y., & Pardos, Z. A. (2024). Bridging large language model disparities: Skill tagging of multilingual

educational content. *British Journal of Educational Technology, 55*(5), 2039-2057. https://doi.org/10.1111/bjet.13465

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher, 35*(7), 3-12. https://www.jstor.org/stable/3876731

Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. (2024). The life cycle of large language models: A review of biases in education. *arXiv*, 2407.11203. https://doi.org/10.1111/bjet.13505

Lee, M. J., & Singh, S. (2021). The landscape and gaps in open source fairness toolkits. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 699. https://doi.org/10.1145/3411764.3445261

Meaney, M., & Fikes, T. (2019). Early adopter iteration bias and research praxia bias in the learning analytic ecosystem. *Learning Analytics and Knowledge 2019: Companion Proceedings of the 9th International Learning Analytics & Knowledge Conference*. 14-20. https://www.researchgate.net/publication/333755926

Ministry of Education of Korea. (2023). Realizing personalized education for all: Strategies for digital-based educational innovation. Ministry of Education of Korea. [Press release] Retrieved from https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&boardSeq=94011&lev=0&searchType=null&statusYN=W&page=1&s=moe&m=020402&opType=N.

Naismith, B., Han, N.-R., Juffs, A., Hill, B., & Zheng, D. (2018). Accurate measurement of lexical sophistication with reference to ESL learner data. *Proceedings of 11th International Conference on Educational Data Mining*, 259–265.
http://educationaldatamining.org/files/conferences/EDM2018/papers/EDM2018_paper_37.pdf

Park, Y., Lee, S. , Hong, S., & Hwang, Y. (2024). A grounded theory on AI-based teacher supporting platform. *Journal of Korean Association for Educational Information and Media, 30*(3), 1005-1034. http://dx.doi.org/10.15833/KAFEIAM.30.3.1005

Perry, A. (2019). AI can disrupt racial inequity in schools, or make it much worse. https://www.tc.columbia.edu/articles/2019/september/ai-can-disrupt-racial-inequity-in-schools-or-make-it-much-worse/

Plaza P., Castro, M., Merino, J., Restivo, T., Peixoto, A., & Gonza, C. (2020). Educational robotics for all: Gender, diversity, and inclusion in steam. *Proceedings of IEEE Learn. MOOCS*, 19-24. https://doi.org/10.1109/LWMOOCS50143.2020.9234372

Radwan, A., Zaafarani, L., Abudawood, J., AlZahrani, F., & Fourati, F. (2024). Addressing bias through ensemble learning and regularized fine-tuning. *ArXiv*, 2402.00910v2. https://doi.org/10.48550/arXiv.2402.00910

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater R automated scoring engine and humans for demographically based groups in the GRER general test. *ETS Research Report Series*, 1-31. https://doi.org/10.1002/ets2.12192

Reich, J., & Ito, M. (2017). From good intentions to real outcomes: Equity by design in learning technologies. *Digital Media and Learning Research Hub.* https://tsl.mit.edu/research/from-good-intentions-to-real-outcomes-2/

Salman, S., & Liu, X. (2019). Overfitting mechanism and avoidance in deep neural networks. *ArXiv*,

abs/1901.06566. https://doi.org/10.48550/arXiv.1901.06566

Sha, L., Rakovic, M., Das, A., Gasevic, D., & Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on learning and Technologies*, *15*(4), 481-492. http://doi.org/10.1109/TLT.2022.3196278

Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V., Gasevic, D., & Chen, G. (2021). Assessing algorithmic fairness in automatic classifiers of educational forum posts. In Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (Eds.) *Lecture Notes in Computer Science*, 12748. Springer. https://doi.org/10.1007/978-3-030-78292-4_31

Smith, M., Todd, L., & Laing, K. (2017). Students' views on fairness in education: The importance of relational justice and stakes fairness. *Research Papers in Education*, *33*(3), 336–353. https://doi.org/10.1080/02671522.2017.1302500

Stray, J. (2023). The AI learns to lie to please you: Preventing biased feedback loops in machine-assisted intelligence analysis. *Analytics*, *2*(2), 350-358. https://doi.org/10.3390/analytics2020020

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 17,* 1-9. https://doi.org/10.1145/3465416.3483305

Unterhalter, E. (2009). What is equity in education? Reflections from the capability approach. *Stud Philos Educ*, *28*, 415-424. https://doi.org/10.1007/s11217-009-9125-7

Varsik, S., & Vosberg, L. (2024). *The potential impact of artificial intelligence on equity and inclusion in education: OECD artificial intelligence papers*. OECD Publishing. https://doi.org/10.1787/15df715b-en

Wang, Z., Zechner, K., & Sun, Y. (2018). Monitoring the performance of human and automated scores for spoken responses. *Language Testing*, *35*(1), 101−120. https://doi.org/10.1177/0265532216679451

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, Q. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing*, *62*, 101071. https://doi.org/10.1016/j.jslw.2023.101071

Wong, D., Allen, K., & Cordoba, B. (2022). Examining the relationship between student attributional style, perceived teacher fairness, and sense of school belonging. *Contemporary Educational Psychology*, *71*, 102113. http://doi.10.1016/j.cedpsych.2022.102113

Yang, Y., Gupta, A., Feng, J., Singhal, P., Yadav, V., Wu, Y., Natarajan, P., Hedau, V., & Joo, J. (2022). Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 813-822. https://dl.acm.org/doi/10.1145/3514094.3534153

Yu, H., Miao, C., Leung, C., & White, T. J. (2017). Towards AI-powered personalization in MOOC learning. *npj Science of Learning*, *2*, 15. https://doi.org/10.1038/s41539-017-0016-3

Zanna, K., & Sano, A. (2024). Enhancing fairness and performance in machine learning models: A multi-task learning approach with Monte-Carlo dropout and Pareto optimality. *arXiv,* 2404.08230. https://arxiv.org/html/2404.08230v1#bib.bib41

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature*, *559*(7714), 324-326. https://doi.org/10.1038/d41586-018-05707-8